WORLD METEOROLOGICAL ORGANIZATION
GLOBAL ATMOSPHERE WATCH

No. 133

WMO/EMEP Workshop on Advanced Statistical Methods and their Application to Air Quality Data Sets

(Helsinki, 14-18 September 1998)

WMO TD - No. 956
# Table of Contents

1. OPENING OF THE WORKSHOP ................................................................. 1

2. OVERVIEW OF GAW AND EMEP ....................................................... 1
   2.1 The relationship between GAW and EMEP ....................................... 1
   2.2 Goals of the Workshop ....................................................................... 1
   2.3 Quality Assurance's role in data analysis ........................................... 3
   2.4 Introduction to EMEP ....................................................................... 3

3. EMEP AND GAW, STATUS AND NEEDS ............................................. 4
   3.1 EMEP measurements and models: Calculations, applications and data analysis ........................................................................ 4
   3.2 Measurement and interpretive software that can be retrieved from NILU ........................................................ 5

4. LECTURES .................................................................................................. 6
   Basic statistical analysis and data interpretation (R. Vet)
   Cluster/principle component analysis/factor analysis (A. Sirois)
   Time series analysis (A. Sirois)
   Spatial interpolation analysis (J. Schaug)
   Evaluation of Kriged and modelled concentrations (L. Tarrason)
   POP's and heavy metal data and analysis (A Ryaboshapko)
   Hy split trajectory model and clustering (R. Draxler)
   Tropospheric ozone data analysis (O. Hov)

5. REPORTS OF INFORMAL WORKING GROUPS ..................................... 6
   Group 1 Mediterranean ........................................................................... 6
   Group 2 South Central Europe ............................................................... 8
   Group 3 Central Europe ......................................................................... 10
   Group 4 Northern Europe ...................................................................... 13

Annex A LIST OF PARTICIPANTS ................................................................. 15
Annex B AGENDA ........................................................................................ 21
Annex C Principal Component Analysis and other Dimensionality Reduction Techniques: a short overview ...................................................... 24
Annex D A brief overview of Cluster Analysis ............................................ 162
Annex E A brief and brazed overview of Time Series Analysis .................... 186
1. OPENING OF THE WORKSHOP

The workshop was opened by Prof. Antti Kumala, Chief of the Air Quality Division of the Finnish Meteorological Institute (FMI). Prof. Kumala introduced the Prof. E. Jatila, FMI director, who welcomed the participants to Helsinki. (See Annexes A and B for the list of participants and the agenda). Prof. Jatila emphasised the importance of the European Monitoring and Evaluation Programme (EMEP) and the Global Atmosphere Watch (GAW) in monitoring and modelling in Europe. Further he expressed his best wishes for a successful meeting. Next Prof. Kumala introduced Dr John Miller, Chief of the WMO’s Environment Division who is responsible for the Global Atmosphere Watch. Dr Miller welcomed the participants in the name of the World Meteorological Organization’s (WMO) Secretary-General. He stated the EMEP and WMO has a long history of cooperation through a series of Workshops starting with the Passau workshop over five years ago. Further he thanked the FMI for hosting the meeting and commented on the excellent GAW/EMEP activities in the Finland. Prof. Kumala introduced Prof. A. Eliassen, Deputy Director of the Norwegian Meteorological Institute who represented the EMEP programme. He welcomed the group in the name of the EMEP Centers.

2. OVERVIEW OF GAW AND EMEP (CHAIR A. KUMALA)

2.1 The Relationship between GAW and EMEP

Dr Miller gave a brief overview of the GAW programme and its relationship to the EMEP. He stated that WMO members have requested that more emphasis in GAW should be placed on air quality and the urban/regional environment which fits even more closely with EMEP activities. Shown in Figure 1 is the relationship between the two measurement programmes. Further the GAW cooperates directly with the two Meteorological Synthesizing Centres East and West (MSC-W, MSC-E) and the EMEP’s Chemical Coordinating Centre (CCC) on both modelling and monitoring.

2.2 Goals of the Workshop

Ms. Sonja Vidic, Chief of the Environment and Boundary Layer Research Division of the Meteorological and Hydrological Service of Croatia, and a member of the Bureau of the EMEP Steering Body, outlined the goals of the workshop. She pointed out that this workshop is a follow-up of two workshops held in 1997 (WMO/EMEP Workshop on Data Analysis, Validation and Reporting in Usti na Labem, Czech Republic April 1997 and WMO Meeting of Experts on
GA W Regional Network in RA VI – Europe in Budapest, Hungary May 1997) where one of the strongest recommendations from participating scientists pointed to the need for encouraging and intensifying the level of effort going into data analysis and interpretation at both national and international levels. Ms. Vidić stressed that in both programmes, GAW and EMEP, a lot of resources are allocated to measurements and data gathering activities. Recognizing that those activities represent only the first step in the overall process: measurements – data analysis – data interpretation – applying this knowledge in understanding environmental change. Thus just collecting the data does not provide the needed information but an effort must be made to perform the appropriate analysis and interpretation. To place a higher priority on this activity, this workshop was organized with the following objectives:

- Present EMEP and GAW measurement programmes with the emphasis on the use and interpretation of measured, analyzed and modelled data/results;

- Basic and advanced statistical methods and their application to air quality data sets, spatial and temporal data analysis tools, trajectory analysis etc;

- Problems in the analysis and interpretation of measurements and results, shortcomings and limitations of methods applied, etc;

- Plans for specific projects in which results from the above exercise would be used with the purpose of enhancing the analysis and interpretation of EMEP/GAW data.

It was stressed that the focus of the workshop was on the training and demonstration of advanced statistical techniques for potential application to measurement data. The first objective was to update/inform scientists on the latest methods for atmospheric data analysis and interpretation techniques that could be used on EMEP/GAW data in the forthcoming years.

The second goal was to open the discussion about the possibilities for scientific cooperation on a smaller, regional or subregional scales in a smaller working groups. Dividing up Europe into natural regions as Mediterranean, South East, Central and Northern Europe could be of help in addressing problems specific for certain regions of Europe. Therefore the idea to go ahead with this approach and continue with a series of three workshops for the period from 1998 to 2002 was presented. The three workshops should be evolutionary, with each providing a forum for planning and initiating the subsequent one.

Prior to the workshop, participants were invited to prepare for the workshop by bringing with them specific topics of interest specifically on data use and experience with data analysis on the national level. Furthermore, to facilitate the work at the workshop, and enable the preparation of the general overview of environmental and monitoring data use, participants were kindly requested to complete a short questionnaire, which was done by most participants. Also, in order to prepare for the discussion, participants were requested to review the following reports:

- EMEP-WMO Workshop on Data Analysis, Validation and Reporting (EMEP/CCC Report 6/97);

- EMEP-WMO Workshop on Strategies for Monitoring of Regional Air Pollution in relation to the need within EMEP, GAW and other international bodies (EMEP/CCC Report 10/97);

- MSC Status Reports, EMEP/MSV-W Report 1/96 or 1/97;
• Report and Proceedings of the Workshop on the Assessment of EMEP Activities concerning Heavy Metals and Persistent Organic Pollutants and their Further Development (GAW Report No. 117, Volumes I and II);

• NILU Data reports.

In summary, the interpretation of EMEP/GAW measurement data should be encouraged and would lead to furthering our understanding of complex interrelated atmospheric processes, anticipating future changes, and supporting authorities in their efforts to develop and uphold sound environmental policies at different scales.

2.3 Quality Assurance’s role in data analysis

Prof. Volker Mohnen, Director of the Quality Assurance Science Activity Centre (QA/SAC) for the Americas presented an overview of the Quality Assurance (QA) and its importance in data analysis. He emphasized that the programmes must satisfy the cost to benefit ratio for our activities imposed on the programmes by the respective political, economic and social communities. To meet these expectations, data of known quality and adequate for their intended use. He stated that before any data set can be accepted for further analysis it must be traced back to the data generator that include EMEP/GAW data centres, sites and information on instrumentation. Using this information, there must also be a clear statement of the data quality objectives, the project plan to include Standard Operating Procedures (SOPs) and the QA programme. The main point was that though much data are produced it is critical to apply them to a better understanding of the atmosphere and its future change.

2.4 Introduction to EMEP

Prof. A. Eliassen outlined the essential elements of EMEP. He emphasized the complementary nature of the measurement and modelling activities, and the important role of the countries in providing measured data of good quality. Assuming however that these basic elements of the programme were already well known to the participants, he went on to outline how EMEP has been supporting the development of various international emission control protocols under the Convention for Long-Range Transboundary Air Pollution (CLRTAP). EMEP has been instrumental in providing a scientific basis for the Second Sulphur protocol, and plays an equally important role in supporting the ongoing negotiations on a Multi-Pollutant, Multi-Effect Protocol. The most important contribution of EMEP in this context is to attribute acid deposition or surface ozone concentrations to the emissions of $\text{SO}_2$, $\text{NH}_3$, $\text{NO}_x$ or VOC in each country. Together with information on critical loads for acid deposition and critical levels for surface ozone, as well as on the abatement costs of emissions in each country, the information from EMEP is used to estimate emission reduction obligations for each country for least cost attainment of agreed environmental goals. The environmental goals can be defined for example by an overall uniform percent reduction of the excess acid deposition (the so-called gap closure approach), or by other internationally accepted reduction patterns for the excess deposition.

The attribution of deposition or concentrations to emission from each country is an essential piece of information for the development of such cost-efficient effect-based protocols in which the environmental goals are attained at a much lower cost than if all countries were to agree on uniform percent emission reductions, for example. Such attribution can only be done using models for long range transport of air pollution. In order to trust the model output, it is absolutely necessary that the model output compares reasonably well with measurement from the EMEP network. This imposes quality requirements on the models as well as on the measurements.
3. EMEP AND GAW STATUS AND NEEDS

3.1 EMEP measurements and models  Calculations, applications and data analysis

Further Prof. Eliassen outlined that the chemical components relevant for acid deposition have been measured since EMEP started in 1977. Ozone measurements have been added later. The early model calculations were carried out with quite simple model versions, and have not been stored. In principle, the following measurement and modelling data are available from EMEP in the area of acid deposition and photooxidants:

1. Measured data (daily) from the EMEP measurement network (1977-present, note that not all stations have measured the mandatory programme);
2. Consistent Lagrangian model runs for acid deposition (1985-1997, deposition and concentrations attributed to the emissions of each country, 150 km resolution);
3. Eulerian model run for 1996 (50 km resolution, otherwise as 2).

All these data contains errors. How serious are these errors? Do they in any way prevent the data from being used according to the intentions? In an attempt to answer this question, at least in part, one may draw advantage of the fact that measurement and modelling errors behave in different ways. Measurement errors can change abruptly in time and space: in time, when for example a new method for sampling or chemical analysis is implemented, or in space, when moving from one nearby station to another, but operated by different laboratories.

When plotting time series of measured versus calculated data there are quite a few examples in which the behaviour of the measurements change abruptly as compared to the model calculations, at a certain point in time. In some of these cases, the reason for this change is not known. In other cases, we know that the changes in behaviour coincides with the implementation of a new method for sampling and analysis.

Sometimes we also see that the relationship between measured and modelled data is significantly different in one other country compared to other countries. In most cases the reason for this situation is not known, but should be investigated.

This shows that the difference in error behaviour between modelled and measured data can be used to identify changes in the error characteristics of the measured data. It is important to find the reason for such changes in error characteristics.

A general impression from the work at MSC-W is that the quality of the EMEP measurements is generally lower in South and East Europe than in North and West Europe, although there are exceptions. This situation has persisted for many years, but with little sign of improvement. Furthermore, the one-layer Lagrangian model of MSC-W, which works reasonably well in North Europe, is probably not adequate for describing pollution transport in South Europe, with its complex terrain and local circulation patterns. This is the main reason for the development of the Eulerian model, which has a real 3-dimensional capability (20 levels) and an improved resolution. The MSC-W intends to co-operate with the Mediterranean countries and other countries with complex topography, in order to investigate the potential of the Eulerian model, with a view to providing data interesting for national scientific institutions as well as environmental authorities. We hope that this also might lead to improvements in the measurement quality.

The discussion that followed focused on how countries and the EMEP centres could cooperate on analysing data from EMEP. From MSC-W, the following data would, inter alia, be available:

• Calculated concentration and deposition patterns for all major pollutants involved in acid rain formation (SO₂, SO₄, NO/NO₂, NO₃, NH₃, NH₄) and broken down into
patterns due to emissions in each European country separately, annually covering 12 years. (Lagrangian model, 150 km spatial resolution, monthly time resolution);

• Calculated changes in excess deposition patterns above the critical loads due to changes in national emissions of SO₂, NOₓ, or NH₃, annually covering 12 years or more. (Lagrangian model);

• Calculated daily concentration fields of surface ozone, for seven summer periods. (Lagrangian photooxidant model);

• Calculated changes in the geographical distribution of surface ozone, AOT 40 or AOT 60 due to changes in the national emissions of NOₓ and VOC, for seven summer periods (six months each, Lagrangian photooxidant model);

• Calculated time series of the above mentioned components at all EMEP stations, daily resolution, as available;

• All emissions data reported by parties to CLRTAP, and expert estimates of emissions for countries/components/years not officially reported, as available.

In the future, calculations with the Eulerian model (50 km resolution) will be made available. This will also include attribution of concentrations and depositions to the emissions of each country, as hitherto carried out with the Lagrangian model.

3.2 Measurement and interpretive software that can be retrieved from NILU

Dr Hov, Director of NILU, described how the international collaboration in EMEP can contribute to the development of regional and national air pollution assessments in those parts of Europe where the national awareness and technical competence still needed to be strengthened.

The EMEP centres can support this development by making available via internet data on chemical measurements, emissions, meteorology, and model calculations. Also software for educational purposes or as starting point for national efforts in data analysis will be made available or informed about in the web sites of the EMEP centres. Such data analysis can be based on both statistical and deterministic models.

Web-address for EMEP-CCC which contains further pointers and links:

http://www.nilu.no/projects/ccc

• EMEP measurements;
• EMEP site description;
• EMEP analytical methods;
• other information.

Other information can be links to

• photochemical box model;
• photodissociation rate coefficient model;
• trajectory calculations;
• model for the global cycling of persistent organic pollutants.
4. **LECTURES**

One of main purposes of the workshop was to present a series of lectures on statistical methods as they apply to GAW/EMEP data analysis. The text of these lectures are available from the WMO Secretariat. Two of the lecture notes on principle component analysis and time series by A. Sirois is reproduced in the Annex 3 and 4.

- Basic statistical analysis and data interpretation (R. Vet)
- Cluster/principle component analysis/factor analysis (A. Sirois)
- Time series analysis (A. Sirois)
- Spatial interpolation analysis (J. Schaug)
- Evaluation of kriged and modelled concentrations (A. Eliassen)
- POP's and heavy metal data and analysis (A. Ryaboshapko)
- HY split trajectory model and clustering (R. Draxler)
- Tropospheric ozone data analysis (O. Hov)

5. **RECOMMENDATION OF THE WORKING GROUPS**

Four working groups were set up during the workshop and were asked to address problems and issues related to GAW/EMEP. Two specific themes were:

- How can EMEP and GAW centres can assist the Member countries in developing emission inventories, improving measurement systems and data flow. To do this, what educational software (statistical packages, programmes for box/trajectory models) and additional products can be provided;

- Define the needs, composition and objectives for regional sub-groups that could be formed during and after the workshop to possibly prepare national and regional assessments using GAW/EMEP measurements and models.

It should be noted that these discussions can only be considered preliminary and informal. The four regions are in a sense arbitrary and not all GAW/EMEP programmes in countries in the regions are represented. However a beginning can be made to describe the needs of a given geographical area and that these suggestions can be discussed and fully developed in future workshops. Further this being a workshop, the participants come as individual scientists and do not necessarily represent their country or their region but only individual scientific views. However to visualize the region, country names will be used.

The following is the results of the groups discussions:

**Group 1: Mediterranean (Chair S. Vidic)**

Region: Countries active in GAW/EMEP rimming the Mediterranean sea with scientists from Croatia, Cyprus, France, Portugal, Slovenia and Spain

The working group has discussed possibilities of cooperation between countries on sub-regional level that share similar problems and needs, concerning both, EMEP and GAW programmes and their implementation. It has been stressed that general recognition of the scientific and operational work is not good enough to be supportive to this work. In many cases, links between different institutions involved are not well established and exchange of information is insufficient. Participants have stressed that if the work, organised in a form of projects would
be carried out under the umbrella of EMEP and GAW, it might be of considerable help. By working on a sub-regional level countries could share knowledge and resources and focus on what each of them could do the best or concentrate on problems that are already well defined at national levels.

There should not be any obstacles in promoting this type of cooperation among countries and that geographical location of the country should not be the exclusive basis for the cooperation. Countries should try to identify their interest in the field of environmental protection and to communicate them in order to define what would be of benefit for a whole region. Common meteorological patterns, environmental issues (acidification, eutrophication, ozone, particulate, etc.), status of the measurement network, modelling results or data analysis and interpretation for a region of interest in general could serve as a different framework for projects and can involve everybody who is interested in. It is clear that this type of cooperation leaves enough space for participation and can be of considerable benefit. It is said that such cooperation, supported and encouraged by WMO/GAW and EMEP would stress the importance of both programmes and could serve as a good basis for their promotion. Consequently, that might help in allocating more financial resources for the improvement of actual situation (monitoring, instrumentation etc.).

It has been perceived that EMEP and GAW could facilitate this work on the basis of what already exists, in terms of software, data, model calculation results, expertise, etc. and with relatively modest financial resources allocated.

It has been considered that interested participants, willing to take part in this sub-regional cooperation could start with gathering and systematising existing information about the status of monitoring, parameters measured, meteorological information, problems identified, issues important for a region, particular country, etc. For now, without actual financial support that would give more flexibility and prospects, this work could be conducted only based on personal commitment, time available and enthusiasm. It has been recognised that it actually means that at this stage it can be expected only that participants to this workshop, or interested scientists could do it by putting more efforts into a regular work to integrate specific additional tasks.

Concerning issues discussed, the working group has proposed:

- To identify EMEP/GAW experts involved in operational and scientific work who could form a pool of experts that could be contacted and invited to participate or lead some activities of common interest; make lists of experts, regularly update it and circulate that information;

- To find a way and help in identifying the existing chains of responsibilities in different countries, especially where actual communication is not well established or understood (members of different bodies under the Convention);

- To promote and enable exchange of results, papers, reports etc. among experts;

- To use internet as one of possibility for communication and exchange of information, results, views in a more coherent way, support and encourage communication through existing EMEP, WMO web pages;

- To list of available software for statistical analysis, data interpretation, and modelling;

- To make available complete information/easy access to the data available for the analysis at national, regional or sub-regional levels;
• To consider possibilities for joint projects supported by WMO GAW and EMEP based on well-defined goals, participants, budgets, etc;

• To name S. Vidic, organiser of the next workshop on the data analysis and interpretation (Dubrovnik, Croatia, October 1999), to serve as a contact person for the exchange of information and facilitation with regard to the ideas about common projects for the countries of the Mediterranean region, but also for all participants that might be willing to take part;

• To compose a letter of recommendations and conclusions from this workshop that should be circulated to the relevant bodies and experts involved in or benefitting from EMEP / GAW work and activities (EMEP Steering Body members, EB Members, Ministries of Environment, Institutes and Meteorological Services etc.).

**Group 2: South Central Europe (Chair V.Cuculeanu)**

Region: Countries active in GAW/EMEP located in the Balkan peninsula and nearby inland areas with participating scientists from Bulgaria, Hungary, Macedonia, Romania, and Yugoslavia

The working group highly appreciates the initiative of organizing this Workshop which, due to the outstanding quality of the lectures, resulted in broadening the scientific knowledge of the participants on the use of statistics in their work. Also the Workshop was an opportunity to establish new contacts and share experience.

The conclusions and recommendations of the working group sessions are set out below:

**Emission inventories:**

• Regarding the emission inventories the group considered it to be a difficult problem because the emission data belong to different authorities, each of them estimating and reporting these data in different ways. A single methodology (IPCC, CORINAIR) with corresponding data sheets would be useful for institutes to estimate and collect the emission data.

**Needs relating to measurement systems and data flow:**

• With regard to the measurement systems the group recommend the initiation of a regional project for the rehabilitation of the monitoring network in the geographical area covered by the countries of this group. In the frame of the project, the upgrading of the existing experimental facilities as well as the implementation of new monitoring systems for the pollutants monitored by EMEP/GAW networks are envisaged. It was suggested that project could be partly financed by European Commission (EC) or/and WMO/EMEP and partly by the participating countries;

• To facilitate government support for the national institutes deploying EMEP/GAW activities it would be advisable that WMO/EMEP notify the sponsoring ministries on the importance of these activities;

• The group members revealed the importance of the periodical training and courses organized by NILU as well as the visits of the WMO/EMEP experts to the national stations in order to analyse all scientific and technical aspects of their activity (expert consultation programme);
• Often the scientists have to apply dispersion models at regional scale to solve environmental issues within one or between countries. In this respect an internationally recognized regional model is needed as common tool. The EMEP centre will be of great help to the national experts if a PC version of the Lagrangian model is prepared to be made available upon request.

Moreover the implementation and run of this model will result in increasing the scientific capacity in each country. Comparison of the measured data with the model simulations could reveal inconsistencies in the measurement methodology.

• In order to unify the necessary computer software (statistics, programs for data base management) used for data analysis, a Communication page on the EMEP internet site should be created. On this page people could look up what products and information are available and/or send their information for common use. It was suggested that even the data measured by the national stations be sent to this EMEP web site.

As a basic statistical software, the use of EXCEL and Statgraphics was recommended.

• The Workshop on quality assurance which is to be held in Czech Republic is expected to contribute significantly to the improvement of the measurement quality;

• The interpretation of the ozone data at the national scale is necessary both for the scientific knowledge and to provide sound information to the governmental authorities. In this respect the group consider to be quite useful that EMEP make available a PC version of the box-model;

• For the research purposes (e.g., dynamical systems analysis) it would be important that EMEP make available certified time series of atmospheric gas concentrations (particularly ozone) with a relevant number of data (>30,000).

Objectives for groups:

• The activity of the regional groups should be organized according to a single scheme proposed by WMO/EMEP;

• The first task of the group should be the evaluation of the present status of the measurement programme and corresponding facilities, quality of data as well as data flow for each country. The existing software should be analysed as well.

Finally the needs for each country have to be emphasized.

• Assess the pollution state at national and regional scales using GAW/EMEP measurements and models (statistical and deterministic). For this purpose the NILU database of emissions should be freely accessed by countries.

Suggestions for the follow up plans

• EMEP to initiate projects for developing the PC version of the mentioned models (Lagrangian model, box-model) as well as for creating certified time series. In order to comply with the EMEP requirements of the measurement programme, the Working Group-2 countries, as Parties to the Convention, have to improve the network functioning at national level. In this respect, it was recommended the initiation of a regional project aiming at the rehabilitation of the monitoring network in this
geographical area. The project led by WMO/EMEP should be financed jointly by one international organization (CE, WMO/EMEP) and by each country from group.

Possible item for Croatia meeting

- Discussion of the geographical distribution of the special stations (POP, heavy metals) which EMEP plans to locate throughout Europe.

Group 3: Central Europe (Chair: R. Gehrig)

Region: Countries with GAW/EMEP programmes in central Europe with participating scientists from Austria, Croatia, Czech Republic, Germany, Slovakia, Slovenia and Switzerland.

The main task of the work was to discuss efficient ways of joint work of several countries in a region on a problem for which a country needs assistance and which is of interest and benefit for a whole region, by making use of the knowledge available in this region. In addition the group was asked to discuss how the EMEP centres can support these activities.

Generally the temporary limited formation of dedicated groups for specific projects was considered to be more efficient than the installation of fixed regional expert groups. The main reason for this is that a project oriented approach to regional problems allows more flexibility and a more efficient use of personal and financial resources.

Examples for possible regional projects are:

- Source attribution on a regional scale;
- Trend analysis on a regional level;
- Assessment of site representativeness of national stations and/or optimisation of national networks using EMEP data;
- Comparison of EMEP models with models used on national/regional level.

The group recommended to plan and realise projects according to the usual rules of scientific project management. The following procedure is proposed:

- Identification of the problem by a country;
- Contacts with other countries in the region which could be interested to work on the problem. Information about project plan forwarded to CCC for insertion into web-list;
- Setting up a detailed project plan, including:
  - Goals of the project;
  - Needed informations, data, tools, resources;
  - Definition of work-packages (who does what);
  - Project schedule;
  - Budget (manpower, equipment, tools).
- Evaluation of the project by one of the centres;
- If financial contributions are needed: Submitting the evaluated project to a suitable body (eg. Government, WMO, ETC/AQ, PTUAQ).
The group gratefully accepted the offer of the centres to provide on request additional assistance. In particular, the following services (which are partly already available) are of interest:

- Additional plausibility checking of submitted data by CCC within due time (<6 months) and access of the laboratories to their submitted data for check of proper transfer to the database;

- Availability of a software set for a basic statistical evaluation of the national EMEP data;

- Guidance for further analysis of the data (e.g. data aggregation, trend analysis, trajectories, box-model);

- Easy access to data for emissions, measurements and meteorology for use in national/regional projects;

- Support of national O₃-alert activities with the O₃-forecast programme;

- Kriging calculations with additional national data input.

The group recommended the following actions to be taken:

Introducing additional informations to the already existing EMEP-CCC web-pages:

- List of EMEP-experts with their fields of special interest and know-how in order to give informations to project leaders for possible partners. Input expected from EMEP-experts;

- Short description of planned regional projects. Input expected from project leaders;

- List of software tools for statistical evaluation and modelling including tutorials and educational programmes available from the centres or from other sources (e.g. internet). Input expected from the centres and from EMEP-experts.

Sending out of a letter (by one of the centres) to all EMEP-experts informing about the workshop recommendations and requesting the informations to be inserted into the web-pages mentioned above (if possible within 1-2 months).

A report by the centres about ongoing activities concerning the workshop recommendations at the next EMEP-workshop in Croatia.

Basic idea:

Joint work of several countries in a region on a problem that:

- a country needs assistance;
- is of interest and benefit for a whole region by making use of the knowledge available in this region and/or asking for assistance from EMEP-centres, WMO, etc.

Formation of ad-hoc project groups is considered to be more efficient than the installation of a fixed regional expert group.

Possible projects:

- Source attribution on a regional scale;
• Trend analysis on regional level;
• Assessment of site representativeness of national stations an/or optimisation of national networks using EMEP data;
• Comparison of EMEP models with models used on national/regional level.

Recommended procedure for regional projects:

• Problem/need identified by a country;
• Contact with other countries in the region which could be interested to work on the problem. Info put on internet list;
• Setting up the project plan, including:
  - Goals of the project;
  - Needed information/data/tools/resources;
  - Definition of work-packages (who does what), project schedule;
  - Budget (manpower, equipment, tools ...).

• Evaluation of the project by the centres;
• If financial contributions are needed: Submitting the evaluated project to a suitable body (Government, WMO, ETC/AQ, PTU/AQ ...).

Assistance from the centres:

The Working Group gratefully accepted the offer of the centre to provide on request additional assistance.

In particular, the following possibilities are of interest:

• Additional plausibility checking of submitted data by CCC within due time (<6 months) and access of the laboratories to their submitted data for check of proper transfer to the data base;

• Availability of a basic set of statistical evaluation of the national EMEP data (in NASA format);

• Guidance for further analysis (e.g. data aggregation, trend analysis, trajectories, box-model);

• Easy access to emission, measurement and meteo-data for use in national/regional projects;

• Support of national O₃ alert activities with the O₃-forecast programme;

• Kriging calculations with additional national data input.

Recommended actions to be taken:

Set-up of EMEP internet pages to be gradually filled with:

• List of EMEP-experts with their fields of special interest;
• Planned regional projects;
• List of scheduled EMEP workshops;
• List of available EMEP reports and EMEP-related publications;
• List of available software for statistical evaluation and modelling (incl. Tutorials/educational programmes).

Sending out of a letter (by one of the centres) to all EMEP experts informing about the important workshop recommendations and requesting information about special fields of interest (interesting software with links) and planned programmes to be introduced into the internet pages (if possible within 1-2 months).

Report by the centres about related activities at the workshop in Croatia.

**Group 4: Northern Europe (Chair T. Ruoho-Airola)**

Region: Countries with GAW/EMEP programmes in Northern Europe with participating scientists from Bulgaria, Denmark, Estonia, Finland, Latvia, Lithuania, and Russia.

The group felt that informal, regional co-operation - from scientist to scientist - would be needed beside the work which is done by the EMEP centres. However, this regional group should not undertake tasks already assigned to the EMEP centres. A contact list with e-mail addresses of scientists working within WMO and EMEP would help the regional group to start its work. Finland was requested to compile a list covering all relevant countries and to circulate it.

It was agreed that the natural composition of a regional group would cover all countries bordering the Baltic Sea plus Norway and Iceland. The historical co-operation, both among the Nordic countries and among countries around the Baltic Sea, offers a good foundation for regional air quality work in this part of Europe, which is reinforced by similarities in climate and ecological sensitivity across the region.

In the first years of co-operation more weight should be given to the harmonisation of measuring methods and regional models. Later the work could focus more on other common regional interests: e.g. assessing how best to analyse the data and report the results to serve the scientific, administrative and public need.

In some countries the measuring methods used in EMEP and WMO monitoring needs to be improved. Basically this is taken care of by the EMEP CCC or WMO, but a part of the training and support could be done bilaterally or regionally. At the beginning of the co-operation, one of the existing EMEP/WMO stations in the region could act as a co-operation station, where scientists could share their experience of applying standard methods. The role of the co-operation station and of the host country running it would need to be formalised within the EMEP and WMO organizations, so that external financing for the training could be applied for.

The group also discussed national needs for assessing long term trends in the air quality data. First, the quality of the earliest data gathered in various monitoring programmes should be investigated. Experience of statistical methods and software appropriate for this kind of work could be discussed among the scientists listed on the contact list. To assist in this, the EMEP CCC was requested to provide advice concerning software problems.

The group appreciated the offer by the EMEP MSC-W to calculate the sum of foreign contributions to the monthly, seasonal and annual depositions for each country. Individual countries could use this information when preparing detailed national or regional assessments with mesoscale models developed for their specific purposes. Russia was requested to collect more information on such models and report back.

Finally, the group expressed the need to analyse the EMEP/WMO data for scientific, administrative and public purposes. The lectures on statistical methods at the workshop were very useful for the participants and the handouts were clear and informative. Training courses of advanced statistical methods, where everyone could practise with their own air quality data,
would be very useful to enhance this knowledge further. The EMEP centres and WMO were encouraged to organize this kind of training.

The goals and possibilities for regional co-operation will develop with time. Future EMEP and WMO workshops would serve as an excellent forum for the regional group to meet and discuss progress and further common interests.

Recommendations from Working Group 4:

• Regional co-operation should be based on informal contacts between scientists;

• One of the existing EMEP/WMO stations in the region could act as co-operation station, where scientists could share their experience of applying standard methods in order to improve the quality of monitoring;

• The quality of early national air quality data should be investigated before long term trends can be calculated. EMEP CCC could probably assist with software problems;

• EMEP and WMO centres should investigate the possibility of organizing training courses on advanced statistical methods, where participants could practise with their own air quality data.
LIST OF PARTICIPANTS

Hayreddin Bacanli
State Meteorological Service
P.O. Box 401
ANKARA
Turkey
Tel: 90 312 302 2675
Fax: 90 312 359 3430
Email: tub@ecmwf.int

Dr Laszlo Bozo
Hungarian Meteorological Service
Institute for Atmospheric Physics
P.O. Box 39
H-1675 BUDAPEST
Hungary
Tel: 361 290 0163
Fax: 361 290 4174
Email: bozo@met.hu

Alica Bajic
Meteorological and Hydrological Service of Croatia
Gric 3, P.O. Box 254
10 000 ZAGREB
Croatia
Tel: 385 1 4585686
Fax: 385 1 431026
Email:bajic@cirrus.dhz.hr

Patrice Coddeville
Ecoles des Mines de Douai
Chemistry and Environment Department
941 rue Charles Bourseul
BO 838
59508 DOUAI Cedex
France
Tel: 33 3 27712636
Fax: 33 3 27712525
Email: Patrice.coddeville@ensm-douai.fr

Dimitrios Balis
Aristotle University of Thessaloniki
Campus Box 149
THESSALONIKI 54006
Greece
Tel: 303 1 248602
Fax: 303 1 998184
Email: balis@ccf.auth.gr

Vasile Cuculeanu
National Institute of Meteorology and Hydrology
Sos. Bucuresti-ploiesti, 97
71552 BUCHAREST
Romania
Tel: 40 1 2303240
Fax: 40 1 2303143

Ekaterina Batchvarova
National Institute of Meteorology and Hydrology
66 Tzarigradsko chaussee
1784 SOFIA
Bulgaria
Tel: 359 2 9753991
Fax: 359 2 884494
Email: ekaterina.batchvarova@meteo.bq

Roland Draxler
NOAA/Air Resources Laboratory
Office of Atmospheric Research
1315 East West Hwy
SILVER SPRING, MD 20910
USA
Tel: 1 301 7130295
Fax: 1 301 7130119
Aristodemos Economides  
Department of Labour  
Apelli 12  
1080 NICOSIA  
Cyprus  
Tel: 357 2 300334  
Fax: 357 2 663788  

Thomas Gauger  
Institut of Navigation Stuttgart University  
Geschwiler Scholl-Strasse 24  
70174 STUTTGART  
Germany  
Tel: 49 711 1214177  
Fax: 49 711 1212755  
Email: gauger@nav.unistuttgart.de  

Anton Eliassen  
DNMI  
P.O. Box 43, Blindern  
N-0113 OSLO  
Norway  
Tel: 47 22 963000  
Fax: 47 22 963050  
Email: anton.eliassen@dnmi.no  

Robert Gehrig  
EMPA  
Uberlandstrasse 129  
Ch-8600 DUBENDORF  
Switzerland  

Zita Ferenczi  
Meteorological Service  
P.O. Box 39  
1675 BUDAPEST  
Hungary  
Tel: 36 1 2900163  
Fax: 36 1 2904174  

Mateja Gjerek  
Hydrometeorological Institute  
Vojkova 1/B  
1000 LJUBLJANA  
Slovenia  

Marina Frohlich  
Umweltbundesamt  
Spittelauer Laende 5  
A-1090 VIENNA  
Austria  
Tel: 43 1 313045862  
Fax: 43 1 313045800  
Email: froehlich@uba.ubavie.gv.at  

Dr M.S. Golovko  
Hydrochemical Laboratory Centre for  
Radiation Control and Environmental  
Monitoring  
GLAVGIROMET  
16 Komsomolskaya Street  
22060 MINSK/50  
Belarus  
Tel: 7-0172 267725  
Fax: 7-0172 640335  

Marina Frolova  
Latvian Hydrometeorological Agency  
165 Maskavas Str.  
RIGA, LV-1019  
Latvia  
Tel: 371 7 112046  
Fax: 371 7 145154  
Email: epcc@meteo.lv  

Anne Hjellbrekke  
NILU  
P.O. Box 100  
N-2007 KJELLER  
Norway  
Tel: 47 63 898000  
Fax: 47 63 898050  
Email: anne.hjellbrekke@nifu.no
Dr Alexey G. Ryaboshapko
Institute of Global Climate and Ecology
Glebovskaya St. 20B
107258 MOSCOW
Russian Federation
Tel: 7 095 1605867
Fax: 7 095 1600831
Email: ryabosha@ra.igce.msk.ru

Alain Sirois
Service de l’Environnement Atmosphérique
2121 Voie de Service Nord,
Suite 500
Route Trans-Canadienne
DORVAL, Quebec
Canada H9P1J3
Tel: 1 514 4214674
Fax: 1 514 4212106
Email: alain.sirois@ec.gc.ca

Dr Jaroslav Santroch
Air Quality Laboratory
Czech Hydrometeorological Institute
Na Sabatce 17
143 06 PRAHA 4
Czech Republic
Tel: 420 2 4016719
Fax: 420 2 4010800
Email: santroch@chmi.cz

Liljana Talevska
Republic Hydrometeorological Institute
St Skubi b.b
P.Box 218
91000 SKOPJE
Republic of Macedonia
Tel: 389 91 397004
Fax: 389 91 397118
Email: life@meteo.gov.mk

Jan Schaug
NILU
P.O. Box 100
N-2007 KJELLER
Norway
Tel: 47 63 898000
Fax: 47 63 898050
Email: jan.schaug@nifu.no

Dalia Sopauskiene
Institute of Physics
Aerosol Research laboratory
A. Gostauto Str. 12
LT-2600 VILNIUS
Lithuania
Tel: 370 2 641227
Fax: 370 2 235162

Ms Radmila Simeva
Republic Hydrometeorological Institute
Street Skupi bb
P.Box 218
91000 SKOPJE
Republic of Macedonia
Tel: 389 91 397004
Fax: 389 91 397118
Email: life@meteo.gov.mk

Milan Vana
Czech Hydrometeorological institute
Kosetice Observatory
394 22 KOSETICE
Czech Republic
Tel: 420 365 498015
Fax: 420 365 498015
Email: vanam@chmi.cz

Mr Vasil Sitak
State Committee of Ukraine
for Hydrometeorology
Zolotorolskaya st., 6
252601 KIEV 34
Ukraine

Nedialko Valkov
National Institute of
Meteorology and Hydrology
66 Tsarigradsko chaussee
1784 SOFIA
Bulgaria
Tel: 359 2 9753991
Fax: 359 2 880494
Email: nedialko.valkov@meteo.bg
Annex A, p 6

Robert Vet
Atmospheric Environment Service
4905 Dufferin Street
DOWNSVIEW, Ontario
Canada M3H5T4
Tel: 1 416 7394853
Fax: 1 416 7395704
Email:robert.vet@ec.gc.ca

Ms Sonja Vidic
Environment and Boundary Layer
Research Division
Meteorological and
Hydrometeorological Service
Gric 3
P.O. Box 254
1000 ZAGREB
Croatia
Tel: 385 1 456 5719
Fax: 385 1 431 026
Email:Vidic@cirus.dhz.hr

Mrs Tanja Vukovic
Federal Hydrometeorological Institute
Bircaninova 6
P.O. Box 604
11001 BELGRADE
Yugoslavia
Tel: 381 11 645 779
Fax: 381 11 646 369
Email:slavko@meteo.yu

Jari Walden
Finnish Meteorological Institute
Air Quality Research
Sahaajankatu 20 E
00810 HELSINKI
Finland
Tel: 358 9 19291
Fax: 358 9 19295403
Email: jari.walden@fmi.fi

Jorg Zimmermann
Deutsche Wetterdienst Forschung
und Entwicklung
Frankfurter Strasse 135
63067 OFFENBACH
Germany
AGENDA

Monday, 14 September

14.00-14.30 OPENING OF THE MEETING
Welcome WMO: Jatila
Welcome EMEP: J M Miller

14.30-17.30 OVERVIEW OF GAW AND EMEP
14.30-14.50 The relationship between GAW and EMEP - J M Miller
14.50-15.10 Goals of the Workshop - S Vidic
15.10-15.40 Quality assurance's role in data analysis - V Mohnen
15.40-16.00 Coffee break
16.00-17.00 Introduction to EMEP: Results, their use and accessibility - A Eliassen
17.00-17.30 Discussion
17.30 -18.30 Get Together Party on the Workshop Premises hosted by FMI

Tuesday, 15 September

09.00-11.00 EMEP AND GAW: STATUS AND NEEDS
09.00-10.00 EMEP measurements and models: Calculations, applications and data analysis - A Eliassen
10.00-11.00 Measurements and Interpreative software hat can be retrieved from NILU - Ø Hov
11.00-11.30 Coffee break

11.30-15.30 LECTURES
11.30-12.30 Basic statistical analysis and data interpretation - R Vet
12.30-13.00 Discussion
13.00-14.00 Lunch
14.00-15.00 Cluster/principle component analysis/factor analysis - A Sirois
15.00-15.30 Discussion
15.30-16.00 Coffee break
16.00-17.30 DISCUSSION AND NATIONAL PRESENTATIONS

Question period concerning lectures, countries experience and short communications

Wednesday, 16 September

09.00-12.30 LECTURES

<table>
<thead>
<tr>
<th>Time</th>
<th>Lecture</th>
<th>Discussion</th>
</tr>
</thead>
<tbody>
<tr>
<td>09.00-10.00</td>
<td>Time series analysis - A. Sirois - or someone from NILU</td>
<td></td>
</tr>
<tr>
<td>10.00-11.00</td>
<td>Spatial interpolation analysis - J. Schaug</td>
<td>Discussion</td>
</tr>
<tr>
<td>11.00-11.30</td>
<td>Coffee</td>
<td></td>
</tr>
<tr>
<td>11.30-12.30</td>
<td>Evaluation of kriged and modelled concentrations - Tarrason</td>
<td>Discussion</td>
</tr>
<tr>
<td>12.30-14.00</td>
<td>Lunch</td>
<td></td>
</tr>
</tbody>
</table>

14.00-15.30 DISCUSSION AND NATIONAL PRESENTATIONS

Question period concerning lectures, countries' experience and short communications - conclusions

<table>
<thead>
<tr>
<th>Time</th>
<th>Activity</th>
</tr>
</thead>
<tbody>
<tr>
<td>15.30-16.00</td>
<td>Coffee</td>
</tr>
<tr>
<td>16.00-17.30</td>
<td>Visit to FMI, Air Quality Research</td>
</tr>
<tr>
<td>20.00-</td>
<td>Workshop dinner hosted by the Ministry of the Environment</td>
</tr>
</tbody>
</table>

Thursday, 17 September

09.00-12.30 LECTURES

<table>
<thead>
<tr>
<th>Time</th>
<th>Lecture</th>
<th>Discussion</th>
</tr>
</thead>
<tbody>
<tr>
<td>09.00-10.00</td>
<td>POPs and heavy metal data and analysis at MSC-E - A. Ryaboshapko</td>
<td>Discussion</td>
</tr>
<tr>
<td>10.00-11.00</td>
<td>HY split trajectory model and clustering – R. Draxler</td>
<td>Discussion</td>
</tr>
<tr>
<td>11.00-11.30</td>
<td>Coffee</td>
<td></td>
</tr>
<tr>
<td>11.30-12.30</td>
<td>Tropospheric ozone data analysis - Ø. Hov</td>
<td>Discussion</td>
</tr>
<tr>
<td>12.30-14.30</td>
<td>Lunch</td>
<td></td>
</tr>
</tbody>
</table>
14.00-15.30 PROBLEMS AND BENEFITS

The best approach to data analysis and interpretation. Application on national, regional and subregional levels. Combining measurements and modelled results - A. Eliassen, ØHov
Discussion

15.30-17.00 COMPREHENSIVE APPROACH TO DATA ANALYSIS AND INTERPRETATION

Resources needed, tools available, cooperation between scientists and countries, etc. - R. Vet, V.Mohnen, ØHov
Discussion

Friday, 18 September

09.00-10.30 Working groups - Outlining conclusions
10.30-11.00 Coffee
11.00-12.30 Wrap up and close of the workshop - ØHov
Principal Component Analysis and Other Dimensionality Reduction Techniques: A Short Overview

Alain Sirois

Atmospheric Environment Service
Environment Canada
2121, Voie de Service Nord, Suite 500
Route Trans-canadienne,
Dorval, Quebec, Canada, H9P 1J3
Notice

This report is the background document for the presentation of the same name given at the WMO-EMEP Workshop on Advanced Statistical Methods and their Application to Air Quality Data Sets, Finnish Meteorological Institute, Helsinki, Finland, 14-18 September, 1998. The author could be reached at the following address: Dr. Alain Sirois (ARQI), Service de l'Environnement Atmosphérique, 2121, Voie de Service Nord, Suite 500, Route Trans-canadienne, Dorval, Quebec Canada, H9P 1J3. Tel: (514) 421-4674; Fax: (514) 421-2106; E-mail: Alain.Sirois@ec.gc.ca
Preface

By definition, an overview should be as comprehensive as possible. Unfortunately, the lack of both time and space has put some physical limits on the extent to which the present overview can meet this requirement. Therefore, the author has had to select only a few of the many available dimensionality reduction techniques for examination. The criteria used in this selection were many, but the main one was whether the technique had already been or could be used in atmospheric chemistry. It was also felt that it was important that the techniques presented should have something in common. As the reader will see if he has the courage to read this overview, all the techniques presented here try to fit indirectly or directly the same conceptual model, one that can be called the factor model. In this conceptual model, the relationships between the observed variables are assumed to be caused by their relationships to unobserved variables. Depending on the method used, those unobserved variables could be independent or not. The reader should always keep that model in mind during his or her reading of this overview.

Six dimensionality reduction techniques will be reviewed here. To facilitate the task of the reader, five of the six sections describing those techniques have all been organized similarly. The exception is the one describing subjective principal component analysis (Section 5). The sections start with an introduction giving a brief description of the methods. They also include a step-by-step guide to the use of the technique. This is followed by a more mathematical section that gives all important technical aspects of the techniques. Questions as to how missing values are handled, statistical testing, and other details are discussed in it. This is followed by one or two examples of the technique. Details not included in the introduction or technical sections are discussed in the remarks section. In that section, possible extensions to the techniques may also be mentioned. A discussion of some of the available computer software related to the technique, and a brief summary of references to textbooks or journal papers related to the technique complete the section.

Following the sections describing the six dimensionality reduction techniques are two technical sections that relate to most of the techniques presented earlier. The first of these is a discussion of rotation. Most of the techniques presented here do not produce an unique solution. Therefore, the solution can be rotated. Many techniques have been developed to help in selecting a rotation that would produce a solution that is easier to interpret. We will discuss briefly some of those rotations in this section. The other technical section discusses various methods of estimating nonparametric confidence intervals for the dimensionality reduction techniques presented earlier. This is an important aspect of this field of research, as in most of the techniques confidence intervals for model parameters can be calculated only in special situations. This is why estimation techniques that are applicable to all the parameters and to most situations are useful tools.

The examples presented in the sections describing the dimensionality reduction techniques were taken from different research fields and are not generally related to atmospheric chemistry. It was therefore decided to present two examples linked to atmospheric chemistry in one section. Note that, because of lack of time, only a partial analysis of the data was made and the results are presented only as an illustration of some of the techniques and not as an analysis per se of the data.
As the reader will realize, the techniques presented in this overview cannot be used as “black boxes" but need the constant intervention of the user during the calculation. To help the reader, a short guideline to the use of those techniques is presented at the end of the overview.

In the last fifty years or more, complete confusion has been introduced into the field of dimensionality reduction techniques by the use of the term factor analysis to describe techniques that in fact use principal component analysis. As we will see, the two techniques are mathematically and conceptually different, although, the factor model applies to both - indirectly for the principal component analysis and directly for the factor analysis. This introduces a dangerous confusion, and in some papers it is impossible to be sure from what is written which of the two techniques has been used. The present author believes that the correct use of terms is a very important aspect of science and has not hesitated to rename techniques to reflect their true nature. He would strongly encourage the reader to do the same.

As mentioned earlier, time for the preparation of this overview was limited. Although the author tried as much as possible to make sure that the equations, examples, and descriptions are correct, errors could easily have crept in, especially in unifying the notation used by the different authors consulted. The author would therefore recommend, in particular, that before using the equations presented here the reader verify them in the references cited.
Contents

Preface iii

Contents v

1. Introduction 1

2. Principal Component Analysis (PCA) 5
   2.1 Introduction 5
   2.2 Technical Details 7
      2.2.1 Principal Components in Population 7
      2.2.2 Principal Components for Observed Data 10
      2.2.3 Principal Components Using Singular Value Decomposition 12
      2.2.4 Correlation Matrix 12
      2.2.5 Robust Principal Components 13
      2.2.6 Missing Data 14
      2.2.7 Confidence Intervals and Tests 15
         2.2.7.1 Confidence Intervals for the Eigenvalues or Latent Roots 15
         2.2.7.2 Confidence Intervals for the Elements of the Loadings Matrix 16
         2.2.7.3 Testing Equality of all Eigenvalues or Latent Roots 16
         2.2.7.4 Testing Equality of the Smallest Eigenvalues or Latent Roots 17
         2.2.7.5 Confidence Regions for the Ratio of the Sum of the Smallest
         Eigenvalues to the Total Sum of the Eigenvalues 18
      2.2.8 Reducing the Number of Principal Components 18
      2.2.9 Communality 19
      2.2.10 Determining the Numbers of Principal Components to Retain 19
         2.2.10.1 Informal Rule of Thumb Techniques 19
         2.2.10.2 Testing for the Statistical Significance of the Eigenvalues 21
         2.2.10.3 Exner's Function 22
         2.2.10.4 Cross Validation 23
         2.2.10.5 Partial Correlation 23
      2.2.11 Rotations 24
         2.2.11.1 Orthogonal Rotations 24
         2.2.11.2 Oblique Rotations 25
   2.3 Example of Principal Component Analysis 25
      2.3.1 Extracting the Principal Components 26
      2.3.2 Number of Principal Components to Retain 27
      2.3.3 Rotation 29
      2.3.4 Plotting for Principal Components 31
         2.3.4.1 Loadings Bar Chart 32
         2.3.4.2 Loadings Scattergrams 32
         2.3.4.3 Scores Scattergrams 34
   2.4 Remarks 34
   2.5 Computer Software 35
   2.6 References 36

3. Absolute Principal Component Analysis (APCA) 37
   3.1 Introduction 37
4. Target Principal Component Analysis (TPCA)
   4.1 Introduction 43
   4.2 Technical Details 45
      4.2.1 Q-mode Principal Component Analysis 45
      4.2.2 R-mode Principal Component Analysis 46
      4.2.3 Relationship between Q-mode and R-mode Principal Component Analysis 46
      4.2.4 Target or Procrustes Rotations 47
      4.2.5 Sources Scaling 49
   4.3 Examples of Target Principal Component Analysis 50
   4.4 Remarks 52
   4.5 Computer Software 53
   4.6 References 54

5. Subjective Principal Component Analysis (SPCA) 55

6. Factor Analysis 57
   6.1 Introduction 57
   6.2 Technical Details 58
      6.2.1 Factor Analysis Model 58
      6.2.2 Estimation of Factor Analysis Model 61
         6.2.2.1 Principal Component Method 61
         6.2.2.2 Principal Factor Method 62
         6.2.2.3 Communalities Estimation 63
         6.2.2.4 Maximum Likelihood Method 64
      6.2.3 Estimation of Scores Matrix 66
         6.2.3.1 The Weighted Least Squares Method 66
         6.2.3.2 The Regression Method 67
      6.2.4 Confidence Intervals and Tests 69
         6.2.4.1 A Large Sample Test for the Number of Common Factors 69
         6.2.4.2 Confidence Intervals for Loadings 70
      6.2.5 Rotations 71
         6.2.5.1 Orthogonal Rotations 71
         6.2.5.2 Oblique Rotations 71
   6.3 Examples of Factor Analysis 72
   6.4 Remarks 76
   6.5 Computer Software 78
   6.6 References 78

7. Positive Matrix Factorization (PMF) 79
   7.1 Introduction 79
   7.2 Technical Details 80
      7.2.1 Mathematical Model 80
11.5 Estimation of the Communalities 118
11.6 Selection of the number of Components or Factors to Retain 118
11.7 Estimation of Factors 119
11.8 Rotation of the Solution? 119
11.9 Estimation of the Factor Scores 119
11.10 Scaling for Sample Masses 119
11.11 Interpretation 120

12. Final Remarks 121

13. References 123
1. Introduction

Most research scientists involved in atmospheric chemistry monitoring have at least a passing acquaintance with the terms principal component analysis. They have probably heard of the technique at least once during their careers and may even have used it once or twice themselves. The terms dimensionality reduction techniques, however, may be much less familiar.

To clarify the meaning of these terms, we consider a simple example. Suppose that we have measured four variables (e.g., ionic concentration for four ions in precipitation) for n (> 4) samples. Those four variables would be generally correlated with each other, as illustrated in Figure 1.1. In that figure, $X_1$, $X_2$, $X_3$, and $X_4$ indicate the four observed variables and $\sigma_{12}$, $\sigma_{23}$, $\sigma_{34}$, $\sigma_{13}$, $\sigma_{24}$, and $\sigma_{14}$ are the correlation coefficients between those variables. In the techniques presented in this text, it is assumed that the relationships between the observed variables are due to latent factors that are not observed. They are indicated as $\mathcal{F}_1$ and $\mathcal{F}_2$ at the left of Figure 1.1. All the dimensionality reduction techniques presented in this overview try to estimate those latent factors. It is assumed in all these techniques that the number of latent factors is smaller than the number of observed variables. This explains the generic name given to those techniques. Another characteristic of the latent factors, that most of these techniques assume, is that they are independent of each other. This is illustrated in Figure 1.1 by the fact that no link exists between the two latent factors. However, this is not a necessary condition, and these techniques can be used even if one looks for correlated latent factors.

Figure 1.1 Schematic illustration of the relationship between observed and latent variables.
An example of possible latent factors in the context of the analysis of ionic concentrations in precipitation would be: a) a sea-salt factor which may influence Na⁺, Cl⁻, Mg²⁺, Ca²⁺ and also some contribution to SO₄²⁻ ionic concentration; and b) an anthropogenic factor which may influence SO₄²⁻, NH₄⁺, and NO₃⁻ ionic concentrations.

In this presentation, we will discuss the following dimensionality reduction techniques:

1. Principal Component Analysis (PCA)
2. Absolute Principal Component Analysis (APCA)
3. Target Principal Component Analysis (TPCA)
4. Subjective Principal Component Analysis (SPCA)
5. Factor Analysis (FA)
6. Positive Matrix Factorization (PMF)

That list represents only some of the techniques that have been developed in different fields of research in the last 100 years.

At this point, the reader is likely to ask:

- What exactly are these techniques?
- Why should I use them?

To try to answer those questions, it is necessary to give some background about the origin of these techniques and especially about the concepts behind them.

The concept of dimensionality reduction was initiated by researchers like Karl Pearson, Charles Spearman, and others in the early twentieth century in the course of trying to define and measure "intelligence". Because of this early association with notions such as intelligence, most of the techniques described here were created and developed primarily by scientists interested in psychometric measurements. This explains some of the terminology used by most of those techniques. It also explains the basic concepts that unify all those techniques.

When the first psychometric measurements were made at the beginning of the twentieth century, researchers accumulated large data sets that covered many persons and included measurements of several variables for each person. It became rapidly evident that some of those variables were correlated and others were not and that the variables could to some extent be clustered (using that characteristic of the correlation matrix) into a smaller number of groups. From that observation, it was postulated that some latent factors (or variables) that could not be observed directly were responsible for the observed correlations between the measured variables. It was also postulated that the number of those latent factors would be lower than the measured variables.

That concept may be written mathematically as follows:

\[ \mathbf{x} = \mu + \Lambda \mathbf{f} + \mathbf{e} \]  

(1.1)

where \( \mathbf{x} \) is a vector of \( p \) observable variables; \( \mu \) is a vector of \( p \) mean values; \( \Lambda \) is a \((p \times m)\) matrix call the loadings matrix; \( \mathbf{f} \) is a vector of \( m \) unobserved factors \((m < p)\); and \( \mathbf{e} \) is a
vector of p random errors. That model is the basic conceptual model for all the techniques described here, although in some cases the link between the technique and model 1.1 is not self evident. Not much could be done before the advent of computers because fitting such a model to data is very labor intensive.

The advent of computers in the 1950s and 1960s provided a practical means of making the necessary calculations and initiated the development of the many techniques described here.

It should now be evident to the reader that these techniques are all tools for finding possible structures between the measured variables. In the context of atmospheric chemistry, these structures would normally be related to physical or chemical processes and to emission source characteristics. Therefore, the use of these techniques may help in understanding these processes and characterizing the emission sources. The examples that will be given later should help the reader to see how these techniques can be useful to atmospheric chemistry researchers.

In the following sections, we will first try to give an idea of what these different techniques are and how they work. Then, we will give some examples of their uses.
2. Principal Component Analysis (PCA)

2.1 Introduction

Principal component analysis (PCA) is the oldest and best known of the dimensionality reduction techniques. It was first introduced by Karl Pearson (1901, 1904) and later developed independently by H. Hotelling (1933).

Principal component analysis is often confused with factor analysis. The confusion arises because many authors in the past and even presently use the term factor analysis to describe techniques that in reality are based on PCA. This is confusing and, in fact, misleading as the two techniques are theoretically and conceptually different, as we will see here. The confusion may be due to the fact that the conceptual equation of both techniques can be expressed by relations similar to 1.1. However, the similarity is only in the look of the equations and not in the concepts behind them. In the present text, we will separate the two techniques.

The purpose of principal component analysis is to find principal components (i.e., latent factors) in order to explain as much of the total variance in the measured data as possible with as few of the principal components as possible. Those principal components or latent factors are linear combinations of the observed variables. The principal components are extracted in order, such that the first component explains more variance than the second; the second more than the third, and so on.

Let us suppose that we observed variables $X_1, X_2, ..., X_p$; the first component ($P_1$) is written as:

$$P_1 = a_{(1)1}X_1 + a_{(1)2}X_2 + ... + a_{(1)p}X_p$$  \hspace{1cm} (2.1)

where the $a_{(1)1}, a_{(1)2}, ..., a_{(1)p}$ are the weights chosen to maximize the ratio of the variance of $P_1$ to the total variation, subject to the constraint that $\sum_{j=1}^{p} a_{(1)j}^2 = a_{(1)}^t a_{(1)} = 1$.

The second principal component ($P_2$) is the combination of the observed variables, which is uncorrelated with the first linear combination and which accounts for the maximum amount of the remaining total variance not already accounted for by $P_1$.

The $m^{th}$ principal component is the weighted linear combination of the observed variables, which has the largest variance of all the linear combinations that are uncorrelated with all of the previously extracted principal components.

The principal components have a geometric interpretation that can be more easily illustrated by a two-dimensional example. Figure 2.1a shows the observations for two observed variables that are correlated. The correlation coefficient between the two variables is equal to the cosine of the angle between the axes in Figure 2.1a. Before determining the two principal components, we first translate the coordinate system so that the mean of each variable is zero, giving Figure 2.1b. The two principal components are the new orthogonal coordinate system presented in Figure 2.1c. The first principal component is the one along which the data have the highest variance. Mathematically, the new coordinate system is related to the old one by the relation:
Figure 2.1 Illustration of principal components in 2 dimensions: (a) observed variables; (b) centered observed variables; (c) principal components.

\[ P = A^T(X - \mu) \]  
\[ X = \mu + AP \]
As one can easily see, this is similar to the general factor analysis relation (see relation 1.1) given earlier. Note that relation 2.3 is exact if all the p principal components are used and inexact only when r (\(<\) p) principal components are retained. We will return to that idea later.

Matrix A in relation 2.3 is called the matrix of loadings and gives the correlation between the observed variables and the new principal components. The matrix P is called the scores matrix and gives the coordinates of the observations in the new coordinate system.

When the principal components have been extracted, the next step is to select the number of components to retain. There is no infallible technique to determine the perfect number of components to retain.

Some of the most frequently used techniques are in fact only rules of thumb based mostly on trial and error arguments. Some more theoretically sound techniques have been developed, but they are only asymptotically exact and postulate that the observed variables follow a multivariate normal distribution. More recently, other types of techniques like cross-validation and partial correlation, which are very calculation intensive, have been put forward. All these techniques give slightly different results. They are briefly described later in this section. The basic principle to keep in mind, however, when selecting the number of components to retain is that they must be interpretable.

The final step in principal component analysis is usually to see if a rotation of the new coordinate system will not produce a more interpretable coordinate system. As we have mentioned earlier, the main criterion used in the extraction of the principal components is the percentage of the total variance explained. Any rotation of the coordinate system will not change this amount. Therefore, one can try to rotate the coordinate system to improve interpretability. Many different techniques of rotation have been developed to accomplish that. They can be grouped into two categories, namely: (1) orthogonal or rigid rotation and (2) oblique rotation. In the first type of rotation, the system of coordinates remains orthogonal; this is not the case in the second type of rotation. We will return to this point later in this section and also in Section 8.

Principal component analysis can thus be summarized by the following three steps:

1. Extract the components.
2. Retain only m (\(<\) p) components.
3. Rotate the principal components to improve interpretability.

2.2 Technical Details

We will now consider how to extract the principal components. We will first discuss the population theory of principal components and then the extraction of the principal components from observed data. Two techniques for the extraction of the principal components will be described here.

2.2.1 Principal Components in Population

Before discussing how the principal components are extracted for an observed data set, it is important to consider some theoretical aspects of principal component analysis.
Suppose that \( \mathbf{X} = [X_1, X_2, \ldots, X_p] \) is a \( p \)-dimensional random variable with mean \( \mu \) and covariance matrix \( \Sigma \). The first principal component is obtained by finding the weights \( a_{(1)}, a_{(2)}, \ldots, a_{(p)} \) of relation 2.1 such that:

\[
\text{var}(P_1) = \text{var}(a_{(1)}^T \mathbf{X}) = a_{(1)}^T \Sigma a_{(1)}
\]

is maximized with the condition: \( a_{(1)}^T a_{(1)} = 1 \).

The standard procedure to solve this type of problem is the Lagrange multipliers technique. In the present situation, the following function is to be maximized:

\[
L(a_{(1)}) = a_{(1)}^T \Sigma a_{(1)} - \lambda (a_{(1)}^T a_{(1)} - 1)
\]

We write therefore:

\[
\frac{\partial L}{\partial a_{(1)}} = 2 \Sigma a_{(1)} - 2 \lambda a_{(1)} = 0
\]

which can be written as:

\[
(\Sigma - \lambda I) a_{(1)} = 0
\]

where \( I \) is the \((p \times p)\) identity matrix. To have a solution for \( a_{(1)} \) other than the null vector, \((\Sigma - \lambda I)\) must be a singular matrix. Therefore, \( \lambda \) must be chosen such that:

\[
\det(\Sigma - \lambda I) = 0
\]

This relation is the same as the one to find the eigenvalues of a rectangular matrix. Generally the covariance matrix \( \Sigma \) will have \( p \) eigenvalues, which would all be nonnegative because \( \Sigma \) is positive semi-defined. If we denote the \( p \) eigenvalues of \( \Sigma \) as \( \lambda_1, \lambda_2, \ldots, \lambda_p \) and assume, for the present, that they are distinct, such that \( \lambda_1 > \lambda_2 > \ldots > \lambda_p \geq 0 \), the question is now: Which one do we choose to get the first component? We have:

\[
\text{var}(a_{(1)}^T \mathbf{X}) = a_{(1)}^T \Sigma a_{(1)}
\]

using relation 2.7 and the fact that \( a_{(1)}^T a_{(1)} = 1 \). Therefore, we choose the highest eigenvalue to get the highest variance for the first principal component. The first principal component is the eigenvector associated with the highest eigenvalue. It is interesting to note that relation 2.9 shows that the variance of a principal component is equal to the eigenvalue associated with it.

Now, if we consider the second principal component, we have to find \( P_2 = a_{(2)}^T \mathbf{X} \) such that \( a_{(2)}^T a_{(2)} = 1 \) and \( P_2 \) is uncorrelated to \( P_1 \). We have that:
\[
\text{cov}(P_2, P_1) = \text{cov}(a_{(2)}^T X, a_{(1)}^T X) \\
= E[a_{(2)}^T (X - \mu)(X - \mu)^T a_{(1)}] \\
= a_{(2)}^T \Sigma a_{(1)} \tag{2.10}
\]

This must be equal to zero. Since \( \Sigma a_{(1)} = \lambda_1 a_{(1)} \) (relation 2.7), \( \text{cov}(P_2, P_1) \) equals 0 only if \( a_{(2)}^T a_{(1)} = 0 \). Using the Lagrange multipliers, we have:

\[
L(a_{(2)}) = a_{(2)}^T \Sigma a_{(2)} - \lambda (a_{(2)}^T a_{(2)} - 1) - \delta a_{(2)}^T a_{(1)} \tag{2.11}
\]

To maximize \( L \), we must have:

\[
\frac{\partial L}{\partial a_{(2)}} = 2 \Sigma a_{(2)} - 2 \lambda I a_{(2)} - \delta a_{(1)} = 0 \tag{2.12}
\]

If we multiply relation 2.12 by \( a_{(1)}^T \), we obtain:

\[
2 a_{(1)}^T \Sigma a_{(2)} - \delta = 0 \tag{2.13}
\]

but from relation 2.10, we also require that \( a_{(2)}^T \Sigma a_{(2)} \) equals zero. Therefore, \( \delta \) must also be zero and relation 2.12 becomes:

\[
(\Sigma - \lambda I)a_{2} = 0 \tag{2.14}
\]

which is similar to relation 2.7. Therefore, we have to chose \( \lambda \) to be the second eigenvalue of \( \Sigma \) and the second principal component to be the second eigenvector. The other principal components are similarly constructed.

There is no difficulty in extending this theory to the case where some of the eigenvalues of \( \Sigma \) are equal. In this case, there is no unique way of choosing the eigenvectors associated with multiple roots, but as long as they are chosen to be orthogonal, then the theory presented earlier is still valid.

In summary, the p principal components for the population are the p eigenvectors of \( \Sigma \) ordered in decreasing order of the eigenvalues. This gives us a technique to extract the principal components when n samples of p variables are available, as we will see in the next section. An important character of the eigenvalues of \( \Sigma \) is obtained as follows. If the covariance matrix of the principal component is denoted by \( \Lambda \), we have:

\[
\Lambda_{(p \times p)} = \begin{pmatrix}
\lambda_1 & 0 & \cdots & 0 \\
0 & \lambda_2 & \cdots & 0 \\
\vdots & \vdots & \ddots & \vdots \\
0 & 0 & \cdots & \lambda_p
\end{pmatrix} \tag{2.15}
\]

Using the fact that:

\[
\text{var}(P) = A^T \Sigma A \tag{2.16}
\]
we have:

$$\Lambda = A^T \Sigma A$$  \hspace{1cm} (2.17)

Therefore, we can write:

$$\Sigma = A \Lambda A^T$$  \hspace{1cm} (2.18)

since $A$ is an orthogonal matrix with $A^T A = A A^T = I$.

We have seen earlier (relation 2.9) that the eigenvalues can be interpreted as the respective variances of the different principal components. The sum of these variances is given by:

$$\sum_{i=1}^{p} \text{var}(P_i) = \sum_{i=1}^{p} \lambda_i = \text{trace}(\Lambda)$$  \hspace{1cm} (2.19)

But, using relation 2.17, we can write:

$$\text{trace}(\Lambda) = \text{trace}(A^T \Sigma A) = \text{trace}(\Sigma A A^T)$$ since $\text{trace}(AB) = \text{trace}(BA)$.

$$= \text{trace}(\Sigma) = \sum_{i=1}^{p} \text{var}(X_i)$$  \hspace{1cm} (2.20)

Thus, we arrive to the important result that the sums of the variances of the original variables and of their principal components are the same. It is therefore convenient to make statements such as "the $ith$ principal component accounts for a proportion $\lambda_i / \sum_{j=1}^{p} \lambda_j$ of the total variance in the original data".

2.2.2 Principal Components for Observed Data

Suppose that we have n samples for p variables. We can write the basic (n x p) data matrix as:

$$X = \begin{pmatrix}
X_{11} & X_{12} & \cdots & X_{1p} \\
X_{21} & X_{22} & \cdots & X_{2p} \\
\vdots & \vdots & \ddots & \vdots \\
X_{n1} & X_{n2} & \cdots & X_{np}
\end{pmatrix}$$  \hspace{1cm} (2.21)

where $X_{ij}$ is the value of variable $j$ obtained for sample $i$.

When the matrix $X$ is used, relation 2.2 becomes:

$$P = (X - M)A$$  \hspace{1cm} (2.22)

where $M$ is the mean matrix given by:
where $\bar{X}_j = (1/n) \sum_{i=1}^{n} X_{ij}$ is the mean for variable $j$. A, called the matrix of standardized loadings, is a $(p \times p)$ matrix such that $A^T A = I$. P, called the scores matrix, is a $(n \times p)$ matrix such that $P^T P$ is a diagonal matrix.

Relation 2.3 becomes:

$$X = M + PA^T$$

(2.24)

It remains now to obtain matrices P and A.

We have seen in the preceding section that the principal components can be obtained by finding the eigenvectors of the population variance-covariance matrix $\Sigma$. Normally that matrix is unknown. However, $\Sigma$ can be estimated using the data matrix X (relation 2.21). An estimate $S$ of $\Sigma$ is equal to:

$$S = \frac{1}{n-1} (X - M)^T (X - M)$$

(2.25)

The theory presented in the preceding section is applicable, and the eigenvalues are obtained in solving the following relation:

$$\text{det}(S - \ell I) = 0$$

(2.26)

for the $p$ eigenvalues if the matrix $S$ is non-singular or for the $r$ ($< p$) non-zero values of the estimated eigenvalues $\ell$. The eigenvector $a_{(i)}$ is found by solving the system of relation:

$$(S - \ell I) a_{(i)} = 0$$

(2.27)

The first principal component is the eigenvector associated with the highest eigenvalue, the second principal component is that associated with the second highest eigenvalue, and so on.

The loadings matrix $A$ of relation 2.24 is equal to:

$$A = \begin{pmatrix}
a_{(1)1} & a_{(1)2} & \cdots & a_{(1)p} \\
a_{(2)1} & a_{(2)2} & \cdots & a_{(2)p} \\
\vdots & \vdots & \ddots & \vdots \\
a_{(p)1} & a_{(p)2} & \cdots & a_{(p)p}
\end{pmatrix}$$

(2.28)

The scores matrix $P$ of relation 2.24 is given by relation 2.22. The scores matrix gives the projection of the data on the principal coordinate system.

An interesting characteristic is that the sign of the vectors $a_{(i)}$ is arbitrary. Therefore, any of the columns of $A$ can be multiplied by -1 without changing the solution. However, it should be noted that the sign of the columns of matrix $A$ influences the sign of the columns of
matrix $P$, as can be seen in relation 2.22. Therefore, any change of sign of any column of $A$ should be reflected by a similar change in $P$.

Note that in the case of principal components from sampled data, as opposed to data from a population, strict equality of the sample eigenvalues is precluded (see Basilevsky, 1994).

2.2.3 Principal Components Using Singular Value Decomposition

The theory presented in the preceding sections can be used to extract principal components. However, for some data matrices, the calculation may be unstable. A better technique for extracting principal components uses singular value decomposition (SVD), introduced by Fisher and Mackenzie in 1923 in the context of ANOVA (analysis of variance) theory. SVD can be described as follows.

Given an arbitrary matrix $X$, of dimension $(n \times p)$, which in the present context will be a matrix of $n$ observations on $p$ variables, $X-M$ can be written as:

$$X-M=ULV^T$$

where:

(i) $U$, $V$ are $(n \times r)$, $(p \times r)$ matrices respectively, each of which has orthonormal columns such that $U^TU=I$ and $V^TV=I$;

(ii) $L$ is a $(r \times r)$ diagonal matrix;

(iii) $r$ is the rank of $X$. $r$ is equal to $p$ if the eigenvalues of $X^TX$ are all greater than 0.

As can easily be seen, relation 2.29 is the same as relation 2.24 if we write:

$$A=V$$

and

$$P=UL$$

The eigenvalue $\lambda_i$ is equal to $\lambda_i^2/(n-1)$ (see Chambers, 1977).

2.2.4 Correlation Matrix

Up to now, we have only used the variance-covariance matrix to extract the principal components. However, this can create some problems if there are large differences in the variance of the variables. Such a situation can result in a solution that would be dominated by only some of the variables. A solution to that problem is the standardization of the variables before the analysis, which results in replacing the variance-covariance matrix by the correlation matrix in relations 2.26 and 2.27.

The standardized matrix of data ($X^*$) is equal to:

$$X^* = (X - M)D^{-1/2}$$

where $D$ is a $(p \times p)$ diagonal matrix equal to:
where the $S_{ii}$ are the diagonal elements of the variance-covariance matrix defined in relation 2.25. The correlation matrix, $R$, is equal to:

$$
R = \frac{1}{n-1} X^* X^* \\
= \frac{1}{n-1} ((X - M)D^{-1/2})^T (X - M)D^{-1/2} \\
= \frac{1}{n-1} D^{-1/2} (X - M)^T (X - M)D^{-1/2} \\
= D^{-1/2} S D^{-1/2} 
$$

(2.33)

Relations 2.26 and 2.27 for the correlation matrix become, respectively:

$$
\text{det}(R - \ell I) = 0 
$$

(2.34)

and

$$
(R - \ell_{(i)} I) a_{(i)} = 0 
$$

(2.35)

The scores matrix $P$ is now equal to:

$$
P = X^* A = (X - M)D^{-1/2} A 
$$

(2.36)

and relation 2.24 becomes:

$$
X = M + PA^T D^{1/2} 
$$

(2.37)

Singular value decomposition can still be used by replacing $(X - M)$ by $X^*$ in 2.29.

An interesting characteristic of the eigenvalues obtained using 2.34 is that:

$$
\sum_{i=1}^p \ell_{(i)} = p 
$$

(2.38)

It is important to note that the principal components obtained using the correlation matrix $R$ are different from those obtained using the variance-covariance matrix $S$. The same is true for the latent roots (or eigenvalues). It is not possible to jump from one solution to the other because principal components are not scale invariant. The choice of which one to use is mainly determined by the nature of the problem under consideration.

2.2.5 Robust Principal Components

We have seen earlier that principal components are obtained using either the variance-covariance matrix or the correlation matrix. These matrices are obtained using the Euclidian distances between the variables. These distances are heavily influenced by outlier data. Therefore, one would expect that the principal components would also be influenced by the
presence of outliers. Outliers would influence both the latent roots (or eigenvalues) and the principal components. Faced with the possibility of outliers in the data, we can choose between two courses of action.

The first strategy is to try to identify the outliers and, in particular, to estimate their influences on the results of the PCA. This can be done by using the influence function derived in Critchley (1985). This technique permits one to verify the presence of outliers and to identify them.

The second strategy is to "robustify" the method used to extract the principal components. This can be done in many ways. The first method would be to use robust estimation of the variance-covariance or correlation matrix in relations 2.26 and 2.27 or 2.34 and 2.35. Devlin et al. (1981) present five robust estimators based on three different approaches. Campbell (1980) proposed using M-estimators to estimate the variance-covariance and correlation matrices (see also Matthews, 1984).

A different approach to the estimation of the principal components is discussed by Gabriel and Odoroff (1983a, 1983b). Their technique relies on the fact that, as we have seen earlier, the principal components can be obtained using singular value decomposition. To find the SVD, a set of equations, involving weighted means of a function of the element of the observation matrix X, are solved iteratively. To obtain robust estimation of the principal components, one replaces the weighted means by robust estimators of locations as the medians or weighted trimmed means.

Another approach, proposed by Li and Chen (1985), uses the statistical technique called projection pursuit. In that technique, the principal components are estimated directly without first finding robustly estimated variance-covariance or correlation matrices.

The reader is referred to the given references for more details about those techniques.

2.2.6 Missing Data

Another problem that occurs frequently in atmospheric chemistry, is the absence of data for some of the variables in some samples. It is therefore important that the problem of how to handle missing data should be discussed. We will present only an overview of the possible techniques here. The reader can find more details in the references cited.

The easiest way to deal with missing data is to eliminate all samples with one or more missing data items. This is generally an acceptable solution when the number of samples so eliminated is small compared with the total data set and especially if the elimination process can be considered random, that is, if it does not create a pattern in the data set like the elimination of every tenth sample. That technique can be wasteful of information in some situations, and other techniques should then be employed.

Another option is to compute the (i,j)th variance-covariance or correlation using all observations for which values of both \( x_i \) and \( x_j \) are available. Unfortunately, that technique cannot be recommended, as it leads to variance-covariance or correlation matrices that are not necessarily positive semidefinite and PCs cannot be calculated.

Another option is to replace the missing values for variable \( x_i \) by the mean value, \( \bar{x}_i \), of the observations for which the value of \( x_i \) is available. Beale and Little (1975) noted that this approach is fairly common and has produced satisfactory results.
Another technique is based on the hypothesis of multivariate normality. The variance-covariances and the correlations are estimated under this assumption using the method of maximum likelihood (see Anderson, 1957; Beale and Little, 1975). de Ligny et al. (1981) presented an iterative algorithm which, they claim, will produce maximum likelihood estimates.

Wiberg (1976) proposed a method based on the singular matrix decomposition, which gives a least squares approximation of rank m to the data matrix X. The same technique is implicitly suggested by Gabriel and Zamir (1979). This method has the advantage that it does not rely on a hypothesis about the distribution of the data.

Finally, Frane (1976) has suggested estimating missing values for a particular observation by means of regression analyses to derive the missing variables from the variables that are present for the given observation.

2.2.7 Confidence Intervals and Tests

If we can assume that the samples constituting the observation matrix X are independent and from a multivariate normal distribution, it is possible to derive relations giving asymptotic confidence intervals for the eigenvalues and loadings in the case of principal components extracted from the variance-covariance matrix. Some approximate tests are also available in this case. In the case of principal components extracted from the correlation matrix, the relations are tractable only in very special cases. Therefore, all results presented in this section are for the principal components extracted from the variance-covariance matrix, unless otherwise indicated. We will present two other techniques to estimate confidence intervals in Section 9.

2.2.7.1 Confidence Intervals for the Eigenvalues or Latent Roots

A two-sided normal 100(1-α) percentage confidence interval for the population eigenvalues or latent roots, \( \lambda_i \), is given by:

\[
\frac{\ell^{(i)}}{1 + \sqrt{2/\nu - 1} Z_{\alpha/2}} \leq \lambda_i \leq \frac{\ell^{(i)}}{1 - \sqrt{2/\nu - 1} Z_{\alpha/2}}
\]

(2.39)

where \( Z_{\alpha/2} \) is the upper 100(1-α/2)th percentile of the standard normal distribution. An alternative approximate confidence interval is obtained by looking at the distribution of \( \ln(\ell^{(i)}) \). An approximate 100(1-α) percentage confidence interval for \( \ln(\lambda_i) \) is given by Jolliffe (1986):

\[
\ln(\ell^{(i)}) - \sqrt{2/\nu - 1} Z_{\alpha/2} \leq \ln(\lambda_i) \leq \ln(\ell^{(i)}) + \sqrt{2/\nu - 1} Z_{\alpha/2}
\]

or for \( \lambda_i \):

\[
\ell^{(i)} e^{-\sqrt{2/\nu - 1} Z_{\alpha/2}} \leq \lambda_i \leq \ell^{(i)} e^{\sqrt{2/\nu - 1} Z_{\alpha/2}}
\]

(2.40)
Bonferroni-type simultaneous 100(1-\(\alpha\))% confidence intervals (see Miller, 1981 for Bonferroni’s simultaneous confidence intervals) for \(m\lambda_i\)'s are obtained by replacing \(Z_{a/2}\) with \(Z_{a/2m}\).

If the last \(p-r\) eigenvalues are equal (see later for test of equality), a 100(1-\(\alpha\))% confidence interval for the common value is:

\[
\frac{\bar{\ell}_q}{1 + \sqrt{q(n-1)} Z_{a/2}} \leq \lambda_i \leq \frac{\bar{\ell}_q}{1 - \sqrt{q(n-1)} Z_{a/2}}
\]

(2.42)

where \(q = p - r\) and

\[
\bar{\ell}_q = \frac{1}{q} \sum_{i=r+1}^{p} \ell_{(i)}
\]

(2.43)

Note that there also seems to be some evidences that those results are asymptotically valid for non-normal populations (Davis, 1977).

2.2.7.2 Confidence Intervals for the Elements of the Loadings Matrix

If \(a_{ij}\) is the \((i,j)\)th element of the standardized loadings matrix \(A\), then we have:

\[
\text{var}(a_{ij}) = \gamma_{ij}^2 = \frac{\ell_{(i)}}{(n-1)} \sum_{s=1 \atop s \neq i}^{p} \frac{\ell_{(s)}}{(\ell_{(s)} - \ell_{(i)})^2} a_{ij}^2
\]

(2.44)

\[
\text{cov}(a_{gi}, a_{hi}) = \frac{\ell_{(i)}}{(n-1)} \sum_{s=1 \atop s \neq i}^{p} \frac{\ell_{(s)}}{(\ell_{(s)} - \ell_{(i)})^2} a_{gi} a_{hi}
\]

(2.45)

and

\[
\text{cov}(a_{gi}, a_{hj}) = \frac{-\ell_{(i)} \ell_{(j)}}{(n-1) (\ell_{(i)} - \ell_{(j)})^2} a_{gi} a_{hj}
\]

(2.46)

for \(i \neq j\). This last relation indicates that the elements of the loadings matrix are correlated between different components even though the different components are orthogonal.

If the distribution of the observed variables is multivariate normal, an asymptotic 100(1-\(\alpha\))% confidence interval for the population element, \(a_{ij}\), of the loadings matrix is:

\[
a_{ij} - \gamma_{ij}Z_{a/2} \leq a_{ij} \leq a_{ij} + \gamma_{ij}Z_{a/2}
\]

(2.47)

where \(\gamma_{ij}\) is given by relation 2.44. To obtain confidence for more than one element of the loadings matrix, Bonferroni’s technique (Miller, 1981) can be used.

2.2.7.3 Testing Equality of All Eigenvalues or Latent Roots
Here we want to test:

\[ H_0 : \lambda_1 = \lambda_2 = \cdots = \lambda_p \]

\[ H_1 : \text{not all equal} \quad (2.48) \]

This is equivalent to testing that the observed variables are independent with equal variance. \( H_0 \) is rejected at a 100(1-\( \alpha \))% confidence level if

\[ \chi^2 = -\left[ n - \frac{1}{6p} (2p^2 + p + 2) \left( \sum_{i=1}^{p} \ell_i(1) \right) + p \ln \left( \frac{1}{p} \sum_{i=1}^{p} \ell_i(1) \right) \right] \quad (2.49) \]

is greater than the 100(1-\( \alpha \))% point of the chi-squared distribution with \( \frac{1}{2} (p + 2)(p - 1) \) degrees of freedom.

Note that when using the correlation matrix, an equivalent test is possible with

\[ \chi^2 = -\left[ n - \frac{1}{6} (2p + 5) \right] \ln \left( \prod_{i=1}^{p} \ell_i(1) \right) \quad (2.50) \]

This is equivalent to test for the independence of the observed variables but not on the equality of their variance.

2.2.7.4 Testing Equality of the Smallest Eigenvalues or Latent Roots

In this section, we want to test the hypothesis that the smallest (p-r) eigenvalues or latent roots are equal. Therefore, we can write:

\[ H_0 : \lambda_{r+1} = \lambda_{r+2} = \cdots = \lambda_p \]

\[ H_1 : \text{not all (p-r) eigenvalues are equal} \quad (2.51) \]

\( H_0 \) can be rejected at a 100(1-\( \alpha \))% confidence level if

\[ \chi^2 = -\left[ n - r - \frac{1}{6q} (2q^2 + q + 2) \left( \sum_{i=1}^{q} \ell_i(1) \right) - q \ln \left( \frac{1}{q} \sum_{i=1}^{q} \ell_i(1) \right) \right] \quad (2.52) \]

where \( q = p - r \), is greater than the 100(1-\( \alpha \))% point of the chi-squared distribution with \( \frac{q}{2} (q + 1) - 1 \) degrees of freedom.

Lawley (1956) has shown that the chi-squared approximation is improved slightly if the following equation for \( \chi^2 \) is used:

\[ \chi^2 = -\left[ n - r - \frac{1}{6q} (2q^2 + q + 2) + \sum_{i=1}^{q} \left( \frac{\ell_i(q)}{q} \right)^2 \left( \sum_{i=1}^{q} \ell_i(1) - q \ln \left( \frac{1}{q} \sum_{i=1}^{q} \ell_i(1) \right) \right) \right] \quad (2.53) \]

where:

\[ \ell_i(q) = \frac{1}{q} \sum_{i=r+1}^{p} \ell_i(1) \quad (2.54) \]
Note that the test using this relation may be conservative (James, 1969). When the correlation matrix is used, $\chi^2$ does not possess an asymptotic chi-squared distribution. However the test is sometimes employed when the last $p - r$ eigenvalues account for a small percentage of the total variance and when $n$ (the number of observations) is large.

2.2.7.5 Confidence Region for the Ratio of the Sum of the Smallest Eigenvalues to the Total Sum of the Eigenvalues

Let

$$\Theta = \frac{\sum_{i=r+1}^{p} \lambda_i}{\sum_{i=1}^{p} \lambda_i} \quad (2.55)$$

be the ratio of the last $(p-r)$ eigenvalues or latent roots. An approximate $100(1-\alpha)\%$ confidence range for $\Theta$ (Anderson, 1984) is:

$$0 < \Theta < \frac{\sum_{i=r+1}^{p} \ell_{(i)} + Z_{2\alpha}}{\sum_{i=1}^{p} \ell_{(i)}} \left[ 2\left( \sum_{i=r+1}^{p} \ell_{(i)} \right) + \sum_{i=1}^{p} \ell_{(i)}^2 \right]^{1/2} \sqrt{n\left( \sum_{i=1}^{p} \ell_{(i)}^2 \right)} \quad (2.56)$$

where $Z_{2\alpha}$ is the upper $100(1-2\alpha)\text{th}$ percentile of the standard normal distribution.

2.2.8 Reducing the Number of Principal Components

Up to now, the number of principal components was equal to the number of observed variables. We will now consider what happens when $m$ ($< p$) principal components are used to describe the data.

When retaining only the first $m$ components, relation 2.24 for the principal components extracted from the variance-covariance matrix becomes:

$$\hat{X} = \hat{M} + \hat{P} \hat{A}^T$$

where $\hat{P}$ is the $(n \times m)$ matrix obtained by retaining only the first $m$ columns of the scores matrix $P$; $\hat{A}$ is the $(p \times m)$ matrix obtained by retaining the first $m$ columns of the loadings matrix $A$. If $m = p$, $\hat{X}$ will be equal to the original matrix $X$; otherwise $\hat{X}$ would differ from $X$ by:

$$E = X - \hat{X} \quad (2.58)$$

Therefore, $X$ can be written as:

$$X = \hat{M} + \hat{P} \hat{A}^T + E \quad (2.59)$$

Although this relation is similar to relation 1.1, there is a fundamental difference between the two. On the one hand, $E$ is a vector of random variables in relation 1.1. On the other hand, $E$ is not random but depends on $m$. But the similarity between the two relations justifies the use
of principal components as a dimensionality reduction technique. We will discuss how to select \( m \) in Section 2.2.10. The main goal in selecting \( m \) is to have it as small as possible while keeping the differences between \( m \hat{X} \) and \( X \) as small as possible.

For principal components extracted using the correlation matrix, we can write:

\[
\hat{X}^* = m \hat{X} = m \sum_{m} A^T
\]

where the matrices \( P \) and \( A \) are obtained using relations 2.34 to 2.36. Relation 2.37 becomes:

\[
\hat{X} = M + m \sum_{m} A^T D^{1/2}
\]

### 2.2.9 Communality

The communality of an observed variable is a concept that was first defined in the context of factor analysis (see Section 6). It is defined as the part of the variance explained by the retained principal components or common factors in the terminology of factor analysis. Using the notation used here, the communality \( h_i^2 \) for observed variable \( i \) is given for principal components extracted from the variance-covariance matrix by:

\[
h_i^2 = \frac{\sum_{m} \ell_{(j) m} a^2_{ij}}{S_{ii}}
\]

where \( a_{ij} \) is the \((i,j)\)th element of the standardized loading matrix \( mA \). \( S_{ii} \) is the \( i \)th diagonal element of the variance-covariance matrix (relation 2.25) and for principal components extracted from the correlation matrix:

\[
h_i^2 = \sum_{j=1}^{m} \ell_{(j) m} a^2_{ij}
\]

By definition \( h_i^2 \) varies between 0 and 1. Note that when \( m = p \), all the communalities are equal to 1. The communalities can be useful in determining how many components to retain.

### 2.2.10 Determining the Number of Components to Retain

Unfortunately, there is no universally accepted method for selecting the number of principal components to retain. In the end, the decision is largely judgmental and a matter of taste. We will present in what follows some of the most often used rules. Some of these rules apply only to analyses with the variance-covariance matrix, while others apply only to analyses with the correlation matrix. These will be indicated specifically. Otherwise, the techniques can be used in either cases.

#### 2.2.10.1 Informal Rule of Thumb Techniques

In this section, we will present rules that are based on experience rather than statistical theory. These are the techniques that are normally tried first, and they very often give satisfactory results. We will summarize them here in no particular order.
Table 2.1 Values of $\ell^*_k$ for $p$ between 2 and 10.

<table>
<thead>
<tr>
<th>$p$</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>0.750</td>
<td>0.250</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>0.611</td>
<td>0.278</td>
<td>0.111</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>0.521</td>
<td>0.271</td>
<td>0.146</td>
<td>0.063</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>0.457</td>
<td>0.257</td>
<td>0.157</td>
<td>0.090</td>
<td>0.040</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>0.408</td>
<td>0.242</td>
<td>0.158</td>
<td>0.103</td>
<td>0.061</td>
<td>0.028</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>0.370</td>
<td>0.228</td>
<td>0.156</td>
<td>0.109</td>
<td>0.073</td>
<td>0.044</td>
<td>0.020</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>0.340</td>
<td>0.215</td>
<td>0.152</td>
<td>0.111</td>
<td>0.079</td>
<td>0.054</td>
<td>0.033</td>
<td>0.016</td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>0.314</td>
<td>0.203</td>
<td>0.148</td>
<td>0.111</td>
<td>0.083</td>
<td>0.061</td>
<td>0.042</td>
<td>0.026</td>
<td>0.012</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>0.293</td>
<td>0.193</td>
<td>0.143</td>
<td>0.110</td>
<td>0.085</td>
<td>0.065</td>
<td>0.048</td>
<td>0.034</td>
<td>0.021</td>
<td>0.010</td>
</tr>
</tbody>
</table>

The first approach is the percentage-of-variance-criterion. In this technique, we first calculate the cumulative percentage of the total variance explained by the first $k$ principal components given by:

$$
\tau_k = \frac{\sum_{i=1}^{k} \ell_i}{\sum_{i=1}^{p} \ell_i}
$$

where $\ell_i$ is the $i^{th}$ eigenvalue or latent root. The first $m$ ($< p$) principal components such that $\tau_m < \tau^*$, where $\tau^*$ is some a priori specified limit (for example between 0.7 and 0.9), are retained. The problem with this technique is that there is no rule for the selection of the cut-off limit, $\tau^*$. The best value for $\tau^*$ will generally become smaller as $p$ increases, or as $n$, the number of observations, increases. Some attempts (e.g., Mandel, 1972; Krzanowski, 1979; and Sugiyama and Tong, 1976) have been made to find the distribution of $\tau_k$ but it is not clear how this information can help in the selection of the number of principal components to retain.

The preceding technique considers the total variance in the data set on the basis of how good the fit would be if only $m$ components were retained. In some circumstances, however, it may be more important to reproduce the variance of some observed variables than others. One way of taking that requirement into consideration is to consider the variation of the communality, $h^2_i$, of each observed variable as the number of components retained is decreased. This approach may help in determining the minimum number of principal components to retain.

The next approach is Kaiser's rule (Kaiser, 1958, 1960). In this rule, one keeps all the principal components having an eigenvalue, $\ell$, greater that the mean of the eigenvalues, $\bar{\ell} = \frac{1}{p} \sum_{i=1}^{p} \ell_i$. Note that if the principal components are from a correlation matrix, $\bar{\ell} = 1$. It has been argued that a cut-off equal to $\bar{\ell}$ retains too few principal components and that a smaller limit should be used. Jolliffe (1972) has suggested, based on simulation studies, that a cut-off limit equal to 0.7 $\bar{\ell}$ should be the correct level.
Another technique is the so-called broken stick model. If we have a stick of unit length, which is broken, at random, into \( p \) segments, then it can be shown that the expected length of the \( k \)th longest segment is:

\[
E_k = \frac{1}{p} \sum_{j=k}^{p} \frac{1}{j}
\]

(2.65)

One way of deciding whether the proportion of variance accounted for by the \( k \)th PC is large enough for the component to be retained is to compare the proportion with \( \ell_k^* \) and retain only the PC with the proportion exceeding that limit. Values for \( \ell_k^* \) for \( p \) between 2 and 10 are given in Table 2.1.

A slightly more complex technique is the "scree test" or "scree graph" of Cattell (1966). The method consists of plotting the eigenvalues (or latent roots) against their rank numbers and observing whether, at some point (eigenvalue), the slope becomes "markedly" less steep (i.e., the eigenvalues tend to be isotropic). In practice, that elbow in the curve is identified by applying, say, a straightedge to the bottom portion of the eigenvalues to see where they form an approximately straight line. The number of components to retain is given by the point at which the components curve above the straight line formed by the smaller eigenvalues. This is illustrated in Figure 2.2. We will discuss that figure later.

This technique is \emph{a priori} relatively simple. However complications can occur. First, there may be no obvious break in the slope, in which case the test is inconclusive. Second, there may be several breaks. In this case, it is sometimes difficult to decide which of the breaks reflects the "correct" number of principal components to retain. Although artificial simulations based on "random" data do indicate that these plots can be useful guides for the "correct" number of components to retain, real data appear to give less clear-cut results (Jolliffe, 1986; Farmer, 1971).

A popular alternative to the scree graph in meteorology is to plot \( \log(\ell_i) \), rather than \( 1/\ell_i \), against \( i \). This is known as the log-eigenvalue (or LEV) diagram. The same technique that is applied for the scree test is then used to try to estimate the number of components to retain.

2.2.10.2 Testing for the Statistical Significance of the Eigenvalues

As we have seen earlier, it is possible to estimate, in some cases, confidence intervals or tests for the eigenvalues if we assume that the data are from a multivariate normal distribution.

For the principal components extracted from the variance-covariance matrix, the null hypothesis that the last \( r \) eigenvalues are equal can be tested against the hypothesis that at least two are different (Section 2.2.7.4). Therefore, by using this test for different values of \( r \), it should be possible to determine the number of principal components to retain. The tests are done sequentially, starting with \( r=2 \), and then the testing is continued until a significant result has been found at \( r=r^* \). The number of components to retain is then equal to \( r^* + 1 \).

One of the disadvantages of this technique is that many tests have to be made, so the overall significance level of the sequence of tests is not the same as the individual significance levels of each test. Furthermore, it is difficult to get an approximate idea of the overall significance level because the number of tests done is random and not fixed, and
the tests are not independent of each other. It should be noted that in nearly all real examples, the method tends to retain more PCs than are really necessary. The same test can also be used when the principal components are extracted from a correlation matrix when the $r$ eigenvalues represent a small percentage of the total variance and the number of observations is large, although the chi-squared distribution used in the test is not the correct distribution in that case. As in the case of the variance-covariance matrix, the test is very conservative and indicates a larger number of components to retain than is really necessary.

Another possible technique is to select a value of $r$ such that the upper limit of the confidence range of the ratio of the sum of the last $(p-r)$ eigenvalues to the sum of all $p$ eigenvalues would be less than a specified value (see Section 2.2.7.5). This technique has the same disadvantage as the preceding one in that many ranges have to be estimated sequentially.

Because the tests presented in this section assume that the data are from a normal distribution and that they are conservative, they are not recommended. They can, however, be used as ad hoc estimators of the number of principal components to retain in the same spirit as the techniques presented in the preceding section.

### Table 2.2 Values of Exner function $v_m$

<table>
<thead>
<tr>
<th>$v_m$</th>
<th>Quality of Fit</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.02</td>
<td>very good agreement</td>
</tr>
<tr>
<td>0.10</td>
<td>good</td>
</tr>
<tr>
<td>0.20</td>
<td>fair</td>
</tr>
<tr>
<td>0.50</td>
<td>upper limit for acceptability</td>
</tr>
<tr>
<td>$&gt;0.5$</td>
<td>unacceptable</td>
</tr>
</tbody>
</table>

2.2.10.3 Exner Function

One possible approach to selecting the number of principal components to retain is to choose a value such that the difference between $X$ and $\mu\hat{X}$ is, in the mean, small. In other words, we want to select $m$ in order to reproduce as much as possible of the original data. The problem is to find a criterion for deciding when the reproduction of the data is good. For principal components from a correlation matrix, Exner (1966) has proposed using the following parameter, $v_m$, to select the number of components to retain:

$$v_m = \left( \frac{\sum_{i=1}^{n} \sum_{j=1}^{p} (x^*_{ij} - \hat{x}^*_{ij})^2}{\sum_{i=1}^{n} \sum_{j=1}^{p} (x^*_{ij} - \bar{x}^*)^2} \right)^{1/2} \frac{m n}{(m n - p)}$$

(2.66)

where $x^*_{ij}$ and $\hat{x}^*_{ij}$ are the elements of the standardized matrices $X^*$ and $\mu\hat{X}^*$ respectively. $\bar{x}^*$ is the grand mean of all the $x^*_{ij}$. Exner (1966) described how the t-test can be used to determine the value of $m$ to employ. Table 2.2 gives a rough guide for acceptable values of $v_m$. Therefore, $m$ should be selected to obtain values of $v_m$ less than about 0.1.
2.2.10.4 Cross Validation

Cross-validation is a resampling scheme, like jackknife and bootstrap techniques (see Section 9). It is based on the idea that a set of data can be subdivided into groups, with the model estimated in one group and then evaluated in the remaining group(s) to verify goodness of fit, forecasting properties, and so forth. The basic procedure for cross-validation is as follows:

1. Parts of the data are deleted.
2. The model is estimated using the remaining data.
3. The deleted data are predicted using the model and compared with the original data.
4. The model that provides the best prediction of those data is selected.

In the case of PCA, cross validation is applied as follows. The rows of the data matrix \( X \) are divided into \( G \) groups. The first group is omitted from the data and the PCs are estimated using the reduced data set. That model is used (relation 2.57) to predict the data in the deleted group (i.e., \( m \hat{X} \)). Then the predicted sum of squares

\[
PRESS(m) = \frac{1}{Np} \sum_{i=1}^{N} \sum_{j=1}^{p} (x_{ij} - m \hat{x}_{ij})^2
\]

is calculated. \( N \) is the number of data components in the omitted group, \( m \) is the number of components retained, and \( x_{ij} \) and \( m \hat{x}_{ij} \) are the elements of \( X \) and \( m \hat{X} \) respectively. The process is repeated for each of the \( G \) omitted groups, and the total predictive sum of squares is then computed as

\[
TPRESS(m) = \sum_{g=1}^{G} PRESS_g(m)
\]

for \( m \) PCs. The number of groups to use is between 4 and 7 (Wold, 1978). Two different cross-validation approaches, that differ in the ways that they estimate \( m \hat{x}_{ij} \) for the omitted data and in the function of \( TPRESS(m) \) that they use as a criterion for choosing \( m \), have been proposed by Wold (1978) and Eastment and Krzanowski (1982). The reader is referred to those papers for more details. One should, however, note that these techniques are very calculation intensive.

2.2.10.5 Partial Correlation

Another technique for selecting nontrivial PCs has been proposed by Velicer (1976). It is based on the partial (residual) correlation after the extraction of \( r < p \) components. We have seen that after the extraction of \( r \) components, one can write:

\[
X = M + P \cdot A^T + E
\]

where \( E \) is the \((n \times p)\) residual matrix after \( r \) \((<p)\) PCs have extracted. Hence:

\[
E = X - M - P \cdot A^T
\]

A similar relation exists for the principal components extracted from the correlation matrix. Let
\[ R^* = D^{-1/2} E^T E D^{-1/2} \]  

(2.71)

be the \((p \times p)\) matrix of partial correlations, where \(D = \text{diag}(E^T E)\). Let \(r_{ij}^*\) represent the off-diagonal element of \(R^*\) and let

\[ f_r = \sum_{i \neq j} \frac{(r_{ij}^*)^2}{p(p-1)} \]  

(2.72)

which is between 0 and 1. Velicer (1976) proposes to accept components up to and including those that correspond to the minimum of \(f_r\), since small values of relation 2.72 indicate that the \(r\) retained PCs are uniformly correlated with most of the variables and in this sense capture nonresidual variation. Velicer also concluded that, for known data in psychology, using this procedure results in a smaller number of PCs than would be retained by the commonly used rules of thumb.

2.2.11 Rotations

The reduced loadings matrix (i.e., the one that is left after keeping only the first \(m\) principal components) is used to interpret the factor. However if the loadings of the principal components on the observed variables do not vary much, the interpretation of the factors becomes difficult. One solution to that problem is to rotate the components to simplify their interpretation. As will be mentioned later (Section 8), there are two types of rotation, namely orthogonal and oblique. In the former, the orthogonal principal components are still orthogonal after rotation; this is not the case for the latter type of rotation.

Thurstone (1947) developed the criterion of "simple structure" as a guide to rotation. The three major points of this criterion are:

1. Any column of the loadings matrix should have mostly small values that are as close to zero as possible.
2. Any row of the loadings matrix should have only a few entries that are far from zero.
3. Any two columns of the matrix should exhibit a different pattern of high and low loadings.

These three points should be taken into account when selecting a rotation.

The effects of rotation on the loadings and scores matrices depend on the type of rotation used.

2.2.11.1 Orthogonal Rotations

An orthogonal rotation is defined by a matrix \(T\), such that \(TT^T = T^TT = I\). Therefore, relation 2.57 becomes:

\[ m \hat{X} = M + m P T T^T m A^T \]

\[ = M + m P T (m A T)^T \]

\[ = M + m P R (m A R)^T \]  

(2.73)
where $m \mathbf{P}^R = m \mathbf{PT}$ and $m \mathbf{A}^R = m \mathbf{AT}$. Therefore, the orthogonal rotation $\mathbf{T}$ does not change the values of $m \mathbf{\hat{X}}$ and the sum of the variance of $m$ principal components. Note that after rotation the eigenvalues associated with the principal components may no longer be in decreasing order. In other words, it may no longer be the case that $l(1) > l(2) > \ldots > l(m)$. Therefore, the principal components should be reordered after the rotation if we want to keep that property.

If the normalized loadings matrix is rotated, $m \mathbf{A}^R$ is still orthogonal and $(m \mathbf{A}^R)^T m \mathbf{P}^R = \mathbf{I}$. However, the scores matrix, $m \mathbf{P}^R$, is no longer orthogonal; that is, $(m \mathbf{P}^R)^T m \mathbf{P}^R$ is not a diagonal matrix. To get an orthogonal matrix for $m \mathbf{P}^R$, we have to replace the loadings matrix, $m \mathbf{A}$, by $m \mathbf{A} \Lambda^{1/2}$ where $\Lambda$ is the $(r \times r)$ matrix:

$$
\Lambda = \begin{pmatrix}
I_{(1)} & 0 & \ldots & 0 \\
0 & I_{(2)} & \ldots & 0 \\
\vdots & \vdots & \ddots & \vdots \\
0 & 0 & \ldots & I_{(r)}
\end{pmatrix}
$$

(2.74)

and $m \mathbf{P}^R$ by $m \mathbf{P}^R \Lambda^{-1/2}$. However, in that case the new rotated loadings matrix $m \mathbf{A}^R$ is not orthogonal anymore. In summary, one cannot have both $m \mathbf{A}^R$ and $m \mathbf{P}^R$ orthogonal at the same time. The usual practice is to use the second technique and keep the scores matrix orthogonal.

2.2.11.2 Oblique Rotations

For an oblique rotation, relation 2.73 is still valid but with $m \mathbf{P}^R = m \mathbf{PT}^{-1}$ and $m \mathbf{A}^R = m \mathbf{AT}^T$.

If $\mathbf{T}$ is an oblique rotation, the principal components are no longer orthogonal and $m \mathbf{A}^R$ will no longer be a diagonal matrix. The following relations exist:

(i) $\mathbf{\hat{X}}^T \mathbf{\hat{X}} = m \mathbf{A}^R \Phi (m \mathbf{A}^R)^T$, where $\Phi = (m \mathbf{P}^R)^T m \mathbf{P}^R = \mathbf{T}^T \mathbf{T}$, is the correlation matrix of the oblique components.

(ii) $\mathbf{\hat{X}}^T \mathbf{\hat{X}} = m \mathbf{P}^R (m \mathbf{A}^R)^T m \mathbf{A}^R (m \mathbf{P}^R)^T$.

(iii) $(m \mathbf{A}^R)^T = (m \mathbf{P}^R)^T m \mathbf{P}^R = (m \mathbf{P}^R)^T m \mathbf{X}$.

(iv) $\mathbf{\hat{X}} = m \mathbf{P}^R (m \mathbf{A}^R)^T$.

Because principal components can be correlated after an oblique rotation, these types of rotation can be useful when it is known that the unobserved factors behind the observed variables may be correlated.

The different possible orthogonal and oblique rotations will be discussed in Section 8.

2.3 Example of Principal Component Analysis

Now that we have explained what principal components are and how to extract them, it is time to consider an example.
Table 2.3 The 1970 census provided regional information on five socioeconomic variables for the Madison, Wisconsin, area (from Johnson and Wichern, 1982)

<table>
<thead>
<tr>
<th>Region</th>
<th>Total population (thousands)</th>
<th>Median school years</th>
<th>Total employment (thousands)</th>
<th>Health services employment (hundreds)</th>
<th>Median home value ($10,000s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>5.935</td>
<td>14.2</td>
<td>2.265</td>
<td>2.27</td>
<td>2.91</td>
</tr>
<tr>
<td>2</td>
<td>1.523</td>
<td>13.1</td>
<td>0.597</td>
<td>0.75</td>
<td>2.62</td>
</tr>
<tr>
<td>3</td>
<td>2.599</td>
<td>12.7</td>
<td>1.237</td>
<td>1.11</td>
<td>1.72</td>
</tr>
<tr>
<td>4</td>
<td>4.009</td>
<td>15.2</td>
<td>1.649</td>
<td>0.81</td>
<td>3.02</td>
</tr>
<tr>
<td>5</td>
<td>4.687</td>
<td>14.7</td>
<td>2.312</td>
<td>2.50</td>
<td>2.22</td>
</tr>
<tr>
<td>6</td>
<td>8.044</td>
<td>15.6</td>
<td>3.641</td>
<td>4.51</td>
<td>2.36</td>
</tr>
<tr>
<td>7</td>
<td>2.766</td>
<td>13.3</td>
<td>1.244</td>
<td>1.03</td>
<td>1.97</td>
</tr>
<tr>
<td>8</td>
<td>6.538</td>
<td>17.0</td>
<td>2.618</td>
<td>2.39</td>
<td>1.85</td>
</tr>
<tr>
<td>9</td>
<td>6.451</td>
<td>12.9</td>
<td>3.147</td>
<td>5.52</td>
<td>2.01</td>
</tr>
<tr>
<td>10</td>
<td>3.314</td>
<td>12.2</td>
<td>1.606</td>
<td>2.18</td>
<td>1.82</td>
</tr>
<tr>
<td>11</td>
<td>3.777</td>
<td>13.0</td>
<td>2.119</td>
<td>2.83</td>
<td>1.80</td>
</tr>
<tr>
<td>12</td>
<td>1.530</td>
<td>13.8</td>
<td>0.798</td>
<td>0.84</td>
<td>4.25</td>
</tr>
<tr>
<td>13</td>
<td>2.768</td>
<td>13.6</td>
<td>1.336</td>
<td>1.75</td>
<td>2.64</td>
</tr>
<tr>
<td>14</td>
<td>6.585</td>
<td>14.9</td>
<td>2.763</td>
<td>1.91</td>
<td>3.17</td>
</tr>
</tbody>
</table>

mean 4.323 14.01 1.952 2.17 2.45
Standard Deviation 2.076 1.33 0.895 1.40 0.71

2.3.1 Extracting the Principal Components

The data set selected to illustrate the principal component analysis is given in Table 2.3. It has noting to do with atmospheric chemistry, but it is a small data set that is easy to handle. Examples related to atmospheric chemistry will be presented later. The mean and standard deviation for each variable are also given.

The variance-covariance matrix (relation 2.25) for this data set is equal to:

\[
S = \begin{pmatrix}
4.308 & 1.683 & 1.803 & 2.155 & -0.253 \\
1.683 & 1.768 & 0.588 & 0.177 & 0.176 \\
1.803 & 0.588 & 0.801 & 1.065 & -0.158 \\
2.155 & 0.177 & 1.065 & 1.970 & -0.357 \\
-0.253 & 0.176 & -0.158 & -0.357 & 0.504
\end{pmatrix}
\]

and the correlation matrix (\(R\)) is:
Table 2.4 Eigenvalues and loadings for principal components extracted from variance-covariance matrix for data in Table 2.3.

<table>
<thead>
<tr>
<th></th>
<th>Component 1</th>
<th>Component 2</th>
<th>Component 3</th>
<th>Component 4</th>
<th>Component 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Eigenvalue ($\ell_k$)</td>
<td>6.9311</td>
<td>1.7851</td>
<td>0.3896</td>
<td>0.2295</td>
<td>0.0142</td>
</tr>
<tr>
<td>Difference ($\ell_{k+1} - \ell_k$)</td>
<td>5.1460</td>
<td>1.3955</td>
<td>0.1601</td>
<td>0.2153</td>
<td></td>
</tr>
<tr>
<td>Proportion</td>
<td>0.7413</td>
<td>0.1909</td>
<td>0.0417</td>
<td>0.0245</td>
<td>0.0015</td>
</tr>
<tr>
<td>Cumulative Proportion</td>
<td>0.7413</td>
<td>0.9323</td>
<td>0.9739</td>
<td>0.9985</td>
<td>1.0000</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Loadings Total population</th>
<th>Loadings Median school years</th>
<th>Loadings Total employment</th>
<th>Loadings Health services employment</th>
<th>Loadings Median home value</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.7812</td>
<td>0.0709</td>
<td>0.0037</td>
<td>-0.5417</td>
<td>-0.3020</td>
</tr>
<tr>
<td></td>
<td>0.3056</td>
<td>0.7639</td>
<td>-0.1618</td>
<td>0.5448</td>
<td>-0.0093</td>
</tr>
<tr>
<td></td>
<td>0.3344</td>
<td>-0.0829</td>
<td>0.0148</td>
<td>-0.0510</td>
<td>0.9373</td>
</tr>
<tr>
<td></td>
<td>0.4260</td>
<td>-0.5795</td>
<td>0.2205</td>
<td>0.6360</td>
<td>-0.1721</td>
</tr>
<tr>
<td></td>
<td>-0.0543</td>
<td>0.2624</td>
<td>0.9617</td>
<td>-0.0513</td>
<td>0.0245</td>
</tr>
</tbody>
</table>

The latent roots (or eigenvalues) and the loadings matrices using the variance-covariance and correlation matrices are given in Tables 2.4 and 2.5 respectively. The first thing that one notices is that the two sets of eigenvalues and principal components are different. The relative importance of the eigenvalues is different, although in both cases more than 97% of the total variance could be explained by three components. The loadings are also different, although they reflect the same general idea in the present examples, which is not usually the case.

For both types of principal component analysis, these results show that it is possible to neglect some of the components associated with the lowest eigenvalues. We will try to find how many to retain in the next section.

2.3.2 Number of Principal Components to Retain

We will now use some of the techniques proposed in Section 2.2.10 to determine the number of principal components to retain. We will illustrate here only the informal rule-of-thumb techniques.

If one uses the percentage-of-variance criteria, $\tau_m$, three components explain more than 95% of the total variance for both the principal components extracted from variance-
Table 2.5 Eigenvalues and loadings for principal components extracted from correlation matrix for data in Table 2.3.

<table>
<thead>
<tr>
<th>Component</th>
<th>Component</th>
<th>Component</th>
<th>Component</th>
<th>Component</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>Eigenvalue($\lambda_k$)</td>
<td>3.0289</td>
<td>1.2911</td>
<td>0.5725</td>
<td>0.0954</td>
</tr>
<tr>
<td>Difference ($\lambda_{k-1} - \lambda_k$)</td>
<td>1.7378</td>
<td>0.7186</td>
<td>0.4771</td>
<td>0.0833</td>
</tr>
<tr>
<td>Proportion</td>
<td>0.6058</td>
<td>0.2582</td>
<td>0.1145</td>
<td>0.0191</td>
</tr>
<tr>
<td>Cumulative Proportion</td>
<td>0.6058</td>
<td>0.8640</td>
<td>0.9785</td>
<td>0.9976</td>
</tr>
</tbody>
</table>

Loadings

<table>
<thead>
<tr>
<th>Variable</th>
<th>Component 1</th>
<th>Component 2</th>
<th>Component 3</th>
<th>Component 4</th>
<th>Component 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total population</td>
<td>0.5584</td>
<td>0.1314</td>
<td>0.0079</td>
<td>-0.5506</td>
<td>-0.6065</td>
</tr>
<tr>
<td>Median school years</td>
<td>0.3133</td>
<td>0.6289</td>
<td>-0.5490</td>
<td>0.4526</td>
<td>0.0067</td>
</tr>
<tr>
<td>Total employment</td>
<td>0.5683</td>
<td>0.0043</td>
<td>0.1173</td>
<td>-0.2681</td>
<td>0.7690</td>
</tr>
<tr>
<td>Health services employment</td>
<td>0.4866</td>
<td>-0.3096</td>
<td>0.4549</td>
<td>0.6480</td>
<td>-0.2013</td>
</tr>
<tr>
<td>Median home value</td>
<td>-0.1743</td>
<td>0.7010</td>
<td>0.6912</td>
<td>-0.0151</td>
<td>0.0142</td>
</tr>
</tbody>
</table>

covariance and correlation matrices (Tables 2.4 and 2.5). For the former, two components are sufficient to explain more than 93% of the total variance.

The communalities for the five observed variables for $m$ between 2 and 4 are given in Table 2.6. Retaining only four components does not modify the communality much. One starts to see differences when only three components are retained. However, the communalities are still greater than 0.95 for both types of principal component analysis. Drastic changes occur when the number of components retained is reduced to two, especially in the case of the principal components extracted from the variance-covariance matrix. In that case, the communality for the median value home variable decreases to 0.2842, which is not surprising as most of the loading of that variable is on the third component (Table 2.4). For the principal components extracted from the correlation matrix, three of the communalities become less than 0.85.

Table 2.6 Variation of the communality, $h_k^2$, for the five observed variables with the number of components retained for the two types of principal components.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Variance-covariance Matrix</th>
<th>Correlation Matrix</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Number of Components retained</td>
<td>Number of Components retained</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td>Total population</td>
<td>0.9997</td>
<td>0.9841</td>
</tr>
<tr>
<td>Median school years</td>
<td>1.0000</td>
<td>0.9615</td>
</tr>
<tr>
<td>Total employment</td>
<td>0.9845</td>
<td>0.9837</td>
</tr>
<tr>
<td>Health services employment</td>
<td>0.9998</td>
<td>0.9526</td>
</tr>
<tr>
<td>Median home value</td>
<td>1.0000</td>
<td>0.9988</td>
</tr>
</tbody>
</table>
The values of $\epsilon$ are equal to 1.8699 and 1 for the principal components extracted from the variance-covariance and correlation matrices respectively. Therefore, if one uses the 0.7 cut-off limit of Jolliffe (1972), the result indicates that only two components should be retained for both types of PCA.

The broken stick model approach gives one and two components to retain for the principal components extracted from the variance-covariance and correlation matrices respectively (Tables 2.3, 2.4, and 2.5). Thus, this approach identifies fewer components to be retained than the two preceding approaches.

To estimate the number of principal components to retain using the scree test, we will consider Figure 2.2. In Figure 2.2a, for the principal components obtained using the variance-covariance matrix, a line can be drawn passing through the smallest three eigenvalues (or latent roots). Therefore, two principal components should be retained in this case. For the principal components obtained using the correlation matrix, the straight line could be passed through only the two smallest eigenvalues (Figure 2.2b). Hence, one should retain three components in this case.

From those results, one can see that the number of principal components to retain varies, depending on the approach used and the type of analysis made. However, using two or three components for both types of analysis should be about correct. Here we will retain three components for the two types of analysis.

2.3.3 Rotation

To illustrate the effect of rotation, an orthogonal varimax rotation was applied to the first three principal components that were retained in our example. The new eigenvalues and loadings matrix are given in Tables 2.7 and 2.8 for the principal components extracted, using the variance-covariance and correlation matrices respectively. In this example the matrix $W \Lambda W^{1/2}$ was rotated. The tables give the values for the first three columns of that matrix before and after the varimax rotation.

One notices first that the eigenvalues for each component as well as their relative importance are changed by the rotation. However, the sum of the eigenvalues is not modified by it. In the present example, there is a redistribution of the variance between the three components. In Table 2.7, one can see that about 13% of the variance is transferred from
Table 2.7 Eigenvalues and loadings for first three principal components extracted from variance-covariance matrix for data in Table 2.3 before and after varimax rotation.

<table>
<thead>
<tr>
<th></th>
<th>Before Rotation</th>
<th>After Rotation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Comp. 1</td>
<td>Comp. 2</td>
</tr>
<tr>
<td><strong>Eigenvalue ($\lambda_i$)</strong></td>
<td>6.9311</td>
<td>1.7851</td>
</tr>
<tr>
<td><strong>Difference ($\lambda_{k-1} - \lambda_k$)</strong></td>
<td>5.1460</td>
<td>1.3955</td>
</tr>
<tr>
<td><strong>Proportion</strong></td>
<td>0.7413</td>
<td>0.1909</td>
</tr>
<tr>
<td><strong>Cumulative Proportion</strong></td>
<td>0.7413</td>
<td>0.9323</td>
</tr>
</tbody>
</table>

Component 1 transferred about 60% to component 2 and about 2% to component 3. In the case of the principal components extracted from the correlation matrix, most of the variance is transferred from components 1 to component 3.

Table 2.8 Eigenvalues and loadings for first three principal components extracted from correlation matrix for data in Table 2.3 before and after varimax rotation.

<table>
<thead>
<tr>
<th></th>
<th>Before Rotation</th>
<th>After Rotation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Comp. 1</td>
<td>Comp. 2</td>
</tr>
<tr>
<td><strong>Eigenvalue ($\lambda_i$)</strong></td>
<td>3.0289</td>
<td>1.2911</td>
</tr>
<tr>
<td><strong>Difference ($\lambda_{k-1} - \lambda_k$)</strong></td>
<td>1.7378</td>
<td>0.7186</td>
</tr>
<tr>
<td><strong>Proportion</strong></td>
<td>0.6058</td>
<td>0.2582</td>
</tr>
<tr>
<td><strong>Cumulative Proportion</strong></td>
<td>0.6058</td>
<td>0.8640</td>
</tr>
</tbody>
</table>

Loadings

<table>
<thead>
<tr>
<th></th>
<th>Loadings</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Total population</strong></td>
<td>0.9718</td>
</tr>
<tr>
<td><strong>Median school years</strong></td>
<td>0.5452</td>
</tr>
<tr>
<td><strong>Total employment</strong></td>
<td>0.9890</td>
</tr>
<tr>
<td><strong>Health services employment</strong></td>
<td>0.8469</td>
</tr>
<tr>
<td><strong>Median home value</strong></td>
<td>-0.3033</td>
</tr>
</tbody>
</table>
The loadings are also modified by the rotation. One should notice as well that the rotation followed Thurstone's simple structure criterion (see Section 2.2.11). For the principal components extracted from the variance-covariance matrix, the first component is highly loaded on four variables before the rotation and on only three after it (Table 2.7). This is also true for the first component of the principal components extracted from the correlation matrix (Table 2.8). In Table 2.7 one can see that before the rotation the second component is loaded on two variables on which the first component is also highly loaded. After rotation, the second component is still loaded on two variables, but one of the variables is different and only one variable is common with the first component. The third component is not changed much by the rotation. For the principal components extracted from the correlation matrix, the second component is loaded on two variables before rotation and on one after. In this case, the third component is quite different before and after rotation.

The loadings matrices after the varimax rotation can be used to try to interpret the three components. The first component is similar for the two types of analysis. It is an population-employment factor. For the second and third components, the two types of analysis give different results. In Table 2.7, one sees that the second component is a population-education factor, and the third a home value factor. In Table 2.8, the second component is the home value factor and the third component mainly an education factor. This example illustrates the fact that the results of component analyses using the variance-covariance matrix are usually different from those using the correlation matrix.

2.3.4 Plotting for Principal Component Analysis

Many different types of plots can be constructed to simplify the interpretation of the results of the principal component analysis. Here, using the example presented in this section, we have plotted bar charts of the loadings and scattergrams of the loadings and scores.
2.3.4.1 Loadings Bar Chart

Figure 2.3 presents bar charts of the normalized loadings, before and after varimax rotation, for the first three components of the analysis using the correlation matrix. The loadings matrix was normalized by normalizing every principal component to a norm of 1. For each component, the bars correspond to the loadings for total population, median school years, total employment, health services employment, and median value home, in that order. Such (or equivalent) plots are useful in the interpretation of the principal components.

2.3.4.2 Loadings Scattergrams

Figure 2.4 Scattergrams of the normalized loadings of the first three components extracted from the correlation matrix before (a, b, and c) and after (d, e, and f) a varimax rotation. 1: Total population; 2: Median school years; 3: Total employment; 4: Health service employment; 5: Median home value.
Figure 2.5 Scattergrams of the scores for the first three principal components extracted from the correlation matrix before (a, b, and c) and after (d, e, and f) a varimax rotation.

Scattergrams of the loadings can be useful in interpreting the principal components, especially if there is a large number of observed variables. A scattergram is created for each possible pair of components. In these graphs, each observed variable is represented by a point.
The main goal in constructing these graphs is to see if clusters of variables can be found in one of more of the scattergrams. The most interesting clusters are the ones that can be found in more than one scattergram.

Figure 2.4 presents scattergrams of the loadings for the data in Table 2.3 for the first three components before and after the varimax rotation. Before the rotation, the first four observed variables seem to form an extended cluster. The varimax rotation helps by separating that cluster into two well-defined clusters: variables 1, 3 and 4 are closely linked, and variable 2 is mostly loaded on component 3. Variable 5 is loaded on component 2.

2.3.4.3 Scores Scattergrams

A second very useful type of graphs is a scattergram of the scores. Such plots for before and after the varimax rotation are shown in Figure 2.5 for the first three principal components extracted from the correlation matrix. The three pre-rotation graphs plot for component 1 against component 2, component 1 against component 3, and component 2 against component 3. Similar post-rotation graphs are also given.

One can see large differences in the distribution of the data before and after the varimax rotation. For component 2 versus component 1, the data are closer to the component 1 axis. The same is true for component 3 versus component 1. For component 3 versus component 2, the data are also mostly along component 2 axis after the varimax rotation.

These scattergrams are often used to try to find clusters that can be used to regroup the rows of the observation matrix. Such clusters do not seem to be present in the present example.

2.4 Remarks

We have tried in the preceding sections to cover the most important aspects of principal component analysis. However, it is impossible to cover such an extended subject completely, and the reader is referred to the references given at the end of this section for more information. Nevertheless, there are points that are important to address before ending our discussion of principal component analysis.

In atmospheric chemistry, it is common to get below-detection-limit data. Usually, although not always, the detection limit is available. In statistics, these data are called left-censored data, meaning that the observed distribution has been truncated to the left or low values. There is presently no technique to handle such data in the context of principal component analysis. If no detection limits are available, the only technique possible is to treat them as missing values or to replace them by a very small value which may be more appropriate. (Note that to replace them by a value of zero may not be the most appropriate solution, especially if the data have to be transformed.) If the detection limits are available, the current practice is to replace the below-detection-limit data by some fraction of the detection limit like 1/2 or 2/3, the first value being the mean for a rectangular distribution and the second for a triangular distribution. The latter value would be more appropriate, especially if the data are log-normally distributed. On one hand, if the number of below-detection-limit data is small, their presence should not have much of an effect on the results of the principal component analysis. On the other hand, if they represent a large fraction of the available data, the value used may influence the results of the analysis. In such cases, a sensitivity analysis
should be made. The uncertainties caused by the presence of the *below-detection-limit* data could be estimated by redoing the calculation many times, each time replacing the *below-detection-limit* data by a random number between 0 and the detection limit using a rectangular or, better, a triangular distribution.

The second important point to discuss is the question of transforming the data before doing the analysis to be closer to normality. This should be done if one wants to use any of the techniques for creating confidence intervals or testing hypotheses. Otherwise, there is no necessity to transform the data. However, if the distributions of the data are very skewed, the large values may completely dominate the analysis and the results may not be representative of the bulk of the distribution. In that case, it may be good practice to transform the data to reach more symmetric distributions or to "robustify" the analysis.

In atmospheric chemistry, the data that we want to analyze are usually time series. Therefore, the samples are not independent and the techniques for obtaining confidence intervals and testing will not be applicable. The presence of temporal structures in the data, such as long-term trends or seasonal variation, does not in itself create a problem, and we will illustrate later how that characteristic of the data can be used. But one should keep in mind that the principal components that we obtain will generally reflect associations between the components of the temporal variation that have the highest variability. For example, the seasonal variations are generally the variations with the highest amplitude, and consequently the principal components will mainly reflect the relationship between the seasonal cycles. Thus, it may be necessary in some cases to eliminate some or all of the temporal variation before extracting the principal components.

The final points that we will address is whether one should use the variance-covariance or the correlation matrix. There is no simple answer to this question. If the ranges of the different variables vary by many orders of magnitude, the use of the correlation matrix is recommended. Otherwise, the choice is mostly a matter of preference and judgment. One important rule to follow is that *the principal components that we get should make physical sense*. Therefore, strange results should be examined carefully before one tries to explain them. This rule should also be used when we try to fix the number of components to retain.

A possible extension of principal component analysis is three-mode principal component analysis (also called three-mode factor analysis, although principal component analysis is used) or parallel principal component (also called parallel factor) analysis. This type of analysis is especially appropriate when a third dimension or classification variable is present. An example of appropriate data would be weekly or monthly observations of many different variables for many years. The data could be seen as a series of t annual 2-dimensional tables or matrices, each table or matrix having weeks (or months) as the row index and variables as the columns index. The third dimension would be the year. The idea of three-mode principal component analysis is to analyze jointly the set of t (nxp) matrices. There are many different ways to do that. Examples are the two-stage principal components method (see Bouroche and Dussaix, 1975), the mapping covariance matrices method (see Escoufier, 1980a and b), and Tucker’s method (see Tucker, 1966, 1967). The reader is referred to Basilevsky (1994) and Kroonenberg (1983a and b) for more details.

2.5 Computer Software
All commercially available statistical analysis systems like SAS®, S-Plus®, and SPSS®, contain programs or functions for estimating principal components and rotating them. Most of them also contain programs or functions for solving the eigenvalue problem and doing singular value decomposition. One should, however, be careful when using them with small data sets because some of them may use 1/n to estimate the variance-covariance in place of 1/(n-1) (relation 2.25). The data in Table 2.3 could easily be used to verify which one is used. Note that these software packages do not usually include the tests and confidence intervals of Section 2.2.7. However, most of those tests and confidence intervals can be calculated using the output from these programs.

For readers who do not have access to statistical analysis programs and who can program in Fortran or C, some good source codes for the eigenvalue problem and for the singular value decomposition are available in Press et al. (1992a, 1992b). A Fortran source code for singular matrix decomposition of complex matrices is available in Businger and Golub (1969).

### 2.6 References


Examples of the use of PCA in meteorology are: Craddock and Flood (1969), Craddock and Flintoff (1970), Blasing (1975), Schickendanz (1977), Rasmusson et al. (1981), Overland and Preisendorfer (1982), Wigley et al. (1984), and Pitchford and Pitchford (1985). Principal component analysis has also been used in atmospheric chemistry. The papers by Henry and Hidy (1979), Heidam (1981), Eder (1989), and Barrie and Barrie (1990) are only a few examples.
3. Absolute Principal Component Analysis (APCA)

3.1 Introduction

The two main goals in using PCA in atmospheric chemistry are: (1) to try to identify different emission sources and (2) to determine the contribution of each source to the observed concentration or mixing ratio. This can be expressed in matrix form as:

\[ C = AS \] (3.1)

where \( C \) is the matrix of observations, \( A \) is the source composition matrix (i.e., \( a_{ij} \) is the fraction of element \( i \) in source \( j \)), and \( S \) is the matrix of source contributions. This relation is similar to relation 2.24, which justifies the use of principal component analysis for identifying the sources and their contributions.

However, in atmospheric chemistry, observed data are measured relative to an absolute zero, which is not the cases for the scores (i.e., the \( P \) matrix in relation 2.24). This is due to the fact that one of the first steps in PCA involves eliminating the mean value or, in other words, aligning the variables (Figure 2.1b). Thus, the scores are not deviations from zero but deviations from the mean instead. This results in the presence of matrix \( M \) in relation 2.24. One would therefore like to be able to define the scores as deviations from an absolute zero by a translation of the principal component axis system of coordinates as illustrated in Figure 3.1.

We have seen in the preceding section that solutions in principal component analysis are not unique, and any solution obtained by rotation from the initial one is also valid. Therefore, there are an infinite number of possible transformations or rotations, most of which produce models that are physically impossible. Henry (1987) defined five natural constraints that any physically realistic transformation must obey. They are:

1. The original data must be reproduced by the model; in other words, the model must explain the observations.
2. The predicted source compositions must be non-negative; that is, a source cannot have a negative percentage of an element. Therefore, the elements of the loadings matrix must be equal to zero or greater.
3. The sum of the predicted elemental mass fractions for each source must be less than or equal to 1; the whole is greater than or equal to the sum of the parts.
4. The predicted source contributions must all be non-negative; a source cannot emit negative mass. Therefore, the elements of the scores matrix must be equal or greater than zero.
5. The sum of the predicted source contributions on a given day cannot exceed the total mass on that day.

Henry (1987) also pointed out that, in general, these constraints are not sufficient to ensure a unique factor model.

The important point to retain from Henry’s rules is that the elements of both loadings and scores matrices should be greater than zero. It should be noted that, in some situations in
atmospheric chemistry, some of the loadings can be negative, as some sources may in fact be sinks for some ions. Therefore, Henry's second rule can be relaxed to some extent.

Absolute Principal Component Analysis (APCA) was developed to try to find a solution with the loadings and scores matrices having both elements greater than or equal to zero.

APCA was introduced by Thurston and Spengler (1985) as an attempt to solve that problem. In this technique, a standard principal component analysis is done first (see preceding section). Then, the scores of the absolute zero of the observed variables are calculated. The absolute zero scores are next subtracted from the scores of the samples to define them relative to an absolute zero in the new coordinate system.

In summary, an absolute principal component analysis consists of the following three steps:

1. Do a standard PCA.
2. Calculate the scores of the absolute zero.
3. Subtract the absolute zero scores from the sample scores.

3.2 Technical Details

The steps in doing a standard PCA are:

1. Extract all the principal components.
2. Select a subset of m (<p) principal components.
3. Rotate the m principal components using a varimax or other rotations.
4. Determine $m \mathbf{P}^R$ and $m \mathbf{A}^R$ in relation 2.73.

Using the notation of the preceding section, one can write for the principal components from the variance-covariance matrix:
\[ \mu \mathbf{P}^R = (\mu \hat{\mathbf{X}}^R - \mathbf{M}) \mu \mathbf{A}^R \] (3.2)

and for the principal component from a correlation matrix:

\[ \mu \mathbf{P}^R = (\mu \hat{\mathbf{X}}^R - \mathbf{M}) \mathbf{D}^{-1/2} \mu \mathbf{A}^R \] (3.3)

where \( \mathbf{D} \) is defined in relation 2.32.

The \((1 \times m)\) scores matrix, \(\mu \mathbf{P}_1^R\), of the absolute zeros of the observed variables can be obtained by replacing \(\mu \hat{\mathbf{X}}^R\) and \(\mathbf{M}\) in relations 3.2 or 3.3 by the following two \((1 \times p)\) matrices:

\[ \mu \hat{\mathbf{X}}^R = (0 \ 0 \ \cdots \ 0) \] (3.4)

\[ \mathbf{M} = (\bar{X}_1 \, \bar{X}_2 \, \cdots \, \bar{X}_p) \] (3.5)

If one defines the \((n \times m)\) matrix \(\mu \mathbf{\Pi}^R\) as:

\[
\mu \mathbf{\Pi}^R = 
\begin{bmatrix}
(\mu \mathbf{P}_0^R)_1 & (\mu \mathbf{P}_0^R)_2 & \cdots & (\mu \mathbf{P}_0^R)_m \\
(\mu \mathbf{P}_0^R)_1 & (\mu \mathbf{P}_0^R)_2 & \cdots & (\mu \mathbf{P}_0^R)_m \\
\vdots & \vdots & \ddots & \vdots \\
(\mu \mathbf{P}_0^R)_1 & (\mu \mathbf{P}_0^R)_2 & \cdots & (\mu \mathbf{P}_0^R)_m 
\end{bmatrix}
\] (3.6)

The Absolute PC Scores (APCS), \(\mu [\mathbf{AP}]^R\), are given by:

\[ \mu [\mathbf{AP}]^R = \mu \mathbf{P}^R - \mu \mathbf{\Pi}^R \] (3.7)

Note that the relations giving \(\mu \hat{\mathbf{X}}^R\) are modified. The relationship for the principal components obtained using the variance-covariance matrix becomes:

\[ \mu \hat{\mathbf{X}}^R = [\mathbf{M} + \mu \mathbf{\Pi}^R (\mu \mathbf{A}^R)^T] + \mu [\mathbf{AP}]^R (\mu \mathbf{A}^R)^T \] (3.8)

For the principal components from the correlation matrix, the relation becomes:

\[ \mu \hat{\mathbf{X}}^R = [\mathbf{M} + \mu \mathbf{\Pi}^R (\mu \mathbf{A}^R)^T \mathbf{D}^{-1/2}] + \mu [\mathbf{AP}]^R (\mu \mathbf{A}^R)^T \mathbf{D}^{-1/2} \] (3.9)

Note that the first term to the right (i.e., term in brackets) in relation 3.8 and 3.9 is equal to zero if \(m = p\).

**Warning:** The following erroneous description of the APCA technique can be found in the literature and should not be used. It seems to result from a misreading of Thurston and Spengler (1985). The description is:

The procedure of APCA consists in including a dummy zero sample in the data matrix which provides the offset from the true zero for the principal component scores, allowing the true principal scores to be calculated.

### 3.3 Example of Absolute Principal Component Analysis

To illustrate the use of APCA, we will continue the example presented in Section 2.3. The eigenvalues and the loadings matrices do not change. Thus, they are given in Table 2.7
and 2.8. The scores do change and their scattergrams are shown in Figure 3.2. These graphs should be compared with Figures 2.5d, 2.5e, and 2.5f. A close examination will reveal that the relative positions of the points to each others do not change but are only translated. Note that for Comp. 3, the range (0 to 12) of the axis in Figure 3.2 is twice its value (-3 to 3) in Figure 2.5, which indicates that the variance of the data is smaller in Figure 3.2 than in Figure 2.5.

### 3.4 Remarks

Because APCA is only a slight modification of PCA, the discussion in Section 2 on obtaining robust principal components, treating missing and below-detection-limit observations, selecting the number of components to retain, and other points applies.

Some authors (e.g., Li and Winchester, 1990) indicate that the original concentrations should be regressed on the absolute principal component scores according to the relation:

\[
C_k' = \alpha_0 + \sum_{j=1}^{m} \beta_j m [\text{AP}]_j^k \quad (3.10)
\]

This regression step is not necessary, as the relationship between the observed variables and the principal components is given by relation 3.8 or 3.9. If one regresses \( C_k \) on \( m [\text{AP}]_k^k \) by least squares, one will obtain:

\[
\alpha_0 = [M_m + \Pi_m^R (m \mathbf{A}^R)^T]_k \quad (3.11)
\]

and

\[
\beta_j = [(m \mathbf{A}^R)^T]_k \quad (3.12)
\]

Note that if one does a regression the confidence interval obtained does not represent the real confidence interval for the regression coefficients, because in regression theory it is assumed that there are no uncertainties in \( m [\text{AP}]_k^k \), which is not the case in APCA.
case in the present situation. The loadings and scores matrices are both subject to sampling errors, and confidence intervals can be estimated for both.

Thurston and Spengler (1985) developed APCA to apportion observed particle masses to different sources. In their analysis, the particulate masses were known, but in the preceding section it was assumed that this information was not available. If that parameter is known, however, the following steps can be added to APCA:

4. Regress the observed particle mass on the absolute principal component scores, \( m[AP]^k \), using the relation:

\[
M_k = \xi_0 + \sum_{j=1}^{m} \xi_j m[AP]^j
\]  

(3.13)

where \( M_k \) is the particle mass recorded during observation \( k \);

5. Regress the concentrations, \( C_k \), on \( S_j = (\xi_j m[AP]^j) \).

More details can be found in Thurston and Spengler (1985). Note that Poissant (1994) has recently argued that Thurston and Spengler's technique might be biased by an unknown amount. If true, that would reduce the usefulness of the technique in practice.

Finally, one should remark that the solution obtained using APCA does not always have all elements of the loadings and/or scores matrix greater than zero. However, it is sometimes possible to find a series of \( p \) simple rotations that will produce the desired solution (see Shen and Israel, 1989). The solution is said to be \( p \)-rotatable. One should notice, however, that in some cases the solution found is not \( p \)-rotatable. In those cases, APCA fails to give the desired solution and another technique should be used.

3.5 Computer Software

The present author does not know of any software available for APCA. However, the calculations can be easily made if a program for PCA is available, since steps 2 and 3 of the APCA are easily made using the relations given in Section 3.2. If the complete analysis is made, steps 4 and 5 could be easily done using regression analysis. Note, as mentioned in Section 3.4, that the absolute principal component scores are subject to sampling uncertainties and therefore no confidence intervals for the regression coefficients can be obtained directly. Resampling methods like the jackknife and bootstrap techniques can be used to estimates of those uncertainties (see Section 9).

3.6 References

The main reference to APCA is Thurston and Spengler (1985). Some interesting comments on analyses of this type can be found in Henry (1987). Another interesting paper on the subject is the one of Poissant (1994).
4. Target Principal Component Analysis (TPCA)

4.1 Introduction

Like absolute principal component analysis, Target Principal Component Analysis (TPCA) is an extension of standard principal component analysis. In fact, TPCA is similar to APCA in its objectives. Chronologically, TPCA was put forward earlier than APCA, but the latter was developed in part because the former is difficult to implement if one does not have access to source codes to extract principal components.

In the literature, TPCA is usually called Target Transformation Factor Analysis (TTFA; Weiner et al., 1970; Albert and Hopke, 1980) or Target Factor Analysis (TFA; Hopper, 1986). However, since principal component analysis and not factor analysis is used in all these studies, the name target principal component analysis is preferable and will be used here.

TPCA differs from standard principal component analysis in many ways. First, the variables are not centered around the mean and therefore the variance-covariance or correlation matrices are calculated relative to an absolute zero and not relative to the means. This is done to obtain principal components that are relative to the absolute zero. The principal components are then extracted as described in Section 2. This is illustrated in Figure 4.1 for a two-dimensional example.

The second difference is the use in many examples of TPCA of so-called Q-mode principal component analysis (see Rummel, 1970). Up to this point, we have extracted the principal components using either the variance-covariance or the correlation matrix between the observed variables (except when using singular value decomposition). These matrices were calculated using relations 2.25 or 2.33 respectively. This type of analysis is called R-mode PCA. As we have seen, the loadings matrix is obtained directly, and the scores matrix can be obtained from the loadings matrix and the original data. In Q-mode principal component analysis, the variance-covariance or correlation matrices are calculated between the samples and not the observed variables. The PCA gives the scores matrix, and the loadings matrix is calculated using the scores matrix and the original data. Theoretically, the two approaches should give the same results. Practically, however, the two PCA modes give different answers in TPCA (see Hwang et al., 1984) when the correlation matrices are used. As pointed out by Heidam and Kronberg (1985), the differences between the two analyses come from differences in scaling (see the Technical Details section). Note that when the number of samples is large, the use of Q-mode TPCA becomes computer intensive and R-mode TPCA should be used instead.

The last difference between TPCA and standard PCA is that the retained principal components are rotated, using what is called target rotation, to fit a priori estimates of the latent factors as closely as possible. This type of rotation was developed in the 1960s within the context of the social science and factor analysis (see Hurley and Cattell, 1962). Henry (1977) was one of the first to use the technique in atmospheric chemistry for the problem of mass balance.

When the total particulate masses for the observation are known (see Section 3), one of the last stages in the analysis is to scale the results to apportion of observed particulate masses.
The five Henry's rules mentioned in Section 3.1 still apply in the case of TPCA. One therefore looks for loadings and scores matrices having both elements greater than or equal to zero. These conditions may limit the possible choices in the initial target loadings matrix for the target rotation step.

The different steps in target principal component analysis can be summarized as follows:

1. Calculate variance-covariance or correlation matrices relative to *absolute zero*. These matrices are calculated between the samples in Q-mode TPCA and between the variables in R-mode TPCA.

2. Extract the principal components. Note that the number of non-zero eigenvalues is the same in both types of analysis.

3. Select the number of principal components to retain, using the techniques described in Section 2.2.10 or others.
4. Calculate the loadings and scores matrices.

5. Define the \emph{a priori} loadings matrix of the factor.


7. Scale the results if the total particulate masses are available.

\subsection*{4.2 Technical Details}

\subsubsection*{4.2.1 Q-mode Target Principal Component Analysis}

If \( X \) is a \((n\times p)\) observation matrix (relation 2.21), the \((n\times n)\) variance-covariance matrix relative to \emph{absolute zero} is:

\[ S_Q = \frac{1}{p-1} XX^T \]

and the \((n\times n)\) correlation matrix is:

\[ R_Q = \frac{1}{p-1} V_Q^{-1} X (V_Q^{-1} X)^T = V_Q^{-1} S_Q V_Q^{-1} \]

where \( V_Q \) is a \((n\times n)\) diagonal matrix with elements given by:

\[ V_{ii} = \frac{1}{p-1} \sum_{j=1}^{p} x_{ij}^2 \]

The principal components can be extracted as indicated in Section 2.2 by replacing \( S \) or \( R \) by \( S_Q \) or \( R_Q \). Note that the smallest \((n-p)\) eigenvalues would be equal to zero. In the present analysis the eigenvectors do not constitute the loadings matrix \( A \) but the scores matrix \( P \). The loadings matrix is related to the scores matrix by the relation

\[ A_Q = X^T P_Q \]

where \( P_Q \) is the standardized scores matrix, if the variance-covariance matrix is used. Note that \( A_Q \) and \( P_Q \) are \((p\times m)\) and \((n\times m)\) matrices respectively, if \( m (\leq p) \) components are retained. If the correlation matrix is used, relation 4.4 becomes:

\[ A_Q = (V_Q^{-1} X)^T P_Q \]

If \( m \) components are retained, the estimated observed matrix, \( \hat{X}_Q \), is equal to

\[ \hat{X}_Q = m P_Q (m A_Q)^T \]

if \( S_Q \) is used, and to

\[ \hat{X}_Q = V_Q m P_Q (m A_Q)^T \]

if the correlation matrix \( R_Q \) is used. Note that in 4.6 and 4.7, the scores matrix, \( m P_Q \), is standardized and not the loading matrix, \( m A_Q \), as is usually the case. This can be changed by replacing \( m P_Q (m A_Q)^T \) by \( m P_Q L^{1/2} (m A_Q L^{-1/2})^T \) where the diagonal matrix \( L \) is equal to:
where the \( a_{ij} \) are the elements of matrix \( m \mathbf{A}_Q \).

### 4.2.2 R-mode Target Principal Component Analysis

In R-mode principal component analysis, the analysis is similar to that presented in Section 2, except that the variables are not centered around the means. For R-mode PCA, relations 4.1 and 4.2 become:

\[
S_R = \frac{1}{n-1} \mathbf{X}^T \mathbf{X}
\]  

and

\[
R_R = \frac{1}{n-1} (\mathbf{V}_R^{-1} \mathbf{X})^T \mathbf{V}_R^{-1} \mathbf{X} = \mathbf{V}_R^{-1} \mathbf{S}_R \mathbf{V}_R^{-1}
\]  

respectively, where \( \mathbf{V}_R \) is a \((p \times p)\) diagonal matrix with elements given by:

\[
v_{ij} = \left[ \frac{1}{n-1} \sum_{i=1}^{n} x_{ij}^2 \right]^{1/2}
\]

If \( \mathbf{S}_R \) is used, relation 2.22 for the scores matrix, \( \mathbf{P} \), becomes:

\[
\mathbf{P}_R = \mathbf{X} \mathbf{A}_R
\]  

where \( \mathbf{A}_R \) is the standardized loadings matrix. If \( \mathbf{R}_R \) is used, one gets:

\[
\mathbf{P}_R = (\mathbf{X} \mathbf{V}_R^{-1}) \mathbf{A}_R
\]

When \( m \) components are retained, relations 2.57 and 2.61 become:

\[
\mathbf{\hat{X}}_R = m \mathbf{P}_R (m \mathbf{A}_R)^T
\]

and

\[
\mathbf{\hat{X}}_R = m \mathbf{P}_R (m \mathbf{A}_R)^T \mathbf{V}_R
\]

respectively.

### 4.2.3 Relationship between Q-mode and R-mode Target Principal Component Analysis

Zhou et al. (1983) have shown that the R- and Q-mode solutions can both be interrelated through a singular value decomposition of the data matrix:

\[
\mathbf{X} = \mathbf{T} \mathbf{D}^T \mathbf{S}^T \mathbf{V}_R = \mathbf{V}_Q^T \mathbf{T} \mathbf{D} \mathbf{S}^T
\]  

46
where $S$ is the $(p \times p)$ matrix of eigenvectors obtained from the $XX^T$ matrix, $T$ is the $(n \times n)$ matrix of eigenvectors obtained from the $X^T X$ matrix, and $D$ is the $(p \times n)$ diagonal matrix containing the square roots of the first $n$ eigenvectors obtained from the $X^T X$ matrix. $V_r$ is equal to $\sqrt{n-1}V_r$ and $V_q$ to $\sqrt{p-1}V_q$.

4.2.4 Target or Procrustes Rotations

As we have seen earlier (Section 2.2.11), it is usual to rotate the principal component coordinate system to help in the physical interpretation of the $m$ factors retained. Two types of rotations are usually employed: (1) orthogonal rotations and (2) oblique rotations. Although these rotations usually help in the interpretation of the retained factors, there is no guarantee that the factors obtained will have a physical interpretation. However, it is often possible from existing knowledge of the physical processes to have some idea of how these factors should look. For example, in atmospheric chemistry knowledge of the relative elemental composition of actual source materials can be used to estimate such factors. Using such knowledge, it is possible to create what are called test vectors. The problem now is to find a procedure to geometrically realign the factor axes of the loadings matrix, $mA$, obtained using either R-mode or Q-mode PCA with these test vectors. The realignment procedure, called target transformation (Weiner et al., 1970), involves finding a rotation vector, $r$, which aligns a column of the loadings matrix, $mA$, with the input test vector, $b$, by least squares maximization of the overlap between a rotated axis of $mA$ and the test vectors.

The rotation vector, $r$, can be found as follows. The difference between the rotated column of the matrix $mA$ and the input vector $b$ is given by:

$$
\varepsilon = Ar - b
$$

writing $mA$ as $A$. In a least squares fit, the quantity to be minimized is $\varepsilon^2$ given by:

$$
\varepsilon^2 = (Ar - b)^T(Ar - b)
$$

$$
= (r^T A^T - b^T)(Ar - b)
$$

$$
= r^T A^T Ar - r^T A^T b - b^T A r + b^T b
$$

The third term in the last equation is a scalar quantity and thus equal to its transpose. Therefore:

$$
\varepsilon^2 = r^T A^T Ar - 2r^T A^T b + b^T b
$$

Taking the derivative and setting it equal to 0 gives:

$$
\frac{\partial \varepsilon^2}{\partial r} = 0 = 2A^T Ar - 2A^T b
$$

Thus, one has:

$$
r = (A^T A)^{-1} A^T b
$$

It has been found that using weighted least squares greatly enhances the ability of the analysis to resolve sources with similar concentration profiles (Roscoe and Hopke, 1981b). The weighted target transformation rotation is given by:

$$
\text{47}
$$
\[ r = (A^TWA)^{-1}A^Twb \]  
\hspace{1cm} (4.22)

where \( W \), the weight matrix, is a diagonal matrix that can have as its diagonal elements the inverse of the variance in the elemental concentrations or the inverse of the squares of the average error of determination for the concentration values.

Using relation 4.21 or 4.22, it is possible to test a suspected particle source by determining if a factor axis can be rotated so as to overlap with the suggested test vector. Another possibility is to analyze uniqueness test vectors that have all but one element set equal to zero and the remaining value set to unity. An element is considered unique if its concentration in the data does not covary with the other measured species. The uniqueness test then determines if the concentration of an element is strongly related to any other element(s). The resulting test vector predicted by the uniqueness test can be used as a normalized source concentration profile for sources whose concentration profiles are unknown.

Roscoe and Hopke (1981b) have found that sources profiles can be obtained by an iterative process from simple uniqueness test vectors. From the uniqueness test vector, relation 4.21 is used to obtain a rotation vector, \( r \), from which

\[ b' = Ar \]  
\hspace{1cm} (4.23)

can be obtained. After setting any negative values to small positive ones, a new value of \( r \) is obtained by using \( b' \) as the initial value in relation 4.21. Then using relation 4.23, \( b'' \) can be calculated. This iterative process can be continued until the average per cent change in the values of \( b' \) and \( b'' \) is less than 10\(^{-2}\), for example.

The search for the profile vectors can be done similarly for the \( m \) postulated sources. The obtained \( b \) and \( r \) vectors can be grouped to produce the \( _mA^R \) and \( R_r \) matrices such that:

\[ _mA^R = _mA^R R_r \]  
\hspace{1cm} (4.24)

An important characteristic of the loadings matrix, \( _mA^R \), that is thus obtained is that it no longer has to be orthogonal, as the rotation, \( R_r \), may be oblique.

Once the rotated loadings matrix and the rotation matrix are known, the rotated scores matrix can be obtained by an inverse rotation applied to the scores matrix. The rotated scores matrix is given by:

\[ _nP^R = _nP (R_r^T)^{-1} \]  
\hspace{1cm} (4.25)

Severin et al. (1983), in an effort to improve the fit between model and observation, developed an alternative procedure to estimate \( _mP^R \). They used the final matrix of source profiles, \( _mA^R \), and the weighting matrix, \( W \), used in relation 4.22, to calculate least squares fits to the columns of the observed data matrix \( X \). \( _mP^R \) is equal to:

\[ _nP^R = ((_mA^R)^T W _mA^R)^{-1}(_mA^R)^T WX \]  
\hspace{1cm} (4.26)

The original data can then be compared point by point with the reproduced concentrations to determine how well the chosen set of profiles represents the actual sources in the system. This comparison may suggest further refinements to the initial loadings matrix of the target rotation. Note that the \( (_mP^R)_{ij} \) should be positive.
Table 4.1 Eigenvalues and loadings extracted from non-central correlation matrix for data in Table 2.3.

<table>
<thead>
<tr>
<th>Component</th>
<th>Component</th>
<th>Component</th>
<th>Component</th>
<th>Component</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>Eigenvalue($\ell_k$)</td>
<td>4.6162</td>
<td>0.3043</td>
<td>0.0572</td>
<td>0.0203</td>
</tr>
<tr>
<td>Difference ($\ell_{k-1} - \ell_k$)</td>
<td>4.3119</td>
<td>0.2471</td>
<td>0.0369</td>
<td>0.0182</td>
</tr>
<tr>
<td>Proportion</td>
<td>0.9232</td>
<td>0.0609</td>
<td>0.0114</td>
<td>0.0041</td>
</tr>
<tr>
<td>Cumulative Proportion</td>
<td>0.9232</td>
<td>0.9841</td>
<td>0.9955</td>
<td>0.9996</td>
</tr>
</tbody>
</table>

Loadings

<table>
<thead>
<tr>
<th></th>
<th>Total population</th>
<th>Median school years</th>
<th>Total employment</th>
<th>Health services employment</th>
<th>Median value home</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.4569</td>
<td>-0.2178</td>
<td>-0.5915</td>
<td>-0.2507</td>
<td>-0.5754</td>
</tr>
<tr>
<td>Total population</td>
<td>0.4524</td>
<td>0.3714</td>
<td>-0.0482</td>
<td>0.8505</td>
<td>-0.0827</td>
</tr>
<tr>
<td>Median school years</td>
<td>0.4592</td>
<td>-0.2588</td>
<td>-0.2916</td>
<td>-0.0745</td>
<td>0.7947</td>
</tr>
<tr>
<td>Total employment</td>
<td>0.4371</td>
<td>-0.5449</td>
<td>0.6938</td>
<td>0.0294</td>
<td>-0.1743</td>
</tr>
<tr>
<td>Health services</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>employment</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Median value home</td>
<td>0.4297</td>
<td>0.6714</td>
<td>0.2854</td>
<td>-0.5314</td>
<td>0.0253</td>
</tr>
</tbody>
</table>

Target rotation, originally developed for use in the social sciences by Hurley and Cattell (1962), is also known as Procustes rotation. This name comes from a character in Greek mythology, the robber Procrustes, who made his victims lie on an iron bed and either stretched their bodies or cut off their legs so that they would fit it exactly. The parallel between Procrustes' methods and the basic concept behind target rotation is clear.

4.2.5 Source Scaling

If the total masses for the samples are known, it is possible to scale the TPCA solution to provide the correct calculated sources and the total measured sample weights. They are obtained by means of a multiple regression analysis of the form:

$$M_j = \sum_{k=1}^{m} s_k (m^P R)_{kj}$$  \hspace{1cm} (4.27)

or

$$M_j = s_0 + \sum_{k=1}^{m} s_k (m^P R)_{kj}$$  \hspace{1cm} (4.28)

where $M_j$ is the measured mass of sample $j$. The coefficients should be positive and statistically significant. The source profiles are scaled by replacing $(m^A R)_{kj}$ by $(m^A R)_{kj}/s_j$. If physically unacceptable values for the scaled source profiles are found, this means that the number of sources selected (i.e., $m$, the number of components retained) is inappropriate and the analysis should be restarted with a better value.
4.3 Examples of Target Principal Component Analysis

As a first example of the use of target principal component analysis, we will recalculate the principal components for the data given in Table 2.3, using the R-mode and the correlation matrix (relation 4.10). The results of the analysis are given in Table 4.1. Those results should be compared to the ones presented in Table 2.5. The principal components obtained when the variables are not centered are very different from the ones obtained after centering the variables. In the present case, the first two principal components explain more than 98% of the total variance. In the analysis presented in Section 2.3.1, three components were necessary to explain 97% of the total variance. We will therefore retain two components for target rotation. Based on the results of Section 2.3, we will use the following target matrix loading:

\[
B_0 = \begin{bmatrix}
0.5774 & 0 \\
0 & 0.7071 \\
0.5774 & 0 \\
0 & 0.7071
\end{bmatrix}
\]

. The first factor is the factor relating total employment and health services employment with total population. The second factor links median home values with median school years. The loadings after a target rotation are given in Table 4.2. Although the target rotation conserves the basic definition of the components or factors, the new ones are no longer orthogonal. The correlation between the two is equal to about 0.3.

For our second example, we will use urban pollution data from Edmonton, Alberta, in November 1978, as presented by Hopper (1986: Table 4.5). The data are given in Table 4.3. The 11 eigenvalues and principal components were extracted using the R-mode and the correlation matrix (relation 4.10). The eigenvalues and principal components are given in Table 4.4. A look at the results indicates that between four and six principal components or factors should be retained. We have decided to retain four factors, although Hopper (1986) used five components. Based on the results of Hopper (1986), we have used the following initial loadings matrix:

Table 4.2 Loading matrix after Target Rotation.

<table>
<thead>
<tr>
<th></th>
<th>Component 1</th>
<th>Component 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total population</td>
<td>0.5022</td>
<td>0.2199</td>
</tr>
<tr>
<td>Median school years</td>
<td>0.1809</td>
<td>0.5853</td>
</tr>
<tr>
<td>Total employment</td>
<td>0.5263</td>
<td>0.1961</td>
</tr>
<tr>
<td>Health services employment</td>
<td>0.6619</td>
<td>0.0001</td>
</tr>
<tr>
<td>Median value home</td>
<td>0.0000</td>
<td>0.7754</td>
</tr>
</tbody>
</table>
Table 4.3 Urban aerosol concentration at Edmonton, Alberta, Canada, in November, 1978, from Hopper (1986)

<table>
<thead>
<tr>
<th>Date</th>
<th>Pb</th>
<th>Fe</th>
<th>Ca</th>
<th>K</th>
<th>Si</th>
<th>Al</th>
<th>Na⁺</th>
<th>NH₄⁺</th>
<th>Cl⁻</th>
<th>NO₃⁻</th>
<th>SO₄²⁻</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nov. 3</td>
<td>0.60</td>
<td>4.30</td>
<td>3.80</td>
<td>1.80</td>
<td>36.00</td>
<td>3.00</td>
<td>0.20</td>
<td>0.50</td>
<td>0.10</td>
<td>0.70</td>
<td>1.20</td>
</tr>
<tr>
<td>4</td>
<td>0.70</td>
<td>1.90</td>
<td>2.40</td>
<td>9.20</td>
<td>17.00</td>
<td>1.40</td>
<td>0.10</td>
<td>0.70</td>
<td>0.50</td>
<td>0.90</td>
<td>1.30</td>
</tr>
<tr>
<td>5</td>
<td>0.10</td>
<td>0.60</td>
<td>0.50</td>
<td>0.20</td>
<td>5.00</td>
<td>0.70</td>
<td>0.20</td>
<td>0.50</td>
<td>0.50</td>
<td>0.90</td>
<td>1.20</td>
</tr>
<tr>
<td>6</td>
<td>0.90</td>
<td>3.10</td>
<td>2.50</td>
<td>1.20</td>
<td>3.00</td>
<td>2.30</td>
<td>0.10</td>
<td>0.50</td>
<td>0.20</td>
<td>0.50</td>
<td>0.60</td>
</tr>
<tr>
<td>7</td>
<td>1.10</td>
<td>2.90</td>
<td>2.50</td>
<td>1.20</td>
<td>30.00</td>
<td>2.80</td>
<td>0.10</td>
<td>0.60</td>
<td>0.20</td>
<td>0.50</td>
<td>0.80</td>
</tr>
<tr>
<td>8</td>
<td>0.20</td>
<td>2.30</td>
<td>0.90</td>
<td>0.50</td>
<td>13.00</td>
<td>0.80</td>
<td>0.10</td>
<td>0.20</td>
<td>0.20</td>
<td>0.20</td>
<td>0.50</td>
</tr>
<tr>
<td>9</td>
<td>0.20</td>
<td>1.20</td>
<td>0.80</td>
<td>0.50</td>
<td>12.00</td>
<td>1.00</td>
<td>0.20</td>
<td>0.50</td>
<td>0.30</td>
<td>0.10</td>
<td>0.80</td>
</tr>
<tr>
<td>10</td>
<td>0.80</td>
<td>2.30</td>
<td>2.40</td>
<td>0.90</td>
<td>21.00</td>
<td>1.90</td>
<td>0.70</td>
<td>1.50</td>
<td>0.70</td>
<td>1.20</td>
<td>1.70</td>
</tr>
<tr>
<td>11</td>
<td>0.40</td>
<td>1.70</td>
<td>1.10</td>
<td>0.60</td>
<td>14.00</td>
<td>1.40</td>
<td>0.50</td>
<td>0.90</td>
<td>0.90</td>
<td>1.60</td>
<td>1.60</td>
</tr>
<tr>
<td>12</td>
<td>0.40</td>
<td>1.20</td>
<td>2.20</td>
<td>0.60</td>
<td>12.00</td>
<td>1.10</td>
<td>0.40</td>
<td>0.60</td>
<td>1.00</td>
<td>1.70</td>
<td>2.50</td>
</tr>
<tr>
<td>13</td>
<td>0.70</td>
<td>1.80</td>
<td>1.40</td>
<td>0.70</td>
<td>19.00</td>
<td>1.80</td>
<td>0.30</td>
<td>1.10</td>
<td>0.90</td>
<td>0.90</td>
<td>1.80</td>
</tr>
<tr>
<td>14</td>
<td>0.60</td>
<td>3.00</td>
<td>3.00</td>
<td>1.20</td>
<td>29.00</td>
<td>2.60</td>
<td>0.40</td>
<td>0.80</td>
<td>0.60</td>
<td>0.80</td>
<td>1.80</td>
</tr>
<tr>
<td>15</td>
<td>0.50</td>
<td>2.00</td>
<td>2.10</td>
<td>0.80</td>
<td>15.00</td>
<td>1.30</td>
<td>0.20</td>
<td>0.60</td>
<td>1.10</td>
<td>1.70</td>
<td>3.50</td>
</tr>
<tr>
<td>16</td>
<td>0.30</td>
<td>1.70</td>
<td>1.10</td>
<td>0.20</td>
<td>14.00</td>
<td>1.30</td>
<td>0.30</td>
<td>0.60</td>
<td>1.30</td>
<td>1.80</td>
<td>4.30</td>
</tr>
<tr>
<td>17</td>
<td>0.30</td>
<td>0.80</td>
<td>0.50</td>
<td>0.30</td>
<td>6.00</td>
<td>0.70</td>
<td>1.10</td>
<td>2.30</td>
<td>0.50</td>
<td>0.30</td>
<td>1.10</td>
</tr>
<tr>
<td>18</td>
<td>0.40</td>
<td>0.50</td>
<td>0.90</td>
<td>0.30</td>
<td>4.00</td>
<td>0.40</td>
<td>1.00</td>
<td>1.90</td>
<td>0.70</td>
<td>0.70</td>
<td>1.70</td>
</tr>
<tr>
<td>19</td>
<td>0.50</td>
<td>0.40</td>
<td>0.40</td>
<td>0.10</td>
<td>4.00</td>
<td>0.40</td>
<td>2.80</td>
<td>4.90</td>
<td>0.90</td>
<td>1.30</td>
<td>2.00</td>
</tr>
<tr>
<td>20</td>
<td>1.20</td>
<td>1.40</td>
<td>1.30</td>
<td>0.60</td>
<td>15.00</td>
<td>1.20</td>
<td>4.70</td>
<td>7.60</td>
<td>1.50</td>
<td>2.70</td>
<td>2.20</td>
</tr>
<tr>
<td>21</td>
<td>2.30</td>
<td>3.40</td>
<td>4.00</td>
<td>1.30</td>
<td>32.00</td>
<td>2.80</td>
<td>5.20</td>
<td>9.50</td>
<td>3.40</td>
<td>5.40</td>
<td>3.80</td>
</tr>
<tr>
<td>22</td>
<td>1.20</td>
<td>2.50</td>
<td>4.20</td>
<td>1.10</td>
<td>24.00</td>
<td>1.70</td>
<td>4.80</td>
<td>7.20</td>
<td>1.80</td>
<td>3.60</td>
<td>3.30</td>
</tr>
<tr>
<td>23</td>
<td>1.00</td>
<td>1.00</td>
<td>1.50</td>
<td>0.40</td>
<td>7.00</td>
<td>0.80</td>
<td>0.70</td>
<td>1.60</td>
<td>1.00</td>
<td>1.30</td>
<td>1.40</td>
</tr>
<tr>
<td>24</td>
<td>1.60</td>
<td>0.90</td>
<td>0.80</td>
<td>0.20</td>
<td>4.00</td>
<td>0.50</td>
<td>0.30</td>
<td>1.70</td>
<td>2.10</td>
<td>1.80</td>
<td>3.20</td>
</tr>
<tr>
<td>25</td>
<td>0.90</td>
<td>0.80</td>
<td>2.20</td>
<td>0.30</td>
<td>6.00</td>
<td>0.80</td>
<td>1.00</td>
<td>2.00</td>
<td>0.90</td>
<td>2.60</td>
<td>1.50</td>
</tr>
<tr>
<td>26</td>
<td>0.20</td>
<td>0.10</td>
<td>0.20</td>
<td>0.10</td>
<td>1.00</td>
<td>0.40</td>
<td>0.20</td>
<td>0.50</td>
<td>1.50</td>
<td>0.90</td>
<td>3.40</td>
</tr>
<tr>
<td>27</td>
<td>0.60</td>
<td>0.80</td>
<td>1.10</td>
<td>0.40</td>
<td>6.00</td>
<td>0.90</td>
<td>1.30</td>
<td>2.70</td>
<td>3.20</td>
<td>2.80</td>
<td>6.70</td>
</tr>
<tr>
<td>28</td>
<td>0.10</td>
<td>0.20</td>
<td>0.20</td>
<td>0.10</td>
<td>1.00</td>
<td>0.30</td>
<td>0.30</td>
<td>0.30</td>
<td>1.00</td>
<td>0.30</td>
<td>2.70</td>
</tr>
<tr>
<td>29</td>
<td>0.30</td>
<td>0.20</td>
<td>0.40</td>
<td>0.10</td>
<td>1.00</td>
<td>0.30</td>
<td>0.30</td>
<td>0.70</td>
<td>3.80</td>
<td>1.80</td>
<td>13.60</td>
</tr>
<tr>
<td>30</td>
<td>0.90</td>
<td>0.20</td>
<td>0.30</td>
<td>0.10</td>
<td>1.00</td>
<td>0.10</td>
<td>0.30</td>
<td>1.40</td>
<td>1.00</td>
<td>4.00</td>
<td>16.20</td>
</tr>
</tbody>
</table>

\[
B_{4} = \begin{bmatrix}
0 & 0 & 0 & 0.7071 \\
0 & 0 & 0.5000 & 0 \\
0 & 0 & 0.5000 & 0 \\
0 & 0 & 0.5000 & 0 \\
0 & 0 & 0.5000 & 0 \\
0 & 0 & 0.5000 & 0 \\
0 & 0 & 0.5000 & 0 \\
0 & 0 & 0.7071 & 0 \\
0 & 0 & 0.7071 & 0 \\
0.5774 & 0 & 0 & 0 \\
0.5774 & 0 & 0 & 0 \\
0.5774 & 0 & 0 & 0 \\
0.5774 & 0 & 0 & 0
\end{bmatrix}
\]

The results of the target rotation are given in Table 4.5. The correlation matrix between the four components is:
Table 4.4 Eigenvalues and loadings for the urban pollution data in Table 4.3, using R-mode principal component analysis and the correlation matrix (relation 4.10)

<table>
<thead>
<tr>
<th>Component</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
</tr>
</thead>
<tbody>
<tr>
<td>Eigenvalues($\lambda_k$)</td>
<td>7.557</td>
<td>1.542</td>
<td>0.869</td>
<td>0.569</td>
<td>0.150</td>
<td>0.146</td>
<td>0.069</td>
<td>0.051</td>
<td>0.026</td>
<td>0.018</td>
<td>0.002</td>
</tr>
<tr>
<td>Difference($\lambda_k-\lambda_{k-1}$)</td>
<td>6.014</td>
<td>0.673</td>
<td>0.300</td>
<td>0.419</td>
<td>0.004</td>
<td>0.077</td>
<td>0.077</td>
<td>0.018</td>
<td>0.025</td>
<td>0.008</td>
<td>0.016</td>
</tr>
<tr>
<td>Proportion</td>
<td>0.687</td>
<td>0.140</td>
<td>0.079</td>
<td>0.052</td>
<td>0.014</td>
<td>0.013</td>
<td>0.006</td>
<td>0.005</td>
<td>0.002</td>
<td>0.002</td>
<td>0.000</td>
</tr>
<tr>
<td>Cumulative Proportion</td>
<td>0.687</td>
<td>0.827</td>
<td>0.906</td>
<td>0.958</td>
<td>0.971</td>
<td>0.985</td>
<td>0.991</td>
<td>0.996</td>
<td>0.998</td>
<td>1.000</td>
<td>1.000</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Loadings</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pb</td>
</tr>
<tr>
<td>Fe</td>
</tr>
<tr>
<td>Ca</td>
</tr>
<tr>
<td>K</td>
</tr>
<tr>
<td>Si</td>
</tr>
<tr>
<td>Al</td>
</tr>
<tr>
<td>Na⁺</td>
</tr>
<tr>
<td>NH₄⁺</td>
</tr>
<tr>
<td>Cl⁻</td>
</tr>
<tr>
<td>NO₃⁻</td>
</tr>
<tr>
<td>SO₄²⁻</td>
</tr>
</tbody>
</table>

This indicates that the four components are not orthogonal after the target rotation and that the rotation was oblique.

The four principal components or factors are very similar to the last components found by Hopper (1986), although the analysis made and the target used are different. We have therefore used the same identification. As Hopper (1986) did not give the total mass of each sample, it was impossible to continue the analysis by doing a source scaling.

4.4 Remarks

TPCA, like APCA, requires only slight modifications to PCA. Therefore, the techniques presented in Section 2 (e.g., for obtaining robust principal components, treating missing and below-detection-limit observations, selecting the number of components to retain, etc.) still apply.

The two major differences between TPCA and PCA are, first, that the data are not centered before the extraction of the principal components and second, that target rotation is used. In practice, these two techniques can be used separately with the rest of the PCA techniques. Hence, one can use any orthogonal or oblique rotation in lieu of the target rotation or employ centered observations in the calculation. Note, however, that in such cases, relations 4.6, 4.7, 4.14, and 4.15 have to be modified.

The target rotation technique is a general technique (see Hurley and Cattell, 1962) and can be used with other dimension reduction techniques like factor analysis (FA) that will be discussed later.
Table 4.5 Loading matrix after target rotation for data in Table 4.3. Loadings greater than .3 are underlined.

<table>
<thead>
<tr>
<th></th>
<th>Sulphate</th>
<th>Traffic</th>
<th>Fly Ash</th>
<th>Refuse</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pb</td>
<td>0.198</td>
<td>0.283</td>
<td>0.234</td>
<td>0.145</td>
</tr>
<tr>
<td>Fe</td>
<td>0.013</td>
<td>0.000</td>
<td>0.519</td>
<td>0.062</td>
</tr>
<tr>
<td>Ca</td>
<td>0.044</td>
<td>0.118</td>
<td>0.418</td>
<td>0.155</td>
</tr>
<tr>
<td>K</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.944</td>
</tr>
<tr>
<td>Si</td>
<td>0.000</td>
<td>0.058</td>
<td>0.482</td>
<td>0.082</td>
</tr>
<tr>
<td>Al</td>
<td>0.040</td>
<td>0.016</td>
<td>0.507</td>
<td>0.056</td>
</tr>
<tr>
<td>Na⁺</td>
<td>0.005</td>
<td>0.620</td>
<td>0.000</td>
<td>0.101</td>
</tr>
<tr>
<td>NH₄⁺</td>
<td>0.063</td>
<td>0.590</td>
<td>0.018</td>
<td>0.125</td>
</tr>
<tr>
<td>Cl⁻</td>
<td>0.512</td>
<td>0.245</td>
<td>0.046</td>
<td>0.087</td>
</tr>
<tr>
<td>NO₃⁻</td>
<td>0.395</td>
<td>0.331</td>
<td>0.091</td>
<td>0.113</td>
</tr>
<tr>
<td>SO₄²⁻</td>
<td>0.731</td>
<td>0.000</td>
<td>0.000</td>
<td>0.063</td>
</tr>
</tbody>
</table>

In practice, Q-mode TPCA is only possible if the number of observations is small. In atmospheric chemistry, one usually has \( n >> p \), and the extraction of the eigenvalues and principal components becomes difficult when the number of observations, \( n \), becomes too large. Therefore, R-mode TPCA is more commonly used.

Originally, TPCA was applied to atmospheric chemistry to help identify the different emission sources affecting an observation site and, more specifically, to obtain an apportionment of the observed particulate masses. This last goal can only be achieved, however, if the total masses for the observations are known, which is not always the case. Nevertheless, the technique can still be used, even if the last step, sources scaling, is omitted.

It should be noted that the requirement in Henry's rules for a physically realistic solution (Section 3.1) greatly complicates the search for the initial target loading matrix. In some cases, although target rotation will produce values greater than zero for all elements of the loading matrix, some elements of the scores may be lower than zero. In those cases, the initial loadings matrix has to be modified somewhat or the solution obtained has to be rotated, if possible, by a series of \( p \) simple rotations to arrive at a situation in which all elements of both the loadings and scores matrices are greater than zero (see Shen and Israel, 1989). However, in some cases, it will be impossible to find such a solution. In those cases, one should either accept the solution as it is or use another technique, such as APCA or PMF.

4.5 Computer Software

Programs or functions for the calculation of TPCA are not available in software packages such as SAS® or S-Plus®. It is therefore necessary to develop functions in the development language of those software packages to implement TPCA. This can be done by implementing the relations given in this section. One should note that target rotation is available in SAS® and S-Plus® under the name of Procrustes rotation. However, it is not clear from the available information whether iteration is possible with these programs or not.
A computer program called FANTASIA was developed in the 1980s by Hopke et al. (1983; see also Hopke and Dharmavaram, 1986). It is not known to the present author if this program is still available.

4.6 References

The principal references for TPCA are the doctoral thesis of Hopper (1986) and the following journal papers (in chronological order): Hopke et al. (1980); Albert and Hopke (1980, 1981); Roscoe and Hopke (1981b); Hopke (1981); Liu et al. (1982); Severin et al. (1983); Hwang et al. (1984); and Hopke (1988). All these papers are related to atmospheric particulate measurements in an urban context.

Target rotation was developed for use in the social sciences by Hurley and Cattell (1962). A program called Procrustes was also created (Hurley and Cattell, 1962; Cattell and Khanna, 1977).
5. Subjective Principal Component Analysis (SPCA)

The concept behind Subjective Principal Component Analysis (SPCA) is the same as that behind target rotation. This technique was developed for cases in which one has some idea of what the components should look like and one wants to obtain principal components that reflect that knowledge. The similarity between the two techniques stops there.

The SPCA technique was developed by Korhonen (1984) in the 1980s. The present author does not know of any application of this technique in atmospheric chemistry or other fields of research. Therefore, we will limit ourselves to a brief description of the technique and refer the interested reader to Korhonen's original paper. One possible explanation of why this technique has not been used is that it needs human intervention to influence the course of the calculations. Thus, one can say that it is an interactive technique. To the knowledge of the author, no software exists that implements Korhonen's ideas. With the advent of MicroSoft's Windows and similar operating systems, however, it should now be easy to write an interactive program to implement SPCA.

Assuming that the data matrix, X, has been centered (each variable has a mean of zero) and standardized (each variable has a variance of 1), relation 2.2 can be written:

\[ P = XA^T \] (5.1)

where \( P, X, \) and \( A \) are \((n \times m), (n \times p)\) and \((m \times p)\) matrices respectively. If one writes \( Y = P \) and \( B = A^T \) to conform with Korhonen's notation, relation 5.1 (relation 3.1 in Korhonen, 1984) becomes:

\[ Y = XB \] (5.2)

In addition, it is assumed that the columns of \( Y \) are uncorrelated (i.e., \( Y^TY \) is diagonal). If matrix \( B \) is scaled such that \( B^TRB = I \), where \( R \) is the correlation matrix = \( X^TX \), then \( Y^TY = B^TX^TXB = B^TRB = I \).

The correlation matrix representing the correlations between the original variables and the linear combinations, \( Y \), can be written:

\[ R_{XY} = X^TY = X^TXB = RB \] (5.3)

Note that \( R_{XY} \) is denoted \( U \) in Korhonen (1984).

The basic concept of SPCA is to find \( m \) components that will maximize some of the \( u_i \) (or \( (x_{xy})_i \)) and are such that the relationship \( Y^TY = I \) is valid. Korhonen’s technique is to find component after component and reduce the possible choices to only the valid ones. The reader is referred to Korhonen (1984) for more details.

Finally, as Korhonen point out, SPCA can be used to rotate the solution of a PCA. In a way, this is similar to target rotation. However, unlike target rotation, SPCA is orthogonal.
6. Factor Analysis (FA)

6.1 Introduction

The essential purpose of factor analysis is to describe, if possible, the covariance relationships among many variables in terms of a few underlying, but unobservable, random quantities called factors.

The basic factor model, for the population, can be written as (see relation 1.1):

\[ \mathbf{x} = \mu + \Lambda \mathbf{f} + \mathbf{e} \]  

where \( \mathbf{x} \) is a p-vector of observed variables; \( \mu \) is a p-vector of the mean of the observed variables; \( \Lambda \) is a (p\times m) matrix called the loadings matrix; \( \mathbf{f} \) is a m-vector of unobservable \textit{common factors}; and \( \mathbf{e} \) is a p-vector of unobservable \textit{specific or unique factors}. It is generally assumed that the components of \( \mathbf{e} \) have zero means and are mutually uncorrelated as well as being uncorrelated with the elements of \( \mathbf{f} \). In other words, the covariance matrix of \( \mathbf{e} \) is a (p\times p) diagonal matrix, \( \mathbf{\Lambda} \), and the cross-covariance matrix between \( \mathbf{f} \) and \( \mathbf{e} \) is null. Furthermore, it is usually assumed that the covariance matrix of the common factors, \( \mathbf{\Phi} \), is equal to the identity matrix. In other words, it is assumed that the common factors are orthogonal and standardized.

Relation 6.1 is called a \textit{factor pattern}, and the loadings matrix, \( \Lambda \), can be referred to as the \textit{pattern loadings matrix}. Often we are interested in the correlation between the observed variables and the common factors. A matrix of such correlations is called a \textit{factor structure matrix}, or more briefly, a \textit{structure matrix}. Both the structure and pattern matrices are needed for a complete solution. Though in general the elements of a structure matrix are different from the coefficients of a pattern matrix, in the case of uncorrelated and standardized factors the two are identical. In such cases, the factor loadings \( \lambda_{ij} \) are the correlations between the jth common factors and the ith observed variables. Such an interpretation cannot hold if the common factors are not orthogonal; in general, the structure and pattern matrices will then be different.

Using these hypotheses, one can write that (see Section 6.2):

\[ \Sigma = \mathbf{xx}^T = \Lambda \Lambda^T + \mathbf{\Delta}. \]  

(6.2)

This relation shows that, in the factor analysis model, one tries to divide the covariance between the observed variables into two components, namely: (1) a part that is common to the variables and (2) a part that is unique to each variable. It is interesting to note that in principal component analysis it is the total variance that we try to explain, whereas in factor analysis it is the covariance between the observed variables. This shows that the two techniques are fundamentally different. As we will see later, relation 6.2 is the basic relation in trying to fit the factor model.

For a sample of \( n \) observations, model 6.1 can be written as:

\[ (\mathbf{X} - \mathbf{M})^T = \Lambda \mathbf{F} + \mathbf{E} \]  

(6.3)

where \( \mathbf{X} \) (relation 2.21) is (n\times p), \( \mathbf{F} \) is (m\times n), \( \mathbf{M} \) (relation 2.23) is (n\times p), and \( \mathbf{E} \) is (p\times n). Relation 6.2 becomes:
where \( S \) is the (p\times p) variance-covariance matrix (relation 2.25). Fitting the model involves finding the pm elements of matrix \( \Lambda \) and the elements of the diagonal matrix \( \Delta \).

One should note first that there is no guarantee that a solution exists to this problem. For a solution to exist, it is necessary but not sufficient that \( m \) should be small compared to \( p \). A second characteristic is that, if \( m > 1 \), the solution is not unique. To illustrate, let \( T \) be a (m\times m) orthogonal matrix (i.e., \( T^T T = TT^T = I \)). We have then:

\[
(\Lambda T)(\Delta T)^T = \Lambda T T^T A^T = \Lambda \Lambda^T.
\]

Therefore, as in principal component analysis, the solution is within a rotation that can be used to improve its interpretability.

Many techniques to fit relation 6.4 have been designed since the introduction offactor analysis. We will limit our review here to only three of those techniques, namely: principal component estimation, principal factor estimation, and maximum likelihood estimation. It should be noted that in the last technique the number of common factors, \( m \), is assumed to be known.

As in principal component analysis, the correlation matrix can be used in place of the variance-covariance matrix in relation 6.4. On the one hand, the two solutions will be different when principal component estimation and principal factor estimation are used. On the other hand, they will be essentially the same when maximum likelihood estimation is used. In other words, the maximum likelihood solution is scale independent.

Another interesting characteristic of factor analysis is that the scores (i.e., the coordinates of the observations on the new coordinate system) are not available directly as they are in principal component analysis. They must be estimated. Two approaches to the scores will be presented here.

When doing a factor analysis of sampled data, one usually proceeds as follows:

1. Estimate the variance-covariance matrix or the correlation matrix from the data.
2. Select the number of common factors, \( m < p \). Usually, that number will be much smaller than the number of observed variables.
3. Select one of the techniques to solve the problem and obtain estimates of \( \Lambda \) and \( \Delta \).
4. Rotate the solution using orthogonal or oblique rotations to obtain a solution that is easier to interpret.
5. Estimate the scores using one of the two methods presented later.

6.2 Technical Details

6.2.1 Factor Analysis Model

Mathematically, the basic factor analysis model for the population can be written as:
\[
\mathbf{x} = \mathbf{\mu} + \mathbf{\Lambda} \mathbf{f} + \mathbf{e}
\]

(6.6)

where

\( x_i \) = jth observed variable

\( \mu_i \) = mean of observed variable i

\( e_i \) = ith unique or specific factor

\( f_j \) = jth common factor

\( \mathbf{\Lambda} = (p \times m) \) matrix of unknown constants called factors loadings,

\[
\mathbf{\Lambda} = \begin{bmatrix}
\lambda_{11} & \lambda_{12} & \cdots & \lambda_{1m} \\
\lambda_{21} & \lambda_{22} & \cdots & \lambda_{2m} \\
\vdots & \vdots & \ddots & \vdots \\
\lambda_{p1} & \lambda_{p2} & \cdots & \lambda_{pm}
\end{bmatrix}
\]

(6.7)

There are p unique factors, and it is assumed that the unique part of each variable is uncorrelated with the unique parts of the other variables or with the common factors. Therefore, the variance-covariance matrix of \( \mathbf{e} \) is a diagonal matrix that is written as:

\[
\mathbf{\Delta} = \begin{bmatrix}
\delta_1^2 & 0 & \cdots & 0 \\
0 & \delta_2^2 & \cdots & 0 \\
\vdots & \vdots & \ddots & \vdots \\
0 & 0 & \cdots & \delta_p^2
\end{bmatrix}
\]

(6.8)

and the covariance between \( \mathbf{e} \) and \( \mathbf{f} \) is zero:

\[
\text{cov}(\mathbf{e}, \mathbf{f}^T) = 0.
\]

(6.9)

It is also assumed that the unique and common factors have zero means. The variance-covariance matrix, \( \Sigma \), of the observed variables \( \mathbf{x} \) is equal to:

\[
\Sigma = (\mathbf{x} - \mathbf{\mu})(\mathbf{x} - \mathbf{\mu})^T
\]

\[
= (\mathbf{\Lambda f} + \mathbf{e})(\mathbf{\Lambda f} + \mathbf{e})^T
\]

\[
= (\mathbf{\Lambda f} + \mathbf{e})(\mathbf{\Lambda f}^T + \mathbf{e}^T)
\]

\[
= \mathbf{\Lambda f f}^T \mathbf{\Lambda}^T + \mathbf{e e}^T + \mathbf{\Lambda f e}^T + \mathbf{e}^T \mathbf{e}
\]

\[
= \mathbf{\Lambda \Phi \Lambda}^T + \mathbf{\Delta}
\]

(6.10)

where \( \Phi = \mathbf{f f}^T \) is the variance-covariance matrix of the common factors. Relation 6.10 is usually simplified by assuming that the common factors are standardized and orthogonal (i.e., \( \Phi = \mathbf{I} \)). Relation 6.10 becomes:

\[
\Sigma = \mathbf{\Lambda} \mathbf{\Lambda}^T + \mathbf{\Delta}
\]

(6.11)
The factor analysis problem becomes the following: given an estimation of \( \Sigma \), one has to find \( \Lambda \) and \( \Delta \) such that relation 6.11 is true.

Before discussing the methods to solve that problem, a few additional concepts and terms have to be introduced. As a consequence of relation 6.11, one has:

\[
\text{variance}(x_i) = \sigma_{ii} = \sum_{j=1}^{p} \lambda_{ij}^2 + \delta_i^2 \quad i = 1, \ldots, p
\]  

(6.12)

where \( \sigma_{ii} \) is the \( i \)th diagonal element of \( \Sigma \). The quantity

\[
h_i^2 = \sum_{j=1}^{p} \lambda_{ij}^2
\]

(6.13)

is called the communality of the \( i \)th variable, while \( \delta_i^2 \) is termed the uniqueness of the \( i \)th variable (\( i = 1, 2, \ldots, p \)). By definition:

\[
\text{total variance} = \text{tr}(\Sigma) = \sum_{i=1}^{p} \sigma_{ii} = \sum_{i=1}^{p} (h_i^2 + \delta_i^2) = \sum_{i=1}^{p} \sum_{j=1}^{p} \lambda_{ij}^2 + \sum_{j=1}^{p} \delta_i^2 = V + \delta^2
\]  

(6.14)

where \( V = \sum_{i=1}^{p} \sum_{j=1}^{p} \lambda_{ij}^2 \) is the total communality and \( \delta^2 = \sum_{j=1}^{p} \delta_i^2 \).

Without any loss of generality, the original variables may be assumed to be standardized (i.e., \( \sigma_{ii} = 1 \) for \( i = 1, 2, \ldots, p \)), so that the variance-covariance matrix, \( \Sigma \), becomes the \((p \times p)\) correlation matrix, \( \Gamma \), and relation 6.11 becomes:

\[
\Gamma = \Lambda \Lambda^T + \Delta
\]

(6.15)

or

\[
1 = h_i^2 + \delta_i^2 \quad \text{for } i = 1, \ldots, p
\]

\[
\rho_{ii} = \sum_{j=1}^{p} \lambda_{ij} \lambda_{ji} \quad \text{for } i \neq l
\]

(6.16)

where \( \rho_{ij} \) is an element of the correlation matrix \( \Gamma \).

The \((p \times p)\) matrix \( \Gamma^* \), whose diagonal elements are the communalities \( h_i^2 \) and off-diagonal elements are the correlation coefficients \( \rho_{ij} \) between the pairs of the observed variables, is called the reduced correlation matrix. This matrix plays an important role in the principal factor method of solving the factor analysis problem. This matrix has the following properties:
i) if every $\delta_i^2 > 0$, the diagonal elements of $\Gamma^*$ are less than 1;

ii) the rank of $\Gamma^*$ = the minimum number of linearly independent factors required for reproducing the correlations among the observed variables = the dimensionality of the factor space;

iii) if $h_i^2 = 1$ for all $i$, the rank of $\Gamma^* = p$, and no reduction of dimensionality is accomplished by factor analysis.

6.2.2 Estimation of Factor Analysis Model

We will describe here in general terms three possible techniques for solving the factor analysis equation (relation 6.11), namely: (1) the principal component method, (2) the principal factor method, and (3) the maximum likelihood method. The first technique is not often used in practice. However, it is a good introduction to the principal factor method. The last two methods are the most widely used and studied techniques for extracting factors. The maximum likelihood method is the only method for factor extraction that currently provides a sound statistical basis for testing the adequacy of the basic common factor analysis model. Another possible technique for solving the factor analysis problem is the minimum residuals (MINRES) method described in Harman and Jones (1966) and Harman (1967, 1977). Interested readers are referred to these references for more details.

6.2.2.1 Principal Component Method

Singular matrix decomposition, as described in Section 2.2.3, can be applied to any matrix and therefore to the variance-covariance matrix $\Sigma$ (estimated by $S$, relation 2.25) or the correlation matrix $\Gamma$ (estimated by $R$, relation 2.33). Because these matrices are symmetric, one has $U = V$. We can therefore write:

$$\Sigma = ULU^T$$

where $L$ is a $(p \times p)$ diagonal matrix whose elements are equal to the eigenvalues of $\Sigma$:

$$L = \begin{bmatrix}
\lambda_1 & 0 & \cdots & 0 \\
0 & \lambda_2 & \cdots & 0 \\
\vdots & \vdots & \ddots & \vdots \\
0 & 0 & \cdots & \lambda_p
\end{bmatrix}$$

where the $\lambda_i$ are the eigenvalues of $\Sigma$ and $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p \geq 0$. $U$ is such that $UU^T = I$. Relation 6.17 can be written:

$$\Sigma = ULL^T \quad \Sigma^T$$

$$= ULL^T (UL^T)^T$$

$$= ULL^T (UL^T)^T + 0$$

$$= \Gamma^* (\Gamma^*)^T + 0$$

61
where $L^* = U L^{1/2}$. Relation 6.21 is identical to 6.15 if one writes $\Lambda = L^*$ and $\Delta = 0$. Although this representation of $\Sigma$ is exact, it is not very useful. It employs as many common factors as there are observed variables and does not allow for any variation in the specific factors $e$ in relation 6.6.

One approach to improve the usefulness of the model is to neglect the contribution of the last ($p-m$) eigenvectors when the corresponding eigenvalues are small. Relation 6.21 becomes:

$$\Sigma = m L^* (m L^*)^T + \Delta$$

where $\Delta$ is a diagonal matrix with elements $\delta_i^2 = \sigma_{ii} - \sum_{j=1}^{m} \hat{\lambda}_{ij}^2$ for $i = 1, 2, \ldots, p$ and $\ell_y$ are the elements of $\Sigma L^*$. In this method, $\hat{\Lambda}$ is estimated by $m L^*$; $\hat{\Delta}$ by the diagonal matrix with elements equal to $\delta_i^2 = \sigma_{ii} - \sum_{j=1}^{m} \hat{\lambda}_{ij}^2$.

The communalities are estimated as:

$$h_i^2 = \sum_{j=1}^{m} \frac{\ell_{ij}^2}{\ell_y^2}.$$  (6.23)

Note that the same technique can be used for the correlation matrix by replacing $\Sigma$ by $\Gamma$.

In summary, the principal component method consists of: (1) extracting the eigenvalues and standardized eigenvectors from the variance-covariance matrix or from the correlation matrix; (2) keeping only the first $m$ components; and (3) estimating the loadings matrix and communalities using the theory given in this section.

Note that the techniques given in Section 2.2.10 to estimate the number of components to retain can be used to determine $m$.

6.2.2.2 Principal Factor Method

The principal factor method is based on the use of the reduced correlation matrix, $\Gamma^*$, $R^*$. Note that a similar argument can be developed if the variance-covariance matrix is used. We will limit the discussion here to the correlation matrix.

Let us suppose that we have estimates for the specific variances, $\hat{\delta}_i^2$. Then, the communalities, $\hat{h}_i^2 = 1 - \hat{\delta}_i^2$, are known and the reduced correlation matrix, $R^*$:

$$R^* = \begin{bmatrix} \hat{h}_1^2 & \hat{r}_{12} & \cdots & \hat{r}_{1p} \\
\hat{r}_{12} & \hat{h}_2^2 & \cdots & \hat{r}_{2p} \\
\vdots & \vdots & \ddots & \vdots \\
\hat{r}_{1p} & \hat{r}_{2p} & \cdots & \hat{h}_p^2 \end{bmatrix}$$  (6.24)

is known. Now, apart from sampling variation, all the elements of the reduced sample correlation matrix $R^*$ should be accounted for by the $m$ common factors. In particular, $R^*$ can be factored as:
\[ \mathbf{R}^* \equiv \mathbf{L} \mathbf{L}^T \]

where \( \mathbf{L}_t \) is the estimated loadings. Therefore, the same technique as in the principal component method is applied to the reduced correlation matrix. New estimates for the specific variances, \( \hat{\lambda}_i^2 \), and the communalities, \( \hat{h}_i^2 \), are obtained.

This procedure can be used iteratively, with the new communality becoming the initial communality in relation 6.24 for the next stage.

Note that if all the communalities are equal to 1, one obtains the solution to the principal component analysis.

If some or all of the diagonal elements of \( \mathbf{R}^* \) are less than 1, then \( \mathbf{R}^* \) need not be positive semi-defined. Hence some of the eigenvalues may be negative, with the consequence that the vectors of factor loadings associated with these negative eigenvalues will be imaginary. In practice, one discards the negative eigenvalues and the associated imaginary vector loadings. However, since the sum of the eigenvalues of \( \mathbf{R}^* \) equals the total communality, the sum of just the positive eigenvalues will exceed the total communality if there are any negative eigenvalues at all. Hence, in extracting the factors, one would not continue until their number \( m \) is as large as the number of positive eigenvalues but, rather, would stop when the sum of the eigenvalues is close to and lower than (or equal to) the total communality. This value of \( m \) can be used as a maximum for the number of factors to retain.

As mentioned earlier, initial values for the communalities must be obtained before the principal factor method can be used. Some of the possible estimates for the communalities are presented in the next section.

6.2.2.3 Communality Estimation

Many techniques can be used to estimate the communalities. Four of the available techniques will be presented here. They are:

1. \( \hat{h}_i^2 = \text{the highest observed positive correlation of variable } i \) with the remaining \((p-1)\) variables. This is equal to the largest positive element in the \( i \)th row (column) of \( \mathbf{R} \).

2. \( \hat{h}_i^2 = \frac{r_{ik} r_{ij}}{r_{jk}} \) where the variables \( k \) and \( j \) are the two variables that have the highest correlation with variable \( i \).

3. \( \hat{h}_i^2 = \text{the average (presumed positive) of the observed correlations of variable } i \) with the other variables = \( \sum_{l=1}^{p} r_{il} / (p-1) \).

4. \( \hat{h}_i^2 = \text{the square of the multiple correlation coefficient of variable } i \) with the other variables = \( 1 - \frac{1}{r_{ii}^2} \), where the \( r_{ij} \) are the elements of \( \mathbf{R}^{-1} \).
By definition for the correlation matrix, the communality $h_i^2$ is between 0 and 1. With the second technique it may happen that the value obtained for $h_i^2$ is greater than 1. The usual procedure in this case is to replace that value by 0.99 or 1.

The fourth estimator is the most popular choice when using the correlation matrix.

6.2.2.4 Maximum Likelihood Method

Thus far no assumption has been made about the distribution of the observed data. In the maximum likelihood method, however, it is assumed that the p-dimensional vectors of observed responses $X$ have a nonsingular p-dimensional normal distribution, with mean vector $\theta$ and variance-covariance matrix $\Sigma = \Lambda \Phi \Lambda^T + \Delta$. The loadings matrix $\Lambda$ has $m$ common factors, where $m$ has been specified before extraction of the factor estimates.

Essentially, given specific distributional assumptions, maximum likelihood estimators are estimators of the parameters most likely to have generated the observed data. The following description by Mulaik (1972) is a good description of the maximum likelihood method:

The idea of a maximum likelihood estimator is this: We assume that we know the general form of the population distribution from which a sample is drawn. For example, we might assume the population distribution is a multivariate normal distribution. But what we do not know are the population parameters which give this distribution a particular form among all possible multivariate normal distributions. In the absence of such knowledge, however, we can take arbitrary values and treat them as if they were the population parameters and ask ourselves what is the likelihood... of observing certain values for the variables on a single observation drawn from such a population. If we have more than one observation, then we can ask what is the joint likelihood of obtaining such a sample of observation vectors? Finally we can ask: What values for the population parameters make the sample observations have the greatest joint likelihood? When we answer this question, we will take such values to be maximum likelihood estimators of the population parameters.

If $S$ is an estimate of the variance-covariance matrix based on the observed data (relation 2.25), the log-likelihood function, $\Phi$, of $\Lambda$, $\Phi$, and $\Delta$ is equal to:

$$
\Phi = -\frac{1}{2} n \left[ \ln(|\Sigma|) + \text{trace}(\Sigma^{-1}S) \right]
$$

$$
= -\frac{1}{2} n \left[ \ln(|\Lambda \Phi \Lambda^T + \Delta|) + \text{trace}\left((\Lambda \Phi \Lambda^T + \Delta)^{-1}S\right) \right]
$$

(6.26)

where $|X|$ indicates the determinant of matrix $X$ and $\Phi$ is the variance-covariance matrix of the common factors. We must now find the values of the elements of $\Lambda$, $\Phi$, and $\Delta$ which maximized $\Phi$. This is equivalent to minimizing

$$
F(\Lambda, \Phi, \Delta) = \ln|\Sigma| + \text{trace}(\Sigma^{-1}S) - \ln|S| - p
$$

(6.27)
where \( n \) times the minimum value of \( F \) gives the likelihood ratio test statistics of goodness-of-fit.

To minimize the function \( F \), we need to take its partial derivatives with respect to the elements of \( \Lambda \) and the diagonal elements of \( \Delta \). The equations, after some simplifications, are:

\[
\frac{\partial F}{\partial \Lambda} = 2 \Sigma^{-1} (\Sigma - S) \Sigma^{-1} \Lambda
\]

and

\[
\frac{\partial F}{\partial \Delta} = \text{diag}((\Sigma^{-1} (\Sigma - S) \Sigma^{-1}))
\]

where \( \text{diag}(\cdot) \) represents the diagonal matrix formed from \( (\cdot) \) by replacing all non-diagonal elements of \( (\cdot) \) with zeros. Note that there are \( p \) equations and we have \( 1/2p(p+1) \) distinct elements in \( \Sigma \). There are, in total, \( pm \) parameters in \( \Lambda \), \( 1/2m(m+1) \) parameters in \( \Phi \), and \( p \) parameters in \( \Delta \). Some of those parameters may be fixed. For a non-trivial solution to the system, the number of free parameters must be less than \( 1/2p(p+1) \). In other words, without some restrictions on the system there is a basic indeterminacy. Specifically, if \( T \) is any non-singular \((m \times m)\) matrix, then:

\[
F(\Lambda T^{-1}, T\Phi T^T, \Delta) = F(\Lambda, \Phi, \Delta)
\]

The problem is that the maximum likelihood estimates of \( \Lambda \) and \( \Phi \) are not unique, since their elements are not independent of one another. Thus, any orthogonal rotation of the factors in the relevant common space will give a new set of factors, which will satisfy \( \Sigma = \Lambda \Phi \Lambda^T + \Delta \). A necessary, though not sufficient, condition is that the number of fixed elements be greater than or equal to \( m^2 \).

One convenient set of restrictions corresponds to:

\[
\Phi = I
\]

and

\[
\Lambda^T \Lambda^{-1} \Lambda = \text{diag}(\Omega)
\]

Relation 6.31 and 6.32 restrict the \( 1/2m(m+1) \) and \( 1/2m(m-1) \) terms respectively. Therefore, a total of \( m^2 \) restrictions is imposed. One consequence of these restrictions is that the elements in \( \Lambda \) can be solved for analytically. Iterative procedures need to be used for only the \( p \) nonzero elements of \( \Lambda^T \Lambda^{-1} \Lambda = \Omega \).

Note that a simple interpretation can be given to the elements of the diagonal matrix \( \Omega \). If the variates are rescaled so that the residual variance of each is unity, then \( \lambda_i^2/\delta_i^2 \), where \( \lambda_i \) is an element of \( \Lambda \), represents the part of the variance of variable \( i \) that is due to the \( r \)th common factor. If we consider all \( p \) variates, the total variance of \( X \) due to the \( r \)th common factor is \( \sum_i \lambda_i^2/\delta_i^2 \). This is the \( r \)th diagonal element of \( \Lambda^T \Lambda^{-1} \Lambda \). Our choice of common factors is such that the first makes a maximum contribution to the variance in \( X \).
while the second makes a maximum contribution subject to being uncorrelated with the first common factor, and so on.

The minimization of $F$ is made in two stages. First, the conditional minimum of $F$ for a given $\Lambda$ is found. The basic equation to solve is:

$$S \Delta^{-1} \Lambda = \Lambda (I + \Omega)$$

(6.33)

where $\Omega = \Lambda^T \Delta^{-1} \Lambda$. Let $\Lambda_0$ be the value of $\Lambda$ that satisfies that equation. The second stage requires that the matrix $\partial F / \partial \Lambda$ be found when evaluated at $\Lambda = \Lambda_0$. The solution can be expressed as:

$$\text{diag}(\Lambda_0 \Lambda_0^T + \Delta - S) = 0$$

(6.34)

or

$$\Delta = \text{diag}(S - \Lambda_0 \Lambda_0^T)$$

(6.35)

Though this last relation seems simple, it can only be solved by a numerical interactive method. More details can be found in Lawley and Maxwell (1971), Jöreskog and Lawley (1968), and Jöreskog (1977).

6.2.3 Estimation of scores matrix

In atmospheric chemistry as in other fields, the user of factor analysis is interested in knowing the position of the observed samples in the space of the common factors. Therefore, one would like to calculate the projection of each observation on each of the common factors. In PCA, as we have seen in Section 2, the factor scores can be calculated directly as linear combinations of the original or standardized variables (relations 2.22 or 2.36). Unfortunately, in the factor model this cannot be done, since there is no exact solution. The reason for this relates to the indeterminacy problem found in the common factor model. Thus, factor scores cannot be calculated directly, but must be estimated. We will describe two techniques that are often used. Of the two methods presented here, neither can be recommended as uniformly superior. A third, more recent technique is based on the EM algorithm (Demster et al., 1977). It is a maximum likelihood iterative algorithm and treats the scores as data which are 100% missing. As this algorithm has not been implemented, to the author’s knowledge, in any well-known, currently available software, we will not comment further on this technique. More details can be found in Basilevsky (1994).

6.2.3.1 The Weighted Least Squares Method

Suppose that the mean vector, $\mu$, the factor loadings matrix, $\Lambda$, and the specific variance, $\Lambda$, are known for the factor model (relation 6.1 or 6.6); i.e., the factor model has been solved:

$$\mathbf{x} - \mu = \Lambda \mathbf{f} + \mathbf{e}$$

(6.36)

If we consider the specific factors, $\mathbf{e}$, as errors, this relation is similar to a regression equation without the intercept. Because the $\text{var}(e_i) = \sigma_i^2$, $i = 1, 2, \ldots, p$, need not be equal, Bartlett
(1937) has suggested that weighted least squares could be used to estimate the common factor values, \( f \).

The weighted least squares estimate of \( f \) for \( j \)th samples, \( \hat{f}_j \), is equal to:

\[
\hat{f}_j = (\hat{\Lambda}^T \hat{\Delta}^{-1} \hat{\Lambda})^{-1} \hat{\Lambda}^T \hat{\Delta}^{-1} (x_j - \bar{x})
\]

(6.37)

where \( \hat{\Lambda}, \hat{\Delta} \) and \( \bar{x} \) are estimations of \( \Lambda, \Delta, \) and \( \mu \) (see Johnson and Wichern, 1982).

If the maximum likelihood method is used to estimate the factor model, we then have that \( \Lambda^T \Delta^{-1} \Lambda = \Omega \), where \( \Omega \) is a diagonal matrix. In that case, relation 6.37 can be written as:

\[
\hat{f}_j = \hat{\Omega}^{-1} \hat{\Lambda}^T \hat{\Delta}^{-1} (x_j - \bar{x}_j).
\]

(6.38)

If the correlation matrix is used to extract the factor, relation 6.38 becomes:

\[
\hat{f}_j = \hat{\Omega}^{-1} \hat{\Lambda}^T \hat{\Delta}^{-1} \mathbf{D}^{-1/2} (x_j - \bar{x}_j)
\]

(6.39)

where \( \mathbf{D} \) is defined by relations 2.25 and 2.32.

The factor scores generated by those relations have mean vector \( \mathbf{0} \) and zero sample covariances.

If the loadings are rotated (i.e., \( \hat{\Lambda}^* = \hat{\Lambda} \mathbf{T} \)), the subsequent factor scores, \( \hat{f}_j^* \), are related to \( \hat{f}_j \) by \( \hat{f}_j^* = \mathbf{T}^T \hat{f}_j \), \( j = 1, 2, \ldots, n \).

Note that, if the principal component method is used, it is customary to generate factor scores using an unweighted (ordinary) least squares procedure. Implicitly, this amounts to assuming that the \( \delta_j^2 \) are equal or nearly equal. In that case, the scores thus obtained are the same as the scaled scores obtained using principal component analysis.

6.2.3.2 The Regression Method

Starting with relation 6.36, we assume initially that the loadings matrix \( \Lambda \) and the specific variance matrix \( \Delta \) are known. When the common factors \( f \) and the specific factors \( e \) are jointly normally distributed with means and covariances given in Section 6.2.1, the linear combination \( x - \mu = \Lambda \mathbf{f} + e \) has p-dimensional normal distribution with mean zero and covariance \( \Lambda \Lambda^T + \Delta \). Moreover, the joint distribution of \( (X-\mu) \) and \( \mathbf{f} \) is a \((m+p)p\)-multidimensional normal distribution with mean \( \mathbf{0} \) (i.e., a \((m+p)p \times p \) matrix of zeros) and covariance matrix:

\[
\Sigma^* = \begin{bmatrix}
\Sigma &=& \Lambda \Lambda^T + \Delta & \Lambda \\
(\Lambda \Lambda^T) & (\Lambda \mu) \\
(\Lambda \mu) & (\mu \mu)
\end{bmatrix}
\]

(6.40)
The conditional distribution of $f \mid x$ is multivariate normal (see Johnson and Wichern, 1982) with:

$$ mean = E(f \mid x) = \Lambda^T \Delta^{-1} (x - \mu) = \Lambda^T \left( \Lambda \Lambda^T + \Delta \right)^{-1} (x - \mu) $$ (6.41)

where $E(.)$ is the expected value of $(.)$; and

$$ covariance = Cov(f \mid x) = I - \Lambda^T \Delta^{-1} \Lambda = I - \Lambda^T \left( \Lambda \Lambda^T + \Delta \right)^{-1} \Lambda. $$ (6.42)

The quantities $\Lambda^T \left( \Lambda \Lambda^T + \Delta \right)^{-1}$ in relation 6.41 are the coefficients in a (multivariate) regression of the factors on the variables. Estimates of these coefficients produce factor scores that are analogous to the estimates of the conditional mean values in multivariate regression analysis. Consequently, given any vector of observations $x_j$ and taking the maximum likelihood estimates $\hat{\Lambda}$ and $\hat{\Lambda}$ as the true values, the scores vector for the $j$th sample is given by:

$$ \hat{f}_j = \Lambda^T \Sigma^{-1} (x_j - \bar{x}) = \hat{\Lambda}^T \left( \hat{\Lambda} \hat{\Lambda}^T + \Delta \right)^{-1} (x_j - \bar{x}) $$ (6.43)

$j = 1, 2, ..., n.$

The calculation of $\hat{f}_j$ in 6.43 can be simplified by using the following matrix identity:

$$ \hat{\Lambda} \left( \hat{\Lambda} \hat{\Lambda}^T + \Delta \right)^{-1} = \left( I + \hat{\Lambda}^T \hat{\Lambda}^{-1} \hat{\Lambda} \right)^{-1} \hat{\Lambda}^T \hat{\Lambda}^{-1}. $$ (6.44)

This identity permits the comparison of the factor scores generated by the regression method (relation 6.43) and the ones generated by the least squares procedure (relation 6.37). Using relation 6.44, we have:

$$ \hat{f}_j^R = \hat{\Lambda}^T \left( \hat{\Lambda} \hat{\Lambda}^T + \Delta \right)^{-1} (x_j - \bar{x}) = \left( I + \hat{\Lambda}^T \hat{\Lambda}^{-1} \hat{\Lambda} \right)^{-1} \hat{\Lambda}^T \hat{\Lambda}^{-1} (x_j - \bar{x}) $$ (6.45)

Therefore,

$$ \hat{\Lambda}^T \hat{\Lambda}^{-1} (x_j - \bar{x}) = \left( I + \hat{\Lambda}^T \hat{\Lambda}^{-1} \hat{\Lambda} \right) \hat{f}_j^R $$ (6.46)

The least squares estimates of the factor scores are (relation 6.37):

$$ \hat{f}_j^{LS} = (\hat{\Lambda}^T \hat{\Lambda}^{-1} \hat{\Lambda}^T \hat{\Lambda}^{-1} (x_j - \bar{x}) = (\hat{\Lambda}^T \hat{\Lambda}^{-1} \hat{\Lambda}) \hat{f}_j^R = (I + \hat{\Lambda}^T \hat{\Lambda}^{-1} \hat{\Lambda}) \hat{f}_j^R $$ (6.47)

For the maximum likelihood estimates, $(\hat{\Lambda}^T \hat{\Lambda}^{-1} \hat{\Lambda})^{-1} = \hat{\Omega}^{-1}$ and the elements of this diagonal matrix are close to zero. The regression and generalized least squares methods will give nearly the same factor scores.

In an attempt to reduce the effects of a (possibly) incorrect determination of the number of factors, practitioners tend to calculate the factors scores in relation 6.43 by using $S$ (the original sample covariance matrix; relation 2.25) instead of $\hat{\Sigma} = \hat{\Lambda} \hat{\Lambda}^T + \hat{\Lambda}$. The factor scores are then given by:

$$ \hat{f}_j = \hat{\Lambda}^T S^{-1} (x_j - \bar{x}) $$ (6.48)
for \( j = 1, 2, \ldots, n \). If the correlation matrix is used, we have:

\[
\hat{f}_j = \hat{\Lambda}^T R^{-1/2} D^{-1/2} (x_j - \bar{x})
\]

(6.49)

where \( R \) is given by relation 2.33 and \( D \) by relations 2.25 and 2.32.

Again, if a rotation is applied to the loadings, the factor scores are also rotated using the relation \( \hat{f}_j^r = T^r \hat{f}_j \).

6.2.4 Confidence Intervals and tests

When using a statistical technique, one would like to have access to estimates of confidence intervals for the different fitted parameters of the model or be able to test their significance. One would also like to be able to test for the validity of the model. In factor analysis, such theories exist only for the maximum likelihood method. Even in this case, only some large-sample results are available. In the remainder of this section, we will assume that the data are from a multivariate normal distribution.

As mentioned in the discussion of PCA, the jackknife and bootstrap techniques can be used to estimate confidence intervals for the parameters of the factor model. One of the advantages in using these techniques is that they can be used for all the different methods of solving factor analysis equations. We will discuss these techniques in Section 9.

6.2.4.1 A Large Sample Test for the Number of Common Factors

In this section, we will assume that the assumption made about the distribution of the data in the maximum likelihood method holds. Suppose also that an \( m \) common factors model is the correct one. In this case, \( \Sigma = \Lambda \Lambda^T + \Delta \), and testing the adequacy of the \( m \) common factor model is equivalent to testing:

\[
H_0: \quad \Sigma = \Lambda \Lambda^T + \Delta \quad \text{versus} \quad H_1: \quad \Sigma \text{ any other positive define matrix .}
\]

(6.50)

(6.51)

When \( \Sigma \) does not have any special form, the maximum of the likelihood function (see Johnson and Wichern, 1982, pages 177-178) is proportional to:

\[
|S_n|^{-n/2} e^{-np/2}
\]

(6.52)

where \( \hat{\Sigma} = S_n = [(n-1)/n]S \) and \( |A| \) is the determinant of matrix \( A \).

Under \( H_0 \), \( \Sigma \) is restricted to the form given by relation 6.50. In this case, the maximum of the likelihood function with \( \hat{\mu} = \bar{x} \) and \( \hat{\Sigma} = \hat{\Lambda} \hat{\Lambda}^T + \hat{\Delta} \), where \( \hat{\Lambda} \) and \( \hat{\Delta} \) are the maximum likelihood estimates of \( \Lambda \) and \( \Delta \) respectively, is proportional to:

\[
\left|\hat{\Sigma}\right|^{-n/2} e^{-\frac{1}{2} \text{tr} \left( \hat{\Sigma}^{-1} \left( \sum_{j=1}^{n} (x_j - \bar{x})(x_j - \bar{x})^T \right) \right)}
\]

69
The likelihood ratio statistic for testing $H_0$ is:

$$U = -2 \ln \left[ \frac{\text{maximized likelihood under } H_0}{\text{maximized likelihood}} \right]$$

$$= n \ln \left( \frac{|\hat{\Lambda}^T + \hat{\Delta}|}{|S_n|} \right) + n[\text{trace}((\hat{\Lambda}\hat{\Lambda}^T + \hat{\Delta})^{-1}S_n) - p] \quad (6.54)$$

with

$$\nu - \nu_0 = \frac{1}{2} p(p + 1) - \left[p(m + 1) - \frac{1}{2} m(m - 1) \right] = \frac{1}{2} [(p - m)^2 - p - m] \quad (6.55)$$

degrees of freedom. If $\hat{\Lambda}$ and $\hat{\Delta}$ are the maximum likelihood estimates, the term in brackets in relation 6.54 is equal to zero. Thus we have:

$$U = n \ln \left( \frac{|\hat{\Lambda}^T + \hat{\Delta}|}{|S_n|} \right) \quad (6.56)$$

Barlett (1954) has shown that the chi-squared approximation to the sampling distribution of $U$ can be improved by replacing $n$ by $(n-1-(2p+4m+6)/6)$. Using this correction, we reject $H_0$ at an $\alpha$ level of significance if:

$$(n - 1 - (2p + 4m + 5)/6) \ln \left( \frac{|\hat{\Lambda}^T + \hat{\Delta}|}{|S_n|} \right) > \chi^2_{[(p-m)^2 - p-m]/2} (\alpha) \quad (6.57)$$

provided $n$ and $n-p$ are large. Since the number of degrees of freedom $\frac{1}{2}[(p-m)^2 - p-m]$, must be positive, it follows that

$$m < \frac{1}{2} (2p + 1 - \sqrt{8p + 1}) \quad (6.58)$$

to be able to apply the test.

Note that in practice this test will often indicate a larger number of common factors than can be interpreted in a meaningful way. Therefore, this test should not be used blindly.

6.2.4.2 Confidence Intervals for Loadings

Theoretically, if one uses the maximum likelihood method to solve the factor equations, one should be able to find confidence intervals for the different parameters of the factor model. However, in practice, only large-sample estimates can be obtained and those only in some cases. This is due to the fact that the likelihood is complex and must be solved by numerical iterative methods. As the available algebraic results are complex, they are not given here. The interested reader is referred to Lawley (1967), Lawley and Maxwell (1971; see also Jennrich and Thayer, 1973, for corrections to some of the equations), Archer and Jennrich (1973), Jenrich (1973), and Basilevsky (1994) for more details.
6.2.5 Rotations

As mentioned earlier, the solution obtained for the factor model can be rotated to help in the interpretation of the common factors. Historically, two types of rotations have been used, namely orthogonal and oblique. We will discuss these rotations in more detail in Section 8. In this section, we will limit ourselves to presenting some characteristics of these rotations. Note that target rotation (Section 4.2.4) and subjective rotation (Section 5) should be added to these classical rotations. Target rotation is generally oblique. Subjective rotation by construction is orthogonal.

6.2.5.1 Orthogonal Rotations

The following characteristics of orthogonal rotations should be noted:

1. Factors resulting from an orthogonal rotation remain orthogonal if the initial factors were orthogonal.

2. Any orthogonal rotation method will not alter the values of the communality estimates. However, the proportion of a variable’s variance accounted for by a given factor will be different.

3. Though the total amount of variance accounted for by the common factors does not change, the percentage of the variance accounted for by an individual factor will, in general, change. Thus, no special significance is accorded to the order of the common factors in factor analysis.

6.2.5.2 Oblique Rotations

Oblique rotation methods present more complexities than orthogonal methods. The following considerations should be noted:

1. An oblique rotation will result in common factors that are not orthogonal, even if the initial common factors were. This means that, in general, the pattern and structures matrices will no longer be identical. The pattern loadings are based on the projections of each variable (point) onto each rotated axis by lines parallel to these two axes and may be interpreted as measures of the unique contribution that each factor makes to the variance of the variables. The structure loadings are
based on the projection of each variable (point) onto each rotated axis by lines perpendicular to these two axes and may be interpreted as the simple correlations of the variables with the oblique factors. This is illustrated in Figure 6.1, which shows both pattern and structure representations of the location of a variable in two-factors oblique space.

2. Two types of oblique axes can be delineated - primary and reference. The primary axes refer to the original oblique solution. Rather than interpret the oblique factor axes directly, we define a new coordinate system by drawing though the origin lines perpendicular to each of the original primary axes. The two systems of axes are connected by the fact that the pattern of the primary axes is the structure of the reference axes, and the structure of the primary axes is the pattern of the reference axes. The relationship between these two types of oblique axes is illustrated in Figure 6.2.

3. Statements concerning communality estimates or variance accounted for are no longer applicable with oblique rotations. This is due to the fact that the sums of the squared elements in the loading matrix are not invariant under oblique transformations. Although they are sometimes useful, oblique rotations produced results that are not always easy to interpret. One of the reasons for this is that the standardized loadings cannot be interpreted as correlation coefficients between observed variables and common factors.

6.3 Examples of Factor Analysis

To illustrate factor analysis and the differences between the solutions obtained by the principal factor method and the maximum likelihood method, an example given by Johnson and Wichern (1982) has been selected. The data are from Linden (1977), who conducted a study of Olympic decathlon scores since World War II. Altogether, 160 complete starts were made by 139 athletes. The scores for each of the 10 decathlon events were standardized and a sample correlation matrix was calculated. It is:
Table 6.1 Loadings matrix and uniqueness variance for the factor model solved using the principal factor method and the maximum likelihood method for the Olympic decathlon scores data.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Principal Factor</th>
<th>Maximum likelihood</th>
<th>Specific variances</th>
<th>Principal Factor</th>
<th>Maximum likelihood</th>
<th>Specific variances</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Estimated Factor loadings</td>
<td></td>
<td>$\hat{\delta}_2^2$</td>
<td>Estimated Factor loadings</td>
<td></td>
<td>$\hat{\delta}_2^2$</td>
</tr>
<tr>
<td></td>
<td>$F_1$ $F_2$ $F_3$ $F_4$</td>
<td></td>
<td>$\hat{\delta}_2^2$</td>
<td>$F_1$ $F_2$ $F_3$ $F_4$</td>
<td></td>
<td>$\hat{\delta}_2^2$</td>
</tr>
<tr>
<td>1. 100-m run</td>
<td>0.695 0.239 -0.484 0.103</td>
<td>0.21</td>
<td>-0.090 0.341 0.830</td>
<td>-0.169 0.16</td>
<td>0.695 0.239 -0.484 0.103</td>
<td>0.21</td>
</tr>
<tr>
<td>2. Long jump</td>
<td>0.750 0.149 -0.097 -0.184</td>
<td>0.37</td>
<td>0.065 0.433 0.595</td>
<td>0.275 0.38</td>
<td>0.750 0.149 -0.097 -0.184</td>
<td>0.37</td>
</tr>
<tr>
<td>3. Shot put</td>
<td>0.704 -0.544 0.058 0.181</td>
<td>0.17</td>
<td>-0.139 0.990</td>
<td>0.000 0.00</td>
<td>0.704 -0.544 0.058 0.181</td>
<td>0.17</td>
</tr>
<tr>
<td>4. High jump</td>
<td>0.614 0.065 0.176 -0.322</td>
<td>0.48</td>
<td>0.156 0.406 0.336</td>
<td>0.445 0.50</td>
<td>0.614 0.065 0.176 -0.322</td>
<td>0.48</td>
</tr>
<tr>
<td>5. 400-m run</td>
<td>0.629 0.565 -0.021 0.334</td>
<td>0.17</td>
<td>0.376 0.245 0.671</td>
<td>-0.137 0.33</td>
<td>0.629 0.565 -0.021 0.334</td>
<td>0.17</td>
</tr>
<tr>
<td>6. 100-m hurdles</td>
<td>0.617 0.018 -0.043 -0.256</td>
<td>0.55</td>
<td>-0.021 0.361 0.425</td>
<td>0.388 0.54</td>
<td>0.617 0.018 -0.043 -0.256</td>
<td>0.55</td>
</tr>
<tr>
<td>7. Discuss</td>
<td>0.595 -0.474 0.106 0.224</td>
<td>0.36</td>
<td>-0.063 0.728</td>
<td>0.030 0.19</td>
<td>0.595 -0.474 0.106 0.224</td>
<td>0.36</td>
</tr>
<tr>
<td>8. Pole vault</td>
<td>0.459 0.036 0.241 -0.213</td>
<td>0.68</td>
<td>0.155 0.264</td>
<td>0.229 0.394</td>
<td>0.459 0.036 0.241 -0.213</td>
<td>0.68</td>
</tr>
<tr>
<td>9. Javelin</td>
<td>0.362 -0.258 0.145 0.077</td>
<td>0.78</td>
<td>-0.026 0.441</td>
<td>-0.010 0.098</td>
<td>0.362 -0.258 0.145 0.077</td>
<td>0.78</td>
</tr>
<tr>
<td>10. 1500-m run</td>
<td>0.133 0.438 0.553 0.164</td>
<td>0.46</td>
<td>0.998 0.059</td>
<td>0.000 0.00</td>
<td>0.133 0.438 0.553 0.164</td>
<td>0.46</td>
</tr>
<tr>
<td>Cumulative proportion of total variance explained</td>
<td>0.34 0.36 0.53 0.58</td>
<td></td>
<td>0.12 0.37 0.55 0.61</td>
<td>0.34 0.36 0.53 0.58</td>
<td></td>
<td>0.12 0.37 0.55 0.61</td>
</tr>
</tbody>
</table>

The correlation coefficients between the variables varied between -0.12 and 0.73. Negative correlations are found only in the case of the 1500-m run and three other variables. The highest correlation is that between discuss throwing and the shot put, both of which are hurling sports.

The resulting loadings matrices for a four common factor solution using the principal factor method and the maximum likelihood method are given in Table 6.1. The specific variances, $\hat{\delta}_2^2$, are also indicated.

As in principal component analysis, multiplying the loadings vectors by -1 does not change the solution to the factor equations. If the principal component method is used, the solution is similar to the one obtained using the principal factor method.
The two methods used here produce very different results. For the principal factor method, all events except the 1500-meter run have large positive loadings on the first factor. We can label this factor general athletic ability. The others cannot be easily interpreted, although the second factor seems to contrast running with throwing. For the maximum likelihood method, the 1500-meter run is the only variable with a large loading on the first factor. Thus, it could be labeled a running endurance factor. The second factor seems to be a strength factor, with discuss and shot put loading heavily on that factor. The third factor may be a running speed factor, as the two short length races as well as the long jump are loaded heavily on that factor. The fourth factor is not very easy to identify, but it may have something to do with leg strength.

In the factor analysis model, as we have seen earlier, one tries to reproduce the variance-covariance or correlation matrix using relation 6.2. It would therefore be interesting to see how good the fit is. One therefore calculates:

$$R = \hat{\Lambda} \hat{\Lambda}^T + \hat{\Delta}$$

(6.59)

where $\hat{\Lambda}$ and $\hat{\Delta}$ are the estimated loadings and uniqueness matrices. In the present examples, the following matrix is obtained for the principal factor solution:

\[
\begin{bmatrix}
.000 \\ .005 .000 \\ .006 .012 .000 \\ .016 -.002 .031 .000 \\ .013 -.007 -.004 -.022 .000 \\ -.028 .003 -.016 .005 .027 .000 \\ .008 -.014 .006 -.011 -.009 .023 .000 \\ .011 -.006 -.039 -.005 -.002 .002 .006 .000 \\ -.178 .005 .022 -.037 .025 -.013 -.031 .064 .000 \\ -.017 .008 .003 .025 .016 -.024 .013 -.005 -.028 .000 
\end{bmatrix}
\]

The residual matrix for the maximum likelihood solution is:

\[
\begin{bmatrix}
.000 \\ .000 .000 \\ .000 .000 .000 \\ .012 .002 .000 .000 \\ .000 -.002 .000 -.033 .000 \\ -.012 .006 -.000 .001 .028 .000 \\ .004 -.025 -.000 -.034 -.002 .036 .000 \\ .000 -.009 -.000 .006 .008 -.012 .043 .000 \\ -.018 -.000 -.000 -.045 .052 -.013 .016 .091 .000 \\ .000 .000 .000 .000 .000 .013 .000 .000 .000 
\end{bmatrix}
\]

Those two residual matrices show that, although the solutions for the loadings and the specific variances are very different, they both fit the original correlation matrix quite well. The residuals for the principal component solution are larger than the ones presented here. The residuals vary between -0.254 and 0.114, compared to a range of -0.178 and 0.091 for the two other methods.

We have mentioned earlier that a rotation of the common factors can usually help in their interpretation. A varimax rotation was applied to the two solutions. The results of these rotations are presented in Table 6.2. One can see that after a varimax rotation the two
Table 6.2 Varimax rotated loadings matrix for the factor model solved using the principal factor method and the maximum likelihood method for the Olympic decathlon scores data.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Estimated Factor loadings</th>
<th>Specific variances</th>
<th>Estimated Factor loadings</th>
<th>Specific variances</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Principal Factor</td>
<td>Maximum likelihood</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>$F_1$ $F_2$ $F_3$ $F_4$</td>
<td>$\hat{\delta}_i^2$</td>
<td>$F_1$ $F_2$ $F_3$ $F_4$</td>
<td>$\hat{\delta}_i^2$</td>
</tr>
<tr>
<td>1. 100-m run</td>
<td>0.141 0.816 0.287 -0.132</td>
<td>0.21</td>
<td>0.167 0.857 0.246 -0.138</td>
<td>0.16</td>
</tr>
<tr>
<td>2. Long jump</td>
<td>0.209 0.473 0.600 0.021</td>
<td>0.37</td>
<td>0.240 0.477 0.580 0.011</td>
<td>0.38</td>
</tr>
<tr>
<td>3. Shot put</td>
<td>0.851 0.177 0.247 -0.102</td>
<td>0.17</td>
<td>0.966 0.154 0.200 -0.058</td>
<td>0.00</td>
</tr>
<tr>
<td>4. High jump</td>
<td>0.201 0.149 0.662 0.121</td>
<td>0.48</td>
<td>0.242 0.173 0.632 0.113</td>
<td>0.50</td>
</tr>
<tr>
<td>5. 400-m run</td>
<td>0.072 0.751 0.194 0.468</td>
<td>0.17</td>
<td>0.055 0.709 0.236 0.330</td>
<td>0.33</td>
</tr>
<tr>
<td>6. 100-m hurdles</td>
<td>0.213 0.280 0.568 -0.048</td>
<td>0.55</td>
<td>0.205 0.261 0.589 -0.071</td>
<td>0.54</td>
</tr>
<tr>
<td>7. Discuss</td>
<td>0.771 0.138 0.164 -0.025</td>
<td>0.36</td>
<td>0.697 0.133 0.180 -0.009</td>
<td>0.46</td>
</tr>
<tr>
<td>8. Pole vault</td>
<td>0.195 0.005 0.493 0.177</td>
<td>0.68</td>
<td>0.137 0.078 0.513 0.116</td>
<td>0.70</td>
</tr>
<tr>
<td>9. Javelin</td>
<td>0.442 0.029 0.164 0.041</td>
<td>0.78</td>
<td>0.416 0.019 0.175 0.002</td>
<td>0.80</td>
</tr>
<tr>
<td>10. 1500-m run</td>
<td>0.047 0.021 0.111 0.726</td>
<td>0.46</td>
<td>- 0.056 0.113 0.990</td>
<td>0.00</td>
</tr>
<tr>
<td>Cumulative</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>proportion of</td>
<td>0.17 0.33 0.49 0.58</td>
<td></td>
<td>0.18 0.34 0.50 0.61</td>
<td></td>
</tr>
<tr>
<td>total variance</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>explained</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

...solutions are very close to each other. Note that the rotated principal component solution is also similar to these two. *Shot put, discuss,* and *javelin* load highly on the first factor. Linden (1977) called this factor *explosive arm strength.* Linden (1977) called the second factor *explosive leg strength,* as it is highly loaded by the *high jump,* the *110-meter hurdles,* the *pole vault,* and, to some extent, the *long jump.* The *100-meter run,* the *400-meter run,* and again to some extent the *long jump* load highly on the third factor, called *running speed* by Linden. The *1500-meter run* and to some extent the *400-meter run* load on the last factor. Linden called this factor *running endurance.*

Figure 6.3 presents plots of the rotated maximum likelihood loadings for factor pairs (1,2) and (1,3). The points that correspond to the variables are generally grouped along the factor axes. The relationships between the variables are also evident on these plots. The plots

Figure 6.3 Rotated maximum likelihood loadings for factor pairs (1,2) and (1,3) for the Olympic decathlon data. The numbers in the figures correspond to the variables.
Table 6.3 Summary of the characteristics of the three methods described in Section 6.2.2

<table>
<thead>
<tr>
<th>Feature</th>
<th>Principal Component Method</th>
<th>Principal Factor Method</th>
<th>Maximum Likelihood Method</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Estimate of communalities</td>
<td>Not required</td>
<td>Required</td>
<td>Not required</td>
</tr>
<tr>
<td>2. Dimensionality of common factor spaces</td>
<td>Inferable from manner of computing</td>
<td>Inferable from manner of computing</td>
<td>Assumed for obtaining a solution but then may be statistically tested for adequacy</td>
</tr>
<tr>
<td>3. Distributional assumptions</td>
<td>None specific</td>
<td>None specific</td>
<td>Multivariate normal</td>
</tr>
<tr>
<td>4. Formal statistical inference status</td>
<td>Same as principal component analysis</td>
<td>Not much is known</td>
<td>Large-sample theory is available</td>
</tr>
<tr>
<td>5. Iteration for obtaining the solution</td>
<td>None</td>
<td>Optional (i.e., not required unless one chooses to estimate communalities iteratively)</td>
<td>Necessary</td>
</tr>
<tr>
<td>6. Convergence of iterative procedure</td>
<td>Not applicable</td>
<td>Good</td>
<td>Depends on technique used. Some more modern ones are good.</td>
</tr>
<tr>
<td>7. Scale “invariance”</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
</tr>
</tbody>
</table>

for rotated principal component loadings and principal factor loadings are very similar.

6.4 Remarks

As in principal component analysis, both the variance-covariance and correlation matrices can be used. However, the solutions obtained will be different, except when the maximum likelihood method is used. This is due to the fact, as we have seen, that the maximum likelihood is scale invariant.

The available methods to fit the factor model will usually give results that differ greatly from one another. As mentioned earlier, there is no method that is always the most efficient. As we have seen in the example presented in this section, the three methods can converge to similar solutions after rotation, even if the initial solutions were very different. It is important to remember that the three methods assume different hypotheses and have different characteristics. These are summarized for the three techniques presented here in Table 6.3.

The reader should keep in mind that the results obtained using the maximum likelihood method may not be appropriate if the observed data deviate strongly from a multidimensional normal distribution. In such cases, one should try either to transfer the data to normality or use a method that does not assume normality. Note that the maximum likelihood solution of the factor problem could still be used as a descriptive solution, even if the data do not follow a multivariate normal distribution. However, the tests or confidence intervals are not valid if the distribution of the data is far from normal. The maximum likelihood solution can be used, even when the data are not normally distributed, because it is equivalent to Rao’s (1955) canonical factor solution and Howe’s solution maximizing the determinant of the partial correlation matrix (Morrison, 1976).
When the data do not follow a normal distribution the data can always be transformed to try to improve their normality. This should be done if the maximum likelihood solution is used and especially if one wants to test the results. Another solution is to use the principal factor method and employ the jackknife or bootstrap methods for testing (see Section 9).

The same suggestions made for treating below-detection-limit data in PCA apply also to factor analysis. Depending on whether or not the detection limit is known, we can either treat below-detection-limit data as missing or replace them by some fraction of the detection limit, such as 1/2 or 2/3. If they represent a large percentage of the data, a sensitivity analysis should be made.

The selection of the number of common factors to use is still difficult, as it was in the case of principal component analysis. Some of the techniques presented in Section 2.2.10 can be adapted for use in factor analysis. The informal rules of Section 2.2.10.1 can in general still be used. Cross validation (Section 2.2.10.4) and partial correlation (Section 2.2.10.5) could also be useful tools. We have also seen that if the maximum likelihood method is used, a test exists to verify that the right number of common factors has been selected. However, as mentioned earlier, this test is conservative and tends to indicate a greater number of common factors than are really necessary. The reader should remember that the golden rule in data analysis is to use common sense and one should apply it to the selection of the number of common factors to use. One should be able to interpret these factors physically. Usually one looks for the smallest number of common factors possible, but that goal should not be reached by creating groups of variables that do not make physical sense. One should look carefully at the results before accepting a factor analysis solution.

As in principal component analysis, outlier observations may strongly influence the results. Therefore, one should "robustify" the factor analysis calculation. To obtain a robust factor analysis solution, the techniques mentioned in Section 2.2.5 for making robust estimates of the variance-covariance or the correlation matrices can be used.

Some of the techniques mentioned in Section 2.2.6 for principal component analysis can also be used in factor analysis when there are missing values in the data matrix. All techniques that either estimate the missing data or estimate the variance-covariance or the correlation matrices can still be used.

The concept behind absolute principal component analysis (APCA) can be extended to factor analysis without difficulty. It is therefore possible to fit an absolute factor analysis (AFA) model to data. The extension of target principal component analysis (TPCA) is more problematic. If the principal component method is used, there should be no problem. That should also be true in the case of the principal factor method, although one should consider the impact of using variance-covariance or correlation matrices that are not calculated on centralized data. This was not considered at present by the author. Therefore, the reader should be careful in using any extension of TPCA to FA calculated by the principal factor method. In the case of the maximum likelihood technique, the answer is easy. Because the likelihood has been calculated for centralized data, it is not applicable to the non-centralized data used in TPCA. Therefore, the likelihood would have to be rewritten before any extension of the TPCA to FA. Note that target rotation can always be used. It is the fact that the data are not centralized that creates the difficulties in extending TPCA to FA.

We have seen in Section 6.2.2.2 that the principal factor method gives the same loading matrix as principal component analysis if the communalities are all equal to 1. It is
therefore tempting to use a factor analysis program employing the principal factor method with communalities equal to 1 to obtain a principal component analysis solution. That would work for the eigenvalues or latent roots and the loadings matrix, but the programs usually use the weighted least squares method to calculate the scores matrix. Consequently, this approach may result in scores that differ from those produced by true principal component analysis. The moral is that the reader should be careful when trying to use short cuts.

6.5 Computer Software

As in the case of principal component analysis all commercially available statistical analysis systems, like SAS®, S-Plus®, and SPSS®, contain programs or functions for estimating factors and rotating them. Most of these programs include the principal factor method (with iteration for the calculation of the communalities) and the maximum likelihood techniques. Usually both the variance-covariance or correlation matrix can be used. Note that except for the $\chi^2$-squared test of Section 6.2.4.1, these programs do not calculate confidence intervals for the parameters of the factor models.

For readers who do not have access to similar programs, we have shown that in the case of the principal component method and the principal factor method only a program calculating the eigenvalues and eigenvectors is necessary. Good algorithm for programs to extract eigenvalues and eigenvectors are given in Press et al. (1992a, 1992b). If such a program is available, the principal component and principal factor solutions can be found using the relations presented in this section. Algorithms for the Minres method (Section 6.2.2) can be found in Harman and Jones (1966) and Harman (1967, 1977). An algorithm for the maximum likelihood method could be found in Jöreskog (1977).

6.6 References

Several books on multivariate analysis, such as Morrison (1976), Gnanadesikan (1977), Chatfield and Collins (1980), Johnson and Wichern (1982), and Dillon and Goldstein (1984), give good introductions to factor analysis. Anderson (1984) presents a good summary of the theoretical aspects of factor analysis. Good introductory books on factor analysis are Harman (1967) and Rummel (1970). In addition, Lawley and Maxwell (1971) and Basilevsky (1994) give a good theoretical and practical overview of the subject. Malinowski and Howery (1980) give a presentation of factor analysis designed for chemists.

Examples of the use of factor analysis in atmospheric chemistry are Heidam (1982, 1984) and Crawley and Sievering (1986).
7. Positive Matrix Factorization (PMF)

7.1 Introduction

We have seen earlier when discussing principal component analysis (Section 2.2) that, if $X$ is a $(n \times p)$ observation matrix of $n$ samples of $p$ variables, the general factor model can be written as (relation 2.59):

$$X = PA + E$$ (7.1)

where $P$ is a $(n \times m)$ scores matrix, $A$ is a $(m \times p)$ loadings matrix, and $E$ is a $(n \times p)$ matrix of random errors. We noted earlier that, in atmospheric chemistry, one would generally want to obtain positive elements in both matrix $P$ and matrix $A$ (see Henry, 1987), as the latter describes the strength of different sources and the former their contributions to the observations. APCA and TPCA are two techniques that were developed with that goal in mind. However, as mentioned in the discussion of these techniques, there is no guarantee that the solutions obtained will satisfy those conditions. We have also mentioned the fact that in some cases, but not in all, it is possible to find a series of rotations that would give positive elements for both matrix $P$ and matrix $A$. If such a series of rotations is found, one says that the solution is p-rotatable. In practice, there is no guarantee that the solution found using APCA or TPCA will be p-rotatable.

Recognizing that problem, Paatero et al. (1991) and Paatero and Tapper (1994) proposed a different technique called positive matrix factorization (PMF) to resolve relation 7.1 based on weighted least squares. Briefly, what one tries to do with the PMF technique (see section 7.2 for a more formal description) is to find values for the elements of matrices $P$ and $A$ such that the weighted sum of the squares of the differences between models and observations, $Q$, is minimal, under the condition that the elements are greater than zero. The weights used are estimations of the measurement errors of the observations in $X$. Note that, in the context of atmospheric chemistry, measurement errors include all sources of uncertainties and not only laboratory measurement errors. One should also notice that the observations are not centralized before the analysis as in target principal component analysis. Therefore, the information about absolute zero is treated in the same way as in TPCA.

As in factor analysis, the number of factors, $m$, has to be specified in advance. In practice, that means that one has to fit the model for different values of $m$ and select the one that makes the most sense and is least affected by changes in the measurement error estimations.

PMF has some very interesting characteristics. First, the solution is scale invariant; in other words, the same basic results are found regardless of the units employed. The solutions will only differ by a normalization factor. Secondly, the solution will, in most cases, be unique, in the sense that it cannot be rotated. In other words, any rotation will increase the weighted sum of the squares of the differences between models and observations, $Q$. However, there is no theoretical warranty that the solution is unique. Therefore, one has to verify that the solution is unique. Thirdly, the factors are not generally orthogonal. In almost every cases they will be oblique. This feature has to be taken into account when interpreting the results.
The last characteristic of PMF is that the problem is symmetric in \( P \) and \( A \). Therefore, either of the two matrices could be normalized. This would only result in changes in the elements of the other.

The PMF analysis can be summarized in the following five steps:

1. Create the matrix of observation \( X \).
2. Define a matrix \( S \) in which the elements \( s_{ij} \) are estimations of the error in measuring \( x_i \).
3. Solve the PMF relations (see section 7.2.2) and obtain matrices \( P \) and \( A \).
4. Normalize matrix \( A \) or \( P \).
5. Verify that the solution is unique. If it is not, the solution can be rotated to improve interpretability.

7.2 Technical Details

7.2.1 Mathematical Model

The basic problem to solve in PMF analysis can be written as follows:

Given observation matrix, \( X \), and measurement error matrix, \( S \), and the value of \( m \) (< \( p \)), can we find matrices \( P \) and \( A \) such that

\[
X = PA + E
\]

\( P : n \times m \), \( A : m \times p \)

\( P_{ij} \geq 0 \), \( A_{ij} \geq 0 \) \hspace{1cm} (7.2)

and \( Q \) defined as

\[
Q = \sum_{i=1}^{n} \sum_{j=1}^{p} \frac{E_{ij}^2}{S_{ij}^2}
\]

(7.3)

is a minimum? This is not an easy task because there are two different non-linearities: inequalities and products of unknowns. Therefore, special techniques have to be used.

7.2.2 Algorithms for PMF

One should mention first that the algorithm used by Paatero and Tapper (1994) has evolved with time, and the one presently in use in the PMF2 program (see Section 7.5) can be somewhat different from the ones presented here.

The original base of the algorithms used to solve the PMF equations is alternating regression (AR; see Karjalainen and Karjalainen, 1985 and 1991). Paatero and Tapper (1994) point out that the original AR algorithm was slow.

The speed at which convergence is obtained can be improved by performing extended \((P,A)\) steps, where both \( P \) and \( A \) are changing simultaneously. The iteration consists in repeating the following three basic steps, starting with \( P = P_0 \) and \( A = A_0 \):
1. Solve for $P = P_0 + \Delta P$ while keeping $A = A_0$.

2. Solve for $A = A_0 + \Delta A$ while keeping $P = P_0 + \Delta P$.

3. Solve for the extension coefficient $\alpha$ in $(P_0 + \alpha \Delta P)(A_0 + \alpha \Delta A) = X + E$,
   where $\Delta P$ and $\Delta A$ are as determined from the first and second step.

   Note that $\alpha$ gets quite large by the end of the iterative process. This is a significant improvement over the basic AR algorithm. This algorithm is fast but tricky. It can also handle very large data sets.

   A third algorithm worked out by Paatero and Tapper (1994) consists of simultaneous solutions of $\Delta P$ and $\Delta A$ by minimization of $\|E\|_F$ (i.e., the Frobenius norm of $E = \sqrt{\sum \sum E_{ij}^2}$) in the equation

   $$(P_0 + \Delta P)(A_0 + \Delta A) = X + E$$

   where the second-order term $\Delta P \Delta A$ can be ignored. This algorithm converges rapidly but requires a large amount of memory for a matrix of size $(n + p) \times n \times m^2$. One advantage of this algorithm is that it is possible to estimate the variance of the factor elements. To the knowledge of the present author, it is a variant of this last algorithm that is in use in the PMF2 program (see later).

   The negative entries in these three algorithms must be dealt with. In the original AR algorithm, the negative values were simply truncated to zero. Paatero and Tapper (1994) used a penalty function which was proportional to the square of the negative value. The penalty coefficients were adjusted dynamically; if a component got a positive value, the corresponding penalty coefficient was decreased toward a very small minimum value. If a component got a negative value, the corresponding penalty coefficient was increased towards a fixed maximum value.

   Note that the function $Q$ (relation 7.3) may have a global minimum and local minima. All three algorithms can find a minimum of $Q(P,A)$, but they do not indicate whether the minimum is local or global. Different sets of pseudorandom numbers can be used as starting values for $(P_0,A_0)$ in order to identify the global minimum of $Q$.

### 7.2.3 Error Estimation

As we have seen earlier, estimates of the standard deviation (or absolute precision) of the data, $S_{ij}$, are necessary in PMF (relation 7.3). If the precision of each data point is not known, Paatero (1998) recommends the use of the following relation:

$$S_{ij} = C_1 + C_2 |x_{ij}| + C_3 |x_{ij}|$$

where $C_1$, $C_2$, and $C_3$ are to be selected in advance. The presence of the constant term, $C_1$, guaranties that the uncertainties do not go to zero as the data values decrease. This model also assumes that the uncertainties increase as the data values increase.

Note that in atmospheric chemistry, the columns of data matrix $X$ represent different ions. Therefore, different values for $C_1$, $C_2$, and $C_3$ would usually be used for each column of matrix $S$. 

---

81
Another possibility is to update the values of $S_j$ during the iterative process by using the estimated values, $\hat{x}_y$, in relation 7.4. Paatero (1998) recommends using the following relation in that case:

$$S_j = C_1 + C_2 \sqrt{\max(|x_y|-|\hat{x}_y|)} + C_3 \max(|x_y|-|\hat{x}_y|)$$  \hspace{1cm} (7.5)

where $C_1$, $C_2$, and $C_3$ remain constant.

### 7.2.4 Missing Data

In PMF, missing data can be handled quite naturally by replacing them by, say, the mean value or median value for that variable and increasing the corresponding error, $S_{ij}$, to a value large enough that the missing data don't influence the results of the analysis.

### 7.2.5 Robust Analysis

In atmospheric chemistry, outliers are often present and may have a large influence on the results. It is therefore important to detect them and take them into account. Once they have been detected, one can reduce the influence of possible outliers by increasing their associated error values, $S_{ij}$. Paatero (1998) recommends defining as outliers data for which:

$$\frac{|x_y - \hat{x}_y|}{S_y} > \alpha$$  \hspace{1cm} (7.6)

where $\hat{x}_y$ are the estimated values and $\alpha$ is a constant. He recommends $\alpha$ equal to 2, 4, or 8.

### 7.2.6 Lognormal Data

Let us assume that each data value, $x_{ij}$, comes from a lognormal distribution with geometric mean equal to the fitted value, $\hat{x}_y$, and log(geometric-standard-deviation) equal to $V_{ij}$. It is further assumed that there is measurement error having standard-deviation $T_{ij}$ in each measured value $x_{ij}$. The matrix $P$ and $A$ are then determined so that $\hat{X} = PA$ maximizes the likelihood of $X$, given matrices $T$ and $V$. The standard deviation $S_{ij}$ is computed as:

$$S_{ij} = \sqrt{T_{ij}^2 + cV_{ij}^2(|\hat{x}_y| + |x_{ij}|)}$$  \hspace{1cm} (7.7)

where $c$ is a constant. One can see, therefore, that in PMF the lognormal data do not have to be transformed to be analyzed but can be fitted using a geometric model.

### 7.2.7 Normalization of Results

We mentioned in the introduction of this section that the PMF solution is scale invariant. Therefore, the matrices $P$ and $A$ can be scaled in different ways without changing the basic solution as long as the other matrix is adequately modified. Many different scaling techniques are possible. Here are some of the possible scalings of the loadings matrix, $A$, put forward by Paatero and Tapper (1994) and Paatero (1998):

1. $\max_j(a_{ij}) = 1$, computed by setting $a_{ij} = a_{ij}/\max_j(a_{ij})$ for $j = 1, 2, ..., m$;
2. $\sum_i \tilde{a}_{ij} = 1$, computed by setting $\tilde{a}_{ij} = \frac{a_{ij}}{\sum_i a_{ij}}$ for $j = 1, 2, \ldots, m$;

3. $\frac{1}{p} \sum_p \tilde{a}_{ij} = 1$, computed by setting $\tilde{a}_{ij} = \frac{1}{p} \sum_p a_{ij}$ for $j = 1, 2, \ldots, m$;

4. $\sum_i \tilde{a}_{ij} = 1 - \phi_j$, computed by setting $\tilde{a}_{ij} = (1 - \phi_j) \frac{a_{ij}}{\sum_i a_{ij}}$ for $j = 1, 2, \ldots, m$,

where $\phi_j$ is the residual variation ratio of column $j$ defined by $\phi_j = \sum_i E^2_{ij} / S^2_{ij} / \sum_i x^2_{ij} / S^2_{ij}$. The quantity $\phi_j$ indicates the amount of variation of column $j$ that is not explained by the factors. The quantity $1 - \phi_j$ may be called the explained weighted variation ratio of variable $j$. The individual scaled values, $\tilde{a}_{ij}$, indicate how much of the variation of variable $j$ is explained by each factor $i$ ($i = 1, 2, \ldots, m$). This scaling indicates the importance of each factor element $a_{ij}$ in explaining the weighted variation of $X$.

Note that the scores matrix $P$ can be scaled in lieu of the loading matrix $A$. The first three possible scaling techniques would also be appropriate in this case. As mentioned earlier, any scaling of $A$ or $P$ must be accompanied by a change in the other such that their product stays the same.

### 7.2.8 Uniqueness of Solution and Rotation

As mentioned earlier, the solution of the PMF analysis is unique most of the time due to the positive constraints imposed on the loadings and scores. But this is not the case all the time. In some cases, the possible rotations of the solution are reduced to only very small modifications of the solution found by the PMF technique and therefore the latter is only affected slightly by the rotations. In other cases, the solution is not unique, and the possible rotations affect the solution markedly. In those cases, we have the same problem as in PCA, APCA, TPCA, and FA. The discussions of Paatero and Tapper (1994) and Paatero (1998) on the subject do not make it clear how to determine whether the solution found by the PMF technique is unique or not.

### 7.3 Example of Positive Matrix Factorization

To illustrate PMF, we will redo the analysis of the urban aerosol concentrations at Edmonton given in Table 4.3. A four-factor model was fitted to the data. The loadings matrix is given in Table 7.1 with estimations of the standard deviation. The loadings vectors have been normalized using the third normalization technique given in Section 7.2.7. The results are somewhat different from the results obtained using TPCA in Section 4.3. If one compares Tables 4.5 and 7.1, one notices: (1) that factor 1 is quite similar to the sulphate component; (2) that factor 2 could be identified with the traffic component, although it is not loaded on NO$_2$; (3) that part of factor 4 is similar to the fly ash component; and that (4) factor 3 is completely different for the last component in Table 4.5.
Table 7.1 Loadings matrix for urban aerosol concentrations at Edmonton (Table 4.3) using the PMF technique. The standard deviations of the estimates are also given.

<table>
<thead>
<tr>
<th></th>
<th>Factor 1</th>
<th>Factor 2</th>
<th>Factor 3</th>
<th>Factor 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pb</td>
<td>0.0220±0.0011</td>
<td>0.0483±0.0020</td>
<td>0.0000±0.0001</td>
<td>0.1133±0.0031</td>
</tr>
<tr>
<td>Fe</td>
<td>0.0016±0.0009</td>
<td>0.0000±0.0023</td>
<td>0.0685±0.0025</td>
<td>0.1604±0.0166</td>
</tr>
<tr>
<td>Ca</td>
<td>0.0198±0.0015</td>
<td>0.0092±0.0024</td>
<td>0.0163±0.0021</td>
<td>0.4116±0.0164</td>
</tr>
<tr>
<td>K</td>
<td>0.0000±0.0003</td>
<td>0.0000±0.0005</td>
<td>0.0070±0.0006</td>
<td>0.1830±0.0052</td>
</tr>
<tr>
<td>Si</td>
<td>0.0000±0.0045</td>
<td>0.0299±0.0212</td>
<td>0.8175±0.0106</td>
<td>0.0000±0.0102</td>
</tr>
<tr>
<td>Al</td>
<td>0.0206±0.0013</td>
<td>0.0000±0.0011</td>
<td>0.0724±0.0014</td>
<td>0.0000±0.0077</td>
</tr>
<tr>
<td>Na+</td>
<td>0.0000±0.0005</td>
<td>0.3066±0.0042</td>
<td>0.0017±0.0002</td>
<td>0.0000±0.0008</td>
</tr>
<tr>
<td>NH₄⁺</td>
<td>0.0190±0.0026</td>
<td>0.5415±0.0099</td>
<td>0.0000±0.0005</td>
<td>0.0574±0.0040</td>
</tr>
<tr>
<td>Cl⁻</td>
<td>0.2369±0.0037</td>
<td>0.0255±0.0034</td>
<td>0.0007±0.0002</td>
<td>0.0000±0.0011</td>
</tr>
<tr>
<td>NO₃⁻</td>
<td>0.2221±0.0043</td>
<td>0.0390±0.0048</td>
<td>0.0000±0.0003</td>
<td>0.0569±0.0034</td>
</tr>
<tr>
<td>SO₄²⁻</td>
<td>0.4598±0.0074</td>
<td>0.0000±0.0019</td>
<td>0.0159±0.0012</td>
<td>0.0174±0.0066</td>
</tr>
</tbody>
</table>

The time series of the scores of the observations on the four factors are shown in Figure 7.1. An interesting result is that the contribution of factors 1 and 2 seems to decrease during the month. The contribution of factor 4, on the other hand, increases during the month. Factor 3 only contributes between about November 16 and November 23.

7.4 Remarks

Positive matrix factorization has many advantages over the other techniques presented here. First, one can be sure of obtaining a solution with positive scores and loadings without having to find a rotation to obtain such a solution. Secondly, the solution is often unique or nearly unique; this is an important feature because it eliminates the arbitrary tendencies of PCA and other similar techniques. Note, however, that the solution is not always unique.

The PMF technique is general. Therefore, the condition regarding zero or negative values can be relaxed and we need require that only some of the loadings and/or scores be greater than zero. This is a very important asset for the PMF technique.

Figure 7.1 Time series of scores for the four factors obtained using the PMF technique.
Note also that the PMF technique may not always converge for certain conditions. This is a disadvantage that is not found in PCA and other related techniques.

*Below-detection-limit* data can be handled in PMF in the same way as in PCA and FA. We therefore refer the reader to the appropriate sections for more details.

It is worth noting as well that the PMF technique can be extended to solve the so-called PARAFAC (parallel factor analysis) model, which has been developed to handle three-way problems. This is another name for the three-mode principal component technique referred to in Section 2.4. The PARAFAC model can be written as:

\[
X_{ijk} = \sum_{h=1}^{m} A_{ih} B_{jh} C_{kh} + E_{ijk}
\]  

(7.8)

where \(m\) is the number of factors in the model. Such a model would be appropriate for use in the case of weekly (or monthly) samples of many ions that had been taken over many years. In such a case, the index \(i\) could be the year; \(j\) the ion, and \(k\) the month of the year.

Finally, we should mention that a similar technique to PMF was developed independently by Treloar (1993). One major difference between Treloar’s technique and PMF is that his technique does not minimize a weighted sum of the differences, as PMF does, but only the differences. Therefore, his technique does not incorporate estimates of the uncertainties of the observations as does PMF.

### 7.5 Computer Software

At present the only available software is the program PMF2 (and PMF3 for the PARAFAC model) that can be bought from P. Paatero, University of Helsinki, Department of Physics, Box 9, FIN-00014 Helsinki/University, Finland. The major problem with this program is that it is not user friendly, and consequently one has to prepare many input files. The documentation is also somewhat lacking, especially in explaining the theoretical aspects. One is not always sure how changing some of the parameters will affect the program and the solutions it produces.

### 7.6 References

8. Rotation

8.1 Introduction

In most of the techniques presented here, it is often necessary to rotate the obtained solution to facilitate the interpretation of the components or factors found. This is due to the fact that in techniques like PCA, APCA, TPCA, or FA the solutions obtained are not unique and can be rotated arbitrarily. The only exception is the PMF technique, which usually produces a unique solution (although in some situations it does produce solution that can be rotated).

The fact that the solutions of PCA or other techniques are not unique is not always a drawback. Rotation can then be used to help in finding a rotated loadings matrix that would be easier to interpret than the original one. That of course raises the question: "What kind of loadings matrix is easy to interpret". As we have seen in Section 2.2.11, Thurstone (1947) developed a three-point criterion to define a simple structure for the loadings matrix. According to this criterion, a simple loadings matrix is one with a few high loadings and many low loadings such that two columns have different patterns of lows and highs.

Starting from the original loadings matrix, the technique presented in this section tries to find a rotation matrix such that the new loadings matrix would have a structure as close as possible to this simple structure. Thus, the general problem of rotation is to find a matrix \( R \) such that:

\[
V = \Lambda R
\]

has a simple structure. Two general types of rotations are possible, namely orthogonal and oblique. In an orthogonal rotation, an orthogonal loadings matrix remains orthogonal after the rotation. In an oblique rotation, this is not the case.

In some situations, we have complete or partial knowledge of the final loadings matrix. The objective then is to find a rotation that could transform the original loadings matrix to one that is identical or close to that target loadings matrix. If we have only partial information on the final loadings matrix, we want the rotated loadings matrix to incorporate that information. We have seen that subjective rotation (Section 5) can be used if one wants to use an orthogonal rotation. If an oblique rotation is acceptable, the target or procrustes rotation can be used (Section 4.2.4).

Finally, rotation could be used when one wants to find a solution such that all elements of both the loadings matrix and the scores matrix are greater than or equal to zero. To find such a rotation, if it exists, we divide that rotation into a series of simple rotations. An algorithm to implement this approach was developed by Shen and Israel (1989). We will not discuss it further here, but interested reader can refer to Shen and Israel (1989) for more details.

8.2 Orthogonal Rotations

As mentioned earlier, orthogonal rotations are used most often because they preserve the orthogonal character of orthogonal loadings matrices. Many orthogonal rotation methods exist. The ones more often encountered are the quartimax, varimax, equamax, parsimax, and
orthomax methods. Here, we will discuss the first three in more details. We will also describe a rotation criterion proposed by McCammon (1966).

8.2.1 The Quartimax Method

The quartimax method attempts to put one major loading on a given factor for a given variable. Although such a goal cannot be accomplished in practice, the quartimax method tries to get as close as possible to it. Another way to describe the quartimax method is to say that it seeks to maximize variance across the factors, that is, to get only one large loading for a variable. Let

\[ Q_i' = \frac{1}{m} \sum_{j=1}^{m} (a_{ij}^2) - \frac{1}{m^2} \left( \sum_{j=1}^{m} a_{ij}^2 \right)^2 \]  

(8.2)

where the \( a_{ij} \) represent the new loadings, and \( i = 1, 2, ..., p \). \( Q_i' \) represents the "variance" of the squared loadings, which in turn represents the contribution to the variance of the variables. The quartimax maximizes

\[ Q = \sum_{i=1}^{p} Q_i' \]  

(8.3)

the sum of variances of the rotated loadings. As the quartimax method attempts to maximize variance across the components, it tends to produce a dominant component. It has, therefore, a tendency to concentrate variance in the first component, which is an undesirable characteristic for atmospheric chemistry applications.

8.2.2 The Varimax method

Introduced by Kaiser (1958), several versions of the varimax method are available; the differences between them relate largely to whether the factors are rotated simultaneously or in pairs. The varimax method is presently the most popular of the rotation methods. It seeks to maximize the variance of the loadings across the variables, rotating the factors so that the variation of the squared factor loadings for a given factor is substantially enlarged. This is accomplished by having large, medium, and small loadings within a particular factor. Let

\[ V_j^* = \frac{1}{p} \sum_{i=1}^{p} (a_{ij}^2) - \frac{1}{p^2} \left( \sum_{i=1}^{p} a_{ij}^2 \right)^2 \]  

(8.4)

for \( j = 1, 2, ..., m \). The \( a_{ij} \) are the new loadings. \( V_j^* \) represents the variance of the (squared) loadings for the \( j \)th factor. Although squared loadings are used to avoid negative signs, they represent the contribution to the variance of the variables. In the varimax method, the sum

\[ V^* = \sum_{j=1}^{m} V_j^* \]  

(8.5)

is maximized. The result of the varimax method is a loadings matrix in which some of elements are made as small as possible and others are made as large as possible in absolute values. Practically, the varimax method tries to obtain factors that have high correlations with some of the variables and no correlation at all with the others.
Since relation 8.4 depends on the total percentage of variance accounted for by the m factors, less reliable variables are given less weight than those whose variance is well explained by the m factors, clearly a desirable feature in statistical estimation.

However, an adjusted criterion is also sometimes used. The normal varimax criterion is defined as

\[ V_j = \frac{1}{p} \sum_{i=1}^{p} \left( \frac{a_{ij}^2}{h_i^2} \right)^2 = \frac{1}{p^2} \left( \sum_{i=1}^{p} \frac{a_{ij}^2}{h_i^2} \right)^2 \]  

(8.6)

where \( h_i^2 \) is the proportion of variance of the ith variable explained by the m factor (i.e., the communality of variable i). \( V = \sum_{j=1}^{m} V_j \) is maximized. By scaling the loadings by the communalties, all variables are given equal weight in the rotation. Note, however, that several authors have argued against such a scaling, especially when some communalties are very small.

8.2.3 The Equamax Method

The equamax method attempts to achieve a simple structure for the loadings matrix with respect to both factors and variables. In the equamax method, the criterion

\[ E = \frac{1}{2} (V + Q) \]  

(8.7)

where V and Q are the varimax and quartimax criteria, respectively.

8.2.3 The McCammon Method

The following orthogonal rotation criterion has been proposed by McCammon (1966). Let \( a_{ij} \) denote the rotated correlation loadings, then minimize the entropy expression

\[ H = -\sum_{i=1}^{p} \sum_{j=1}^{m} a_{ij}^2 \ln(a_{ij}^2). \]  

(8.8)

The advantage of using this relation over the varimax criterion appears to be that it produces a higher proportion of coefficients whose absolute values are closer to zero. This rotation method is not often used, however.

8.3 Oblique Rotations

Although the factors do not remain orthogonal after an oblique rotation, oblique rotations often produce more meaningful factors than any orthogonal rotation can produce. Since orthogonal factors are usually more a matter of convenience than necessity, one should not reject the use of oblique rotations, although they do introduce difficulties in the interpretation of the rotated factors.

Many oblique rotations exist. We will discuss eight of them here.
8.3.1 The Oblimax Method

The oblimax method seeks to rotate the factors so that the number of high and low loadings is increased by reducing the number of medium loadings. The criterion to maximize is equal to:

\[
K = \frac{\sum_{i=1}^{p} \sum_{j=1}^{m} a_{ij}^4}{\left( \sum_{i=1}^{p} \sum_{j=1}^{m} a_{ij}^2 \right)^{4/3}}.
\]  

(8.9)

where the \( a_{ij} \) are the new loadings.

If, in the factor space, the variable-points fall into clusters close to each axis, then the oblimax method will determine a highly satisfactory oblique structure. If, however, the data are more complex (i.e., the variable-points are spread throughout the factor space without clear breaks), the oblimax method may be unsatisfactory. It may not even converge to a solution in the case of a complex structure.

8.3.2 The Quartimin Method

The quartimin method minimizes the sum of the inner products of the (reference) structure loadings. The function to be minimized is:

\[
Q = \sum_{\ell=1}^{p} \sum_{q=1}^{m} a_{\ell j}^2 a_{q j}^2
\]

(8.10)

where the \( a_{ij} \) are the new loadings, \( \ell \) and \( q \) are the \( \ell \)th and \( q \)th factors. The quartimin technique does not have difficulty in giving a solution for complex data. However, quartimin produces oblique factors that are biased toward high intercorrelations in such situations, and that remains a serious problem.

8.3.3 The Covarimin Method

The covarimin method is an extension to oblique rotation of Kaiser's varimax rotation. In this technique, one tries to minimize the function

\[
C = \sum_{\ell=1}^{p} \sum_{q=1}^{m} \left\{ p \sum_{j=1}^{m} \left( \frac{a_{\ell j}^2}{h_j^2} \right) \left( \frac{a_{q j}^2}{h_j^2} \right) - \left( \sum_{j=1}^{m} \frac{a_{\ell j}^2}{h_j^2} \right) \left( \sum_{j=1}^{m} \frac{a_{q j}^2}{h_j^2} \right) \right\}
\]

(8.11)

where the \( a_{ij} \) are the new loadings, and \( \ell \) and \( q \) are the \( \ell \)th and \( q \)th factors. The difficulty with the covarimin method is that it is biased in an opposite direction to the quartimin rotation. It tends to give factors with intercorrelations that are too low. The factors found with the covarimin method are usually very close to the ones obtained using the varimax method.

8.3.4 The Biquartimin Method

The biquartimin method is a compromise algorithm falling somewhere between the quartimin and covarimin methods. The resulting function to minimize is:

\[
B = Q + \frac{C}{p}
\]

(8.12)

90
where $Q$ and $C$ are given by relations 8.10 and 8.11 respectively. The biquartimin method generally gives a simple structure solution that is more satisfactory in terms of interfactor correlations and factor loadings than either covarimin or quartimin solutions.

8.3.5 The Oblimin Method

The oblimin method like the biquartimin method is a combination of the quartimin and covarimin methods. The function to be minimized is written as:

$$B^* = \beta_1 Q + \frac{\beta_2 C}{p}$$  \hspace{1cm} (8.13)

where $\beta_1$ and $\beta_2$ are two constants. Writing $\gamma = \beta_1/(\beta_1 + \beta_2)$, $B^*$ can be written as:

$$B^* = \sum_{l<q=1}^{p} \left\{ p \sum_{j=1}^{p} \frac{a_{lj}^2}{h_j^2} \left( \frac{a_{lj}^2}{h_j^2} \right) - \gamma \left( \sum_{j=1}^{p} \frac{a_{lj}^2}{h_j^2} \right) \left( \sum_{j=1}^{p} \frac{a_{jq}^2}{h_j^2} \right) \right\}$$  \hspace{1cm} (8.14)

where the $a_{ij}$ are the new loadings, and $l$ and $q$ are the $l$th and $q$th factors.

The oblimin class of rotation is dependent on the value of $\gamma$. If $\gamma = 0.5$, the oblimin method is the same as the biquartimin method. The covarimin method can be obtained using $\gamma = 1.0$, and the quartimin method with $\gamma$ equal zero. Different in-between values of $\gamma$ allow for any weighted combination of covarimin and quartimin criteria.

8.3.6 The Binormamin Method

The binormamin method was developed in trying to find an appropriate balancing value for $\gamma$. The criterion to minimized is equal to:

$$B^{**} = \sum_{l<q=1}^{p} \left\{ \frac{p \sum_{j=1}^{p} \left( a_{lj}^2 / h_j^2 \right) \left( a_{lj}^2 / h_j^2 \right)}{\sum_{j=1}^{p} a_{lj}^2 / h_j^2 \left( \sum_{j=1}^{p} a_{jq}^2 / h_j^2 \right)} \right\}$$  \hspace{1cm} (8.15)

where the $a_{ij}$ are the new loadings, and $l$ and $q$ are the $l$th and $q$th factors.

Binormamin appears to be best for data with very clear or very complex simple structures. For data in between, the biquartimin method appears to be better.

8.3.7 The Promax Method

The promax method was proposed by Hendrickson and White (1964). First, the loadings matrix is rotated using an orthogonal varimax method and then the orthogonality of the factors is relaxed so that they will fit a simple structure better.

Starting with the orthogonally rotated matrix, an ideal loadings matrix is constructed in which high loadings are made higher and low loadings smaller. This is done by normalizing the orthogonal matrix by rows and columns and taking the kth power of each loading. The final step is to find the least squares model to fit the ideal loadings matrix, using the target or Procrustes rotation (Section 4.2.4) without iteration.
The only problem remaining is the selection of the kth power. Based on experimentation on some well-known data set, Hendrickson and White recommend using $k = 4$ for most data sets. A lower power should be used for very “cleanly” structured data sets.

8.3.8 The Crawford-Ferguson Method

The Crawford-Ferguson method (see Crawford and Ferguson, 1970) consists of minimizing the following:

$$CF = K_1 \sum_{i=1}^{p} \sum_{l=1}^{m} a_{il}^2 a_{in}^2 + K_2 \sum_{i=1}^{m} \sum_{l=1}^{m} a_{il}^2 a_{ln}^2$$

(8.16)

where the $a_{ij}$ are the new loadings and $K_1$ and $K_2$ define different members of the rotation family.

8.4 Computer Software

All of the major computer statistical systems like SAS®, S-Plus®, and SPSS® contain programs or functions that implement the rotational methods presented in this section. Most also contain other techniques not presented here. Computer algorithms for the varimax rotation are available in Kaiser (1959) and Horst (1965).

8.5 References

Descriptions of the rotational methods presented here and others can be found in Harman (1967) and Rummel (1970). The varimax method is described in detail in Kaiser (1958, 1959) and Horst (1965). McCammon’s method is described in McCammon (1966). The quartimin rotation method is covered in Carroll (1953), the covarimin method in Kaiser (1958), the biquartimin method in Carroll (1957), and the binormamin method in Kaiser and Dickman (1959).
9. Nonparametric Confidence Intervals

9.1 Introduction

One of the aims of researchers in using statistics is to obtain some measure of the uncertainties, due to sampling variation, of their estimates of the fitted model parameters. We have seen that in PCA, APCA, TPCA, and FA one can obtain estimates of those uncertainties only if the data are assumed to be random samples from a multivariate normal distribution. Even then, most of the available results are large-sample approximations. Estimates are not even available for many of the model parameters. Only in the case of PMF, can one get estimates of the variability of the scores and loadings. It is therefore important to find a way to estimate the precision of all the estimated model parameters for the other techniques, even if the observations cannot be assumed to be random samples from a multivariate normal distribution.

Two procedures based on resampling that can be used to estimate parameter precision in PCA, FA, and other techniques are the jackknife and the bootstrap techniques. Another resampling technique that we have already discussed is the cross-validation method used to estimate the number of components to retain in PCA (see Section 2.2.10.4). Resampling techniques are based on the idea that the data can be used to simulate the sampling variability. In all these techniques, the samples (i.e., rows of the observations matrix) are assumed to be independent random samples from the same multivariate distribution. Note that the presence of long-term trends and/or seasonal cycles and/or auto-correlation will result in an increase in the estimated variance given by the jackknife and bootstrap estimators.

Generally, the jackknife and bootstrap techniques have the same characteristics. First, both require extensive computation. Therefore, their use can be somewhat limited if the number of observations and/or variables gets large. Secondly, they can give nonparametric estimates of the precision of all the estimated model parameters. Thirdly, if one wants to test for the statistical significance of one or more estimated model parameters, one has to assume (approximate) normality.

9.2 The Jackknife Technique

Jackknife is a nonparametric statistical technique developed in the late 1940s to estimate bias. Later the technique was used to estimate the bias and variance of statistical estimators. Before describing how to apply the jackknife method to PCA, FA, or other dimensionality reduction techniques discussed here, we will describe how it is used in a simple case.

Suppose that we have a random sample \( x_1, x_2, \ldots, x_n \) taken from an unknown population. We want to estimate some parameter, \( \theta \), from this population. The jackknife estimation is a procedure for obtaining an estimate, \( \hat{\theta}_n \), of \( \theta \) together with its standard deviation. From the original sample, we create \( n \) subsamples, each of size \( n-1 \), by deleting each observation in turn. Let \( \hat{\theta}_i \) be an estimate of \( \theta \) obtained in using the complete data set and \( \hat{\theta}_i \) (\( i = 1, 2, \ldots, n \)) be estimates of \( \theta \) using subsample \( i \). A set of \( n \) different values (called pseudovalues) of the statistic can be obtained as
The jackknife estimate of $\theta$ is

$$\hat{\theta}_i^* = n\hat{\theta}_n - (n-1)\hat{\theta}_i$$

(9.1)

The jackknife estimate of $\tilde{\theta}$ is

$$\tilde{\theta}^* = \frac{1}{n} \sum_{i=1}^{n} \hat{\theta}_i^*$$

(9.2)

or the mean of the n pseudovalues. The variance of the pseudovalues is

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (\hat{\theta}_i^* - \tilde{\theta}^*)^2$$

(9.3)

and the variance of $\tilde{\theta}^*$ is obtained as $(s')^2 = s^2/n$.

As is easily seen, this simple device can be applied to PCA or to the other techniques discussed here. We will only illustrate its application to PCA. For the other techniques, the application should be straightforward.

Let $X$ be the samples matrix. One first generates n submatrices, $X_n$, by systematically eliminating each row in turn. A PCA, retaining m principal components, is made of each matrix $X_n$, giving n sets of eigenvalues, loadings matrices, and scores matrices. Using the n loadings matrices, $A_n$, the n pseudo-loadings matrices, $A_i^*$, are calculated using the relation:

$$A_i^* = nA_n - (n-1)A_i$$

(9.4)

where $A_n$ is the loadings matrix obtained using all the data.

The jackknife estimate of the loadings matrix is:

$$A^* = \frac{1}{n} \sum_{i=1}^{n} A_i^*$$

(9.5)

with a variance given by:

$$S^2 = \frac{1}{n(n-1)} \sum_{i=1}^{n} (A_i^* - A^*)^2.$$ 

(9.6)

Similar relations can be written for the eigenvalues and the scores matrices. A rotation of the loadings matrix can also be made during the PCA step. This technique can be generalized to other dimensionality reduction methods such as APCA, TPCA, and FA.

When trying to use this technique there is one important technical problem that should be taken into account. We have seen in Section 2 that the principal components are extracted by order of magnitude of the eigenvalues. If two eigenvalues for the complete data set are close to each other, two eigenvectors may not be extracted in the same order in $A_i$ and $A_n$. Those two columns must be permuted before using relations 9.4 to 9.6. This permutation of the columns may also occur during the rotation step of the analysis. It is therefore important to find a way to make sure that the loadings matrices $A_i$ are aligned so that each column corresponds to those in matrix $A_n$. We will discuss an algorithm for realigning the matrices later in this section. Note, however, that, at the same time as the columns of the loadings matrices are realigned, the order of the corresponding eigenvalues and corresponding columns of the scores matrices must be changed. The effect of not realigning the principal components is to increase the variance of the jackknife estimator unduly.
Finally, it should be noticed that the jackknife can be generalized by eliminating more than one sample at a time.

9.3 The Bootstrap Technique

The bootstrap estimator is another resampling procedure which is also a "nonparametric" technique for estimating parameters and their standard errors within a single sample. The concept behind the technique is to reproduce the process of selecting many samples of size $n$ by duplicating each sample $N$ times, mixing the resulting $Nn$ samples and randomly selecting a sequence of independent samples, each of size $n$. This method produces a set of independent estimates for the parameter(s), which allows an overall mean estimate and the calculation of the standard deviation(s). As can easily be seen, the bootstrap method can be readily adapted to PCA, APCA, FA, and other dimensionality reduction techniques. As in the case of the jackknife estimator, approximate normality is required for significance testing.

When the bootstrap technique is used in the contest of PCA, FA, and other similar techniques, the loadings matrices and other parameters must be aligned before an overall mean estimate and the standard deviation can be obtained. A technique for realigning the principal components or factors is presented in the next section.

When using the bootstrap technique, it is also easy to include the uncertainties caused by the presence of below-detection-limit data. One has only to replace those data by random numbers between zero and the detection limit, using a rectangular or a triangular random distribution. As we know, each below-detection-limit value will be used $N$ times in the calculation. If a random selection of the value is made, these values will be different and the uncertainties in the real values will be included in the estimated overall uncertainties.

9.4 Realignment of Loadings Matrices

As mentioned earlier, to be able to use the jackknife and bootstrap techniques, the loadings matrices obtained from the reduced samples must be aligned with those obtained from the complete sample. This is because the principal component or factor loadings are determined only up to permutations and/or changes in sign of the columns. The following algorithm described by Clarkson (1979) can be used to align the loadings matrices obtained using the reduced data set with the one obtained using the complete data matrix.

To align the columns of the loadings matrix $A_1$ with $A_n$ (see Section 9.2), one first creates the $(m \times m)$ matrix $M$ with elements equal to:

$$m_{ij} = \min\left(\sum_{i=1}^{p}(a_{ik}^* - a_{kj})^2, \sum_{i=1}^{p}(a_{ik}^* + a_{kj})^2\right)$$

(9.7)

where $a_{ij}^*$ and $a_{ij}$ are the elements of $A_i$ and $A_n$, respectively.

If $M^*$ is a matrix obtained by permuting the two or more columns of $M$, its trace is equal to:

$$\text{trace}(M^*) = \sum_{i=1}^{n} m_{ij}^*$$

(9.8)
Table 9.1 Estimated loadings matrices using the jackknife and bootstrap estimators for urban aerosol data in Table 4.3. The estimated standard deviations for the loadings are also given.

<table>
<thead>
<tr>
<th></th>
<th>Jackknife Estimations</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Comp. 1</td>
<td>Comp. 2</td>
<td>Comp. 3</td>
<td>Comp. 4</td>
</tr>
<tr>
<td>Pb</td>
<td>0.538±0.175</td>
<td>0.862±0.249</td>
<td>0.467±0.214</td>
<td>-0.584±0.621</td>
</tr>
<tr>
<td>Fe</td>
<td>0.959±0.026</td>
<td>0.060±0.116</td>
<td>-0.164±0.096</td>
<td>0.102±0.021</td>
</tr>
<tr>
<td>Ca</td>
<td>0.869±0.643</td>
<td>0.361±0.178</td>
<td>0.021±0.076</td>
<td>0.170±0.028</td>
</tr>
<tr>
<td>K</td>
<td>-0.423±0.715</td>
<td>-0.075±0.104</td>
<td>-0.046±0.052</td>
<td>1.804±0.837</td>
</tr>
<tr>
<td>Si</td>
<td>0.914±0.064</td>
<td>0.155±0.128</td>
<td>-0.126±0.099</td>
<td>0.215±0.114</td>
</tr>
<tr>
<td>Al</td>
<td>0.974±0.021</td>
<td>0.114±0.136</td>
<td>-0.118±0.102</td>
<td>0.049±0.057</td>
</tr>
<tr>
<td>Na⁺</td>
<td>0.098±0.144</td>
<td>0.983±0.023</td>
<td>0.101±0.151</td>
<td>-0.144±0.141</td>
</tr>
<tr>
<td>NH₄⁺</td>
<td>0.129±0.161</td>
<td>1.012±0.022</td>
<td>0.188±0.136</td>
<td>-0.213±0.211</td>
</tr>
<tr>
<td>Cl⁻</td>
<td>-0.099±0.192</td>
<td>0.368±0.406</td>
<td>0.932±0.357</td>
<td>-0.135±0.098</td>
</tr>
<tr>
<td>NO₃⁻</td>
<td>0.215±0.173</td>
<td>0.733±0.261</td>
<td>0.792±0.472</td>
<td>-0.409±0.404</td>
</tr>
<tr>
<td>SO₄²⁻</td>
<td>-0.244±0.088</td>
<td>-0.186±0.206</td>
<td>0.897±0.057</td>
<td>-0.144±0.114</td>
</tr>
</tbody>
</table>

The next step is to find the permutation of the columns of \( M \) that minimizes the trace. Note that the trace will be minimal for original matrix \( M \) if the loadings matrix, \( A \), is already aligned. The aligned matrix, \( A^* \), is the matrix obtained from \( A \), by applying to it the permutation of the columns of \( M \) that minimizes the trace.

Note that if the number of retained principal components or factors is large, this technique becomes difficult to use, as the number of possible permutations is large (being equal to \( m! \)). The present author has used the technique in PCA and FA analyses in which up to 10 principal components were retained.

9.5 Examples

To illustrate the application of the jackknife and bootstrap estimators, the urban aerosol data given in Table 4.3 were used. The loadings matrices obtained for the jackknife PCA estimator and the bootstrap PCA estimator with \( N = 100 \) are given in Table 9.1. Note that in the present case the centralized correlation matrix was used. The estimated standard deviations for the loadings are also given.
Since the number of data item is small (equal to 28), the two estimates (and particularly the jackknife estimate) will be uncertain. Although the results may not be very precise, the reader should remember that they are given only as illustrations of the techniques discussed in this section and not as part of a study of those data.

One should first notice that the estimated loadings are generally closed to each other. The highest differences occur in component 4. The standard deviations for the bootstrap estimators are usually slightly smaller than those for the jackknife method. Both techniques indicate a very large imprecision for the loadings of K on Components 1 and 4. This may be related to the presence of the 9.2 value on November 4. This value, about 5 times greater than the second highest value, may be an outlier and may have distorted the result of the analysis. One can see from this example that the estimation of the variability in the loadings due to sampling variability may be a very useful exercise.

9.6 Computer Software

To the knowledge of the author, no commercially available statistical software package includes options for calculating jackknife or bootstrap estimates in PCA or FA. It will therefore be necessary to program these techniques. With some well-known statistical software packages such as SAS® and S-Plus®, this can be done with the internal program language. The coding of the realignment technique, however, may be difficult in some cases.

9.7 References

A complete description of the jackknife technique can be found in Gray and Schucany (1972). Efron and Tibshirani (1993) give a complete description of the bootstrap technique. A general and simple introduction to the latter technique can be found in Diaconis and Efron (1983). Descriptions of both techniques and the link between them are presented in Efron (1982). Examples of the uses of the jackknife estimator in PCA can be found in Reyment (1982), Gibson et al. (1984), and McGillivray (1985). Pennell (1972) and Clarkson (1979) are examples of the application of the jackknife technique to factor analysis. Stauffer et al. (1985) is an example of the use of the bootstrap estimator in PCA. An example of the use of jackknife estimation in atmospheric chemistry is Roscoe and Hopke (1981a).
10. Examples of the Use of the Dimensionality Reduction Techniques

We will illustrate the use of the techniques presented here with two examples. Note that not all of the techniques introduced in the text will be used.

10.1 Arctic Lower Tropospheric Aerosol at Alert, Canada

This data set has been studied recently by Sirois and Barrie (1999). More than 15 years (from 1980 to 1995) of weekly measurements of 18 ions were available for the analysis. Because most of the summer data are almost always under the detection limit for some ions, only data collected between November and May will be used, as recommended by Sirois and Barrie (1999). We will start by applying the principal component analysis. This will help us in determining the number of components to retain. The correlation matrix will be used for all relevant analysis.

The 18 eigenvalues are given in Table 10.1. One can see that at least nine components are needed to explain more than 90% of the total variance. If we apply Joliffe's 0.77 cut-off limit (Section 2.2.10.1), however, only seven principal components would be necessary. Figure 10.1 presents the scree graph for those data. Based on that plot, nine or ten components should be necessary, although it is possible to argue for a slightly lower value.

If we now consider the changes in communality with the number of principal components retained (Table 10.2), nine components explain more than 85% of the variance of all the ions. Table 10.2 illustrates the effect of keeping too few components. One can see from Tables 10.1 and 10.2 that an increase in the proportion of the total variance due to an increase in the number of components retained is usually not distributed over all the ions but is concentrated on only one or two of them. It is therefore important to also consider the communalities when we select the number of component to retain.

On the basis of these results, it was decided to retain ten components in the following analysis.

Table 10.3 gives the loadings matrix obtained using PCA followed by a varimax rotation of the correlation matrix. The eigenvalues of the rotated principal components are also given. The results for the same analysis followed by an oblimin rotation are given in Table 10.4.

For both, the sampling variability of the eigenvalues and loadings was estimated using the bootstrap method (Section 9.3) with N = 500.

Let us consider first the results of Table 10.3. Some of the components are easily identified. The first one is a photo-sulphur component, with high loadings of SO$_2^-$, H$^+$, NH$_4^+$, Pb, V, and Zn. The second is a sea-salt component, with most of the loadings on Na, Cl, K$^+$, and

---

Figure 10.1 Scree graph for the Alert Data.
Mg. The third is the soil component, with high loadings on V, Mg, Ca, and Al. The fourth component, loaded mainly on Zn and Cu, is a smelter component. Four of the remaining components are loaded mostly on one each of the following Br, MSA, I, and NO₃⁻. Component 8 is especially interesting. Although it is highly loaded on Mn, this load is not statistically significant. A close examination of the estimated standard deviation of the estimated loadings for that ion shows that, this component is either merged with the soil component or independent, depending on the sample. A similar result occurs for H⁺ for component 10, which is merged with the photo-sulphur component.

We now compare those results to the ones obtained in using the oblique oblimin rotation (Table 10.4). Some of the components are very similar to those in Table 10.3. We can still easily identify the sea-salt, smelter, NO₃⁻, MSA, bromine, and iodine components (components 2, 4, 6, 8, 9, and 10 respectively). Components 1, 3, and 5 are components obtained by redistributing the loadings of components 1, 3, and 10 in Table 10.3. As it did with the varimax rotation, Mn seems to oscillate between being linked with another component (component 1 in the present case) and having its own component.

One should remember that the principal components are orthogonal in Table 10.3 but not in Table 10.4. The correlations between the components in this last case are:

### Table 10.1 Eigenvalues for the Alert data set using the correlation matrix.

<table>
<thead>
<tr>
<th>Component</th>
<th>Eigenvalue</th>
<th>Difference</th>
<th>Proportion</th>
<th>Cumulative Proportion</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>5.162</td>
<td>0.287</td>
<td>0.287</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>3.711</td>
<td>0.206</td>
<td>0.493</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>2.472</td>
<td>0.137</td>
<td>0.630</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>1.435</td>
<td>0.080</td>
<td>0.710</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>1.272</td>
<td>0.070</td>
<td>0.781</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>0.779</td>
<td>0.043</td>
<td>0.823</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>0.708</td>
<td>0.039</td>
<td>0.863</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>0.610</td>
<td>0.034</td>
<td>0.897</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>0.483</td>
<td>0.027</td>
<td>0.924</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>0.337</td>
<td>0.019</td>
<td>0.943</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>0.274</td>
<td>0.015</td>
<td>0.958</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>0.196</td>
<td>0.011</td>
<td>0.969</td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>0.172</td>
<td>0.010</td>
<td>0.978</td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>0.140</td>
<td>0.008</td>
<td>0.986</td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>0.093</td>
<td>0.005</td>
<td>0.991</td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>0.076</td>
<td>0.004</td>
<td>0.995</td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>0.053</td>
<td>0.003</td>
<td>0.998</td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>0.028</td>
<td>0.002</td>
<td>1.000</td>
<td></td>
</tr>
</tbody>
</table>

### Table 10.2 Change in the communality of each ion as the number of components retained increased.

<table>
<thead>
<tr>
<th>Ion</th>
<th>Component Retained</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>6</td>
</tr>
<tr>
<td>SO₄²⁻</td>
<td></td>
</tr>
<tr>
<td>H⁺</td>
<td></td>
</tr>
<tr>
<td>Br</td>
<td></td>
</tr>
<tr>
<td>NH₄⁺</td>
<td></td>
</tr>
<tr>
<td>NO₃⁻</td>
<td></td>
</tr>
<tr>
<td>Na</td>
<td></td>
</tr>
<tr>
<td>Cl</td>
<td></td>
</tr>
<tr>
<td>K⁺</td>
<td></td>
</tr>
<tr>
<td>Pb</td>
<td></td>
</tr>
<tr>
<td>V</td>
<td></td>
</tr>
<tr>
<td>Mg</td>
<td></td>
</tr>
<tr>
<td>Zn</td>
<td></td>
</tr>
<tr>
<td>Cu</td>
<td></td>
</tr>
<tr>
<td>Ca</td>
<td></td>
</tr>
<tr>
<td>Mn</td>
<td></td>
</tr>
<tr>
<td>I</td>
<td></td>
</tr>
<tr>
<td>Al</td>
<td></td>
</tr>
<tr>
<td>MSA</td>
<td></td>
</tr>
</tbody>
</table>
Table 10.3 Eigenvalues and loadings for PCA followed by a varimax rotation of the Alert data. Loadings over 0.5 are indicated in bold. The standard deviations were estimated using the bootstrap method. A shaded area indicates that the loading is statistically different from zero at a 95% family-wise confidence interval. Note that the loadings are assumed to follow approximately a normal distribution.

<table>
<thead>
<tr>
<th>Component</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Eigenvalue</td>
<td>3.929</td>
<td>2.956</td>
<td>2.839</td>
<td>1.602</td>
<td>1.091</td>
<td>1.050</td>
<td>1.002</td>
<td>0.913</td>
<td>0.812</td>
<td>0.753</td>
</tr>
<tr>
<td>±0.240</td>
<td>±0.141</td>
<td>±0.122</td>
<td>±0.042</td>
<td>±0.025</td>
<td>±0.020</td>
<td>±0.030</td>
<td>±0.080</td>
<td>±0.244</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Ions</th>
<th>Loadings</th>
</tr>
</thead>
<tbody>
<tr>
<td>SO$_4^{2-}$</td>
<td>0.787 ±0.040</td>
</tr>
<tr>
<td></td>
<td>-0.035 ±0.023</td>
</tr>
<tr>
<td></td>
<td>-0.032 ±0.027</td>
</tr>
<tr>
<td></td>
<td>-0.100 ±0.027</td>
</tr>
<tr>
<td></td>
<td>0.115 ±0.115</td>
</tr>
<tr>
<td></td>
<td>-0.232 ±0.022</td>
</tr>
<tr>
<td></td>
<td>0.006 ±0.062</td>
</tr>
<tr>
<td>H$^+$</td>
<td>0.574 ±0.231</td>
</tr>
<tr>
<td></td>
<td>-0.271 ±0.049</td>
</tr>
<tr>
<td></td>
<td>0.198 ±0.032</td>
</tr>
<tr>
<td></td>
<td>0.043 ±0.041</td>
</tr>
<tr>
<td></td>
<td>-0.041 ±0.016</td>
</tr>
<tr>
<td></td>
<td>-0.167 ±0.027</td>
</tr>
<tr>
<td>Br</td>
<td>0.127 ±0.022</td>
</tr>
<tr>
<td></td>
<td>-0.030 ±0.026</td>
</tr>
<tr>
<td></td>
<td>0.107 ±0.037</td>
</tr>
<tr>
<td></td>
<td>0.004 ±0.040</td>
</tr>
<tr>
<td></td>
<td>0.034 ±0.070</td>
</tr>
<tr>
<td></td>
<td>0.025 ±0.025</td>
</tr>
<tr>
<td></td>
<td>0.022 ±0.022</td>
</tr>
<tr>
<td>NH$_4^+$</td>
<td>0.183 ±0.024</td>
</tr>
<tr>
<td></td>
<td>-0.027 ±0.049</td>
</tr>
<tr>
<td></td>
<td>0.943 ±0.124</td>
</tr>
<tr>
<td></td>
<td>0.204 ±0.024</td>
</tr>
<tr>
<td></td>
<td>0.013 ±0.002</td>
</tr>
<tr>
<td>NO$_3^-$</td>
<td>0.048 ±0.026</td>
</tr>
<tr>
<td></td>
<td>-0.023 ±0.039</td>
</tr>
<tr>
<td></td>
<td>0.183 ±0.021</td>
</tr>
<tr>
<td></td>
<td>0.024 ±0.020</td>
</tr>
<tr>
<td></td>
<td>0.020 ±0.019</td>
</tr>
<tr>
<td></td>
<td>0.019 ±0.028</td>
</tr>
<tr>
<td>Na</td>
<td>0.869 ±0.067</td>
</tr>
<tr>
<td></td>
<td>-0.074 ±0.117</td>
</tr>
<tr>
<td></td>
<td>0.061 ±0.075</td>
</tr>
<tr>
<td></td>
<td>0.185 ±0.082</td>
</tr>
<tr>
<td></td>
<td>-0.002 ±0.136</td>
</tr>
<tr>
<td></td>
<td>0.197 ±0.037</td>
</tr>
<tr>
<td>Cl</td>
<td>0.023 ±0.010</td>
</tr>
<tr>
<td></td>
<td>-0.030 ±0.021</td>
</tr>
<tr>
<td></td>
<td>-0.008 ±0.016</td>
</tr>
<tr>
<td></td>
<td>-0.006 ±0.027</td>
</tr>
<tr>
<td></td>
<td>0.002 ±0.003</td>
</tr>
<tr>
<td>K$^+$</td>
<td>0.053 ±0.055</td>
</tr>
<tr>
<td></td>
<td>0.107 ±0.048</td>
</tr>
<tr>
<td></td>
<td>0.017 ±0.011</td>
</tr>
<tr>
<td></td>
<td>-0.229 ±0.045</td>
</tr>
<tr>
<td></td>
<td>0.045 ±0.119</td>
</tr>
<tr>
<td></td>
<td>0.199 ±0.054</td>
</tr>
<tr>
<td>Pb</td>
<td>0.060 ±0.055</td>
</tr>
<tr>
<td></td>
<td>0.024 ±0.035</td>
</tr>
<tr>
<td></td>
<td>0.107 ±0.037</td>
</tr>
<tr>
<td></td>
<td>0.037 ±0.037</td>
</tr>
<tr>
<td></td>
<td>0.139 ±0.097</td>
</tr>
<tr>
<td>V</td>
<td>0.098 ±0.054</td>
</tr>
<tr>
<td></td>
<td>0.034 ±0.043</td>
</tr>
<tr>
<td></td>
<td>0.030 ±0.030</td>
</tr>
<tr>
<td></td>
<td>0.029 ±0.029</td>
</tr>
<tr>
<td></td>
<td>0.016 ±0.016</td>
</tr>
<tr>
<td></td>
<td>0.003 ±0.003</td>
</tr>
<tr>
<td>Mg</td>
<td>0.098 ±0.052</td>
</tr>
<tr>
<td></td>
<td>0.029 ±0.027</td>
</tr>
<tr>
<td></td>
<td>0.026 ±0.025</td>
</tr>
<tr>
<td></td>
<td>0.022 ±0.022</td>
</tr>
<tr>
<td></td>
<td>0.011 ±0.003</td>
</tr>
<tr>
<td>Zn</td>
<td>0.109 ±0.067</td>
</tr>
<tr>
<td></td>
<td>0.046 ±0.015</td>
</tr>
<tr>
<td></td>
<td>0.023 ±0.020</td>
</tr>
<tr>
<td></td>
<td>0.025 ±0.014</td>
</tr>
<tr>
<td></td>
<td>0.011 ±0.003</td>
</tr>
<tr>
<td>Cu</td>
<td>0.066 ±0.041</td>
</tr>
<tr>
<td></td>
<td>0.020 ±0.067</td>
</tr>
<tr>
<td></td>
<td>0.043 ±0.043</td>
</tr>
<tr>
<td></td>
<td>0.024 ±0.027</td>
</tr>
<tr>
<td></td>
<td>0.077 ±0.009</td>
</tr>
<tr>
<td></td>
<td>0.026 ±0.048</td>
</tr>
<tr>
<td>Ca</td>
<td>-0.005 ±0.004</td>
</tr>
<tr>
<td></td>
<td>0.901 ±0.034</td>
</tr>
<tr>
<td></td>
<td>-0.014 ±0.000</td>
</tr>
<tr>
<td></td>
<td>-0.036 ±0.015</td>
</tr>
<tr>
<td></td>
<td>0.015 ±0.008</td>
</tr>
<tr>
<td></td>
<td>0.058 ±0.011</td>
</tr>
<tr>
<td>Mn</td>
<td>-0.032 ±0.037</td>
</tr>
<tr>
<td></td>
<td>0.014 ±0.018</td>
</tr>
<tr>
<td></td>
<td>0.037 ±0.025</td>
</tr>
<tr>
<td></td>
<td>0.029 ±0.029</td>
</tr>
<tr>
<td></td>
<td>0.019 ±0.009</td>
</tr>
<tr>
<td>I</td>
<td>0.011 ±0.007</td>
</tr>
<tr>
<td></td>
<td>0.512 ±0.053</td>
</tr>
<tr>
<td></td>
<td>-0.075 ±0.035</td>
</tr>
<tr>
<td></td>
<td>-0.033 ±0.006</td>
</tr>
<tr>
<td></td>
<td>0.041 ±0.027</td>
</tr>
<tr>
<td></td>
<td>0.025 ±0.015</td>
</tr>
<tr>
<td></td>
<td>0.020 ±0.020</td>
</tr>
<tr>
<td></td>
<td>0.025 ±0.025</td>
</tr>
<tr>
<td></td>
<td>0.011 ±0.003</td>
</tr>
<tr>
<td>Al</td>
<td>-0.050 ±0.003</td>
</tr>
<tr>
<td></td>
<td>0.012 ±0.007</td>
</tr>
<tr>
<td></td>
<td>0.011 ±0.006</td>
</tr>
<tr>
<td></td>
<td>0.175 ±0.063</td>
</tr>
<tr>
<td></td>
<td>-0.050 ±0.059</td>
</tr>
<tr>
<td>MSA</td>
<td>0.025 ±0.016</td>
</tr>
<tr>
<td></td>
<td>0.017 ±0.012</td>
</tr>
<tr>
<td></td>
<td>0.022 ±0.033</td>
</tr>
<tr>
<td></td>
<td>0.015 ±0.008</td>
</tr>
<tr>
<td></td>
<td>0.033 ±0.009</td>
</tr>
<tr>
<td></td>
<td>0.196 ±0.100</td>
</tr>
</tbody>
</table>

| 1.000     |
| 0.098 ±0.000 |
| -0.118 ±0.064 |
| 0.023 ±0.040 |
| 0.105 ±0.091 |
| 0.104 ±0.067 |
| 0.052 ±0.050 |
| 0.056 ±0.017 |
| 0.031 ±0.016 |
| -0.006 ±0.057 |
| -0.009 ±0.011 |
Table 10.4: Eigenvalues and loadings for PCA followed by an oblimin rotation of the Alert data. Loadings over 0.5 are indicated in bold. The standard deviations were estimated using the bootstrap method. A shaded area indicates that the loading is statistically different from zero at a 95% family-wise confidence interval. Note that the loadings are assumed to follow approximately a normal distribution.

<table>
<thead>
<tr>
<th>Component</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Eigenvalues</td>
<td>2.117</td>
<td>2.009</td>
<td>1.667</td>
<td>1.547</td>
<td>1.236</td>
<td>1.115</td>
<td>1.061</td>
<td>1.046</td>
<td>1.041</td>
<td>2.009</td>
</tr>
<tr>
<td>Loadings</td>
<td>0.099</td>
<td>-0.099</td>
<td>±0.039</td>
<td>±0.025</td>
<td>±0.073</td>
<td>±0.029</td>
<td>±0.046</td>
<td>±0.026</td>
<td>±0.026</td>
<td>±0.011</td>
</tr>
<tr>
<td></td>
<td>±0.088</td>
<td>±0.057</td>
<td>±0.021</td>
<td>±0.054</td>
<td>±0.043</td>
<td>±0.029</td>
<td>±0.037</td>
<td>±0.024</td>
<td>±0.024</td>
<td>±0.024</td>
</tr>
<tr>
<td>Ion Loadings</td>
<td>±0.012</td>
<td>±0.019</td>
<td>±0.028</td>
<td>±0.013</td>
<td>±0.027</td>
<td>±0.015</td>
<td>±0.014</td>
<td>±0.013</td>
<td>±0.012</td>
<td>±0.009</td>
</tr>
<tr>
<td>SO_{4}^{2-}</td>
<td>±0.067</td>
<td>-0.091</td>
<td>±0.056</td>
<td>±0.010</td>
<td>±0.017</td>
<td>-0.011</td>
<td>±0.088</td>
<td>±0.026</td>
<td>±0.030</td>
<td>±0.004</td>
</tr>
<tr>
<td>H^{+}</td>
<td>±0.098</td>
<td>±0.003</td>
<td>±0.029</td>
<td>±0.011</td>
<td>±0.014</td>
<td>±0.024</td>
<td>±0.024</td>
<td>±0.013</td>
<td>±0.012</td>
<td>±0.009</td>
</tr>
<tr>
<td>Br</td>
<td>±0.004</td>
<td>±0.004</td>
<td>±0.002</td>
<td>±0.005</td>
<td>±0.001</td>
<td>±0.004</td>
<td>±0.004</td>
<td>±0.003</td>
<td>±0.003</td>
<td>±0.003</td>
</tr>
<tr>
<td>NH_{4}^{+}</td>
<td>±0.030</td>
<td>±0.094</td>
<td>±0.114</td>
<td>±0.026</td>
<td>±0.014</td>
<td>±0.026</td>
<td>±0.011</td>
<td>±0.031</td>
<td>±0.036</td>
<td>±0.044</td>
</tr>
<tr>
<td>NO_{3}^{-}</td>
<td>±0.036</td>
<td>±0.009</td>
<td>±0.007</td>
<td>±0.002</td>
<td>±0.004</td>
<td>±0.032</td>
<td>±0.024</td>
<td>±0.006</td>
<td>±0.005</td>
<td>±0.005</td>
</tr>
<tr>
<td>Na</td>
<td>±0.018</td>
<td>±0.035</td>
<td>±0.026</td>
<td>±0.083</td>
<td>±0.105</td>
<td>±0.024</td>
<td>±0.006</td>
<td>±0.005</td>
<td>±0.005</td>
<td>±0.005</td>
</tr>
<tr>
<td>Cl</td>
<td>±0.027</td>
<td>±0.028</td>
<td>±0.137</td>
<td>±0.044</td>
<td>±0.026</td>
<td>±0.072</td>
<td>±0.027</td>
<td>±0.012</td>
<td>±0.013</td>
<td>±0.012</td>
</tr>
<tr>
<td>K^{+}</td>
<td>±0.026</td>
<td>±0.024</td>
<td>±0.029</td>
<td>±0.020</td>
<td>±0.042</td>
<td>±0.030</td>
<td>±0.028</td>
<td>±0.017</td>
<td>±0.016</td>
<td>±0.020</td>
</tr>
<tr>
<td>Pb</td>
<td>±0.013</td>
<td>±0.016</td>
<td>±0.089</td>
<td>±0.078</td>
<td>±0.425</td>
<td>±0.213</td>
<td>±0.0056</td>
<td>±0.062</td>
<td>±0.020</td>
<td>±0.008</td>
</tr>
<tr>
<td>V</td>
<td>±0.029</td>
<td>±0.049</td>
<td>±0.060</td>
<td>±0.045</td>
<td>±0.112</td>
<td>±0.055</td>
<td>±0.099</td>
<td>±0.036</td>
<td>±0.035</td>
<td>±0.030</td>
</tr>
<tr>
<td>Mg</td>
<td>±0.028</td>
<td>±0.035</td>
<td>±0.081</td>
<td>±0.003</td>
<td>±0.012</td>
<td>±0.093</td>
<td>±0.103</td>
<td>±0.028</td>
<td>±0.014</td>
<td>±0.010</td>
</tr>
<tr>
<td>Zn</td>
<td>±0.017</td>
<td>±0.016</td>
<td>±0.042</td>
<td>±0.070</td>
<td>±0.113</td>
<td>±0.032</td>
<td>±0.070</td>
<td>±0.023</td>
<td>±0.029</td>
<td>±0.031</td>
</tr>
<tr>
<td>Cu</td>
<td>±0.016</td>
<td>±0.015</td>
<td>±0.013</td>
<td>±0.012</td>
<td>±0.018</td>
<td>±0.013</td>
<td>±0.049</td>
<td>±0.021</td>
<td>±0.011</td>
<td>±0.016</td>
</tr>
<tr>
<td>Ca</td>
<td>±0.003</td>
<td>±0.008</td>
<td>±0.002</td>
<td>±0.002</td>
<td>±0.011</td>
<td>±0.002</td>
<td>±0.010</td>
<td>±0.004</td>
<td>±0.008</td>
<td>±0.003</td>
</tr>
<tr>
<td>Mn</td>
<td>±0.019</td>
<td>±0.023</td>
<td>±0.015</td>
<td>±0.020</td>
<td>±0.014</td>
<td>±0.021</td>
<td>±0.048</td>
<td>±0.039</td>
<td>±0.032</td>
<td>±0.032</td>
</tr>
<tr>
<td>I</td>
<td>±0.001</td>
<td>±0.005</td>
<td>±0.037</td>
<td>±0.012</td>
<td>±0.050</td>
<td>±0.003</td>
<td>±0.004</td>
<td>±0.009</td>
<td>±0.008</td>
<td>±0.008</td>
</tr>
<tr>
<td>Al</td>
<td>±0.017</td>
<td>±0.015</td>
<td>±0.013</td>
<td>±0.012</td>
<td>±0.018</td>
<td>±0.014</td>
<td>±0.015</td>
<td>±0.017</td>
<td>±0.017</td>
<td>±0.017</td>
</tr>
<tr>
<td>MSA</td>
<td>±0.036</td>
<td>±0.020</td>
<td>±0.078</td>
<td>±0.000</td>
<td>±0.030</td>
<td>±0.065</td>
<td>±0.209</td>
<td>±0.078</td>
<td>±0.078</td>
<td>±0.078</td>
</tr>
</tbody>
</table>

The correlations between the principal components are small (less than 0.4). Therefore, the principal components are close to orthogonality, even though an oblique rotation was used.

This example illustrates that the results obtained by using different rotations may differ in only some of the components. It is always interesting, therefore, to compare different rotations. Such an analysis will help to point out the strongest components.

The results for the factor analysis are given in Table 10.5. Only the first four factors are identifiable. They may be identified as photo-sulphur, sea-salt, soil, and smelter factors respectively. Factors 5 and 6 might be identified as bromine and iodine factors respectively, but the uncertainties are so large that the loadings are not statistically significant. Factor 7 seems to be some H^{+} factor, but here also the loadings are not statistically significant from the principal components.
Table 10.5 Variance explained and loadings for FA followed by a varimax rotation of the Alert data. Loadings over 0.5 are indicated in bold. The standard deviations were estimated using the bootstrap method. A shaded area indicates that the loading is statistically different from zero at a 95% family-wise confidence interval. Note that the loadings are assumed to follow approximately a normal distribution.

<table>
<thead>
<tr>
<th>Factor</th>
<th>Commu-</th>
<th>Variance Explained</th>
<th>Loadings</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>nality</td>
<td>±0.623</td>
<td>±0.585</td>
</tr>
<tr>
<td>1</td>
<td></td>
<td>±0.623</td>
<td>±0.439</td>
</tr>
<tr>
<td>2</td>
<td></td>
<td>±0.305</td>
<td>±0.056</td>
</tr>
<tr>
<td>3</td>
<td></td>
<td>±0.056</td>
<td>±0.275</td>
</tr>
<tr>
<td>4</td>
<td></td>
<td>±0.275</td>
<td>±0.394</td>
</tr>
<tr>
<td>5</td>
<td></td>
<td>±0.394</td>
<td>±0.327</td>
</tr>
<tr>
<td>6</td>
<td></td>
<td>±0.327</td>
<td>±0.327</td>
</tr>
<tr>
<td>7</td>
<td></td>
<td>±0.327</td>
<td>±0.327</td>
</tr>
<tr>
<td>8</td>
<td></td>
<td>±0.969</td>
<td>±0.969</td>
</tr>
<tr>
<td>9</td>
<td></td>
<td>±0.969</td>
<td>±0.969</td>
</tr>
<tr>
<td>10</td>
<td></td>
<td>±0.969</td>
<td>±0.969</td>
</tr>
</tbody>
</table>

zero. The remaining factors are not identifiable. In this analysis, only the strongest factors were extracted.

On the one hand, these results for FA suggest that perhaps too many factors were assumed. On the other hand, the low communalities obtained for MSA, I, and Mn and the high variability in the communality for bromine may indicate that factor analysis may not be the best technique to use in the present case. One should also note that another rotation method might give better results.

This example illustrates an interesting characteristic of the bootstrap estimator of variability. Some or all of the high values obtained for the standard deviations may be due to the possibility that the method's assumption that the samples are independent has not been...
Table 10.6 Loadings for PMF. A shaded area indicates that the loading is statistically different from zero at a 95% family-wise confidence interval. Note that the loadings are assumed to follow approximately a normal distribution.

<table>
<thead>
<tr>
<th>Ion</th>
<th>Loadings</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>SO$_4^{2-}$</td>
<td>±0.0123</td>
<td>±0.0866</td>
<td>±0.8299</td>
<td>±0.4773</td>
<td>±0.1551</td>
<td>±0.8415</td>
<td>±1.1763</td>
<td>±1.0325</td>
<td>±0.0962</td>
<td>±0.7623</td>
<td></td>
</tr>
<tr>
<td>H$^+$</td>
<td>±0.0180</td>
<td>±0.0001</td>
<td>±0.0066</td>
<td>±0.0033</td>
<td>±0.0000</td>
<td>±0.0008</td>
<td>±0.0024</td>
<td>±0.0004</td>
<td>±0.0000</td>
<td>±0.0004</td>
<td></td>
</tr>
<tr>
<td>Br$^-$</td>
<td>±0.0017</td>
<td>±0.0001</td>
<td>±0.0014</td>
<td>±0.0009</td>
<td>±0.0003</td>
<td>±0.0017</td>
<td>±0.0005</td>
<td>±0.0017</td>
<td>±0.0005</td>
<td>±0.0013</td>
<td></td>
</tr>
<tr>
<td>NH$_4^+$</td>
<td>±0.0066</td>
<td>±0.0065</td>
<td>±0.4945</td>
<td>±0.0224</td>
<td>±0.0006</td>
<td>±0.0049</td>
<td>±0.1763</td>
<td>±0.0825</td>
<td>±0.0128</td>
<td>±0.0003</td>
<td>±0.0052</td>
</tr>
<tr>
<td>NO$_2^-$</td>
<td>±0.0154</td>
<td>±0.00126</td>
<td>±0.0568</td>
<td>±0.0496</td>
<td>±0.00169</td>
<td>±0.1049</td>
<td>±0.1704</td>
<td>±0.9280</td>
<td>±0.0360</td>
<td>±0.0069</td>
<td>±0.0701</td>
</tr>
<tr>
<td>Na$^+$</td>
<td>±0.0003</td>
<td>±0.0326</td>
<td>±0.0863</td>
<td>±0.0396</td>
<td>±0.0297</td>
<td>±0.0508</td>
<td>±0.1268</td>
<td>±0.0058</td>
<td>±0.0606</td>
<td>±0.0117</td>
<td></td>
</tr>
<tr>
<td>Cl$^-$</td>
<td>±0.0067</td>
<td>±0.0092</td>
<td>±0.0067</td>
<td>±0.0277</td>
<td>±0.0089</td>
<td>±0.0711</td>
<td>±0.1156</td>
<td>±0.0173</td>
<td>±0.0071</td>
<td>±0.0402</td>
<td></td>
</tr>
<tr>
<td>K$^+$</td>
<td>±0.0111</td>
<td>±0.0359</td>
<td>±0.1250</td>
<td>±0.0227</td>
<td>±0.0212</td>
<td>±0.0513</td>
<td>±0.0931</td>
<td>±0.0328</td>
<td>±0.0115</td>
<td>±0.0961</td>
<td></td>
</tr>
<tr>
<td>Pb</td>
<td>±0.0032</td>
<td>±0.0131</td>
<td>±0.0450</td>
<td>±0.0102</td>
<td>±0.0006</td>
<td>±0.0413</td>
<td>±0.0780</td>
<td>±0.0006</td>
<td>±0.0046</td>
<td>±0.0059</td>
<td></td>
</tr>
<tr>
<td>V</td>
<td>±0.0018</td>
<td>±0.0122</td>
<td>±0.0022</td>
<td>±0.0010</td>
<td>±0.0014</td>
<td>±0.1668</td>
<td>±0.0119</td>
<td>±0.0350</td>
<td>±0.0162</td>
<td>±0.0004</td>
<td></td>
</tr>
<tr>
<td>Mg$^{2+}$</td>
<td>±0.0002</td>
<td>±0.0002</td>
<td>±0.0001</td>
<td>±0.0000</td>
<td>±0.0001</td>
<td>±0.0017</td>
<td>±0.0005</td>
<td>±0.0004</td>
<td>±0.0004</td>
<td>±0.0015</td>
<td></td>
</tr>
<tr>
<td>Zn</td>
<td>±0.0001</td>
<td>±0.0001</td>
<td>±0.0001</td>
<td>±0.0002</td>
<td>±0.0001</td>
<td>±0.0015</td>
<td>±0.0017</td>
<td>±0.0000</td>
<td>±0.0000</td>
<td>±0.0000</td>
<td></td>
</tr>
<tr>
<td>Mn</td>
<td>±0.0001</td>
<td>±0.0001</td>
<td>±0.0001</td>
<td>±0.0001</td>
<td>±0.0001</td>
<td>±0.0001</td>
<td>±0.0001</td>
<td>±0.0001</td>
<td>±0.0001</td>
<td>±0.0001</td>
<td></td>
</tr>
<tr>
<td>Al</td>
<td>±0.0001</td>
<td>±0.0001</td>
<td>±0.0001</td>
<td>±0.0001</td>
<td>±0.0001</td>
<td>±0.0001</td>
<td>±0.0001</td>
<td>±0.0001</td>
<td>±0.0001</td>
<td>±0.0001</td>
<td></td>
</tr>
</tbody>
</table>

met in this case. In fact, the Alert data are not independent, as Sirois and Barrie (1999) have shown. The presence of long-term trends and cycles and auto-correlation tend to increase the bootstrap estimation of the variance. It is important to keep this in mind when using these standard deviation estimates.

The final analysis of these data was done using PMF. The results for the loadings matrix are given in Table 10.6. The PMF analysis seems to give a clearer picture than the other analyses, with better defined factors. Some of the factors are similar to those found earlier, such as sea-salt, MSA, iodine, and bromine, but in most cases the factors are better defined. Thus, we can say that for this example the PMF analysis seems to be better suited to the data.

Using these results, we can identify the ten factors:

1. a sulphuric acid factor
2. a sea-salt factor
Figure 10.2 Time series of PMF scores for the 10 factors.

3. a nitrate factor
4. an MSA factor
5. mixed photo-sulphur and sea-salt factor
6. an iodine factor
7. a bromine factor
8. an anthropogenic factor
9. a soil factor
10. a smelter factor

If we compare the components or factors obtained from PCA and FA with the PMF factors above, it is apparent that the PCA analysis was closer. FA found only about four of these components. This example illustrates quite well that the different techniques presented in this overview will generally produce results that are very different. In this particular example, the results from the PMF analysis make the most sense. The only change that could be made to improve the PMF analysis would be to allow a negative loading for $H^+$ on the sea-salt factors, as in the PCA with varimax rotation.

As mentioned in Section 7, the factors extracted using PMF analysis do not have to be orthogonal. It would therefore be interesting to consider the correlation matrix of the ten factors. This matrix is:

\[
\begin{pmatrix}
1.00 & 0.199 & 0.039 & 0.992 & 0.986 & 0.924 & 0.849 & 0.983 & 0.258 & 0.994 \\
0.199 & 1.00 & 0.075 & 0.190 & 0.274 & 0.196 & 0.186 & 0.189 & 0.102 & 0.241 \\
0.039 & 0.075 & 1.00 & 0.107 & 0.018 & 0.011 & 0.021 & 0.030 & 0.060 & 0.072 \\
0.992 & 0.190 & 0.107 & 1.00 & 0.980 & 0.957 & 0.850 & 0.982 & 0.258 & 0.987 \\
0.986 & 0.274 & 0.018 & 0.980 & 1.00 & 0.118 & 0.842 & 0.975 & 0.102 & 0.995 \\
0.924 & 0.196 & 0.011 & 0.957 & 0.118 & 1.00 & 0.804 & 0.970 & 0.060 & 0.936 \\
0.849 & 0.186 & 0.021 & 0.850 & 0.842 & 0.804 & 1.00 & 0.831 & 0.258 & 0.847 \\
0.983 & 0.189 & 0.030 & 0.982 & 0.975 & 0.970 & 0.831 & 1.00 & 0.088 & 0.981 \\
0.258 & 0.186 & 0.060 & 0.258 & 0.804 & 0.975 & 0.831 & 1.00 & 0.000 & 0.000 \\
0.994 & 0.241 & 0.072 & 0.987 & 0.995 & 0.936 & 0.847 & 0.981 & 0.000 & 1.000 \\
\end{pmatrix}
\]

The large correlations, higher than 0.8, are all associated with very large values of the loadings for $SO_2^-$. We can see that for most cases, however, the correlations are not statistically significantly different from zero. Therefore, the correlations between factors 1 and 5 will be much smaller. The suspected values are underlined. If one replaces all nine statistically significant loadings in Table 10.6 by zero, the correlation matrix between the factors becomes:

\[
\begin{pmatrix}
1.00 & 0.000 & 0.000 & 0.000 & 0.000 & 0.000 & 0.000 & 0.000 & 0.000 & 0.000 \\
0.000 & 1.00 & 0.000 & 0.000 & 0.000 & 0.000 & 0.000 & 0.000 & 0.000 & 0.000 \\
0.000 & 0.000 & 1.00 & 0.000 & 0.000 & 0.000 & 0.000 & 0.000 & 0.000 & 0.000 \\
0.000 & 0.000 & 0.000 & 1.00 & 0.000 & 0.000 & 0.000 & 0.000 & 0.000 & 0.000 \\
0.981 & 0.088 & 0.000 & 0.000 & 1.00 & 0.000 & 0.000 & 0.000 & 0.000 & 0.000 \\
0.000 & 0.000 & 0.000 & 0.000 & 0.000 & 1.00 & 0.000 & 0.000 & 0.000 & 0.000 \\
0.000 & 0.000 & 0.000 & 0.000 & 0.000 & 0.000 & 1.00 & 0.000 & 0.000 & 0.000 \\
0.000 & 0.000 & 0.000 & 0.000 & 0.000 & 0.000 & 0.000 & 1.00 & 0.000 & 0.000 \\
0.000 & 0.000 & 0.000 & 0.000 & 0.000 & 0.000 & 0.000 & 0.000 & 1.00 & 0.000 \\
0.000 & 0.000 & 0.000 & 0.000 & 0.000 & 0.000 & 0.000 & 0.000 & 0.000 & 1.000 \\
\end{pmatrix}
\]

One can see that, except for factors 1 and 5, the factors are orthogonal.
Figure 10.2 presents the scores for the ten factors as a time series. The continuous lines are kernel smoother estimations of the long-term trends (see Sirois, 1993). One should first notice that the factors seem to be log-normally distributed. Secondly, there are large decreasing long-term trends in the sulphuric acid, MSA, anthropogenic, and smelter factors (i.e., factors 1, 4, 8, and 10 respectively). The sea salt factor (factor 2) indicates only an overall small decrease, having decreased more strongly in the first part of the period but increased again after 1988. The nitrate factor (factor 3) oscillated around a mean value during the period. The mixed photo-sulphur and sea salt factor and the iodine factor (factors 5 and 6 respectively) both showed a net increase during the period, although they seem to have decreased during the last few years. The bromine factor (factor 7) has slowly increased during the period. Although the soil factor (factor 9) has decreased during most of the period, it has increased slightly during the last few years. The reader is referred to Sirois and Barrie (1999) for more details.

This example shows how dimensionality reduction techniques can be useful in extracting information about emission sources and also in studying their changes over time.

10.2 Sulphate Concentration in Precipitation in Eastern North America

As a second example of the use of dimensionality reduction techniques we will consider 28-day mean sulphate concentration values at 57 sites in eastern North America between January 1988 and December 1994. Ninety-one samples were available for 57 sites (or variables). Figure 10.3 gives the locations of the sites.
Table 10.7 Eigenvalues for PCA using the correlation matrix for the sulphate concentration in the precipitation example.

<table>
<thead>
<tr>
<th>Principal Component Number</th>
<th>Eigenvalue</th>
<th>Difference</th>
<th>Proportion</th>
<th>Cumulative Proportion</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>19.939</td>
<td>0.349</td>
<td>0.349</td>
<td>0.349</td>
</tr>
<tr>
<td>2</td>
<td>6.211</td>
<td>13.727</td>
<td>0.109</td>
<td>0.459</td>
</tr>
<tr>
<td>3</td>
<td>3.695</td>
<td>2.516</td>
<td>0.065</td>
<td>0.524</td>
</tr>
<tr>
<td>4</td>
<td>2.389</td>
<td>0.306</td>
<td>0.037</td>
<td>0.602</td>
</tr>
<tr>
<td>5</td>
<td>1.629</td>
<td>0.454</td>
<td>0.029</td>
<td>0.631</td>
</tr>
<tr>
<td>6</td>
<td>1.503</td>
<td>0.126</td>
<td>0.026</td>
<td>0.567</td>
</tr>
<tr>
<td>7</td>
<td>1.392</td>
<td>0.110</td>
<td>0.024</td>
<td>0.681</td>
</tr>
<tr>
<td>8</td>
<td>1.383</td>
<td>0.009</td>
<td>0.025</td>
<td>0.706</td>
</tr>
<tr>
<td>9</td>
<td>1.185</td>
<td>0.198</td>
<td>0.021</td>
<td>0.726</td>
</tr>
<tr>
<td>10</td>
<td>1.139</td>
<td>0.046</td>
<td>0.020</td>
<td>0.746</td>
</tr>
<tr>
<td>11</td>
<td>1.081</td>
<td>0.057</td>
<td>0.019</td>
<td>0.765</td>
</tr>
<tr>
<td>12</td>
<td>0.918</td>
<td>0.163</td>
<td>0.016</td>
<td>0.782</td>
</tr>
<tr>
<td>13</td>
<td>0.838</td>
<td>0.080</td>
<td>0.015</td>
<td>0.796</td>
</tr>
<tr>
<td>14</td>
<td>0.783</td>
<td>0.055</td>
<td>0.014</td>
<td>0.810</td>
</tr>
<tr>
<td>15</td>
<td>0.748</td>
<td>0.034</td>
<td>0.013</td>
<td>0.823</td>
</tr>
<tr>
<td>16</td>
<td>0.718</td>
<td>0.031</td>
<td>0.013</td>
<td>0.836</td>
</tr>
<tr>
<td>17</td>
<td>0.644</td>
<td>0.074</td>
<td>0.011</td>
<td>0.847</td>
</tr>
<tr>
<td>18</td>
<td>0.620</td>
<td>0.024</td>
<td>0.011</td>
<td>0.858</td>
</tr>
<tr>
<td>19</td>
<td>0.590</td>
<td>0.030</td>
<td>0.010</td>
<td>0.868</td>
</tr>
<tr>
<td>20</td>
<td>0.590</td>
<td>0.030</td>
<td>0.010</td>
<td>0.868</td>
</tr>
</tbody>
</table>

We will first look at the results given by PCA. A (57×57) correlation matrix was used for the analysis. The scree graph for the eigenvalues is given in Figure 10.4. The first 20 eigenvalues are given in Table 10.7, as well as the proportion and the cumulative proportion. The reader should keep in mind that there are 57 eigenvalues in all. One notices that after the first five or six eigenvalues the decrease is very slow. After ten components, only 72% of the total variance has been explained. This increases to about 86% after another ten components. Another interesting graph is given in Figure 10.5. In this plot, the variations of the minimum communality (Section 2.2.9) with the number of components retained are shown. It is easy to
Figure 10.6 Spatial variation of the rotated loadings for the 10 principal components for the sulphate concentration in the precipitation example.

see that the slope of the increase of the minimum communality decreases after about ten components. Note that the minimum communality at this level is greater than 0.55.

Examination of the scree graph (Figure 10.4) and Table 10.7 indicates that ten components should be about the right number of principal components to retain.

The ten principal components were rotated using a varimax rotation. Note that the loadings were normalized so that the length of each principal component would be equal to the eigenvalue associated with it. The varimax rotation changes the eigenvalues as well as the loadings. However, the sum of the ten eigenvalues remains the same. The new eigenvalues
Figure 10.6 Continued.

are given in Table 10.8. If we compare those values to the first ten in Table 10.7, one sees that the varimax rotation has decreased the differences between the eigenvalues, with the highest decreases occurring in the differences between the first few eigenvalues.

The spatial distribution of the loadings for the ten components is shown in Figure 10.6. To help in interpreting those plots, the spatial distribution of SO$_2$ emissions in Eastern North America in 1985 is shown in Figure 10.7. Although the map in Figure 10.7 is based on

Table 10.8 Eigenvalues for PCA using the correlation matrix after varimax rotation for the sulphate concentration in the precipitation example.

<table>
<thead>
<tr>
<th>Principal Component Number</th>
<th>Eigenvalue</th>
<th>Difference</th>
<th>Proportion</th>
<th>Cumulative Proportion</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>12.805</td>
<td></td>
<td>0.225</td>
<td>0.225</td>
</tr>
<tr>
<td>2</td>
<td>7.558</td>
<td>5.248</td>
<td>0.133</td>
<td>0.357</td>
</tr>
<tr>
<td>3</td>
<td>4.929</td>
<td>2.629</td>
<td>0.086</td>
<td>0.444</td>
</tr>
<tr>
<td>4</td>
<td>3.595</td>
<td>1.331</td>
<td>0.063</td>
<td>0.507</td>
</tr>
<tr>
<td>5</td>
<td>2.955</td>
<td>0.642</td>
<td>0.052</td>
<td>0.559</td>
</tr>
<tr>
<td>6</td>
<td>2.251</td>
<td>0.704</td>
<td>0.039</td>
<td>0.598</td>
</tr>
<tr>
<td>7</td>
<td>1.984</td>
<td>0.267</td>
<td>0.035</td>
<td>0.633</td>
</tr>
<tr>
<td>8</td>
<td>1.956</td>
<td>0.028</td>
<td>0.034</td>
<td>0.667</td>
</tr>
<tr>
<td>9</td>
<td>1.864</td>
<td>0.092</td>
<td>0.033</td>
<td>0.700</td>
</tr>
<tr>
<td>10</td>
<td>1.510</td>
<td>0.355</td>
<td>0.026</td>
<td>0.726</td>
</tr>
</tbody>
</table>
emissions for an earlier period, the spatial distribution has not changed much in the intervening years, so the map can be used as a rough guide.

In the first principal component, the loadings are large within and just west of the region of highest emissions in the United States. This factor seems, therefore, to describe the impact of those emissions. The second component has its highest loadings in southern Ontario and Quebec in Canada. Judging by the pattern, it seems that this factor is related to the effects of Canadian emissions in southern Ontario and some of the emissions in the Midwest of the United States. The third component seems to be related to emissions in Manitoba and Saskatchewan in Canada. This component indicates that these sources are the major contributors to sulphate concentrations north of Lake Superior. The fourth component shows mainly the effect of western emission sources in the United States. The fifth and seventh components seem to describe the effects of local emissions in two adjacent (even overlapping) regions of the southeastern United States. The sixth component is interesting, as it seems to reflect sources in southern Ontario as well as local sources in Maine in the United States and Nova Scotia in Canada. The remaining components are more difficult to interpret, although it seems that part of component 8 reflects the effects of emissions in the United States on Maine, New Brunswick, and southeastern Quebec. Note that at present these interpretations are somewhat speculative, as time limitations have prevented the author from investigating the results in more detail.

Another interesting variable to consider is the scores. They can be used, as we have seen in the preceding examples, to show the temporal variation of the different components. Graphs of the time series for scores are given in Figure 10.8. The lines are estimates of the long-term trends using a kernel smoother (Sirois, 1993).
Figure 10.8 Temporal variation of scores for PCA for the sulphate concentration in the precipitation example.

The first thing we can notice is that the different components present different long-term trends. Only a small decrease at the beginning of the period is present in the first component. This fits with our interpretation that the first component reflects emissions in Pennsylvania, Ohio, and Indiana, because it is known that those emissions did not change much during the second half of the 1980s and the first half of the 1990s. A similar interpretation applies to the fact that the second component did not change during the monitoring period.
The third and fourth components show important decreases during the period considered. As these two components seem to be related to emissions in the western United States and western Canada respectively, they may point to a reduction of those emissions between 1988 and 1994.

The fifth, sixth, and seventh components seem not to have changed much during the sampling period. This is also true for the tenth components. The ninth component indicates a net decrease during the period, while the eighth component shows a decrease at the beginning.
of the period but an increase at the end. These behaviors are hard to interpret because of the difficulty, as we indicated earlier, of identifying these components at present.

The author would at this stage like to remind the reader that these results are only preliminary and more research will be necessary to arrive at more definitive interpretations.

Other dimensionality reduction techniques could (or should) be used to analyze these data.

To illustrate that other methods may produce different results, we can consider the solution obtained using PMF. The reader should notice that in PMF there is no order to the factors obtained. The author would also like to point out that the solution presented here may not be the best one, as only very rough estimations of uncertainties in the data were used. Because most of the data are composite samples, better estimation should be used. Another point is that although ten factors were used to be consistent with the PCA analysis, this may not be the best number of factors in this instance. Notwithstanding that, the results obtained are quite interesting when compared to those obtained from PCA.

The spatial distribution of the loadings for the ten factors found using PMF is shown in Figure 10.9, and the temporal variations of the scores are given in Figure 10.10. We have tried when ordering the factors to give them the same number as the PCA components to which the spatial distribution of the loadings is similar.
One notices to begin with that the first five factors are similar to the first principal component factors. Although the fit between the solutions is far from perfect, it indicates a similar relationship with the SO₂ emissions sources of Figure 10.7. It is also interesting to note that the estimated long-term trends of those five factors are similar to the ones found for the first five principal components.
The identification of the remaining five factors is more difficult. An interesting one is factor 10, which may be related to emissions from Wisconsin and Chicago.

This comparison of the results of the PCA and PMF analyses has shown that both techniques can extract the most important components or factors. They are not exactly the same, but they are close. This would also be true for most of the other dimensionality reduction techniques. The techniques would differ mainly in connection with components or factors contributing less of the total variance.

As a conclusion to this example, the author would like to repeat that this and the preceding example are only presented as illustrations of the techniques and not as a complete analysis of the data. More calculations and comparisons would be necessary to verify and arrive at a better interpretation of these results.
11. Short Guideline to the Use of Dimensionality Reduction Techniques

One of the main conclusions that the reader should draw from this short overview of dimensionality reduction techniques is that these methods cannot be used as "black boxes" that produce perfect results without human intervention. On the contrary, the researcher using them has many decisions and judgments to make. Because of that, it is customary in good books on the subjects to present some guidelines for the use of these methods. Without pretending that this short overview is part of this elite group of publications, we nevertheless present such a short guideline. Our guideline is a personal adaptation of one given by Rurumel (1970). Rurumel, in fact, uses his guideline as the basis for the structure of his book.

A flow chart of the guideline is given in Figure 11.1. This general plan can be used for any of the techniques described in the present overview, with only small adaptations having to be made. The rest of this section will present more questions than answers, its main goal being to formulate some of the questions a researcher has to ask during the analysis of his or her data. In most cases there are no universally valid answers. Most of the time it will be up to the researcher to determine the most satisfactory approach.

11.1 Research Goals

Before using any statistical technique, it is important to define what it is that we want to achieve in doing the analysis. Do we want just to explore the data, to have a feeling for the numbers without having any a priori conception about them? Or do we want to verify some theory that was put forward independently? Our intentions will influence the selection of the data and, even more so, the interpretation of the results.

11.2 Data Selection

This is a very important stage in any data analysis because of its influence on the results. One important question the researcher has to ask at this stage is: Are the data appropriate for the analysis I want to use? Not all data are appropriate for every analysis. The fact that the data are multivariate is not enough to justify the use of any of the techniques presented here. Indeed, there is no value
in using most of them if the variables are not correlated with each other. PCA, for example, gives the original variables when they are not correlated. Therefore, a consideration of the correlation matrix of the original variables should be the first step in using any of these methods.

11.3 Data Preprocessing

This stage of the analysis involves more than just looking to see if the data have to be transformed. First, one has to see if there are missing data. If there are gaps, how are we going to treat them? The same question applies to below-detection-limit data. Next we must ask if there are outliers? If yes, should we use a robust technique? The outliers may simply be the result of the way the data distribution is skewed. If so, would it not be better to try to find a transformation of the data that would result in a more symmetrical distribution? These are questions that have to be answered before starting any analysis. The answers will influence the next stage and, in fact, also be influenced by it.

11.4 Selection of the Method to Use

This is one of the more difficult steps. Unfortunately, no single techniques is best for all situations. Perhaps, the most satisfactory solution is to try a variety of methods and compare the results. Usually this will give a better understanding of the data. In most cases, there will be similarities between the different results, and these may help to distinguish the basic components or factors from those that may be due to sampling variations and/or errors in the data. We will return to this point later. Another decision that one should make at this point is whether or not to estimate the sampling uncertainties in the results of the analysis.

11.5 Estimation of the Communalities

This step is only necessary if one uses the principal factor method to fit the factor analysis model. As we have seen in Section 6.2.2.3, many different estimators are available. Iterative estimation of the communalities seems the most popular technique.

11.6 Selection of the Number of Components or Factors to Retain

In most of the techniques presented in this overview, this step is related to the next one, the estimation of factors. This is true for all techniques related to PCA. It is only in the case of factor analysis using the maximum likelihood method and PMF that one must choose the number of components to use in advance. The methods presented in Section 2.2.10 for selecting the number of components or factors to retain can be used or adapted for most of the methods reviewed here. For the methods necessitating an a priori selection of the number of components or factors to retain, it is necessary to fit the model for different choices and decide which of the possibilities is the most appropriate. Some of the techniques described in Section 2.2.10 can also help in this case. The one important point to remember is that the final components or factors must make physical sense to the researcher. If not, there is something wrong in the analysis, and it should be started again or it should be concluded that the method of analysis chosen was not appropriate for the data.
11.7 **Estimation of Factors**

Depending on the method selected, this step may either be straightforward or require further decisions. In most of the techniques, for example, one has to decide whether the variance-covariance matrix or the correlation matrix should be used. As we have seen, the two matrices will usually yield different results. The choice may be directed by the fact that the ranges (i.e., the differences between the maximum and minimum values) of the different variables are very different. In such a case, the use of the correlation matrix would be recommended. Another example can be taken from factor analysis, where it is necessary to determine which of the available methods for solving the factor analysis equation should be used. In cases where different methods can be used, the usual practice is to try a variety of methods and compare the results to see which is most satisfactory.

11.8 **Rotation of the Solution?**

As we have seen, most of the techniques presented in this overview produce solutions that can be rotated. Solutions should not be rotated automatically without being examined beforehand to see if the procedure is necessary. However, the solutions produced by PCA, FA, and other dimensionality reduction techniques are usually difficult to interpret and thus generally require rotation. If rotation is necessary, the researcher must then decide which method to use. Varimax rotation (Section 8.2.2) is the most popular. However, one should also consider testing other orthogonal rotations. Although the results of oblique rotational methods are more difficult to interpret, they should not be ignored either. As mentioned earlier, there is no reason, especially when working with physical processes, to assume that the hypothetical factors will be orthogonal. Even if we use dimensionality reduction techniques to try to estimate emission sources, as in the example in Section 10.1, there is good reason to expect that some of the sources will be correlated with each other. In summary, it is important to compare the results obtained using different rotational methods, both orthogonal and oblique. As we have seen in the example in Section 10.1, this helps to identify the most important factors or sources.

11.9 **Estimation of the Factor Scores**

Except in FA, this is straightforward, and in the case of PCA the scores can be easily calculated using the formula given in Section 2. Corresponding formulae were also given earlier for APCA and TPCA. In the case of PMF analysis, the scores are obtained at the same time as the loadings matrix. In the case of factor analysis, the scores, as we have seen in Section 6.2.3, have to be estimated, because they cannot be calculated directly from the observed data and the loadings as they can in PCA. Many techniques exist. We have presented two related methods here. The first of these, the weighted least squares method (Section 6.2.3.1) is quite general and can be applied to any method used to solve the factor analysis equation. The second, the regression method (Section 6.2.3.2), uses the maximum likelihood solution which made that method appropriate for this later solution.

11.10 **Scaling for Sample Masses**

This procedure applies only to APCA and TPCA and only in cases where the total sample masses are known. Usually it involves two linear regressions. One should note that
the usual estimations of uncertainties of the regression parameters do not apply in this case because it is assumed that the independent variables are not subject to uncertainties. This is not the case for the estimated loadings of the factors, however, because they are subject to sampling uncertainties to which measurement uncertainties would be added. Two techniques that can be used in this case are the jackknife and bootstrap methods (Section 9).

11.11 Interpretation

This is usually the most difficult part of any statistical analysis. Usually, a good understanding of the data helps at this stage. Also, one should have some ideas about the physical processes that generate and/or modify the values of the observed variables. But the golden rule is to use common sense, and any acceptable solution should make sense within the context of the processes studied.

It may also be useful in some cases to use more than one technique and compare the results obtained.
12. **Final Remarks**

At this stage, most readers would ask: "Which of these techniques should I use?" Unfortunately, there is no answer to that question because none of the methods presented here is satisfactory for all cases. Although the author has not used the PMF technique very often, it usually gives reasonable results. It cannot be recommended completely, however, because in some cases the author has encountered some problems of convergence. Also, the only program presently available is not easy to use and the documentation is limited.

The main point to keep in mind is that none of the techniques presented here should be used as some kind of black box into which we enter the data and from which comes the perfect solution. To do that is asking for trouble. Dimensionality reduction techniques are tools that can give a better understanding of observed data, but they must be used carefully. As with other data analysis techniques, the user must interacts with them but he or she must not abuse these techniques to try to prove some *a priori* theory. These techniques may be used to determine whether the data are compatible with some particular theory, but one should always keep in mind that many hypotheses about the data are made and that the data should respect them if the results obtained are to be valid.

The techniques presented here are only some of the dimensionality reduction methods currently available. Other methods like biplots (see Jolliffe, 1972; Everitt, 1978; and du Toit *et al.*, 1986 for more details) and multivariate scaling (see Kruskal and Wish, 1978; Davison, 1983; Dillion and Goldstein, 1984; and du Toit *et al.*, 1986 for more details) could have been added to the list presented here. But, as both time and space are limited, we leave readers the pleasure of discovering these techniques on their own. The author only hopes that this overview has increased the interest of the readers in these techniques, because they are versatile and can help in reaching a better understanding of the data. However, like everything else in life, they do not deliver results without work.
13. References


Hopper, F.J., 1986. *Analysis of atmospheric aerosol composition*. Thesis submitted to the faculty of graduate studies and research in partial fulfilment of the requirements for the degree of doctor of philosophy in meteorology, Department of Geography, The University of Alberta, Edmonton, Alberta, Canada.


Paatero, P., 1998. User’s guide for Positive Matrix Factorization programs PMF2.EXE and PMF3.EXE. University of Helsinki, Department of Physics, Box 9, FIN-00014 Helsinki/University, Finland.


Pearson, K., 1901. On lines and planes of closest fit to systems of points in space. *Phil. Mag.*, 2, 559-572.


hence $S$ is symmetrical and is normal in the limit as $n \to \infty$. Under $H_0$, the mean and variance of $S$, if there is ties in $x$, are equal to:

$$E[S] = 0$$

(4.4)

$$Var[S] = \frac{n(n-1)(2n+5) - \sum t(t-1)(2t+5)}{18}$$

(4.5)

where $t$ is the extent of any given tie (number of $x$'s involved in a given tie) and the summation in relation 4.5 is over all ties. Both Mann and Kendall derive the exact distribution of $S$ for $n \leq 10$ and show that even for $n = 10$ the normal approximation is excellent, provided one uses a continuity correction of one unit. One computes the standard normal variate $Z$ by:

$$Z = \begin{cases} 
\frac{S - 1}{(Var(S))^{1/2}} & \text{if } S > 0 \\
0 & \text{if } S = 0 \\
\frac{S + 1}{(Var(S))^{1/2}} & \text{if } S > 0 
\end{cases}$$

(4.6)

In a two-sided test, the hypothesis $H_0$ should be accepted if $|Z| \leq Z_{\alpha/2}$, where $F_N(Z_{\alpha/2}) = \alpha/2$, $F_N$ being the standard normal cumulative distribution function and $\alpha$ being the size of the significance level for the test. A positive value for $S$ indicates an increasing monotonic long-term trend and a negative $S$ a decreasing long-term trend.

To illustrate the use of the Mann-Kendall test, we will test for a monotonic trend using the 28-day mean $SO_4^{2-}$ concentration in precipitation at Montmorency, Quebec. The data are shown in Figure 4.1. Using relations 4.2 and 4.3, one obtains $S = -523$ and $Var(S) = 85083$. The $P$-value for the test using the normal approximation is equal to 0.074. Therefore, one cannot reject the hypothesis of no long-term trend at a 95% confidence level.

4.2.2.2 Farrell Test

In atmospheric chemistry, as in other fields, it often happens that a seasonal cycle is present in the data and that the long-term trend also varies with the season. Responding to that problem, Farrell (1980), following Sen (1968a), proposed the following test.

First subtract the monthly (or seasonal) average from each of the corresponding months (or seasons) in the $n$ years of data; i.e., calculate $x_i - \bar{x}_i$ for year $i = 1, 2, \ldots, n$ and month $j = 1, 2, \ldots, m$, and Figure 4.1 28-day precipitation-weighted-mean $SO_4^{2-}$ concentrations in precipitation at Montmorency, Quebec.
where $x_j = \frac{1}{n} \sum_{i=1}^{n} x_{ij}$. Then, rank all the differences from 1 to $nm$. This produces the matrix:

$$\begin{array}{cccccc}
1 & 2 & 3 & \cdots & m & \text{mean} \\
R_{11} & R_{12} & R_{13} & \cdots & R_{1m} & R_1 \\
R_{21} & R_{22} & R_{23} & \cdots & R_{2m} & R_2 \\
\vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\
n & R_{n1} & R_{n2} & R_{n3} & \cdots & R_{nm} & R_n \\
\text{mean} & R_1 & R_2 & R_3 & \cdots & R_m & R_\text{mean}
\end{array}$$

where $R_{ij} = \text{rank of } (x_{ij} - x_j)$ among the $nm$ differences. Ties are dealt with in the usual manner; to each of the $t$ tied values, assign the average of the next $t$ ranks. $R_i$ is the average rank for each year, equal to $\frac{1}{m} \sum_j R_{ij}$, and $R_j$ is the average rank for each month, equal to $\frac{1}{n} \sum_i R_{ij}$. We then calculate the statistic $T$, defined as:

$$T = \left( \frac{12m^2}{n(n+1)\sum_{i=1}^{n} \sum_{j=1}^{m} (R_{ij} - R_j)^2} \right)^{1/2} \left( \sum_{i=1}^{n} \left( i - \frac{n+1}{2} \right) \left( R_i - \frac{nm+1}{2} \right) \right)$$

As $m$ becomes larger the distribution of $T$ tends toward normality, with mean 0 (under the null hypothesis of no trend) and variance 1. The resulting statistical test is, therefore, straightforward: reject the hypothesis of no trend if $|T|$ exceeds a prespecified percentile of the normal distribution. Some limited Monte Carlo studies indicate that the normal approximation is reasonable even for small samples.

The period described by the second index may be a month, week, or season. Note that this test does not allow for missing data. Farrell (1980) suggests filling in missing data with least squares estimates. That is, one estimates, via least squares, the parameter of the following model:

$$x_{ij} = u + a_i + b_j + e_{ij}$$

with $i = 1, 2, \ldots, n$ and $j = 1, 2, \ldots, m$ and

$$\sum_i a_i = \sum_j b_j = 0$$

and then uses those estimates to calculate values for the missing $x_{ij}$. Note that this procedure is not recommended by van Belle and Hughes (1984) if many values are missing, because the variance estimate implicit in relation 4.7 is then too small.
The reader should also note that this test assumes that the trends for the \( m \) months or seasons are homogeneous (i.e., the trend is the same in each month or season). Therefore, the test is invalid if the trends are heterogeneous.

4.2.2.3 Seasonal Kendall Test

Hirsch \textit{et al.} (1982) have extended the Mann-Kendall test to take into account any seasonality in the data. They called this test the \textit{seasonal Kendall test}.

In their test, the hypothesis to test is modified as follows. Let

\[ X = (X_1, X_2, \ldots, X_{12}) \tag{4.10} \]

and

\[ X_i = (x_{i1}, x_{i2}, \ldots, x_{i_n}) \tag{4.11} \]

where \( X \) is the entire sample made up of subsamples \( X_1, X_2, \ldots, X_{12} \) (one for each month), and each subsample contains the \( n_i \) annual values from month \( i \). Note that \( n_i \) may be different from \( n_j \) (\( i \neq j \)). Therefore, missing values are allowed. We will discuss a modification to the test to allow more than one sample per month. Even if the test is presented here for 12 monthly values per year, it can also be used for 52 weekly values or 4 seasonal values or other combinations. The null hypothesis for the seasonal Kendall test is:

\( H_0: \) \( X \) is a sample of independent random variable \( (x_i) \) and \( X_i \) is a subsample of independent and identically distributed random variables \( i = 1, 2, \ldots, 12. \)

and

\( H_1: \) for one or more months there is a monotonic increasing or decreasing trend.

We define \( S_i \) as

\[ S_i = \sum_{k=1}^{n_i-1} \sum_{j=k+1}^{n_i} \text{sgn}(x_{ij} - x_{ik}) \tag{4.12} \]

Under \( H_0 \), the subsample \( X_i \) satisfies the null hypothesis \( H_0 \) of the Mann-Kendall test. Therefore, one has (Section 4.2.2.1):

\[ E[S_i] = 0 \tag{4.13} \]

\[ \text{Var}[S_i] = \frac{n_i(n_i - 1)(2n_i + 5) - \sum_{i=1}^{12} t_i(t_i - 1)(2t_i + 5)}{18} \tag{4.14} \]

where \( t_i \) is the extent of a given tie in month \( i \). The distribution of \( S_i \) is normal in the limit as \( n_i \to \infty \). We then define \( S' = \sum_{i=1}^{12} S_i \) and can derive its expectation, variance, and limit distribution.

\[ E[S'] = \sum_{i=1}^{12} E[S_i] = 0 \tag{4.15} \]
\[ \text{Var}[S'] = \sum_{i=1}^{12} \text{Var}[S_i] + \sum_{i=1}^{12} \sum_{j=1, j \neq i}^{12} \text{cov}(S_i, S_j) \]  

(4.16)

Now \( S_i \) and \( S_j \) (\( i \neq j \)) are functions of independent random variables \( (S_i = f(X_i), S_j = f(X_j)) \), and \( X_i \cap X_j = \emptyset \) because \( X_i \) and \( X_j \) are the data from months \( i \) and \( j \) respectively, and all elements of \( X \) are independent), so \( \text{cov}(S_i, S_j) = 0 \). Note that this is only true if there is no autocorrelation in the time series. We will discuss an extension to the test that includes autocorrelation in the next section. Thus, we have:

\[ \text{Var}[S'] = \sum_{i=1}^{12} \text{Var}[S_i] = \sum_{i=1}^{12} \left\{ \frac{n_i(n_i - 1)(2n_i - 5) - \sum_{t_i} t_i(t_i - 1)(2t_i + 5)}{18} \right\} \]  

(4.17)

For using the normal approximation we define the standard normal deviate \( S' \) as:

\[ Z' = \begin{cases} 
\frac{S' - 1}{(\text{Var}(S'))^{1/2}} & \text{if } S' > 0 \\
0 & \text{if } S' = 0 \\
\frac{S' + 1}{(\text{Var}(S'))^{1/2}} & \text{if } S' < 0 
\end{cases} \]  

(4.18)

which is adequate for \( n_i \geq 3 \) for all \( i \).

When there are multiple observations for each month (or period), the preceding relation can be modified. The basic idea is to treat the data in the same month as tied observations with respect to their time of occurrence. Relation 4.12 becomes:

\[ S_i = \sum_{k=1}^{n_i - 1} \sum_{j=i+1}^{n_i} \sum_{t_{ij}} \sum_{t_{ij}} \text{sgn}(x_{ij1} - x_{j2}) \]  

(4.19)

where \( x_{ijk} \) is observation \( k \) for month \( i \) and year \( j \) and \( n_{ij} \) is the number of observations in month \( i \) and year \( j \). Note that the \( n_{ij} \) can be different, so data can be missing without affecting the results. Because of the ties in time the variance of \( S_i \) must be modified. It is equal to (Kendall, 1975):

\[ \text{Var}[S_i] = \frac{n_i(n_i - 1)(2n_i - 5) - \sum_{t_i} t_i(t_i - 1)(2t_i + 5) - \sum_{k=1}^{n_i} n_{ik}(n_{ik} - 1)(2n_{ik} + 5)}{18} \]

\[ + \frac{\left\{ \sum_{t_i} t_i(t_i - 1)(t_i - 2) \right\} \left\{ \sum_{k=1}^{n_i} n_{ik}(n_{ik} - 1)(n_{ik} - 2) \right\}}{9 n_i(n_i - 1)(n_i - 2)} \]

\[ + \frac{\left\{ \sum_{t_i} t_i(t_i - 1) \right\} \left\{ \sum_{k=1}^{n_i} n_{ik}(n_{ik} - 1) \right\}}{2 n_i(n_i - 1)} \]  

(4.20)
$S'$ is still defined as $\sum_{i=1}^{12} S_i$ and $\text{Var}[S'] = \sum_{i=1}^{12} \text{Var}[S_i]$. Relation 4.18 can still be used to test the significance of the monotonic long-term trend.

Like the Farrell test, this test is not valid if the long-term trends in the different months or seasons are heterogeneous.

Using the data presented in Figure 4.1 and discussed in Section 4.2.2.1, the seasonal Kendall test gives us $S' = -84$ and $\text{Var}[S'] = 707.33$, and the P-value is equal to 0.0018. We can thus conclude that a decreasing monotonic trend that is statistically significant at a 95% confidence level is present in the data.

4.2.2.4 Hirsch-Slack Test

As mentioned in the preceding section, the seasonal Kendall test is incorrect when the data are correlated because the term $\sum_{i=1}^{12} \sum_{j=1}^{12} \text{cov}(S_i, S_j)$ in relation 4.16 was put equal to 0.

Hirsch and Slack (1984) generalized the seasonal Kendall test to correlated data using a consistent estimator for $\text{cov}(S_i, S_j)$ developed by Dietz and Killeen (1981). It can be written as:

$$\text{cov}(S_g, S_h) = K_{gh} \text{sgn}(x_i, -x_i) / 3 + (n^3 - n) r_{gh} / 9$$ (4.21)

where

$$K_{gh} = \sum_{i<j} \text{sgn}[(x_{ij} - x_{ik})(x_{jh} - x_{jh})]$$ (4.22)

$$r_{gh} = \frac{3}{n^3 - n} \sum_{i,j,k} \text{sgn}[(x_{ij} - x_{ik})(x_{ji} - x_{jh})]$$ (4.23)

If there are no ties and no missing values, $r_{gh}$ is the Spearman's correlation coefficient (Lehmann, 1975) for seasons or months $g$ and $h$. If there are no missing values, relation 4.21 reduces to:

$$\text{cov}(S_g, S_h) = \left[ K_{gh} + 4 \sum_{i=1}^{n} R_{ig} R_{ih} - n(n+1)^2 \right] / 3$$ (4.24)

where

$$R_{ig} = \left[ n + 1 + \sum_{i=1}^{n} \text{sgn}(x_{ig} - x_{ig}) \right] / 2$$ (4.25)

If missing data are present, the definition of the $\text{sgn}$ function in relation 4.3 is extended. We define $\text{sgn}(x_{ig} - x_{ig})$ to be zero if either $x_{ig}$ or $x_{ig}$ is missing. Relation 4.25 becomes:

$$R_{ig} = \left[ n + 1 + \sum_{i=1}^{n} \text{sgn}(x_{ig} - x_{ig}) \right] / 2$$ (4.26)

where $n_g$ is the number of nonmissing observations for season $g$. Now the ranks of the nonmissing observations are unchanged, and each missing value is assigned the average or
midrank value \((n_{s}+1)/2\). The Mann-Kendall test statistic \(S_{g}\) (relation 4.12) is unchanged and its variance remains the same, namely,

\[
\text{Var}[S_{s}] = \frac{n_{s}(n_{s} - 1)(2n_{s} + 5) - \sum t_{s}(t_{s} - 1)(2t_{s} + 5)}{18} \tag{4.27}
\]

In relation 4.24, \(K_{gh}\) remains unchanged, but \(r_{gh}\) takes on a new value so that

\[
\text{cov}(S_{g},S_{h}) = \left[ K_{gh} + 4 \sum_{i=1}^{n} R_{ig} R_{ih} - n(n_{s} + 1)(n_{s} + 1) \right]/3 \tag{4.28}
\]

When there are below-detection-limit data, they may be replaced arbitrarily by a small value which is less than the detection limit, because the nonparametric tests are based on ranks rather than magnitudes. Therefore, all censored values (i.e., below-detection-limit values) may be viewed as sharing the same rank. Therefore, the problem of dealing with censored values reduces to the problem of dealing with ties. If there are ties but no missing values, \(S_{s}\) is computed using relation 4.12 and \(S'\) is the sum of the \(S_{s}\) values. The variance of \(S_{s}\) is given by relation 4.14. The formula for \(\text{cov}(S_{g},S_{h})\) remains the same, except that the midranks are used in assigning the values of \(R_{ig}\) for relation 4.24. Thus, if there are \(t_{j}\) censored values, they all have rank \(t_{j}(t_{j}-1)/2\).

When ties and missing values are both present, one must combine the modifications described for missing values and for ties. Note that, for this test there is only one observation per month (or season) of each year.

The original seasonal Kendall test (Section 4.2.2.3) is more powerful than the present test, but the significance level of the test can be seriously in error if there is serial correlation. The present test requires some sacrifice of power but offers a more nearly exact statement of significance for a wide variety of cases. Thus, choosing between the seasonal Kendall test and the Hirsch-Slack's test involves a trade-off.

Like the original seasonal Kendall test, the Hirsch and Slack test assumes that the trends in the different months or seasons are homogeneous. Thus both are invalid if that hypothesis is not fulfilled.

4.2.2.5 Dietz-Killeen Test

The test proposed by Dietz and Killeen (1981) for monotonic long-term trends in multivariate data can also be adapted for use with monthly (or seasonal) data covering many years. As in the case of the seasonal trend, let

\[
X = (X_{1}, X_{2}, \ldots, X_{m}) \tag{4.29}
\]

and

\[
X_{i} = (x_{i1}, x_{i2}, \ldots, x_{im}) \tag{4.30}
\]

\(X\) is the entire sample made up of subsamples \(X_{1}, X_{2}, \ldots, X_{p}\) (one for each period); \(x_{ij}\) is the observation for year \(i\) and period \(j\). Note that \(m\) equals 12 for monthly data. The null hypothesis is that the \(p\)-vectors are randomly ordered versus the alternative hypothesis that there is a monotonic trend in one or more of the \(m\) periods.

Let
an m-vector, and
\[
S = (S_1, \ldots, S_m)^T
\]  
(4.31)

an (mm) matrix, where
\[
\Sigma = (\sigma_{gh})
\]  
(4.32)

and
\[
S_g = \sum_{i<j} \text{sgn}(x_{ji} - x_{ij}) \quad g = 1, \ldots, m
\]  
(4.33)

\[
\sigma_{ss} = n(n-1)(2n+5)/18 \quad g = 1, \ldots, m
\]  
(4.34)

and
\[
\sigma_{gh} = \frac{1}{3} \left[ \sum_{i<j} \text{sgn}(x_{ji} - x_{ij})(x_{ij} - x_{ji}) + \sum_{(j,i,k)} \text{sgn}(x_{jk} - x_{kj})(x_{kj} - x_{jk}) \right]
\]  
(4.35)

where g ≠ h. This last relation gives the estimated covariances of \(S_g\) and \(S_h\). The statistic \(T\) defined as:
\[
T = S^T \Sigma^{-1} S
\]  
(4.36)

is asymptotically \(\chi^2_q\) (i.e., chi-squared with \(q\) degrees of freedom), where \(\Sigma^{-1}\) is any generalized inverse of \(\Sigma\) and \(q \leq m\) is the rank of \(\Sigma\). In the presence of ties, relation 4.34 becomes
\[
\sigma_{ss} = \frac{n_s(n_s-1)(2n_s+5) - \sum t_i(t_i-1)(2t_i+5)}{18} \quad g = 1, \ldots, m
\]  
(4.37)

and the remaining relations do not change. Therefore, the hypothesis of no monotonic trend in the data can be rejected if \(T\) is greater than \(\chi^2_q\).

Note that this test is very conservative. Hirsch and Slack (1984) argued, on the basis of very limited Monte Carlo simulations, that it is probably applicable only to monthly data where at least 40 years of data are available.

Note that, unlike the seasonal tests presented earlier, this test is valid even if the trends in the different seasons are heterogeneous.

4.2.2.6 van Belle-Hughes' Test

As we have seen in the descriptions of the Farrell, seasonal Kendall, and Hirsch-Slack tests, these tests are invalid if the trends in the different months or seasons are heterogeneous. Only the Dietz-Killeen test is valid in this situation, but it may only be applicable if a very long-time series is available. All the other tests of seasonal data mentioned above will be misleading if the trends are not homogeneous among months or seasons, especially when there are opposing trends in different seasons.

van Belle and Hughes (1984) developed a 2-way, anova-like, nonparametric test which can test first for the homogeneity of the trend at different locations and different seasons and then test for the presence of trends if the hypothesis of homogeneity cannot be rejected. For a single location, the test is constructed as follows.
Let the variable \( Z_j \) be defined as

\[
Z_j = \frac{S_j}{(\text{Var}(S_j))^{1/2}}
\]  

(4.38)

where \( S_j \) and \( \text{Var}(S_j) \) are defined in relation 4.12 and 4.14 respectively. Under the hypothesis of no trend in season \( j \), \( Z_j \) has, approximately, a chi-squared distribution with 1 degree of freedom (this follows from the asymptotic normality of \( S_j \); see above). In addition, if the seasonal observations are far enough apart, then the \( Z_j \) will be nearly independent, which is assumed in what follows. The overall statistic is

\[
\chi^2_{\text{total}} = \sum_{j=1}^{m} Z_j^2
\]  

(4.39)

which is, approximately, a \( \chi^2_m \) under the null hypothesis of no trend in any of the seasons. Note that each season’s value is normalized before the total statistic is calculated. This normalization is equivalent to weighting each monthly or seasonal statistic with a weight that is inversely proportional to its variance. Therefore, the months or seasons with high variances contribute less to the total statistic.

A large value of \( \chi^2_{\text{Total}} \) is not really meaningful because it still fails to distinguish heterogeneity between the individual \( Z_j \)'s from the overall trend. To do that, we will partition \( \chi^2_{\text{Total}} \) in the following manner:

\[
\chi^2_{\text{Total}} = \chi^2_{\text{homogeneous}} + \chi^2_{\text{trend}}
\]  

(4.40)

where \( \chi^2_{\text{trend}} = m\bar{Z}^2 \) and \( \bar{Z} = \frac{1}{m} \sum_{j=1}^{m} Z_j \). \( \chi^2_{\text{homogeneous}} \) is obtained by subtracting \( \chi^2_{\text{trend}} \) from \( \chi^2_{\text{Total}} \).

Under the hypothesis that there is no trend in any of the months or seasons, \( \chi^2_{\text{homogeneous}} \) and \( \chi^2_{\text{trend}} \) have, respectively, chi-squared distributions, with \( m-1 \) and 1 degrees of freedom. Thus homogeneity of trend may be tested by comparing \( \chi^2_{\text{homogeneous}} \) to tables of \( \chi^2_{m-1} \) distribution. If \( \chi^2_{\text{homogeneous}} \) is not significant, then a valid test for a common trend is possible by referring \( \chi^2_{\text{trend}} \) to tables of the \( \chi^2_1 \) distribution. If \( \chi^2_{\text{homogeneous}} \) is significant, then evaluation of \( \chi^2_{\text{trend}} \) is not appropriate. Trend tests for each season may still be obtained from the individual \( Z_j \) if desired.

In cases where the seasonal trends are different, it may be interesting to determine if there is a pattern to the heterogeneity. For example, one may want to know if the trends are similar for the winter months. To accomplish that, we use what is called in statistical literature a contrast. A contrast is defined by a linear expression of the form:

\[
\sum_{j=1}^{m} c_j S_j = c_0
\]  

(4.41)

where the \( c_j, j = 1, 2, \ldots, m \), are constants such that \( \sum_j c_j = 0 \). The variance of the contrast is given by:
For example, to test whether the trend in December is different from the trends in January and February, one can define the contrast as $S_{12} - 0.5(S_1 + S_2)$. We then compare

$$\frac{S_{12} - 0.5(S_1 + S_2)}{Var(S_{12}) + 0.25(Var(S_1) + Var(S_2))}$$

(4.43)

to the chi-squared distribution with 1 degree of freedom. If the contrast is suggested by the data, Fleiss (1981) recommends using two degrees of freedom. One can also use techniques such as the Newman-Keuls method (see Milliken and Johnson, 1984) to try to place the trends in homogeneous groups.

The presence of more than one value per season can be handled in the same way as in the seasonal Kendall test (see Section 4.2.2.3). Missing data can also be handled as they are in the seasonal Kendall test.

As mentioned earlier, the van Belle-Hughes test can be used where there are many sites. To test for trend, one must first test for homogeneity between the sites, secondly for homogeneity between seasons, and then for any possibility of interactions between sites and seasons. If we find that the trends are homogeneous across seasons and sites and no interactions are present, the test for the statistical significance for trend can be made. In these tests, it is assumed that the data are independent across seasons and sites. The reader is referred to van Belle and Hughes (1984) or Gilbert (1987) for more details.

4.2.3 Slope Estimation

In the preceding section, we have described some nonparametric tests for detecting monotonic trends. Once we have determined that a trend exists, we would like to be able to estimate that trend. For the two techniques presented here, it is assumed that $f(t)$ in relation 4.1 is equal to

$$f(t) = \alpha t + \beta$$

(4.44)

Using these techniques, we will get nonparametric estimates of the coefficients of this linear relationship. The first estimator assumes that no seasonal cycle is present in the data, and the second allows for such cycles.

4.2.3.1 Sen's Slope Estimator

A consistent nonparametric estimator for the coefficients of a linear regression (relation 4.44) was proposed by Theil (1950). It is based on Kendall’s Tau (see Kendall, 1938, 1975) and was modified by Sen (1968b) to include the possibility of ties in the $t_i$.

If one has $n$ values for the pair $(t_i, C_i)$, the coefficients of the linear relationship (relation 4.44) can be estimated as follows:

1. The $N$ values $A_{ij} = (C_i - C_j)/(t_i - t_j)$ for $i = 1, 2, \ldots, n$ and $j = 1, 2, \ldots, n$; $j > i$ and $t_i \neq t_j$. The estimate of $\alpha(\alpha^2)$ is given by the median of the $A_{ij}$. The estimate of $\beta$ is defined as (after ordering the $A_{ij}$):

$$\text{median}(A_{ij}) \times \frac{\sum (t_i - t_j)}{n}$$
\[ a^s = \begin{cases} 
A_{i(N+1)/2} & \text{if } N \text{ is odd} \\
\frac{1}{2} (A_{iN/2} + A_{i(N+2)/2}) & \text{if } N \text{ is even}
\end{cases} \] (4.45)

2. To obtain an estimate of \( \beta \), we first calculate the \( n \) values \( C_{i} - a^s t_i \) (Theil, 1950). The median value gives an estimate of \( \beta(b^3) \).

A 100(1-\( \varepsilon \))% two-sided confidence interval about the true slope for \( a^s \), which is valid if \( n \) is greater than 10 and if there are not many ties, can be obtained as follows (note that a nonparametric technique is given in Theil (1950) and Sen (1968b)):

1. Choose the desired confidence coefficient \( \varepsilon \) and get \( Z_{1-\varepsilon/2} \) where \( Z_p \) is the upper 100(1-\( \varepsilon \))th percentile of the standard normal distribution.

2. Compute \( C_s = Z_{1-\varepsilon/2}[\text{Var}(S)]^{1/2} \), where \( \text{Var}(S) \) is computed using relations 4.5 or 4.20, depending on whether there are multiple observations for each time. Note that the latter relation must be modified to account for the fact that there is only one season in the present case.

3. Compute \( M_1 = (N-C_s) / 2 \) and \( M_2 = (N+C_s) / 2 \).

4. The lower and upper limits of the confidence interval are the \( M_1 \)th largest and \( (M_2+1) \)th largest of the \( N \) ordered \( A_{i}\) respectively.

4.2.3.2 Seasonal Kendall’s Slope Estimator (Gilbert, 1987)

The Sen’s slope estimator can be extended to take into account the presence of seasonality as follows:

1. First, compute for each season \( k \) ( \( k=1, 2, ..., m \) ) as in the preceding section the \( N_k \) values \( A_{i|k} = (C_{i|k} - C_{i,k}) / (t_{i|k} - t_{i,k}) \), where \( C_{i,k} \) is the concentration for year \( i \) and season \( k \).

2. Then rank the \( N = N_1 + N_2 + ... + N_m \) individual slope estimates and find their median. This median is the seasonal Kendall slope estimator.

Note that the slopes for the different seasons are assumed to be homogeneous. It is a straightforward exercise to extend this technique to cases in which there is more than one observation per season.

A 100(1-\( \varepsilon \))% two-sided confidence interval about the true slope can be obtained as in the preceding section.

1. Choose the desired confidence coefficient \( \varepsilon \) and get \( Z_{1-\varepsilon/2} \) where \( Z_p \) is the upper 100(1-\( \varepsilon \))th percentile of the standard normal distribution.

2. Compute \( C_s = Z_{1-\varepsilon/2}[\text{Var}(S')]^{1/2} \), where \( \text{Var}(S') \) is equal to \( \sum_{i=1}^{n} \text{Var}(S_i) \) and the \( S_i \) are computed as in the preceding section. Note that the data are assumed to be independent as in the seasonal Kendall’s test.

3. Compute \( M_1 = (N-C_s) / 2 \) and \( M_2 = (N+C_s) / 2 \).
4. The lower and upper limits of the confidence interval are the $M_1$th largest and $(M_2+1)$th largest of the $N$ ordered $A_{ij}$ respectively.

4.2.4 Discussion

The techniques described in this section all have the advantage that no assumption is made about the distribution of the residuals (relation 4.1), except that it does not change with time (i.e., the variance is constant). It is also assumed for most of these tests that the residuals are independent. As most observed data in atmospheric chemistry have seasonal cycles, the Mann-Kendall test is limited to annual data only.

Although not having to make assumptions about the shape of the distribution of the residuals is an advantage, the nonparametric tests still present many disadvantages. First, the trends have to be assumed monotonic for the tests or a straight line for the estimators. This limits their applicability, because in many cases these hypotheses cannot be fulfilled in atmospheric chemistry. We have also seen that, except for the Dietz-Killeen test, the seasonal tests all assume that the trends in the different seasons are homogeneous (i.e., the same) which is not necessarily true.

Therefore, the applicability of these nonparametric tests in atmospheric chemistry may be limited. However, they may be useful in some situations.

It should be noted that the nonparametric tests presented here are only a few of the many possible tests that can be used in special situations. An interesting example of the other tests that are available is the Jonckheere test (Jonckheere, 1954; Lehmann, 1975), which uses the Mann-Whitney statistic to test for step trends in data (see Hirsch and Gilroy, 1985; Blanchard et al., 1996).

To the knowledge of the author, only the seasonal Kendall test, of all the nonparametric tests presented here, has been used in atmospheric chemistry. Examples of its use can be found in Schertz and Hirsch (1983), Sisterton et al. (1990), Pollack et al. (1993), and Baier and Cohn (1993).

4.3 Smoothers

4.3.1 Introduction

We have seen in Section 2.8 how smoothers can be useful in the exploratory analysis of a time series. The problem in using smoothers to estimate trends is that, for most smoothing techniques, it is not possible to test the statistical significance of any estimated trend. Even for the few techniques for which confidence intervals for the estimated trend can be constructed, these intervals are difficult to estimate, valid only for large-samples, and subject to theoretical criticisms.

The basic conceptual model behind the smoothing techniques is the same as for the nonparametric techniques presented in the preceding section (relation 4.1) and is:

$$C_i = f(t_i) + e_i$$  \hspace{1cm} (4.46)

It is assumed that $f(t)$ is a continuous function. The $e_i$'s are unobserved random variables representing error terms. Unless otherwise specified, it is assumed that the error terms are
uncorrelated, with mean zero and constant variance, $\sigma^2$. The goal of the smoother techniques presented in this section is to obtain an estimate of $f(t)$ if $n$ values of $C_i$ are available.

In this overview, we will try to summarize some of the most popular smoothing techniques. Programs or functions for most of these techniques can be found in most statistical packages, such as SAS®, S-Plus®, and other similar programs. Where appropriate, we will discuss briefly the possibility of constructing confidence intervals for the estimated trends.

We will begin with two simple smoothing techniques, namely, (1) the *running mean* and (2) the *running median*. Then we will discuss more complex techniques like the *loess* smoother or regression, the *kernel* smoother or regression, and finally the *spline* smoother technique.

Basic to all these techniques, except the last, is the notion of *span* or *bandwidth*. Its definition can be given as follows. If we want to estimate $f(t)$ at time $t_0$, the span of the smoother, $h$, is defined so that only the observations in the interval $[t_0-h, t_0+h]$ are used in the calculation of $\hat{f}(t_0)$. We will denote the number of valid observations in that interval as $N_h(t_0)$. This is illustrated in Figure 4.2. All the data in the shaded area will be used in the calculation of the smoother estimate at $t_0$. Note that the span or bandwidth is sometimes specified as a fraction $\alpha$ ($0 < \alpha < 1$) of the total number of data items. In those cases, one first calculates a constant value for the number of data in the interval. This is equal to the integer part of $\alpha n$. Then, for each time $t_0$, a value of $h$ is found such that $N_h(t_0)$ is equal to that number. Note that if there are missing data and/or the data are not sampled at equal intervals, $h$ will be a function of $t_0$.

In most of these techniques, a problem occurs at both ends of the time series. This is called the *border* or *edge* effect. If we define $t_{\text{min}}$ and $t_{\text{max}}$ as the lowest and highest values for $t$, then the border effect occurs when $t_0 < t_{\text{min}} + h$ and $t_0 > t_{\text{max}} - h$. As $t_0$ gets closer to $t_{\text{min}}$ or $t_{\text{max}}$, the number of data available, $N_h(t_0)$, for the calculation will decrease and the uncertainties in the estimate, $\hat{f}(t_0)$, will increase. For most of the estimators there is no way to compensate for that problem. Where such a technique does exist, we will mention it. In most cases, the best solution is either not to estimate $f(t)$ in those regions or to indicate that large uncertainties exist in them.

4.3.2 *Running Mean Smoother*

The simplest of the smoothing techniques is the running mean, also called the *window estimate* or the *moving average*. The running mean estimate of $f(t)$ at $t_0$ is:

![Figure 4.2 Illustration of the definition of span or bandwidth of a smoother.](image)
In other words, the estimate of \( f(t) \) is the mean value of the point within the interval \([t_0-h, t_0+h]\). The smoothness of \( \hat{f}_h(t) \) depends on \( h \). Note that the choice of \( h \) must be large enough to keep \( N_h(t_0) \) from becoming too small. The choice of \( h \) is critical to the result obtained. This is illustrated in Figure 4.3, where different values of \( h \) are used to estimate the long-term trend of ambient SO\(_2\) air concentrations at Algoma. For \( h \) less than a year, the smoother followed the seasonal cycles quite well. At \( h = 1 \) year, the seasonal cycles are smoothed out and one starts to detect the long-term trend. For \( h \) equal 2 and 3 years, the estimates of the long-term trend are very similar and they are smoother than for \( h = 1 \) year. One can conclude that in this case, a value of \( h \) equal to about 2 years should be the right value. Note that the border or edge effects seems to be small in this case.

Missing data do not present any problem, as we have only to ignore them. If they represent a large fraction of the total number of data, they may influence the selection of the span and therefore the smoothness of \( \hat{f}_h(t) \).

The running mean estimator, like the mean, is very sensitive to the presence of outliers. One technique that can be used to "robustify" it is to employ methods like the \( \alpha \)-trimmed estimator or the \( \alpha \)-Winzorized estimator of the means (see David, 1981; Barnett and Lewis, 1984). If \( n \) is the number of observations (= \( N_h(t_0) \) in the present context), let \( r \) equal the integer part of \( \alpha n \); then the \( \alpha \)-trimmed means is equal to:

\[
\hat{f}_h(t_0) = \frac{1}{N_h(t_0)} \sum_{i \text{ such that } |t_i-t_0| \leq h} C_i
\]

(4.47)

Figure 4.3 Running mean smoother trend estimates for four different bandwidths for ambient SO\(_2\) air concentration at Algoma, Ontario.
\[ T_n(r) = \frac{1}{n-2r} \sum_{i=r+1}^{n-r} C_i \]  

(4.48)

where \( C_i \) denotes the observations sorted in increasing order. The \( \alpha \)-Winzorized mean is defined as:

\[ W_n(r) = \frac{1}{n} \left\{ (r+1)(C_{[r+1]} + C_{[n-r]}) + \sum_{i=r+2}^{n-r-1} C_i \right\} \]

(4.49)

for \( 0 < r < \frac{1}{2}(n-1) \) and

\[ W_n(r) = C_{[(n+1)/2]} \]

(4.50)

if \( n \) is odd and \( r = \frac{1}{2}(n-1) \). Although these equations are for symmetric \( \alpha \)-trimmed and \( \alpha \)-Winzorized mean estimators, they can be modified for trimming or Winzorizing only large values, which would be more appropriate in some situations (see Barnett and Lewis, 1984). Other robust estimators for the mean include L-estimator, M-estimator, and R-estimator (see Huber, 1981; Barnett and Lewis, 1984; Hampel et al., 1986).

Outliers may be due to many causes. They may only be apparent in the sense that the distribution of the \( e_i \)'s in 4.46 may not be symmetrical but skewed to the left as in the log-normal or gamma distributions. In these cases, a simple transformation (see Section 2.9) before the use of the smoother may eliminate these outliers. The reader should keep in mind that use of the running mean is not recommended if the distribution of the residual is very skewed.

Border effects are very important in this technique and cannot be reduced.

Finally, there is no technique to fix the span or bandwidth. We are thus reduced to trying many values and selecting the value that seems to give a smooth curve without eliminating all variations.

**Figure 4.4**: Running median smoother trend estimates for four different bandwidths for ambient \( \text{SO}_2 \) air concentration at Algoma, Ontario.
4.3.3 Running Median Smoother

This technique is very similar to the preceding method. The only difference is that the median of the data in the window defined by the span is used in lieu of the mean. Therefore, we can define the estimate of $f(t)$ at $t_0$ as:

$$\hat{f}_h(t_0) = \text{med} \left( \frac{1}{2} \left( C(t_i) \right) \right)$$

(4.51)

As in the case of the running mean smoother, the smoothness of $\hat{f}_h(t)$ is a function of $h$. This is illustrated in Figure 4.4. If one compares these results with those for the running mean smoother, it will be noticed that they are quite similar, especially with $h \geq 1$ year. Similar results cannot be expected in all cases, however, especially if the distribution of the data is skewed toward high values. As with the running mean smoother, a bandwidth of around 2 years seems to be right.

Unlike the running mean, however, the running median is quite resistant to the presence of outliers and can be used for skewed distributions. Missing values can be ignored, but they may influence the selection of the span if they represent a large fraction of the data.

Border effects are also important for this smoother, as with the running mean the selection of the span is arbitrary. Therefore, the user has to try many values and select one that gives a good compromise between smoothness and detail.

The great disadvantage of the running median compared with the running mean is that it is very computer calculation intensive, because for each value of $t_0$, the $N_h(t_0)$ data have to be sorted in order before calculation of the median. This may take a long time if $\hat{f}_h(t)$ is estimated at many values of $t_0$. Friedman and Stuetzle (1982) have presented an algorithm that improves the speed of running median smoothing.

4.3.4 Robust Locally Weighted Regression (Loess) Smoother

The idea of estimating $\hat{f}_h(t_0)$ by the mean of the data in the interval $[t_0-h, t_0+h]$ can be extended by using the estimated value at $t_0$ of a polynomial of order $d$ fitted to the data (within the interval) by least squares. This estimator can be written as:

$$\hat{f}_h(t_0) = \beta_0 + \beta_1 t_0 + \beta_2 t_0^2 + \cdots + \beta_d t_0^d$$

(4.52)

where the coefficient $\beta_i$ denotes the coefficients that minimize:

$$\sum_{i=1}^{N_h(t_0)} \left( C(t_i) - \beta_0 - \beta_1 t_i - \beta_2 t_i^2 - \cdots - \beta_d t_i^d \right)^2$$

(4.53)

where the summation is on all the observations in the interval $[t_0-h, t_0+h]$. Note that the conceptual model 4.46 is also valid in this case. A better estimate can be obtained by using a weighted least squares (Cleveland, 1979) defined as:

$$\sum_{i=1}^{N_h(t_0)} w_i(t_i) \left( C(t_i) - \beta_0 - \beta_1 t_i - \beta_2 t_i^2 - \cdots - \beta_d t_i^d \right)^2$$

(4.54)

where
\[
\omega_n(t_i) = W\left(\frac{t_i - t_0}{h}\right)
\]  

(4.55)

where \(W(x)\) is a function such that

1. \(W(x) > 0\) for \(|x| < 1\)
2. \(W(-x) = W(x)\)
3. \(W(x)\) is a nonincreasing function for \(x \geq 0\)
4. \(W(x) = 0\) for \(x < 1\)  

(4.56)

An example of \(W(x)\) is the tricube function given by:

\[
W(x) = \begin{cases} 
(1 - |x|^3)^3 & \text{for } |x| < 1 \\
0 & \text{for } |x| \geq 1 
\end{cases}
\]  

(4.57)

One problem with this smoother is that it is not resistant to the presence of outliers. It can be "robustified" in the following way (Cleveland, 1979).

Let's

\[
e_i = C(t_i) - \hat{f}_h(t_i)
\]  

(4.58)

for \(i = 1, 2, \ldots, n\). Let \(s\) be the median of the \(|e_i|\). Define robust weights by:

\[
\delta_i = B(e_i / 6s)
\]  

(4.59)

where \(B(.)\) is the bisquare weight function that is defined by:

\[
B(x) = \begin{cases} 
(1 - |x|^2)^2 & \text{for } |x| < 1 \\
0 & \text{for } |x| \geq 1 
\end{cases}
\]  

(4.60)

The next step in the analysis is to recalculate the coefficients of the weighted regression (relation 4.54) by replacing \(w_{i_0}(t_i)\) by \(\delta_i w_{i_0}(t_i)\) and then deriving a new estimate of \(\hat{f}_h(t_0)\) using relation 4.52. New values for \(\delta_i\) can be obtained and the process can be iterated \(m\) times.

The use of robust locally weighted regression smoothing techniques can be summarized in the following algorithm:

1. Put \(\delta_i = 1\) for \(i = 1, 2, \ldots, n\).
2. For \(i = 1, 2, \ldots, n\), define the interval \([t_i - h, t_i + h]\) around \(t_i\), where \(h\) is the span.

2.1 For each observation \(j\) in the interval \([t_i - h, t_i + h]\), calculate \(w_{i_j}(t_j)\) using relation 4.52.
2.2 Estimate the correlation coefficient \(\beta_k\) by weighted least squares by minimizing the relation

\[
\sum_{j=1}^{N_i(t_i)} \delta_i w_{i_j}(t_j) \left( C(t_j) - \sum_{k=0}^{d} \beta_k t_j^k \right)^2
\]
2.3 Estimate \( \hat{f}_n(t) \) using relation 4.52.

3. Estimate the new \( \delta_i \) using relations 4.58 to 4.60.

4. Loop through steps 2 and 3 \( m \) times.

Four parameters, namely \( h \), \( d \), \( W \), and \( m \), have to be selected before using that algorithm. As with the other smoothers mentioned so far, the choice of \( h \) is somewhat arbitrary. Many values may have to be tried before an acceptable value can be found. For \( d \), Cleveland (1979) recommends 1, because for \( d \geq 2 \), the calculations get more complex and the need for computational convenience overrides the need for flexibility. \( W(x) \) can be any function that satisfies relation 4.56. In addition, it is desirable for it to decrease smoothly to 0 as \( x \) goes from 0 to 1. Finally, \( m \) could be selected indirectly by iterating until some convergence criterion is satisfied. On the basis of some experimentation, Cleveland (1979) concluded that this was unnecessary and that two or three iterations should be sufficient.

Examples of the use of loess smoothing are presented in Figure 4.5. As with the two preceding smoothers, the smoothness of the estimated trend improves as the bandwidth increases. Depending on the smoothness of the estimate, one should use a larger bandwidth (about 3 years) for this smoother than for the preceding two.

As with the running mean and running median smoothers, missing values can be ignored. As in the case of the running mean, it may also be useful to transform the data so that the residuals are symmetric around the assumed function \( f(t) \). Even though the loess smoother gives some protections against outliers, prior transformation of the data should give better results in situations where the distribution of the residuals is skewed toward high values.

\[ \text{Figure 4.5} \] Loess smoother trend estimation for four different bandwidths for ambient \( \text{SO}_2 \) air concentration at Algoma, Ontario.
An interesting property of the loess smoother is that the border effects are strongly reduced (see Hart, 1997). This is due to the fact that it implicitly includes the normalization that reduces those effects. They are, however, not totally eliminated as can be seen in Figure 4.5.

4.3.5 Kernel Smoothers

Kernel smoothers are a somewhat more sophisticated version of the running mean smoother. In these smoothers the mean value is replaced by a weighted mean. Usually, the weight is increased as the data points move closer to $t_0$. Several different versions of kernel smoothers exist. We will mention the three most important. The first is the Nadaraya-Watson estimator (Nadaraya, 1964 and Watson, 1964). It is defined as:

$$ \hat{f}^{NW}_h(t_0) = \frac{\sum_{i=1}^{n} C(t_i) K \left( \frac{t_i - t_0}{h} \right)}{\sum_{i=1}^{n} K \left( \frac{t_i - t_0}{h} \right)} \quad (4.61) $$

where $K(.)$ is a function called, the kernel, and $h$ is the span or bandwidth or smoothing parameter. We will discuss possible kernel functions later in this section. It is interesting to note that if we define the rectangular kernel, $K_R$, as

$$ K_R(u) = \frac{1}{2} I_{(-1,1)}(u) \quad (4.62) $$

where $I_A$ denotes the indicator function for the set $A$, i.e.,

$$ I_A(x) = \begin{cases} 1 & x \in A \\ 0 & x \notin A \end{cases} \quad (4.63) $$

$\hat{f}^{NW}_h(t_0)$ is equal to the running mean.

The second type of kernel smoother is the Priestley and Chao (1972) estimator defined as:

$$ \hat{f}^{PC}_h(t_0) = \frac{1}{h} \sum_{i=1}^{n} (t_i - t_{i-1}) C(t_i) K \left( \frac{t_i - t_0}{h} \right) \quad (4.64) $$

The third type of kernel smoother is the Gasser-Müller (1979) estimator, defined as

$$ \hat{f}^{GM}_h(t_0) = \frac{1}{h} \sum_{i=1}^{n} C(t_i) \int_{s_{i-1}}^{s_i} K \left( \frac{t_0 - u}{h} \right) du \quad (4.65) $$

where $s_0 = 0$, $s_i = (t_i + t_{i+1})/2$, $i = 1, 2, \ldots, n-1$, and $s_n = 1$. Another way of writing this estimator is:

$$ \hat{f}^{GM}_h(t_0) = \frac{1}{h} \int_{0}^{1} C_n(u) K \left( \frac{t_0 - u}{h} \right) du \quad (4.66) $$

where $C_n(.)$ is the piecewise constant function.
Table 4.1 Some of the most often used kernels.

<table>
<thead>
<tr>
<th>Kernel</th>
<th>Formula</th>
</tr>
</thead>
<tbody>
<tr>
<td>a) Rectangle</td>
<td>$K_R(u) = \frac{1}{2} I_{(-1,1)}(u)$</td>
</tr>
<tr>
<td>b) Triangular</td>
<td>$K_T(u) = (1-</td>
</tr>
<tr>
<td>c) Epanechnikov</td>
<td>$K_E(u) = \frac{3}{4} (1-u^2) I_{(-1,1)}(u)$</td>
</tr>
<tr>
<td>d) Quartic</td>
<td>$K_Q(u) = \frac{15}{16} (1-u^2)^2 I_{(-1,1)}(u)$</td>
</tr>
<tr>
<td>e) Rogosinski</td>
<td>$K_{Ro}(u) = (0.5 + \cos(\pi/5) \cos(\pi u) + \cos(2\pi/5) \cos(2\pi u)) I_{(-1,1)}(u)$</td>
</tr>
<tr>
<td>f) Gaussian</td>
<td>$K_G(u) = \frac{1}{\sqrt{2\pi}} e^{-\frac{u^2}{2}}$</td>
</tr>
</tbody>
</table>

where $I_A(x) = \begin{cases} 1 & x \in A \\ 0 & x \notin A \end{cases}$

$C_n(u) = \sum_{i=1}^{n} C(t_i) I_{[u_i,u_j]}(u)$ (4.67)

One should note the similarity of types between the last two smoothers. An interesting characteristic of the Gasser-Müller estimator is that it tends to the function $C_n(.)$ when the bandwidth tends to zero. By contrast, the Nadaraya-Watson estimate is not well defined for a sufficiently small $h$ when $K$ has finite support. As a result the latter is more unstable than the former for small values of $h$.

Chu and Marron (1991) have shown that when the observations are approximately evenly spaced, there is very little differences between the types of estimators. However, there may be great differences when the observations are not equally spaced.

One usually asks that the kernel function $K(.)$ be symmetric and have a unique maximum at 0. Another condition, that one wants to put on the kernel function, is that $\int K(u) du = 1$. A popular way of ensuring that these three conditions are satisfied is to take $K$ to be a probability density function (pdf) that is unimodal and symmetric about zero. Doing so also guarantees a positive regression estimate for positive data. This is an important property when it is known that $f(t) > 0$. However, there are some useful kernels that take negative values.

Table 4.1 lists some of the most often used kernel functions. The Rogosinski kernel differs from the others in that it can take negative values. The Gaussian Kernel is defined for
Kernels are typically designed either to reduce the variance or to optimize some property of the smoothers. More kernel functions are given in Gasser et al. (1985).

It is important to note that the three types of kernel smoothers presented here are biased in the sense that as $n \to \infty$, $h \to 0$ and $nh \to \infty$, and $\hat{f}_h(t)$ does not $\to f(t)$. The bias depends on many factors and is asymptotically proportional (see Jennen-Steinmetz and Gasser, 1988; Gasser and Engel, 1990) to $h^2 f''(t)$ for the Gasser-Müller smoother where $f''(t)$ is the second derivative of $f(t)$. In the case of the Nadaraya-Watson smoother, a second term of the same order of magnitude is added to the preceding term. This second term is proportional to $h f'(t)$, where $f'(t)$ is the first derivative of $f(t)$, and to the distribution (i.e., design) of the observations, thus making the bias very complex. From these results, it might seem that one should make the bandwidth as small as possible to reduce the bias. However, in reducing the bandwidth, the variance of the estimator is increased because the estimator variance for both types of smoothers is proportional to $(nh)^{-1}$ and to the variance of the data around $f(t)$ and depends also on the observation design (see Jennen-Steinmetz and Gasser, 1988; Gasser and Engel, 1990). Therefore, the choice of bandwidth involves a compromise between decreasing the bias and keeping the variance of the estimator small. Note that for both the bias and the variance some of the constant terms entering in these relationships depend on factors like $\int u^2 K(u) du$ and $\int \{K(u)\}^2 du$ and thus on the kernel used. These relations could therefore be used in trying to develop kernels that would give the best compromise between bias and variance for a given value of the bandwidth (see Hart, 1997).

The fact that the bias for the Gasser-Müller smoother is independent of the distribution of the design should make it more useful than the Nadaraya-Watson smoother. However, in time series analysis the distribution of the observations, which is usually regular, should not have much influence on the bias of the latter method.

The most important facts to retain from this discussion are:

- Because the bias is proportional to $f''(t)$, it will be highest when $f(t)$ presents well-marked peaks or troughs. This can be compensated for by decreasing the bandwidth, but that will result in an increase in the variance of the smoother, rendering it more unstable. The effect of the bias on the shape of $f(t)$ can be seen in Figure 4.2. When $t_0$ is near the peak in $f(t)$ and the bandwidth is large, data at the peak and far from the peak will both be used in the calculation of $\hat{f}_h(t_0)$, thus reducing its value. This problem can be corrected by reducing the bandwidth and thus increasing the variability of the estimator. It is interesting to note that the bias would be negative for a peak in $f(t)$ and positive for a trough.

- The variance of the smoother is decreased when the amount of data increases, but the bias is not affected.

- The higher the variability of the data around $f(t)$, the higher the variance of the smoother.

- Both bias and variance are dependent on the kernel used, although this effect is not usually important.

From the first point, we may conclude that when $f(t)$ presents marked peaks and troughs it may be advantageous not to use a constant bandwidth but to vary it according to on the shape of $f(t)$. We will discuss this point later in Section 4.3.8.
Figure 4.6 presents an example of the kernel smoother using the Nadaraya-Watson estimator. The same data as in the preceding examples were used. One should note that as for the running mean, running median, and loess smoothers, the smoothness of the estimation increases with the bandwidth. That means that, as we have seen, the bias will also increase, although \( f_{\theta}(t) \) will always be small in the present example and will thus limit the bias. The increase in the bandwidth will reduce the variance of the estimator. In the present case a bandwidth of about 2 years should be appropriate.

Like the smoothers described in the preceding sections, the kernel smoother has edge effects. The variance of the smoother increases as to the number of points available decreases as one gets to the ends of the time series. Techniques have been developed to correct in part for these edge effects. We will discuss those for the Nadaraya-Watson and Gasser-Müller smoothers.

A correction for edge effects in the Nadaraya-Watson smoother was proposed by Rice (1984). The modification proposed by Rice is based on the fact that near the borders or edges of the time series the bias is proportional to \( h \) and not to \( h^2 \) as we have seen before. To obtain an estimator that has a bias of the same proportionality relative to the bandwidth, one should subtract the term proportional to \( h \) from the estimate \( \hat{f}_h^{NW}(t_0) \) in 4.61. This new estimate can be written as:

\[
\hat{f}_h^{NW}(t_0) = \hat{f}_h^{NW}(t_0) + \beta [\hat{f}_h^{NW}(t_0) - \hat{f}_{oh}^{NW}(t_0)]
\]

where

\[
\begin{align*}
\beta &= \frac{h^2}{h^2} \\
\end{align*}
\]
\[ \beta(\rho) = \frac{R(\rho)}{\alpha R(\rho/\alpha) - R(\rho)} \quad (4.69) \]

with \( R(u) = K_1(u)/K_0(u) \) and

\[ K_j(u) = \int_{-1}^{u} K(v) dv \quad \text{for } i = 0, 1, 2. \]

In these equations, \( \rho \) is equal to \( t_0/h \). For the parameter \( \alpha \), Rice (1984), after discussing different possibilities, recommends making it equal to \( 2 - \rho \).

Relation 4.68 is applied only in the region \( t_0 \in [0,h] \) and \( [t_{\text{max}}-h,t_{\text{max}}] \). For other values of \( t_0 \), relation 4.61 is used. An example of the use of the modified Nadaraya-Watson smoother is presented in Figure 4.7. In that example, the differences between the modified and unmodified curves are small, but this is not always the case.

In the case of the Gasser-Mülller smoother, Gasser and Müller (1979) proposed the use of boundary kernels. One way of constructing such kernels is to define \( K_\rho \) for each \( \rho = t/h \in [0,1) \) by

\[ K_\rho(u) = t_\rho(u) K(u) I_{[-1,\rho]}(u) \quad (4.70) \]

where \( t_\rho \) is a function with the properties

\[ \int_{-1}^{\rho} t_\rho(u) K(u) du = 1 \]

and

\[ \int_{-1}^{\rho} u t_\rho(u) K(u) du = 0 \quad (4.71) \]

At \( t_0 = \rho h \), the \( \hat{\gamma}_h(t_0) \) is estimated by

\[ \frac{1}{h} \sum_{i=1}^{n} C(t_i) \int_{x_{i-1}}^{x_i} K_\rho \left( \frac{t_0 - u}{h} \right) du \]

\[ (4.72) \]

A similar boundary adjustment can be done for \( t_0 \) near \( t_{\text{max}} \) by simply writing \( \rho = (t_{\text{max}} - t_0)/h \).
A simple choice for \( t_p \) is

\[
  t_p(u) = a_p + b_p u.
\]  

(4.73)

To ensure that relations 4.71 are satisfied, one takes:

\[
  a_p = \frac{I_{p,2}}{I_{p,2}I_{p,0} - I_{p,1}^2}
\]

(4.74)

and

\[
  b_p = \frac{-I_{p,1}}{I_{p,2}I_{p,0} - I_{p,1}^2}
\]

(4.75)

where

\[
  I_{p,i}(u) = \int_{-1}^{u} v^i K(v) \, dv
\]

(4.76)

for \( i = 0, 1, 2 \).

Another possible boundary kernel is (Müller, 1991):

\[
  K_p(u) = 6(1 + u)(\rho - u) \frac{1}{(1 + \rho)^2} \times \left( 1 + 8 \left( \frac{1 - \rho}{1 + \rho} \right)^2 + 10 \frac{1 - \rho}{(1 + \rho)^2} u \right) I_{(-1,1)}(u)
\]

(4.77)

At \( \rho = 1 \), \( K_p(u) \) becomes simply the Epanechnikov kernel. To ensure a smooth transition, it would thus been sensible to use that kernel in the edge regions and the Epanechnikov kernel for the interior points.

When making estimates in statistical analysis, one would like to be able to create confidence intervals for the parameter estimated. For the estimation of \( f(t) \) by kernel smoothers, some large-sample confidence intervals have been proposed, either for pointwise confidence intervals or for simultaneous confidence bands for the entire function \( \hat{f}_h(t_0) \).

An asymptotically valid \((1 - \alpha)100\%\) confidence interval for the Gasser-Müller smoother, \( \hat{f}_h^{GM}(t_0) \), is given by Hart (1997):

\[
  \left( \hat{f}_h^{GM}(t_0) - (Z_{1/2} + \hat{B}_{C,\rho}) \hat{\sigma} \right) \left[ \sum_{i=1}^{n} \left\{ \int_{k_{i-1}}^{k_i} K \left( \frac{t_0 - u}{h} \right) \, du \right\}^2 \right]^{1/2},
\]

(4.78)

\[
  \hat{f}_h^{GM}(t_0) - (Z_{1/2} + \hat{B}_{C,\rho}) \hat{\sigma} \left[ \sum_{i=1}^{n} \left\{ \int_{k_{i-1}}^{k_i} K \left( \frac{t_0 - u}{h} \right) \, du \right\}^2 \right]^{1/2}
\]

where \( \hat{\sigma} \) is an estimate of the variance of the residuals and

\[
  \hat{B}_{C,\rho} = \frac{C^4 n^{-3/2} \sigma_k^2}{2 \left( \sum_{i=1}^{n} \left\{ \int_{k_{i-1}}^{k_i} K \left( \frac{n^{1/5}(t_0 - u)}{C} \right) \, du \right\}^2 \right)^{1/2}} \hat{f}_h^{(2)}(t_0)
\]

(4.79)
where $C$ is a constant that depends on the distribution of the observations,

$$\sigma^2 = \int_0^1 u^2 K(u) \, du,$$

and $\hat{f}^{(2)}(t_0)$ is an estimate of the second derivative of $f(t)$, which can be estimated using the following kernel estimator for the $k$th derivatives:

$$\hat{f}^{GM}_{k} (t_0) = \frac{1}{h^{k+1}} \sum_{j=1}^{n} C(t_j) \int_{h^{-1}} M \left( \frac{t_0 - u}{h} \right) \, du$$

(4.80)

where $M$ is a kernel with support (-1,1). This kernel is fundamentally different from the ones used in estimating $\hat{f}^{GM}(t_0)$ (see Hart, 1997). An example of a valid kernel $M$ for the second derivative is:

$$M(u) = \frac{105}{16} (-5u^4 + 6u^3 - 1) I_{(-1,1)}(u)$$

(4.81)

See Gasser et al. (1985) for more details about these types of kernels.

This technique has been criticized because one has to choose a completely new set of parameters for a smoother to estimate the uncertainty in the smoothed regression. Details for a bootstrap approach to obtaining a bias-adjusted interval as in relation 4.78 are provided by Härdle and Bowman (1988). One should note that this technique is complex and only gives asymptotically valid confidence intervals.

A number of methods have been proposed for constructing simultaneous confidence bands for the entire function $f(t)$. As in the case of pointwise confidence intervals, one has to take into account the bias of the nonparametric smoother in order to guarantee the validity of the corresponding interval(s). Each of the techniques that has been proposed provides a means of dealing with this issue. The reader is referred to the following papers for more details: Knafl et al. (1985), Hall and Titterington (1988), Härdle and Bowman (1988), Li (1989), Härdle and Marron (1991), and finally Eubank and Speckman (1993).

Although we have mainly concentrated on the Gasser and Müllér technique, corresponding relations can be derived for the Nadaraya-Watson and other kernel smoothers.

An interesting application of kernel smoothers is related to the use of probability bands to test the adequacy of a parametric model. For example, if one wants to test the adequacy of the straight line model for the regression of $f(t)$, one can use the residual of a kernel smoother like the Gasser-Müller smoother to construct simulated confidence bands that can then be used to test the validity of the straight line model. See Hart (1997) and Hart and Wehrly (1992) for more details.

### 4.3.6 Smoothing Splines

A spline is a piecewise polynomial constructed in such a way that it is smooth at the observations, called knots, at which two polynomials are pieced together. To impose the smoothness criterion, one usually asks that the polynomials of order $k$ be used and that their first $k-1$ derivatives be continuous. Most commonly, $k$ equals 3. In that case, we speak of cubic splines. For example, a cubic spline is defined as:

$$y = S(x) = P_j(x) = a_j + b_j x + c_j x^2 + d_j x^3$$

(4.82)

with the conditions:
where the $\xi_j$ (j = 1, 2, ..., n) are the knots, and

$$P_j^{(k)}(\xi_j) = P_{j+1}^{(k)}(\xi_j); \quad k = 0, 1, 2; \quad j = 1, 2, \ldots, n$$

(4.84)

where $P^{(k)}$ is the kth derivative of $P(x)$.

Splines can be used to approximate virtually any smooth function, at least if a sufficiently large number of knots is used.

As defined in the preceding paragraphs, the spline will pass through all the observations and therefore follow all the variations in the data. However, we would like to filter out the rapid variations to obtain a smooth estimate of $f(t)$, in other words, a smooth spline. A smooth spline, $\hat{f}_\lambda(t)$, has the following properties:

1. It has knots at $t_1, t_2, \ldots, t_n$.
2. It is a cubic polynomial on each interval $[t_i, t_j]$, i = 2, ..., n.
3. It has two continuous derivatives.
4. Finally, it minimizes the function:

$$E_\lambda(f) = \frac{1}{n} \sum_{i=1}^{n} (C(t_i) - f(t_i))^2 + \lambda \int \left[ f''(u) \right]^2 du$$

(4.85)

where $\lambda$ is a positive constant between 0 and $\infty$.

In relation 4.85, the first term assures fidelity to the data in a least squares sense, and the second term is a roughness penalty imposed so that, for example, the curve $\hat{f}_\lambda(t)$ will not have too many bends. The parameter $\lambda$ adjusts the relative weighting given to the error sum of squares and the roughness penalty: a small $\lambda$ means that fit is more important, whereas a large $\lambda$ means that smoothness of function is more important than fit. It therefore controls the degree of smoothing.

For $\lambda = 0$, $\hat{f}_\lambda(t)$ is the (unique) minimizer of $\int \left[ f''(u) \right]^2 du$, subject to the constraint that $f(t_i) = C_i$, i = 1, 2, ..., n. This spline is a so-called natural spline interpolant of $C_1, C_2, \ldots, C_n$. This is the cubic spline mentioned earlier. When $\lambda \rightarrow \infty$, $\hat{f}_\lambda(t)$ becomes the least squares straight line fit to the data.

Note that there is a relationship between smoothing splines and kernel smoothers. Silverman (1984) showed that, asymptotically, minimizing relation 4.85 cause it to behaves like a variable-bandwidth kernel smoothing method.

The use of smoothing splines is illustrated in Figure 4.8 for four values of $\lambda$. For small values of $\lambda$ (Figure 4.8a), the smooth curve follows the seasonal cycles. As $\lambda$ increases, these variations are filtered out and the smooth curve increasingly shows the long-term trend. However, as mentioned earlier, if $\lambda$ becomes very large, the smooth curve becomes equal to the least squares straight-line fit to the data as illustrated in Figure 4.8d.

This example illustrates the importance of the selection of the smoothing parameter $\lambda$. Different techniques have been proposed to help in the choice of $\lambda$. The first is the use of
cross-validation. We will discuss the general principles of this technique later (Section 4.3.8); the reader is referred to that section for more details. Craven and Wahba (1979) have introduced a small variation on cross-validation, called generalized cross-validation, which minimizes the average squared error at the observation points. This technique is a very popular choice for selecting the smoothing parameter, and is used by the ISML package. Silverman (1985) has introduced an approximation to that technique which involves much less computation.

As mentioned earlier, one likes to be able to draw inferences (e.g., create confidence bands) in statistical data analysis. However, inferences about \( f(t) \) from relations 4.1 and 4.85 are not possible because relation 4.85 is a data-smoothing criterion; it does not define a statistical model. Therefore, this relation has to be rewritten to become a statistical model. This can be done using the Bayesian approach, in which we assume a prior distribution for \( f(t) \) and estimate a \( f(t) \) given that prior curve and the observed data. Wahba (1978, 1983) has shown that under certain conditions and fixed \( \lambda \), the posterior mean of \( f(t) \) is the smoothing spline that minimizes relation 4.85. An interval estimate can be obtained using the posterior variance. Silverman (1985), however, has put forward some qualifications regarding the use of that technique. Other techniques have been proposed by Wecker and Ansley (1983) and Silverman (1985).

As for the other smoothing techniques presented in this section, the missing data are simply ignored in the calculation.

In the case of time series, the possibility of serially correlated errors (i.e., the \( \epsilon_i \) values are not independent) may be strong.

Figure 4.8 Smoothing spline trend estimates for four different values of \( \lambda \) for ambient SO\(_2\) air concentration at Chalk River, Ontario.
Several ideas for extensions in this direction are given by Silverman (1985). Kohn and Ansley (1987) have also proposed a new algorithm for spline smoothing that deals with the possibility of serially correlated errors.

The reader who would like to investigate smoothing splines in more detail should refer to Wold (1974), de Boor (1978), Wegman and Wright (1983), Eubank (1984, 1988), and Silverman (1985). Computer algorithms to fit smoothing splines can be found in Herriot and Reinsch (1973) and Duris (1980).

4.3.7 Periodic Smoothers

As mentioned earlier, one difficulty in applying smoothing techniques is the problem of edge or border effects. Some corrective procedures have been mentioned for some of the smoothing techniques. A special case in which these effects can be easily corrected arises when \( f(t) \) in relation 4.1 is assumed to be periodic, with a period equal to \( t_p = t_{\text{max}} - t_{\text{min}} \). Mathematically this can be written as:

\[
 f(t + t_p) = f(t)
 \]

(4.86)

Such a data set might occur, for example, if observations were made over the course of a year and one wanted to estimate the seasonal cycle. A periodic data set can sometimes be created by folding the time series into a new series that extends from zero to \( t_p \). An example of such a process is the folding of many years of observations into one annual period. To do that, there should be no long-term trend in the data, and the seasonal cycles should be the same for every year. Another example of such a folding would be the folding of hourly data for many days or weeks into a one-day period to study the diurnal cycles.

If \( h \) is the bandwidth of a running mean, running median, loess, or kernel smoother, a periodic smoother can be created by first constructing a new data time series, \( C^\ast(t) \), for \( t \in [-h, t_p + h] \), as follows:

1. For \( t \in [-h, 0) \), \( C^\ast(t) = C(t + t_p) \).
2. For \( t \in [0, t_p] \), \( C^\ast(t) = C(t) \).

![Figure 4.9: Creation of the new time series for periodic smoothing.](image)

![Figure 4.10: Seasonal cycles for detrended ambient SO2 air concentrations at Chalk River, Ontario.](image)
3. For \( t \in (t_p, t_p+h] \), \( C^*(t) = C(t-t_p) \).

Then, these data are analyzed for \( t \in [0, t_p] \). This ensures that \( \hat{f}(t_p) \) will be equal to \( \hat{f}(0) \). This process is illustrated in Figure 4.9. The shaded areas are the regions where the data are extended. Note that the normal smoothers are used in periodic smoothing. As an example, Figure 4.10 presents the seasonal cycles of ambient \( \text{SO}_2 \) air concentrations at Chalk River, Ontario, using a periodic kernel Nadaraya-Watson smoother. Note that the long-term trend has been eliminated using a kernel Nadaraya-Watson smoother.

4.3.8 Selecting the Smoothing Parameters.

We mentioned earlier that the result of smoothing techniques are highly dependent on the bandwidth in the case of running mean, running median, loess, and kernel smoothers and on the smoothing constant in the case of smoothing splines. The method used to select these parameters thus far has been very subjective, in the sense that several values of the bandwidth or the smoothing constant were used and the one giving the "best" result in some undefined sense was selected. If one wants only to explore the data, that technique should be enough. In other cases, the characteristics of the time series can help in fixing the bandwidth. For example, if one wants to estimate the long-term trends of a multiyear time series, a bandwidth of 3 to 4 years should filter out the seasonal cycle and any long-term cycles with periods between about 2 and 5 years. In other circumstances, one would like to have a less subjective method of fixing the bandwidth or the smoothing constant. We also mentioned earlier that for the running mean, running median, loess, and kernel smoothers, one would like to have a variable span or bandwidth, especially if there appear to be large peaks and troughs in \( f(t) \). That can be handled only if a non-subjective way of selecting the span or bandwidth is available. We will discuss next the two most important methods of doing this. The reader is referred to Hart (1997) and to the references given therein for other methods as well as more details on the methods discussed here.

4.3.8.1 Cross-validation

Let \( S \) designate any smoothing parameter. \( S \) can stand for the bandwidth or the smoothing parameter or any other parameter that has to be fixed. The cross-validation criterion \( CV(S) \) is defined as:

\[
CV(S) = \frac{1}{n} \sum_{i=1}^{n} (C(t_i) - \hat{f}_i(t_i; S))^2
\]  

(4.87)
where $\hat{f}(t_i; S)$ is the estimated $\hat{f}_S(t_i)$ obtained when the observation $t_i$ has been left out of the calculation. The cross-validation smoothing parameter is that value of $S$ that minimizes $CV(S)$.

This technique can be used to determine either an overall value for $S$ or a local value. In the latter case, the procedure has to be applied for each value $t_i$ at which $f(t)$ has to be estimated by $\hat{f}(t)$. An example of a smoother using local cross-validation is the so-called super-smoother of Friedman and Stuetzle (1982; see also Friedman, 1984). This smoother is a variable bandwidth loess smoother using a polynomial of order 1. An example of results produced by the super-smoother estimator is shown in Figure 4.11 for ambient SO$_2$ air concentrations at Algoma, Ontario.

4.3.8.2 Risk Estimation

As an alternative to cross-validation as a way of selecting smoothing parameters, one can minimize a risk function. A risk function is a function of the smoothing parameters that describes some characteristic of the smoothing process, such as the expected shared errors. An example is the mean average square error (MASE), which is defined as

$$MASE(S) = E \left[ \frac{1}{n} \sum_{i=1}^{n} (\hat{f}(t_i; S) - f(t))^2 \right]$$

where $\hat{f}(t_i; S)$ is the smoothed estimate and $E[.]$ is the expected value of $[.]$.

To illustrate the use of this technique, let us assume that the smoothing function, $\hat{f}(t_i; S)$, can be written as

$$\hat{f}(t_i; S) = \sum_{i=1}^{n} C(t_i)w_i(t_i; S)$$

where the weights $w_i(t_i; S)$ are constants for a fixed value of $S$. This relation applies in the case of the running mean and kernel smoothers. The MASE of this estimator can be written as

$$MASE(S) = E \left[ \frac{1}{n} RSS(S) \right] + 2 \frac{\sigma^2}{n} \sum_{i=1}^{n} w_i(t_i; S) - \sigma^2$$

where $\sigma^2$ is the variance of the residuals and

$$RSS(S) = \sum_{i=1}^{n} (C(t_i) - \hat{f}(t_i; S))^2$$

The quantity $M(S) = MASE(S) + \sigma^2$ has the same minimizer as $MASE(S)$, and one can estimate $M(S)$ as

$$\hat{M}(S) = \frac{1}{n} RSS(S) + 2 \frac{\hat{\sigma}^2}{n} \sum_{i=1}^{n} w_i(t_i; S)$$

where $\hat{\sigma}^2$ is a model free estimator of the variance. One can use, for example, a difference-based estimator such as
The value of the smoothing parameter to use is the value of $S$ that minimizes $\hat{M}(S)$ and therefore the mean average squared error.

4.3.9 Discussion.

The smoothing techniques presented here are only examples of some of the different techniques that have been proposed. Other techniques include the use of polynomials fitted by least squares or the use of a finite series of functions or even a combination of the two (see Eubank and Speckman, 1993). Other forms of orthogonal functions, like Chebyshev’s polynomials, could also be used. The use of polynomial and/or trigonometric functions to fit a curve to data is related to fitting a model by least squares or by maximum likelihood, which we will discuss in Section 4.6.

In this overview, we have illustrated the use of smoothing techniques for time series. These techniques can also be used to estimate the relationship between two variables. For that purpose, relation 4.1 becomes:

$$y(x) = f(x) + \varepsilon$$

and $n$ pairs of observations $(x_i, y_i)$ are available. An example of such an application would be the use of a smoothing technique to estimate the relationship between ionic concentration and precipitation amount in the case of ionic concentration in precipitation.

All of the techniques presented earlier can be used. The first step in the analysis is to order the pairs in increasing order for $x$ such that $x_1 \leq x_2 \leq \ldots \leq x_n$. Note that modifications may have to be made to some of the techniques (e.g., smoothing splines) if there are ties in the $x$ variables.

All the smoothing techniques presented here are very useful, especially when they are used to get a feel for the data, as was illustrated in Section 2.8. However, all of them, except a special version of the smoothing spline technique, assume that the residuals are independent, which is an assumption that does not hold for most time series encountered in atmospheric chemistry. Therefore, if one constructs approximate confidence bands for the smoothing curves, these will only be indicative at best, as they incorporate the assumption that the residuals are not correlated. However, they may still be used to indicate the possible significance of the trend, and, especially in the case of long-term trends, whether the trend is a straight line or monotonic. This can help determine other techniques that could be used.

Smoothing techniques can also be very useful if the temporal variation of many ions at one or several sites has to be analyzed. They can be used to screen for those variations that are more important or significant and that are therefore good candidates for further analysis with more complex techniques.

4.4 Spectral Analysis

4.4.1 Introduction
We have seen in Section 2.7 that spectral analysis can be very useful in isolating the different components contributing to the temporal variations observed in a time series. Spectral analysis can also be very useful in two other ways. First, the smoothed power spectrum can be used to test which components are statistically significant. Secondly, spectral analysis can be used to filter out the high frequency variations in the time series, and thus isolating the long-term variations. We will discuss these two applications in this section.

**4.4.2 Testing for the significance of temporal variation**

The smoothed power spectrum of a time-series can be calculated by using the following relation (see Appendix I; Chatfield, 1984):

\[
\hat{f}(\omega) = \frac{1}{\pi} \left\{ \lambda_0 c_0^m + 2 \sum_{k=1}^{M} \lambda_k c_k^m \cos(2\pi \omega k) \right\}
\]  

(4.95)

where \(\omega (0 \leq \omega \leq 0.5)\) is the frequency in cycles per hour, day, or other period, depending on the data observation frequency; \(c_k^m\) is the auto-covariance coefficient of lag \(k\); \(M (< N)\) is the truncation point that has to be chosen (the lower \(M\) is, the smoother the spectrum is); \(\lambda_k\) is the smoothing window; and \(N\) is the number of observations. Many smoothing windows exist, the most popular being the Tukey and the Parzen windows (see Jenkins and Watts, 1968; Bloomfield, 1976; Priestley, 1981; Chatfield, 1984; Fuller, 1996).

It can be shown that a \((1-\alpha)100\%\) approximate confidence band for \(\hat{f}(\omega)\) is (see Jenkins and Watts, 1968; Priestley, 1981; Chatfield, 1984; Brockwell and Davis, 1987):

![Figure 4.12 Examples of power spectra with 95% confidence bands around the assumed spectrum.](image-url)
where the number of degrees of freedom of the lag window, \( v \), is equal to:

\[
\frac{v \hat{f}(\omega)}{\frac{\nu^2}{(\nu/2)} \cdot \frac{\nu^2}{(1-\nu/2)}}
\] (4.96)

and \( \chi^2_v(\alpha) \) is the chi-squared distribution with \( v \) degrees of freedom.

Although one can use relation 4.96 to draw a confidence band around \( \hat{f}(\omega) \), this will not be very useful. A better use for that relation is as follows.

The power spectrum of random noise is given by (see Jenkins and Watts, 1968; Priestley, 1981; Chatfield, 1984; Brockwell and Davis, 1987; Fuller, 1996):

\[
f(\omega) = \sigma^2 \frac{1}{2\pi}
\] (4.98a)

for \(-0.5 \leq \omega \leq 0.5\) or

\[
f(\omega) = \sigma^2 \frac{1}{\pi}
\] (4.98b)

for \(0 \leq \omega \leq 0.5\), where \( \sigma^2 \) is the variance of the random noise.

If we assume that data are random noise, we can estimate the variance using Parseval's relation (Priestley, 1981) on the smoothed power spectrum of the observations or:

\[
\hat{\sigma}^2 = 2\pi \int_0^{0.5} \hat{f}(\omega) \, d\omega
\] (4.99)

Using that estimate, one can create the power spectrum for a random process with the same variance as that observed in the data. A confidence band for that spectrum can be obtained.

**Figure 4.13** Examples of power spectra with 95% confidence intervals for the assumed spectrum. The observed spectrum is for ambient SO\(_2\) air concentration at Chalk River, Ontario.
using relation 4.96 for the power spectrum of the random process. Figure 4.12a presents the results of such an analysis if the observations are from a normal random distribution. The constant solid line is the spectrum for the random process with the same variance as the data and the dashed lines are the 95% confidence bands. One can see that, overall, the spectrum for the observations is within the confidence band.

Figure 4.13a presents the spectrum for ambient SO$_2$ air concentrations at Chalk River, Ontario. On the same graph the power spectrum and the 95% confidence band of a white noise process with the same variance as the observed data are shown. One can clearly see that the observed spectrum is not from a random process. Significant long-term and seasonal variations are present. Also one notices that for high frequencies the observed power spectrum is lower than the lowest confidence limit. This indicates that auto-correlations are present in the observed data.

The analysis can now be continued by using the power spectrum of an auto-regressive models of different orders and coefficients to improve the fit between model and observations.

Let us write the auto-regressive model of order $p$ as:

$$ y_i = \sum_{k=1}^{p} a_k y_{i-k} + \varepsilon_i $$

(4.100)

Then the power spectrum of this auto-regressive model is (see Jenkins and Watts, 1968; Priestley, 1981; Brockwell and Davis, 1987; Fuller, 1996):

$$ f(\omega) = \frac{\sigma^2}{2\pi} \left[ \frac{1}{\left( 1 - \sum_{j=1}^{p} a_j e^{i2\pi\omega j} \right) \left( 1 - \sum_{j=1}^{p} a_j e^{-i2\pi\omega j} \right)} \right] $$

(4.101a)

for $-0.5 \leq \omega \leq 0.5$ or

$$ f(\omega) = \frac{\sigma^2}{\pi} \left[ \frac{1}{\left( 1 - \sum_{j=1}^{p} a_j e^{i2\pi\omega j} \right) \left( 1 - \sum_{j=1}^{p} a_j e^{-i2\pi\omega j} \right)} \right] $$

(4.101b)

for $0 \leq \omega \leq 0.5$ where $\sigma^2$ is the variance of the $\varepsilon_i$ in relation 4.100.

For an auto-regressive model of order 1, 4.101b becomes:

$$ f(\omega) = \frac{\sigma^2}{\pi} \frac{1}{(1 + a^2 - 2a \cos(2\pi\omega))} $$

(4.102)

It is interesting to note that for $a > 0$, $f(\omega)$ is a decreasing function of $\omega$ and an increasing function if $a > 0$.

For a second order auto-regressive model, relation 4.101b becomes:

$$ f(\omega) = \frac{\sigma^2}{\pi} \frac{1}{(1 + a_1^2 + a_2^2 - 2a_1(1-a_2)\cos(2\pi\omega) - 2a_2 \cos(4\pi\omega))} $$

(4.103)
which is more complex than that for a first order auto-regressive model.

Using the same technique as was used for white noise, one can plot relation 4.102 using Parseval’s relation to estimate the variance. Figure 4.12b illustrates the power spectrum of a time series created with a first order auto-regressive model (\(a = 0.4\)) with normal errors. The theoretical power spectrum for an auto-regressive model of order 1 with the same variance as the “observed” data and a coefficient equal to 0.4 is also shown. The graph also includes the 95% confidence band. As in the case of white noise, the power spectrum for the “observed” data varies around the theoretical spectrum and remains generally within the confidence bands.

Figure 4.12b presents again the power spectrum for ambient \(\text{SO}_2\) air concentrations at Chalk River, Ontario. As a theoretical model, we have now used a first order auto-regressive model with coefficient equal to 0.25. One notices first that the shape of the theoretical power spectrum seems correct but is too high. The explanation for this is simple. The estimation of variance obtained using Parseval’s relation includes the variance due to all the different types of temporal variations present in the observed data. Therefore, it includes the contributions of the long-term trend and the seasonal cycle. Therefore, this estimate is too high if one uses only the auto-regressive model. An easy way to estimate only the variance due to the auto-regression in the data is to start the integration at a higher frequency to remove the influence of lower frequency temporal variations. The result obtained if one integrates for \(\omega\) between 0.02 and 0.5 is shown in Figure 4.14. One can see that the theoretical power spectrum is centered for the high frequencies in the middle of the observed power spectrum. This plot indicates that, in addition to auto-regression, there are statistically significant contributions to the total variance due to a seasonal cycle and longer period variations. These latter could include a long-term trend and long-term cycles.

This example illustrates how the power spectrum of the observed data can be used to test the statistical significance of the different temporal variations that may contribute to the total variance.

### 4.4.3 Linear Filters

A linear filter is defined by the relation

\[
\hat{C}_t = \sum_{k=r}^s g_k C_{t-k}
\]

(4.104)
where \( C_t \) is the observation at time \( t \) and \( \hat{C}_t \) is the filter data at time \( t \). The \( g_k \)'s are the coefficients of the filter. Usually the filter would be symmetric and \( r = -p \) and \( s = p \).

An interesting property of linear filters can be deduced if one assumes that the data are sinusoidal, i.e.,

\[
C_t = R \cos(\omega t + \phi) = Re\left(Re e^{i(\omega t + \phi)}\right)
\]

where \( Re(\cdot) \) indicates the real part of \( \cdot \). If one replaces \( C_t \) in 4.104, one obtains:

\[
\hat{C}_t = \sum_{k=r}^{s} g_k \, Re\left(Re e^{i(\omega t + \phi)}\right)
= Re\left(Re e^{i(\omega t + \phi)} \sum_{k=r}^{s} g_k e^{-i\omega k}\right)
= Re\left(Re e^{i(\omega t + \phi)} G(\omega)\right)
\]

where \( G(\omega) \) is defined as

\[
G(\omega) = \sum_{k=r}^{s} g_k e^{-i\omega k}
\]

The function \( G(\omega) \) is called the transfer function of the filter, since it indicates the way in which a sinusoid of frequency \( \omega \) is transferred by the linear filter. This is valid for all frequencies. If one knows the weights of the linear filter, the transfer function can be obtained. The squared magnitude of the transfer function, \( |G(\omega)|^2 \), is called the power transfer function. In the case of a symmetric filter, one gets

\[
G(\omega) = g_0 + 2 \sum_{k=1}^{n} g_k \cos(\omega k)
\]

If one multiplies each side of relation 4.106 by \( e^{i\omega j} \) and integrates \( \omega \) from \( -\pi \) to \( \pi \), one gets:

\[
\int_{-\pi}^{\pi} G(\omega) e^{i\omega j} d\omega = \sum_{k=r}^{s} g_k \left(\int_{-\pi}^{\pi} e^{-i\omega k} e^{i\omega j} d\omega\right)
= \sum_{k=r}^{s} g_k \left(\int_{-\pi}^{\pi} e^{-i\omega (k-j)} d\omega\right)
= 2\pi \sum_{k=r}^{s} g_k \delta_{kj}
\]

where \( \delta_{kj} = \int_{-\pi}^{\pi} e^{-i\omega (k-j)} d\omega \) is the Kronecker delta and is defined as:

\[
\delta_{mn} = \begin{cases} 
0 & \text{if } m \neq n \\
1 & \text{if } m = n
\end{cases}
\]

Therefore, relation 4.108 can be written as:
\[ g_j = \frac{1}{2\pi} \int_{-\pi}^{\pi} G(\omega) e^{i \omega j} d\omega \quad (4.110) \]

Relations 4.106 and 4.110 show that the transfer function, \( G(\omega) \), and the constant \( g_k \) are Fourier transformed from one another. Thus, one can define a transfer function, and then using relation 4.110 obtain the coefficients \( g_k \) that can be used in relation 4.104, or, one can do the opposite.

An example of a symmetric linear filter is the moving-average obtained by writing:

\[ g_k = \frac{1}{2p+1} \text{ for } -p \leq k \leq p \quad (4.111) \]

Thus, the moving-average filter is:

\[ \hat{C}_i = \frac{1}{2p+1} \sum_{k=-p}^{p} C_{i-k} \quad (4.112) \]

and the transfer function, using relation 4.107, is:

\[
G(\omega) = \frac{1}{2p+1} + \frac{2}{2p+1} \sum_{k=1}^{p} \cos(\omega k)
= \frac{2}{2p+1} \left( \frac{1}{2} + \sum_{k=1}^{p} \cos(\omega k) \right)
= \frac{2}{2p+1} \frac{\sin((p+1/2)\omega)}{2\sin(\omega/2)}
= \frac{1}{2p+1} \frac{\sin((p+1/2)\omega)}{\sin(\omega/2)}
= D_{2p+1}(\omega) \quad (4.113)
\]

The function \( D_n(\omega) \) is called the Dirichlet kernel.

The transfer function, \( G(\omega) \), of a moving average filter is illustrated in Figure 4.15 for different values of \( p \). From that figure, one can see: first, that the transfer function can be negative; second, that it is maximum at \( \omega = 0 \); third that it presents a decreasing oscillations around zero; and fourth, that the gain is more...
concentrated around zero as \( p \) is increased. This last characteristic shows that in choosing \( p \) large enough, one can isolate the low frequency contribution to the temporal variations in the time series. In other words, by using the moving average with a large value of \( p \) one can isolate the long-term trend.

In the preceding example, we first specified the weights, \( g_k \), and then we calculated the transfer function. As mentioned earlier, one can also do the opposite, first choosing a transfer function and then calculating the weight using relation 4.11.

We have seen in Section 2.7, that low frequencies are usually associated with long-term trends and high frequencies with random noise and autocorrelation. The seasonal and diurnal cycles are between the two. If the data have been gathered on a daily basis, the frequency associated with the seasonal cycle is equal to about 0.0027 cycles day\(^{-1}\). Thus if we want to isolate the long term variation, we would like to use a transfer function similar to the one in Figure 4.16 with \( \omega_c \) lower than 0.0027 cycles day\(^{-1}\). These types of filters are called low-pass filters.

The weights of the low-pass filter presented in Figure 4.16 can be calculated using relation 4.11. They are:

1. \( g_k = \frac{1}{2\pi} \int_{-\omega_c}^{\omega_c} e^{i\omega k} d\omega \)

2. \( g_k = \frac{1}{2\pi} \int_{-\omega_c}^{\omega_c} (e^{i\omega k} - e^{-i\omega k}) d\omega \)

3. \( g_k = \frac{1}{\pi} \int_{0}^{\omega_c} \cos(k\omega) d\omega \)

The \( g_k \) are then:

\[ g_k = \frac{\sin(k\omega_c)}{\pi k} \quad |k| \geq 1 \quad (4.114) \]

and

\[ g_0 = \frac{\omega_c}{\pi} \quad k = 0 \quad (4.115) \]

Note that for this filter, \( r \) and \( s \) (relation 4.106) are respectively equal to \(-\infty\) and \(+\infty\). In practice, the summation has to be truncated to reasonable values for \( r \) and \( s \). If we do that, the gain function obtained by using those weights in relation 4.106 is different from the ideal...
low-pass filter. Ideally, we would like to minimize the difference between the ideal transfer function and the latter. This requirement can be presented mathematically as follow:

Find the coefficient $g_k \{ r \leq k \leq s \}$ such that

$$
\int_{\omega} |H(\omega) - G(\omega)|^2 d\omega = \int_{\omega} |H(\omega) - \sum_{k=r}^{s} g_k e^{-j\omega k}|^2 d\omega
$$

(4.116)

is minimized. $H(\omega)$ is the ideal transfer function. The solution to the minimization problem is

$$
g_k = \frac{1}{2\pi} \int_{\omega} H(\omega) e^{j\omega k} d\omega \quad r \leq k \leq s \quad (4.117)
$$

which is identical to relation 4.110.

The transfer function associated with these weights is equal to:

$$G_{r,s}(\omega) = \sum_{k=r}^{s} g_k e^{-j\omega k}$$

$$= \sum_{k=r}^{s} \left( \frac{1}{2\pi} \int_{\omega} H(\lambda) e^{j\omega k} d\lambda \right) e^{-j\omega k}$$

$$= \frac{1}{2\pi} \int_{\omega} H(\lambda) \left( \sum_{k=r}^{s} e^{-j\omega k} e^{j\omega k} \right) d\lambda$$

$$= \frac{1}{2\pi} \int_{\omega} H(\lambda) \left( \sum_{k=r}^{s} e^{-j(\omega - \lambda)k} \right) d\lambda$$

$$= \frac{s - r + 1}{2\pi} \int_{\omega} H(\lambda) D_{s-r+1}(\omega - \lambda) e^{-j(\omega - \lambda)k} d\lambda$$

(4.118)

where $D_s(.)$ is the Dirichlet kernel. If $H(\omega)$ is real and symmetric and we set $r = -p$ and $s = p$, we get $G_{r,s}(\omega)$ equal to:

$$G_p(\omega) = \frac{2p+1}{2\pi} \int_{\omega} H(\lambda) D_{2p+1}(\omega - \lambda) d\lambda$$

(4.119)

An example of the use of that relation is presented in Figure 4.17 for the transfer function presented in Figure 4.16. The two other curves, obtained using 4.119, are for $G_5(\omega)$ and $G_{20}(\omega)$ and $\omega_c = 0.3$. Both curves show pronounced ripples and one of them shows a pronounced overshoot on either sides of the cutoff frequency. This is known as Gibb's phenomenon.
This overshoot and accompanying ripples may be greatly reduced as follows. It may be shown that the wavelength of the ripples is \( \delta = 4\pi/(2p+1) \). Therefore, the smoothed function

\[
\bar{G}_s(\omega) = \frac{1}{\delta} \int_{-\delta/2}^{\delta/2} G_s(\lambda) d\lambda \quad (4.120)
\]

would have far smaller ripples, since the integration is over one complete cycle.

Using relation 4.107, relation 4.120 can be written as:

\[
\bar{G}_s(\omega) = \frac{1}{\delta} \left[ \int_{-\delta/2}^{\delta/2} \left( g_0 + 2 \sum_{k=1}^{p} g_k \cos(k\lambda) \right) d\lambda \right]
\]

\[
= g_0 + \frac{1}{\delta} \left[ 2 \sum_{k=1}^{p} g_k \int_{-\delta/2}^{\delta/2} \cos(k\lambda) d\lambda \right]
\]

\[
= g_0 + 2 \sum_{k=1}^{p} g_k \frac{\sin(k\delta/2)}{k\delta/2} \cos(k\omega) \quad (4.121)
\]

This corresponds to replacing \( g_k \) by:

\[
g_k \frac{\sin(k\delta/2)}{k\delta/2} = g_k \sin\left(\frac{2\pi k}{2p+1}\right) \left/ \frac{2\pi}{2p+1} \right.
\]

The multipliers

\[
\sigma_{pk} = \sin\left(\frac{2\pi k}{2p+1}\right) \left/ \frac{2\pi k}{2p+1} \right. \quad (4.123)
\]

are examples of converge factors.

Figure 4.18 presents \( \bar{G}_s(\omega) \) and \( \bar{G}_{20}(\omega) \). One sees that the ripples have been substantially reduced. The overshoots in \( \bar{G}_{20}(\omega) \) have also been reduced.

It is usually desirable for a filter to pass a constant term (i.e., a component of frequency zero) without changing it. This requires that the transfer function be equal to 1 at \( \omega = 0 \). The \( g_k \) can easily be modified to insure that the transfer function follows that requirement by simply normalizing them so that their sum equals 1.

Figure 4.19: Illustration of the transfer function in relation 4.122 for \( m = 2, 4, \) and 6.

Figure 4.20: Weights corresponding to the transfer function in relation 4.122 with \( m = 4 \) and \( \omega_c = 0.5 \).
An example of the use of such filters in atmospheric chemistry can be found in Li et al. (1993).

Another way to define a transfer functions without ripples and overshoots is to use an exponential function like (see Thoning et al., 1989):

$$ G(\omega) = e^{-c_1 \left( \frac{\omega}{\omega_c} \right)^n} \quad (4.124) $$

The term $c_1$ is usually equal to $\ln(2)$, so that $G(\omega) = 0.5$ at $\omega = \omega_c$. The transfer function is symmetric around zero. This transfer function is illustrated for three values of $m$ ($m = 2, 4, \text{and } 6$) in Figure 4.19.

The weights associated with this transfer function can be obtained by using relation 4.110 and are given by:

$$ g_k = \frac{1}{\pi} \int_0^\pi e^{-c_1 \left( \frac{\omega}{\omega_c} \right)^n} \cos(\omega k) d\omega \quad (4.125) $$

An example of the weights for a transfer function with $\omega_c = 0.002$ cycles day$^{-1}$ used to estimate the long-term trend of the artificial data of Figure 2.20. The highest weights are concentrated around $k = 0$. The reader should note that the weights may be negative for some of the $k$'s.

Another way to create complex filters is to take a convenient set of simple filters, like moving averages (relation 4.112) of various lengths, and use them as building blocks to assemble a more complex filter. An example of such a filter is the Kolmogorov-Zurbenko filter, $KZ_{m,p}$, used by Rao and Zurbenko (1994), Rao et al. (1995), Flaum et al. (1996), and Milanchus et al. (1998).

The first step in creating the $KZ_{m,p}$ filter is to apply a moving average to the data, defined as:

$$ \tilde{C}_i = \frac{1}{m} \sum_{j=-k}^{k} C_{i+j} \quad (4.126) $$

where $m = 2k+1$. The second step is to replace $C_i$ by $\tilde{C}_i$ in 4.126 and apply the filter a second time. The filtering operation is applied $p$ times.

An adaptive version of the Kolmogorov-Zuchero filter has been developed to help in the identification of step changes in the data (see Zurbenko et al., 1996). The filter is adaptive because the parameter $m$ is modified locally to help identify the step changes. Another special adaptation of the Kolmogorov-Zuchero filter is the spatial version used by Zurbenko et al. (1995).
The technique used to create the Kolmogorov-Zuchero filter is a standard one and thus can be used to define other complex filters that have been made from simple ones.

To illustrate the use of filters, a low-pass filter constructed using relations 4.112, 4.113, and 4.119 was applied to the artificial data of Figure 2.20. A cut-off frequency equal to 0.002 cycles day\(^{-1}\) was used. The weights used for the low-pass filter are shown in Figure 4.21. The resulting estimated long-term trend is presented in Figure 4.22. The long-term trend was estimated only for the region outside the borders, defined so that no areas with missing data were used in relation 4.104. The reader should remember that the long-term trend used in creating the data was a straight line. If we assume that the small variation are not significant (some of the ripples could be eliminated by using a wider window for the filter), the low-pass filter is quite good at finding the long-term trend.

4.4.4 Discussion

We have shown in the preceding section that spectral analysis can be a useful tool for studying the temporal variations present in a time series. The testing techniques presented in Section 4.4.2 can be used to test for the significance of the long-term trend or other components, and the filters can be used to estimate their shape. These two techniques have to be used together to get a complete understanding of the temporal variation in a time series.

There are, however, some limitations to the use of spectral analysis, especially in the use of filters. Data gaps introduce problems in using the spectral analysis technique. Although the power spectrum of the time series can still be estimated using the techniques presented in Appendices II and III, the estimated power spectrum obtained from these techniques can be uncertain if the amount of missing data is large. That uncertainty increases as the number of
missing data items increases. The main problem occurs in the use of filters. In the description of the filter technique presented in Section 4.4.3, it was assumed that no missing data were present. When data are missing, the only solution is to ignore them when using relation 4.104, which is equivalent to assuming a value of zero for these data. Therefore, it is a good idea to center the time-series around zero before analyzing it by subtracting from the data either the mean or the median of the data. Another useful practice is to make sure that the data are also symmetric around the mean or zero. Thus, the data may have to be transformed using a logarithmic or other transformation before the analysis. If these two suggestions are followed, the effect of missing data on the filtering process may be reduced.

Another difficulty related to the missing data problem is the edge effect. This will occur when trying to use relation 4.104 with the index i less than r or greater than (n-s), where n is the total number of data items (note that missing data are considered as valid data in the calculation of n). If i is less than r, there would be (r-i) missing data to the left of i. This introduces an imbalance into the filter equation that may results in a large bias in the estimated trend. Unfortunately, there are no techniques, like those used to solve similar edge effect problems in kernel smoothing (Section 4.3.5), that can be used to compensate for these effects. The same problem occurs also if i is situated within the last s points at the end of the time series.

If the reader has compared relation 4.104 for the linear filters with relations 4.61, 4.64, or 4.65 for the different types of kernel smoothing, he or she will have noticed that these equations are the same, that is, they all have the following general form:

$$\tilde{C}_i = \sum_{j=r}^{s} w_j C_j$$

(4.127)

They differ only in the way that the weights w_i are defined. One can therefore consider linear filtering as kernel smoothing with the kernel defined by spectral analysis. An example of the correspondence between the two is the moving average filter, which is identical to a Nadaraya-Watson kernel smoother with a rectangular kernel (see Section 4.3.5). Note, however, that the two techniques differ when data are missing because they do not treat data gaps in the same ways. Although both techniques ignore missing data, the weights in the kernel smoothing are renormalized so that their sum for the non-missing data is equal to 1. This is not done in the moving average filter.

4.5 Anocov and Anova Techniques

4.5.1 Introduction

The techniques presented in this section have in common the fact that they group data into classes. We have already encountered such groupings in the nonparametric techniques presented in Section 4.2.2 (e.g., the seasonal Kendall test). In the case of the analysis of covariance (anocov) techniques, the grouping is used to define indicator variables that are used to create a linear model. In the analysis of variance (anova) techniques, the data are grouped into two-way tables and an analysis of variance is made. These models are fitted using standard statistical procedures, namely, linear regression in the case of the anocov model and anova technique in the case of the anova model. We will therefore concentrate on describing the models and evoking some of their principal characteristics.
These two techniques assume that the residuals after the elimination of the model are independent and are from a normal distribution with constant variance. This is quite different from the approach used by the techniques presented in the preceding sections. In addition, we have to make an assumption about the exact shape of the distribution of the residuals. Usually, if the amount of data is large, the exact distribution of the residuals may differ slightly from normality. The tolerance of the technique to such variations differs from method to method. We will return to this matter when discussing the two techniques.

4.5.2 Anocov Technique

The anocov technique has been used in recent years by Lynch et al. (1995, 1996) and Sirois et al. (1996). A complete description of the anocov techniques can be found in Neter and Wasserman (1974).

If \( s \) is the number of “seasons” used to describe the data, an analysis of covariance model can be written as:

\[
C_t = \alpha + bt + \sum_{i=1}^{s-1} b_i I_{i}^{S}(t) + \varepsilon_t
\]  

(4.128)

where the residuals, \( \varepsilon_t \), are assumed to be independent random variables from a normal distribution with zero mean and constant variance, \( \sigma^2 \). The \( I_i^S(t) \) are indicator variables. At time \( t \), if the observation is in “season” \( \ell \), we get, if \( i < s \):

\[
I_{i}^{S}(t) = \begin{cases} 
1 & i = \ell \\
0 & i \neq \ell 
\end{cases}
\]

(4.129)

and if \( i = s \), all the \( I_i^S(t) \) are set equal to zero. The seasons could be the usual seasons (\( s = 4 \)) or bi-monthly seasons (\( s = 6 \)) or months (\( s = 12 \)) or even weeks (\( s = 52 \)).

In atmospheric chemistry, ion concentrations and other observed variables are often not normally distributed but log-normally distributed. An appropriate model in that case would be:

\[
\log(C_t) = \alpha + bt + \sum_{i=1}^{s-1} b_i I_{i}^{S}(t) + \varepsilon_t
\]  

(4.130)

if \( s \) seasonal indicators are used. This is the model used by Lynch et al. (1995, 1996) on bi-monthly data (\( s = 6 \)). In the case of ion concentrations in precipitation, a more appropriate model would be:

\[
\log(C_t) = \alpha + bt + c\log(P(t)) + \sum_{i=1}^{s-1} b_i I_{i}^{S}(t) + \varepsilon_t
\]  

(4.131)

where \( P(t) \) is the precipitation amount associated with the sample at time \( t \). Note that, depending on the length of the sampling period, the parameter \( c \) may not be statistically significant.

Although a straight line has been assumed thus far for the long-term trend in relations 4.128, 4.130, and 4.131, any polynomial and/or trigonometric function could be used to describe the long-term trend.
Once a model has been postulated, it is fitted to the data using the usual linear regression technique, and the statistical significance of the long-term trend is verified using Student-t tests. In the case of a straight line long-term trend, this means testing the hypothesis that $b = 0$ against the alternative that $b \neq 0$.

To illustrate the use of the anocov technique, the model described by relation 4.130 was fitted to 28-day $SO_4^{2-}$ ionic concentration in precipitation from Penn State, Pennsylvania. The resulting long-term trends are shown in Figure 4.23. The P-value for the test of $b = 0$ against $b \neq 0$ is equal to 0.0725. Therefore, the hypothesis of no trend cannot be rejected at a 95% confidence level. In other words, the trends presented in Figure 4.23 are not statistically significant.

An interesting feature of the models described by relations 4.128, 4.130, and 4.131, is that they require us to fit not only one straight line to the data but $s$ straight lines with the same slope but with different intercepts corresponding to the different seasons (see Figure 4.23). Note that some of these lines may not be statistically different. It is also assumed that the variance of the residuals is the same in each season.

These models can be extended to take into account the possibility that the variance and the slope in each season could be different by separately fitting a straight line in each season. If we assume that the slopes all have the same sign, we can test for the significance of the overall trend by testing whether the sum of the slopes is different from zero. Using the hypothesis that the slopes all have the same sign, their sum will be equal to zero only if they are all equal to zero. The description of this test can be found in Loftis et al. (1989).

A better way is to test the different slopes simultaneously for significance. This make it possible to see in which season the long-term trends are the most important. Techniques for multiple simultaneous tests can be found in Miller (1981) and Hochberg and Tamhane (1987).
One advantage of the anocov method is that missing data are easily treated because one can simply ignore them.

### 4.5.3 Anova Technique

To apply the anova technique, we group the data into a two-way table with year as the rows index and "season" as the columns index. If we let $C_{ik}$ be the kth observation for year i, and "season" j, a general two-way table will look like this:

<table>
<thead>
<tr>
<th>Year</th>
<th>Month</th>
<th>1</th>
<th>2</th>
<th>.</th>
<th>.</th>
<th>m</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>$C_{11}$</td>
<td>$C_{12}$</td>
<td>$C_{13}$</td>
<td>.</td>
<td>.</td>
<td>$C_{1m1}$</td>
</tr>
<tr>
<td></td>
<td>$C_{1n1}$</td>
<td>$C_{1n2}$</td>
<td>.</td>
<td>.</td>
<td>$C_{1nm1}$</td>
<td>$C_{1nm2}$</td>
</tr>
<tr>
<td>2</td>
<td>$C_{21}$</td>
<td>$C_{22}$</td>
<td>$C_{23}$</td>
<td>.</td>
<td>.</td>
<td>$C_{2m1}$</td>
</tr>
<tr>
<td></td>
<td>$C_{2n1}$</td>
<td>$C_{2n2}$</td>
<td>.</td>
<td>.</td>
<td>$C_{2nm1}$</td>
<td>$C_{2nm2}$</td>
</tr>
<tr>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>n</td>
<td>$C_{n1}$</td>
<td>$C_{n2}$</td>
<td>$C_{n3}$</td>
<td>.</td>
<td>.</td>
<td>$C_{nm1}$</td>
</tr>
<tr>
<td></td>
<td>$C_{n(n-1)}$</td>
<td>$C_{n(n-2)}$</td>
<td>.</td>
<td>.</td>
<td>$C_{nnm1}$</td>
<td>$C_{nnm2}$</td>
</tr>
</tbody>
</table>

where $n_{ij}$ is the number of data in year i and "season" j, n is the number of years, and m is the number of "seasons". The definition of the "seasons" is arbitrary. They can be the 4 usual seasons, or 12 months, or even 52 weeks. Note that the $n_{ij}$ can be different.

This two-way table is quite general. The simplest case occurs when there is only one observation per cell. The two-way table then becomes:

<table>
<thead>
<tr>
<th>Year</th>
<th>Month</th>
<th>1</th>
<th>2</th>
<th>.</th>
<th>.</th>
<th>m</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>$C_{11}$</td>
<td>$C_{12}$</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>$C_{1m}$</td>
</tr>
<tr>
<td>2</td>
<td>$C_{21}$</td>
<td>$C_{22}$</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>$C_{2m}$</td>
</tr>
<tr>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>n</td>
<td>$C_{n1}$</td>
<td>$C_{n2}$</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>$C_{nm}$</td>
</tr>
</tbody>
</table>
A possible linear model for these data would be:

\[ C_{ij} = \mu + \alpha_i + \beta_j + \epsilon_{ij} \]  

(4.132)

with the restrictions

\[ \sum_{i=1}^{I} \alpha_i = 0, \quad \sum_{j=1}^{J} \beta_j = 0 \]  

(4.133)

where \( \mu \) is the overall mean, the \( \alpha_i \) (\( i = 1, 2, \ldots, I \)) are the year effects, the \( \beta_j \) (\( j = 1, 2, \ldots, J \)) are the seasons effects, and the \( \epsilon_{ij} \) are independent random errors from a normal distribution with equal variance \( \sigma^2 \). A more general model would include interactions between seasons and years and could be written as:

\[ C_{ij} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \epsilon_{ij} \]  

(4.134)

with the additional restrictions:

\[ \sum_{i=1}^{I} \gamma_{ij} = 0 \text{ for all } j, \quad \sum_{j=1}^{J} \gamma_{ij} = 0 \text{ for all } i \]  

(4.135)

where \( \gamma_{ij} \) is the interaction between year \( i \) and month \( j \).

These models can be easily fitted to the data using regular anova analysis (see Scheffé, 1959; Neter and Wasserman, 1974; Lehmann, 1986; and Miller, 1986). The parameters \( \mu, \alpha_i \) (\( i = 1, 2, \ldots, I \)) and \( \beta_j \) (\( j = 1, 2, \ldots, J \)) can be estimated. The variations of the \( \alpha_i \)'s are an estimate of the long-term trend and those of the \( \beta_j \)'s are an estimate of the seasonal cycle. The statistical significance of the long-term trend can be checked by testing the following hypothesis:

\[ H_0: \alpha_1 = \alpha_2 = \ldots = \alpha_n = 0 \]  

(4.136)

against

\[ H_1: \text{any two of } \alpha_i \text{ are different.} \]  

(4.137)

The statistical significance of the seasonal cycle can also be verified using a similar test for the \( \beta_j \)'s. Note that these tests are resistant to non-normality in the distribution of the residual if the size of the design (i.e., \( IJ \)) is not too small (Miller, 1986).

A resistant and nonparametric way of estimating the parameters of model 4.132 is the median polish (see Emerson and Hoaglin, 1983). This is an iterative technique using the median. The reader is referred to Emerson and Hoaglin (1983) for more details. The technique is quite resistant to outliers, and missing data are allowed. The only disadvantage is that the statistical significance of the variation observed in the \( \alpha_i \)'s cannot be tested. An example of the use of median polish for long-term trend analysis can be found in Wu and Zidek (1989). The median polish technique can be extended for the model described by relation 4.134 when the interactive term can be written as proportional to the product of the two other terms, that is, when model 4.134 can be written as (see Emerson and Wong, 1985):

\[ C_{ij} = \mu + \alpha_i + \beta_j + \kappa \alpha_i \beta_j + \epsilon_{ij} \]  

(4.138)
The relations 4.132 and 4.134 are easily extended to the case where there are \( n \) observations (i.e., all \( n_{ij} = n \) in the general two-way table) for each season and year combination. Relation 4.132 becomes:

\[
C_{jk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \epsilon_{ijk} \tag{4.139}
\]

for \( k = 1, 2, \ldots, n \). The restrictions in relation 4.133 and 4.135 still apply. This model can be easily fitted to the data using the anova technique analysis (see Scheffé, 1959; Neter and Wasserman, 1974; Lehmann, 1986; and Millier, 1986), and hypotheses such as

\[
H_0: \gamma_{ij}=0, \quad H_0: \alpha_i=0 \quad \text{and/or} \quad H_0: \beta_j=0 \tag{4.140}
\]

can easily be tested. As in the case where only one observation is available per cell, these tests are resistant to non-normality of the distribution of the residuals if the size of the design (i.e. \( n_{ij} \)) is not too small (Millier, 1986).

If the number of observations in each cell differs (i.e., the \( n_{ij} \) are not all equal), the analysis becomes more difficult. In most cases the best technique to use to fit relation 4.139 is to employ multiple regressions (see Draper and Smith, 1981, and Milliken and Johnson, 1984). Unfortunately, even if multiple regressions are used, the analysis is murkier than in the balanced case. One of the characteristics of such an analysis is that the sequence in which the hypotheses are tested in relation 4.140 makes a difference. The effect of non-normality of the distribution of the residuals on the tests is also more important in the case of an unbalanced design. These effects increase as the differences between the \( n_{ij} \) increase.

To illustrate the utilization of the anova technique, we will use the observed \( \text{SO}_4^{2-} \) ionic concentration in precipitation at Chalk River, Ontario, between 1979 and 1996. The time series is shown in Figure 4.24. One can detect on this plot a net decrease during the 18 years. A two-way table was generated by calculating the median concentration for each month. The two-way table thus obtained is given in Table 4.1. These data are also illustrated using a three dimensional plots in Figure 4.25. One can easily detect a net decrease in the

---

**Figure 4.24** Observed \( \text{SO}_4^{2-} \) ionic concentrations in precipitation at Chalk River, Ontario.
Table 4.1 Monthly median $\text{SO}_4^{2-}$ ionic concentrations in precipitation at Chalk River, Ontario.

<table>
<thead>
<tr>
<th>Year</th>
<th>J</th>
<th>F</th>
<th>M</th>
<th>A</th>
<th>M</th>
<th>J</th>
<th>J</th>
<th>A</th>
<th>S</th>
<th>O</th>
<th>N</th>
<th>D</th>
</tr>
</thead>
<tbody>
<tr>
<td>1979</td>
<td>27.1</td>
<td>42.7</td>
<td>87.5</td>
<td>47.9</td>
<td>104.1</td>
<td>121.8</td>
<td>176.9</td>
<td>152.0</td>
<td>127.0</td>
<td>72.9</td>
<td>60.4</td>
<td>53.1</td>
</tr>
<tr>
<td>1980</td>
<td>48.7</td>
<td>29.2</td>
<td>45.8</td>
<td>69.8</td>
<td>68.7</td>
<td>106.2</td>
<td>60.4</td>
<td>77.0</td>
<td>72.9</td>
<td>60.4</td>
<td>69.8</td>
<td>31.2</td>
</tr>
<tr>
<td>1981</td>
<td>86.4</td>
<td>66.6</td>
<td>41.6</td>
<td>97.9</td>
<td>81.2</td>
<td>137.4</td>
<td>143.7</td>
<td>93.7</td>
<td>90.6</td>
<td>43.7</td>
<td>54.1</td>
<td>53.1</td>
</tr>
<tr>
<td>1982</td>
<td>27.1</td>
<td>30.2</td>
<td>81.2</td>
<td>124.9</td>
<td>125.9</td>
<td>81.2</td>
<td>88.5</td>
<td>43.7</td>
<td>60.4</td>
<td>73.9</td>
<td>52.1</td>
<td>41.6</td>
</tr>
<tr>
<td>1983</td>
<td>28.1</td>
<td>18.7</td>
<td>136.2</td>
<td>35.2</td>
<td>54.1</td>
<td>58.1</td>
<td>77.4</td>
<td>44.1</td>
<td>71.1</td>
<td>38.3</td>
<td>33.5</td>
<td>27.5</td>
</tr>
<tr>
<td>1984</td>
<td>21.2</td>
<td>58.9</td>
<td>47.1</td>
<td>33.3</td>
<td>65.3</td>
<td>92.4</td>
<td>63.7</td>
<td>81.0</td>
<td>55.4</td>
<td>71.4</td>
<td>48.9</td>
<td>21.1</td>
</tr>
<tr>
<td>1985</td>
<td>10.2</td>
<td>18.9</td>
<td>44.0</td>
<td>54.1</td>
<td>68.1</td>
<td>46.7</td>
<td>48.5</td>
<td>60.2</td>
<td>51.4</td>
<td>41.9</td>
<td>29.4</td>
<td>17.8</td>
</tr>
<tr>
<td>1986</td>
<td>32.8</td>
<td>23.0</td>
<td>70.4</td>
<td>54.9</td>
<td>49.6</td>
<td>41.4</td>
<td>19.2</td>
<td>63.5</td>
<td>51.4</td>
<td>73.7</td>
<td>74.3</td>
<td>17.5</td>
</tr>
<tr>
<td>1987</td>
<td>13.0</td>
<td>4.8</td>
<td>51.2</td>
<td>40.2</td>
<td>80.9</td>
<td>76.7</td>
<td>60.2</td>
<td>47.1</td>
<td>45.8</td>
<td>47.2</td>
<td>20.8</td>
<td>19.0</td>
</tr>
<tr>
<td>1988</td>
<td>56.4</td>
<td>14.7</td>
<td>40.8</td>
<td>52.3</td>
<td>54.3</td>
<td>24.9</td>
<td>29.2</td>
<td>71.4</td>
<td>112.7</td>
<td>35.3</td>
<td>53.5</td>
<td>16.5</td>
</tr>
<tr>
<td>1989</td>
<td>29.9</td>
<td>3.8</td>
<td>27.3</td>
<td>74.4</td>
<td>40.8</td>
<td>62.3</td>
<td>49.8</td>
<td>71.8</td>
<td>43.9</td>
<td>27.3</td>
<td>22.9</td>
<td>9.8</td>
</tr>
<tr>
<td>1990</td>
<td>32.2</td>
<td>26.9</td>
<td>56.0</td>
<td>58.1</td>
<td>36.6</td>
<td>63.1</td>
<td>32.5</td>
<td>87.9</td>
<td>50.4</td>
<td>38.2</td>
<td>41.9</td>
<td>23.3</td>
</tr>
<tr>
<td>1991</td>
<td>12.7</td>
<td>18.1</td>
<td>28.7</td>
<td>34.0</td>
<td>72.7</td>
<td>102.1</td>
<td>41.4</td>
<td>17.1</td>
<td>33.6</td>
<td>40.0</td>
<td>38.2</td>
<td>13.6</td>
</tr>
<tr>
<td>1992</td>
<td>19.6</td>
<td>13.4</td>
<td>56.2</td>
<td>126.9</td>
<td>55.2</td>
<td>59.6</td>
<td>19.0</td>
<td>29.4</td>
<td>57.9</td>
<td>39.6</td>
<td>35.6</td>
<td>30.4</td>
</tr>
<tr>
<td>1993</td>
<td>11.5</td>
<td>17.9</td>
<td>27.6</td>
<td>82.7</td>
<td>59.3</td>
<td>42.7</td>
<td>30.8</td>
<td>43.7</td>
<td>27.1</td>
<td>59.3</td>
<td>35.2</td>
<td>26.6</td>
</tr>
<tr>
<td>1994</td>
<td>7.6</td>
<td>7.7</td>
<td>43.1</td>
<td>38.1</td>
<td>37.1</td>
<td>54.5</td>
<td>53.9</td>
<td>48.9</td>
<td>38.5</td>
<td>28.9</td>
<td>30.4</td>
<td>21.9</td>
</tr>
<tr>
<td>1995</td>
<td>11.5</td>
<td>22.7</td>
<td>67.1</td>
<td>71.0</td>
<td>41.9</td>
<td>86.5</td>
<td>24.6</td>
<td>36.7</td>
<td>33.3</td>
<td>16.9</td>
<td>14.8</td>
<td>15.2</td>
</tr>
<tr>
<td>1996</td>
<td>13.4</td>
<td>20.0</td>
<td>33.8</td>
<td>39.8</td>
<td>32.2</td>
<td>40.1</td>
<td>39.0</td>
<td>64.4</td>
<td>43.0</td>
<td>28.6</td>
<td>18.3</td>
<td>25.8</td>
</tr>
</tbody>
</table>

The median concentration with time in both Table 4.1 and in Figure 4.25. The concentrations are noticeably higher between 1979 and 1981 than between 1994 and 1996.

The model given by relation 4.132 was fitted to the data using both an anova analysis and the median polish technique. The anova analysis indicates statistically significant effects for both the monthly and annual factors. For both tests, P-values less than 0.00000 were obtained.

![Figure 4.25 Variation with month and year of the monthly median $\text{SO}_4^{2-}$ ionic concentration in precipitation at Chalk River, Ontario.](image-url)
The anova and median polish estimates of the overall mean (parameter $\mu$ in relation 4.132) are 49.9 and 48.5 respectively. These two estimates are very close to each others.

The year effects are shown in Figure 4.24 for both the anova analysis and the median polish analysis. The two results are again very close to each others. Overall the median polish estimates give a slightly lower reduction in the median concentration than the anova analysis.

The month effects are shown for both techniques in Figure 4.27. As with the year effects, the results obtained using the two techniques are quite close to each other, although the median polish results are slightly lower than the anova results.

This example shows that the anova method can be used not only for testing for long-term trends in the data but also for estimating those long-term trends. Unlike the anocov technique discussed in the preceding section, the anova technique does not require a hypothesis about the exact shape of the trend.

4.5.4 Discussion

As we mentioned earlier, the anocov techniques have been used in atmospheric chemistry by a number of authors. However, the assumptions made by these techniques are quite stringent. First, the long-term trends are assumed to be straight lines with the same slope. Even if one can assume that the long-term trends are straight lines, there is a high probability that their slopes will be different in the different seasons. Therefore, the reader should verify before using these techniques that there is a basis for assuming that the long-term trends for each season followed a straight line and that their slopes are the same. This can be done, for example, by using the smoothing techniques presented in Section 4.3. The reader should always remember that polynomials of higher degree than 1 can be used and that different long-term trends can be fitted for different seasons.
Three other hypotheses used in the anocov techniques should be checked before accepting the results of the analysis. The anocov technique is quite tolerant of non-normality in the residuals if the number of data items is large. Moderate to large variation across season and with time in the variance of the residuals will create problems, however. In such circumstances, it may be necessary to transform the data to stabilize the variance. The presence of large correlations between the residuals may also create problems. If such autocorrelations are present it may be prudent to use the iterative procedure described in Section 4.6.3.1.

To the knowledge of the author, the anova technique has not been used before to study data in atmospheric chemistry, although it may be a useful tool as the simple example presented here illustrates. On the one hand, the technique is versatile, as it is not necessary to assume a priori any shape for the long-term trend. On the other hand, there is no guarantee that the additive model will be appropriate for explaining the observations. However, one can always use an interaction term in the model to improve its fit to the data. The most important problem is the fact that in most cases the design will be strongly unbalanced (i.e., there will be large differences in the number of data items in the cells), and this will create problems with the testing and interpretation of the results.

As we have seen, the anova technique is quite resistant to the non-normality of the residuals if the size of the design is large. The only exception occurs when the design is strongly unbalanced. As in the case of the anocov technique, variability in the residual variance and/or serial correlations of the residuals will create problems if they are important. A small variability in the variance of the residuals will not drastically affect the result of the analysis. The same is true for serial correlations in the residuals.

An interesting extension of the anova technique involves using a three-way table in place of a two-way table. The third dimension could be used for sites, for example, if we wanted to study temporal variation on a regional scale. The median polish technique can also be extended. For the latter, the reader is referred to Cook (1985) for a complete description of the technique. This technique has been used in the past by Wu and Zidek (1989).

A nonparametric version of the anocov and anova techniques can be created by replacing the observations by their ranks (Conover, 1980). This technique can be used to test for the statistical significance of the trend but not to estimate it.

4.6 Fitting models by Least squares or Maximum Likelihood Techniques

4.6.1 Introduction

One of the most popular techniques for analyzing temporal variation in a time series involves fitting a more or less general linear model to the data using the least squares or maximum likelihood techniques. The most popular model can be written as:

\[ C_t = \alpha + \beta t + \varepsilon_t \]  \hspace{1cm} (4.141)

or, when one wants to model the seasonal cycles,

\[ C_t = \alpha + \beta t + \gamma \sin(2\pi t + \phi) + \varepsilon_t \]  \hspace{1cm} (4.142)

where \( \varepsilon_t \) are independent random errors from a normal distribution with mean zero and constant variance \( \sigma^2 \) (e.g., MAP3s/RAINE, 1982). This model is often slightly modified to
take into account the fact that the distributions of ionic concentrations are closer to log-normality than to normality. Relation 4.142 is then written as:

$$\log(C_i) = \alpha + \beta t + \gamma \sin(2\pi t) + \epsilon_i$$  
(4.143)

or

$$\log(C_i) = \alpha + \beta t + \gamma_1 \sin(2\pi t) + \gamma_2 \cos(2\pi t) + \epsilon_i$$  
(4.144)

(e.g., Dana and Easter, 1987). In the case of ionic concentration in precipitation, a further improvement is necessary to take into account the possible relationship between concentration and precipitation amount. This results in the following model:

$$\log(C_i) = \alpha + \delta \log(P_i) + \beta t + \gamma_1 \sin(2\pi t) + \gamma_2 \cos(2\pi t) + \epsilon_i$$  
(4.145)

where $P_i$ is the precipitation amount associated with the sample at time $t$ (e.g., Berge, 1988). Buishand et al. (1988) have proposed the following model:

$$C_i = \alpha + \delta f(P_i) + \beta t + \gamma \cos(2\pi t - \phi) + \epsilon_i$$  
(4.146)

where $f(P_i)$ could be either $\sqrt{P_i}$ or $1/\sqrt{P_i}$.

All these models are special cases of the following general model for hourly data (e.g., Sirois, 1993):

$$g(C_i) = C_0 + f^p(P_i) + f^T(t) + f^C(t) + f^S(t) + f^0(t) + f^R(t) + f^T(t)$$  
(4.147)

where the different terms are:

1. $g(C_i)$ is a transformation of the original data usually to normalize them and/or to eliminate variation in the variance in the error term. The most popular forms for this function are: (1) $g(C_i) = C_i$ and (2) $g(C_i) = \log(C_i)$. Note that other transformations could also be used.

2. $C_0$ is a constant or intercept.

3. The term $f^p(P_i)$ describes the relationship between ionic concentration and precipitation amount. It is usually written as $f^p(P_i) = C^p P_i$, $f^p(P_i) = C^p \sqrt{P_i}$ or $C^p/\sqrt{P_i}$ when $g(C_i) = C_i$ and $f^p(P_i) = C^p \log(P_i)$ when $g(C_i) = \log(C_i)$. Note that $f^p(P_i) = 0$ for other types of data than ionic concentrations in precipitation.

4. $f^T(t)$ describes the long-term variations or long-term trend. It will usually be written as a combination of a polynomial of order $N_T$ and a series of $M_T$ trigonometric functions with long (i.e., more than 6 to 10 years) periods, $\lambda_i$ (in years). In most cases, the best fit will be obtained using only a polynomial or a series of trigonometric functions. The general form of $f^T(t)$ can be written as:

$$f^T(t) = C_1^T t + C_2^T t^2 + \cdots + C_{N_T}^T t^{N_T} + \sum_{m=1}^{M_T} C_m^\sin \left( \frac{2\pi t}{T \lambda_m} \right) + C_m^\cos \left( \frac{2\pi t}{T \lambda_m} \right)$$  
(4.148)
where \( T \) is equal to \( 24 \times 365.25 \) for hourly data, 365.25 for daily data, and so on, and \( t \) is in hours from some original time before or at the beginning of the time series.

5. The function \( f^C(t) \) describes the possible long-term cycles with periods greater than one year but less than about half the length of the time series (so they can be identified). Generally, \( f^C(t) \) is written as a sum of \( M_c \) trigonometric terms with periods equal to \( \omega_i \) (in years) as follows:

\[
f^C(t) = \sum_{m=1}^{M_c} \left[ C^C_m \sin \left( \frac{2\pi m T}{T \omega_i} \right) + C^C_m \cos \left( \frac{2\pi m T}{T \omega_i} \right) \right]
\]  

(4.149)

6. The seasonal cycle is described by the term \( f^S(t) \) and is usually approximated by a series of \( M_s \) trigonometric functions as follows:

\[
f^S(t) = \sum_{m=1}^{M_s} \left[ C^S_m \sin \left( \frac{2\pi m T}{T} \right) + C^S_m \cos \left( \frac{2\pi m T}{T} \right) \right]
\]  

(4.150)

7. The term \( f^D(t) \) describes the daily variations and can usually be approximated by a series of \( M_d \) trigonometric functions with frequencies, \( \tau_i \) (in days), as:

\[
f^D(t) = \sum_{m=1}^{M_d} \left[ C^D_m \sin \left( \frac{2\pi m T}{24 \tau_i} \right) + C^D_m \cos \left( \frac{2\pi m T}{24 \tau_i} \right) \right]
\]  

(4.151)

8. \( f^H(t) \) is the diurnal cycle, described using a series of \( M_h \) trigonometric functions as:

\[
f^H(t) = \sum_{m=1}^{M_h} \left[ C^H_m \sin \left( \frac{2\pi m T}{24} \right) + C^H_m \cos \left( \frac{2\pi m T}{24} \right) \right]
\]  

(4.152)

9. The last term in relation 4.147, \( f^R(t) \), is the random error term. If the errors are assumed to be independent, this term will be equal to \( f^R(t) = \varepsilon_t \), where the \( \varepsilon_t \) are independent random errors from a normal distribution with mean zero and constant variance \( \sigma^2 \). If the errors are assumed to be correlated, an auto-regressive model of order \( K \) is used to describe the error term. Its general form will be:

\[
\varepsilon_t = a_1 \varepsilon_{t-1} + a_2 \varepsilon_{t-2} + \cdots + a_K \varepsilon_{t-K} + \eta_t
\]  

(4.153)

where the \( \eta_t \) are independent random errors from a normal distribution with mean zero and constant variance \( \sigma^2 \).

Note that for data with sampling periods greater than an hour, only the terms that are applicable will be used. For example, for daily data the term \( f^D(t) \) and \( f^H(t) \) will not be used. If the time series covers less than a year, the term \( f^C(t) \) and \( f^S(t) \) will not usually be used. In summary, this general model has to be adapted for the data under study.
The model given by relation 4.147 is the most general model and includes as special cases all the others mentioned earlier in this section. We will discuss how that model is fitted to the data later, but we must first mention how the parameters \(N_T, M_T, M_C, M_8, M_D, M_H\) and \(K\) can be selected. These parameters have to be chosen before trying to fit the model to the data. On the one hand, they can be selected when writing down the model and not modified later. On the other hand, as in Sirois (1993, 1997), many combinations of the parameters can be tried and the one for which the fit of the model to the data is the best would be selected. The latter method requires more work and time but reduces the number of \textit{a priori} assumptions about the shape of the different temporal variations present in the time series.

One may often want to put some of the parameters (e.g., \(C_0, C_1^T, C_1^D, C_1^{TC}, C_1^{SS}, \ldots\)) of the model equal to zero before trying to fit the model. This can be accomplished easily as we will see later. This permits more flexibility in the construction of the model.

When these arbitrary parameters have been selected, it is possible to fit the model to the observed data. The method used depends on whether the errors are assumed to be independent or serially correlated. In the former case, either the least squares technique or the maximum likelihood technique can be used, and both give the same solution. In the latter case, two different techniques can be used. We will describe them later. In all cases the first step in fitting the model to the data is to write down four matrices using the observations. We will summarize this process first before discussing the different techniques to fit the model.

If we assume that \(n\) observations, \(C_i (i = 1, 2, \ldots, n)\) are available at time \(t_i (i = 1, 2, \ldots, n)\), we can write the \((n \times 1)\) observation matrix, \(Y\), as:

\[
Y = \begin{pmatrix}
g(C_{i_1}) \\
g(C_{i_2}) \\
\vdots \\
g(C_{i_n})
\end{pmatrix}
\]  

(4.154)

If we then write

\[
n_p = 1 + N_T + 2M_T + 2M_C + 2M_8 + 2M_D + 2M_H - \text{(number of parameters put equal to zero)}
\]  

(4.155)

the \((n_p \times 1)\) parameter matrix, \(b\), is:

\[
b = \begin{pmatrix}
C_0 \\
C_1^T \\
C_1^D \\
C_1^{TC} \\
C_1^{SS} \\
\vdots \\
C_1^{IC} \\
C_1^{HIC}
\end{pmatrix}
\]  

(4.156)
Only the parameters not equalized to zero appear in $b$. Finally, we defined the $(nxn_p)$ matrix $X$ as:

$$
X = \begin{pmatrix}
1 & t_1 & t_1^2 & \cdots & \sin\left(\frac{2\pi t_1}{T}\right) & \cos\left(\frac{2\pi t_1}{T}\right) & \cdots & \cos\left(\frac{2\pi M_1 t_1}{T}\right) & \cdots & \cos\left(\frac{2\pi M_1 h t_1}{T}\right) \\
1 & t_2 & t_2^2 & \cdots & \sin\left(\frac{2\pi t_2}{T}\right) & \cos\left(\frac{2\pi t_2}{T}\right) & \cdots & \cos\left(\frac{2\pi M_2 t_2}{T}\right) & \cdots & \cos\left(\frac{2\pi M_2 h t_2}{T}\right) \\
\vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\
1 & t_n & t_n^2 & \cdots & \sin\left(\frac{2\pi t_n}{T}\right) & \cos\left(\frac{2\pi t_n}{T}\right) & \cdots & \cos\left(\frac{2\pi M_n t_n}{T}\right) & \cdots & \cos\left(\frac{2\pi M_n h t_n}{T}\right)
\end{pmatrix}
$$

(4.157)

Note that the function of $t$ associated with the parameters put equal to zero does not appear in the matrix $X$. It is important that the matrices $b$ and $X$ be consistent. If one also writes the following $(nx1)$ matrix $E$ as

$$
E = \begin{pmatrix}
f^R(t_1) \\
f^R(t_2) \\
\vdots \\
f^R(t_n)
\end{pmatrix}
$$

(4.158)

relation 4.147 can be written as a matrix equation:

$$
Y = Xb + E
$$

(4.159)

To help the reader understand how these matrices are obtained, we will present two examples. The first one will be for the model described by relation 4.145. The different matrices are:

$$
Y = \begin{pmatrix}
\log(C_{t_1}) \\
\log(C_{t_2}) \\
\vdots \\
\log(C_{t_n})
\end{pmatrix}, \quad b = \begin{pmatrix}
\alpha \\
\delta \\
\beta \\
\gamma_1 \\
\gamma_2
\end{pmatrix}, \quad E = \begin{pmatrix}
e_1 \\
e_2 \\
\vdots \\
e_n
\end{pmatrix}
$$

(4.160)

and

$$
X = \begin{pmatrix}
1 & \log(P_{t_1}) & t_1 & \sin(2\pi t_1) & \cos(2\pi t_1) \\
1 & \log(P_{t_2}) & t_2 & \sin(2\pi t_2) & \cos(2\pi t_2) \\
\vdots & \vdots & \vdots & \vdots & \vdots \\
1 & \log(P_{t_n}) & t_n & \sin(2\pi t_n) & \cos(2\pi t_n)
\end{pmatrix}
$$

(4.161)
As a second example, we will use the following more complex model for daily data (i.e., $T = 365.25$):

$$
\log(C_t) = C_0 + C^P \log(P_t) + C^T t + C^T t^3 + C^{\sin} \left( \frac{2\pi t}{15T} \right) + C^{\cos} \left( \frac{2\pi t}{3T} \right) + C^{\sin} \left( \frac{2\pi t}{T} \right) + C^{\cos} \left( \frac{4\pi t}{T} \right) + \epsilon_t
$$

(4.162)

The four matrices for that model are:

$$
\begin{align*}
Y &= \begin{pmatrix}
\log(C_{t_1}) \\
\log(C_{t_2}) \\
\vdots \\
\log(C_{t_n})
\end{pmatrix},
\quad b = \begin{pmatrix}
C_0 \\
C^P \\
C^T \\
C^{\sin} \\
C^{\cos}
\end{pmatrix},
\quad E = \begin{pmatrix}
\epsilon_{t_1} \\
\epsilon_{t_2} \\
\vdots \\
\epsilon_{t_n}
\end{pmatrix}
\end{align*}
$$

(4.163)

\[
X = \begin{pmatrix}
1 & \log(P_{t_1}) & t_{1} & t_{1}^3 & \sin\left( \frac{2\pi t_{1}}{15T} \right) & \cos\left( \frac{2\pi t_{1}}{15T} \right) & \sin\left( \frac{2\pi t_{1}}{T} \right) & \cos\left( \frac{4\pi t_{1}}{T} \right) \\
1 & \log(P_{t_2}) & t_{2} & t_{2}^3 & \sin\left( \frac{2\pi t_{2}}{15T} \right) & \cos\left( \frac{2\pi t_{2}}{15T} \right) & \sin\left( \frac{2\pi t_{2}}{T} \right) & \cos\left( \frac{4\pi t_{2}}{T} \right) \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
1 & \log(P_{t_n}) & t_{n} & t_{n}^3 & \sin\left( \frac{2\pi t_{n}}{15T} \right) & \cos\left( \frac{2\pi t_{n}}{15T} \right) & \sin\left( \frac{2\pi t_{n}}{T} \right) & \cos\left( \frac{4\pi t_{n}}{T} \right)
\end{pmatrix}
\]  

(4.164)

4.6.2 Without Auto-correlation in the Residuals

If the residuals are assumed to be independent random samples from a normal distribution with mean zero and constant variance $\sigma^2$, the least squares estimate of $b$, $\hat{b}$, is:

$$
\hat{b} = (X'X)^{-1}X'Y
$$

(4.165)

where $X'$ is the transpose of $X$ and $A^{-1}$ is the inverse of $A$ (see Nater and Wasserman, 1974; Draper and Smith, 1981; Johnson and Wichern, 1982). An estimate of the variance $\sigma^2$, $\hat{\sigma}^2$, is:

$$
\hat{\sigma}^2 = \frac{1}{n-n_p} (Y-X\hat{b})' (Y-X\hat{b})
$$

(4.166)

Note that the maximum likelihood estimate for these two parameters is also given by 4.165 and 4.166.
The estimate \( \hat{b} \) is distributed following a multivariate distribution with mean, \( \beta \) (i.e., the true values for the regression parameter), and standard deviation \( \sigma^2(X'X)^{-1} \). It is therefore possible to test the statistical significance of the regression parameters or to create confidence intervals for their true values (see Neter and Wasserman, 1974; Draper and Smith, 1981; Johnson and Wichern, 1982).

As an example of the use of least squares regression where the residuals are assumed to be independent, we will present the results of such an analysis for the concentration of \( \text{SO}_4^{2-} \) in precipitation at Chalk River, Ontario. The time series is shown in Figure 4.24. The following model was fitted to the data using the least squares technique:

\[
\log(C_i) = C_0 + C_1 \log(P_i) + C_2 i + C_3 i^2 \\
+ C_4 \cos\left(\frac{2\pi t}{365.25}\right) + C_5 \cos\left(\frac{4\pi t}{365.25}\right) + C_6 \sin\left(\frac{4\pi t}{365.25}\right) + \epsilon_t
\]  

Figure 4.28: Estimated long-term trend for ionic concentration of \( \text{SO}_4^{2-} \) in precipitation at Chalk River, Ontario

Although this model may not be the best one for these data (as no efforts were made to find the best model for the present example), the long-term trend and the seasonal cycle should be close to the results that the best model would produce. The long-term trend is shown in Figure 4.28 and the seasonal cycle in Figure 4.29. The model explains about 28% of the total variance, with 2% explained by the relationship between ionic concentration and the precipitation amount, 10% by the long-term trend, and 16% by the seasonal cycle. A visual study of Figures 4.28 and 4.29 suggests that the fit between temporal model and observations is very good. Figure 4.28 indicates that most of the decrease occurred before 1990. Fitting a straight line to these data would give the wrong impression of the long-term variation in the data, especially for the last few years, because the mean ionic concentrations have not changed much in the last five years of the sampling period. The use of a simple sine/cosine function to describe the seasonal cycle would also be misleading, although it would be less so than the use of a straight line in the case of the long term trend (see Figure 4.28). If one considers the residuals, one will notice that serial correlations exist between them. One should therefore use the techniques described in the next section.
4.6.3 With Auto-correlation in the Residuals

Serial correlations or auto-correlations in the residuals result in part from the fact that some important temporal variations in the data are not included in the model used to describe the temporal variation in the observations. This can be seen if one fits a series of models of increasing complexity to the data. If one examines the auto-correlations after fitting each model, one will notice that they decrease after the inclusion of significant temporal variations in the model. In summary, part of auto-correlations in the residuals reflects the fact that not all significant temporal variations present in the observations have been explained by the model.

The main effect of the presence of auto-correlations in the residuals of the least squares estimates of the temporal model is either an under- or an over-estimation of the variance of the coefficients of the model, depending on the exact nature of the auto-correlation. This has an impact on the tests of statistical significance of these coefficients. It is important to note that the least-squares or maximum likelihood estimates of the regression coefficients are still unbiased. However, as mentioned earlier, the estimation of the variance can be quite inaccurate, thus rendering any test of significance meaningless.

It is therefore important to take auto-correlations into account. One way is to develop a comprehensive model for the temporal variation in the data. However, it is not always possible to do this. It is therefore necessary to find techniques to fit models described by relation 4.147 with a random term $f^R(t)$ as defined by relation 4.153 if we are to include the presence of serial-correlations in the data.

We will discuss two types of techniques that may be used to fit such a model. In the first, an ordinary least squares estimate is obtained. This is followed by fitting an auto-regressive model to the residuals and then, after correcting the data for the auto-regressive model, making a new least squares estimation of the regression coefficient. This process can be done iteratively until the coefficients of the temporal variation model and the auto-regressive model do not change. In the second technique, one writes for the observed data the maximum likelihood for the regression model with auto-correlations in the residuals and then...
maximizes it. We will discuss these two types of techniques in more details in the remaining pages of this section.

4.6.3.1 Iterative Regression Method

We will start by discussing the simplest case of an auto-regressive model of order one. Relation 4.153 becomes:

\[ \varepsilon_i = \phi \varepsilon_{i-1} + \eta_i \]  

(4.168)

Let relation 4.147 be written as:

\[ Y = Xb + E \]  

(4.169)

where \( Y, b \) and \( X \) are defined in relations 4.154, 4.156, and 4.157 respectively. The vector \( E \) is defined as:

\[ E = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix} \]  

(4.170)

with the \( \varepsilon_i \) defined by relation 4.168.

If \( \phi \) is known, the generalized least squares estimates of \( b \), \( \hat{b} \), is equal to (see Tiao et al., 1990; Seber and Wild, 1989):

\[ \hat{b} = (X'V^{-1}_\phi X)^{-1}X'V^{-1}_\phi Y \]  

(4.171)

with:

\[ V^{-1}_\phi = \frac{1}{1 - \phi^2}R_\phi^T R_\phi \]  

(4.172)

where \( R_\phi \) is a \((n \times n)\) matrix given by:

\[ R_\phi = \begin{pmatrix} \sqrt{1-\phi^2} & 0 & 0 & \cdots & 0 & 0 \\ -\phi & 1 & 0 & \cdots & 0 & 0 \\ 0 & -\phi & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & -\phi & 1 \end{pmatrix} \]  

(4.173)

The covariance matrix of the \( \hat{b} \)'s is equal to

\[ \text{cov}(\hat{b}) = \frac{\sigma^2_\eta}{1 - \phi^2}(X'V^{-1}_\phi X)^{-1} \]  

(4.174)

where \( \sigma^2_\eta \) is the variance of the \( \eta_i \) in relation 4.168.
It is now necessary to estimate the auto-correlation coefficient. If the coefficient vector $b$ is known, $\phi$ could be estimated using the following relation

$$
\hat{\phi} = \frac{\sum_{t=2}^{n} y'_{t-1} y'_{t}}{\sum_{t=2}^{n} (y'_{t-1})^2}
$$

where $Y' = Y - Xb$, if the observations are distributed uniformly and there are no missing data. If there are observations missing, the technique developed by Dunsmuir (1984) could be used (see Appendix II). An estimation of $\phi$ is given by:

$$
\hat{\phi} = \frac{c_i^\phi}{c_o^\phi}
$$

where the $c_i^\phi$ are calculated using relations A5 to A10 in Appendix II.

One can therefore see that to estimate the vector $b$, one needs to know $\phi$, and to estimate the latter one needs to know the former. This problem can be solved by the following iterative procedure.

1. Assuming that $\phi = 0$, obtain a first estimation for $b$ using ordinary least squares.
2. Calculate the residuals, i.e., $Y' = Y - Xb$.
3. Calculate an estimate of $\phi$ using relation 4.175 or 4.176.
5. Go back to step 2.
6. Loop through steps 2 to 5 until the values obtained for $b$ and $\phi$ do not change.

Other techniques for solving that problem are presented in Seber and Wild (1989).

Seber and Wild (1989) present extensions of these techniques to auto-regressive models of order $K$. The reader is referred to this reference for more details. An example of the use of this technique can be found in Li et al. (1993).

4.6.3.2 Maximum Likelihood Method

The model to fit to the data can be written as:

$$
Y = Xb + E
$$

where $E$ is defined by relation 4.170 with the $\varepsilon_i$ following an auto-regressive model of order $K$ given by:

$$
\varepsilon_i = a_1 \varepsilon_{i-1} + a_2 \varepsilon_{i-2} + \cdots + a_K \varepsilon_{i-K} + \eta_i.
$$

If $n$ observations for $y$ are available and the random errors $\eta_i$ are assumed to be independent and drawn from a normal distribution with zero mean and constant variance $\sigma^2$, the full log-likelihood function for this model is:
\[
\log \{ L(y; a, b, \sigma^2) \} = -\frac{1}{2} n \log(2\pi) - \frac{1}{2} n \log(\sigma^2) - \frac{1}{2} \log(|V|) - \frac{1}{2} \sigma^2 (Y - Xb)' V^{-1} (Y - Xb)
\]

(4.179)

where the \((nxn)\) matrix \(V\) is defined as \(E(\varepsilon \varepsilon') = \sigma^2 V\) and thus \(V\) is a function of the parameters \(a\), \(b\) and \(\sigma^2\) that maximize the log-likelihood function. The easiest way to solve this problem is to rewrite relation 4.176, using the state space formulation of the time series model (Fuller, 1996). As the equations become quite complex, we will not go into the details here but refer the interested reader to Harvey and Philips (1979) and Kohn and Ansley (1985) for a complete description. If there are no missing data, the algorithm of Ansley (1979) can be used to maximize the likelihood. If some of the observations are missing, one has to use the Kalman filter (see Kohn and Ansley, 1986).

As we have mentioned, the errors, \(\eta_i\), are assumed to be normally distributed. However, their distribution can deviate from normality when the amount of data is large. In that case, the technique gives unbiased estimates of the model parameters. This results from the central limit theorem (see Silvey, 1975). For large \(n\), it can be shown (Silvey, 1975) that the random vector \(\theta\) (i.e., \(\theta' = (a', b', \sigma^2)\)) follows a multidimensional normal distribution with mean \(\theta\) and covariance \(\Theta\) equal to:

\[
\{\Theta\}_\theta = \left\{ - \left[ \frac{\partial^2 \log(L)}{\partial \theta_i \partial \theta_j} \right]_{\theta^*} \right\}^{-1}
\]

(4.180)

even if the distribution of the \(\eta_i\) is not normal. \(\theta^*\) is the vector for the parameters of the model that maximizes the log-likelihood. If one wants to test or create confidence intervals for more than one parameter, simultaneous inferences techniques, like Bonferroni’s multiple testing technique, should be used. See Miller (1981) or Hochberg and Tamhane (1987) for more details.

Examples of the use of this technique can be found in Sirois (1997) and Sirois and
A Brief Overview of Cluster Analysis

Alain Sirois

Atmospheric Environment Service
Environment Canada
2121, Voie de Service Nord, Suite 500
Route Trans-canadienne,
Dorval, Quebec, Canada, H9P 1J3
Notice

This report is the background document for the presentation of the same name given at the WMO-EMEP Workshop on Advanced Statistical Methods and their Application to Air Quality Data Sets, Finnish Meteorological Institute, Helsinki, Finland, 14-18 September, 1998. The author can be reached at the following address: Dr. Alain Sirois (ARQI), Service de l'Environnement Atmospherique, 2121, Voie de Service Nord, Suite 500, Route Trans-canadienne, Dorval, Quebec Canada, H9P 1J3. Tel: (514) 421-4674; Fax: (514) 421-2106; E-mail: Alain.Sirois@ec.gc.ca
Preface

Along with principal component analysis, cluster analysis has been one of the most frequently used statistical analysis techniques in the field of atmospheric chemistry. It has been employed mainly as a method to group atmospheric chemistry measurements by air mass origin using associated air mass back-trajectories. References to some of these analyses are given in the overview.

Although the subject of cluster analysis is an important and vast one, the author has had to limit himself to only touch lightly on the topic because of lack of space and time. However, he has tried to give a comprehensive, although brief, overview of the basic concepts of cluster analysis.

After defining the aim of cluster analysis in the introduction, the concept of distance or similarity measurements are discussed. These is followed by a brief description of available basic clustering techniques. A brief mention is then made of graphical techniques that can be used if the number of available measurements to classify is not too large. Finally, an example of the use of cluster analysis in the context of atmospheric chemistry is presented at the end of the overview. Many references to more comprehensive books on cluster analysis are given throughout the overview to compensate for its brevity.

As mentioned earlier, time for the preparation of this overview was limited. Although the author tried as much as possible to make sure that the equations, examples, and descriptions are correct, errors could easily have crept in, especially in unifying the notation used by the different authors consulted. The author would therefore recommend, in particular, that before using the equations presented herein, the reader verify them in the references cited.
Contents

Preface iii
Contents v
1. Introduction 1
2. Definition of Distance or Similarity Measurements 2
3. Clustering Techniques 3
   3.1 Hierarchical Techniques 3
      3.1.1 Agglomerative Methods 3
         3.1.1.1 Single Linkage or the Nearest-Neighbor Method 4
         3.1.1.2 Complete Linkage or the Furthest-Neighbor Method 4
         3.1.1.3 Centroid Method 5
         3.1.1.4 Average Linkage 5
         3.1.1.5 Ward’s Error Sum of Squares Method 5
         3.1.1.6 Lance and Williams Flexible Method 6
      3.1.2 Divisive Methods 6
   3.2 Partitioning Methods 7
      3.2.1 K-Means Clustering 7
      3.2.2 Methods Based on Trace 8
4. Graphical Methods 8
   4.1 Glyphs and Metroglyphs 9
   4.2 Fourier Series 10
   4.3 Chernoff Faces 11
5. Application to Atmospheric Chemistry: An example 11
6. Final Remarks 16
7. References 17
1. Introduction

The starting point of cluster analysis is the same as in principal component analysis (PCA) or other reduction techniques; i.e., one has n observations of p variables that can be put into matrix form as:

\[
X = \begin{pmatrix}
  x_{11} & x_{12} & \cdots & x_{1p} \\
  x_{21} & x_{22} & \cdots & x_{1p} \\
  \vdots & \vdots \\
  x_{n1} & x_{n2} & \cdots & x_{np}
\end{pmatrix}
\]  

(Cl. 1.1)

Cluster analysis differs from principal component analysis and other similar techniques, however, in that its aim is to group the observations (referred to as objects) into clusters, rather than to find links between the p variables. Cluster analysis can be summarized by the following four steps:

1. Collect n observations of p variables.
2. Create a similarity or distance matrix.
3. Form the clusters, which can either a) be mutually exclusive or b) form a hierarchy.
4. Look at the cluster profile.

The main goal of cluster analysis is to arrive at clusters in which the variation within clusters is smaller than that between clusters. This is illustrated in Figure 1.1 for a hypothetical two-dimensional case. In this example, it is clear that the variation between the objects in each cluster is smaller than the variation between objects in different clusters.

When doing a cluster analysis, one has to make two key decisions that will dominate the analysis. They are:

1. decide on a measure of inter-object distance or similarity.
2. define a procedure to define the clusters.

Many definitions of distance or similarity have been used in the past and many cluster algorithms have been developed. In the present overview, we will only mention the most important ones.

There is no generally accepted definition of a cluster. The meanings of such terms as cluster or distance (similarity) depends largely on individual judgments.

If we imagine each object or observation as a point in a p-dimensional space (p being the number of variables measured), we can describe clusters as continuous regions with a high density of points separated by regions having a low density of points (Figure 1.1). This defines what is known as natural clusters and does not impose any a priori restrictions on the structure of the data and clusters. As such, the data guide the analysis.

Figure 1.1 Examples of three clusters
Another definition is based on the criterion of closeness. This type of definition invokes the premise that objects or observations in a cluster should be closer to each other than to objects in other clusters. Note, however, that it is difficult to identify clusters that are not spherical when using this criterion.

In the following sections, we will first discuss the definition of distances and then present different techniques that have been proposed for estimating clusters.

2. Definition of Distance or Similarity Measurements

Since the variables measured in atmospheric chemistry are metric, we will limit our discussion of distance to these types of data. For more information on the definition of similarity for other types of data, the reader is referred to Dillon and Goldstein (1984).

The most common definition of distance between two objects or observations is the Euclidean distance between two objects \( i \) and \( j \), defined as:

\[
d_{ij} = \left( \sum_{k=1}^{p} (x_{ik} - x_{jk})^2 \right)^{1/2}
\]  

(2.1)

Another interesting distance is the absolute or city-block distance, defined as:

\[
d_{ij} = \sum_{k=1}^{p} |x_{ik} - x_{jk}|
\]  

(2.2)

These two types of distance are special cases of the Minkowski metric, defined as:

\[
d_{ij} = \left( \sum_{k=1}^{p} |x_{ik} - x_{jk}|^r \right)^{1/r}
\]  

(2.3)

The reader should notice that the distances defined in 2.1 to 2.3 are not scale invariant. Therefore, if the unit of measurement of a variable is changed, the distance between the observations is also changed.

Using relations 2.1 to 2.3 has the disadvantage that the distance is dominated by the variable(s) with the largest range of variation (i.e., the greatest spread). For example, if most of the variables vary between 0 and 10 but one varies between 0 and 10000, the distance would be dominated by the latter. In these cases, it is recommended that the variables be standardized (by dividing each variable by its standard deviation) before the distances are calculated.

Another way of scaling the variables so that no one variable dominates the distance calculation is to use the Mahalanobis distance, given by:

\[
(X_j - X_i) S^{-1} (X_j - X_i)
\]  

(2.4)

where \( X_i \) and \( X_j \) are the respective vectors of observations \( i \) and \( j \) and \( S \) is the pooled within-group variance-covariance matrix. This distance measure has the advantage of explicitly accounting for any correlations that might exist between the variables. Note that if \( S = I \) (i.e., the identity matrix), the Mahalanobis distance reduces to the Euclidean distance. Figure 2.1 presents the geometry of different definitions of distances. In the case of the Euclidean distance, all the observations lying on a circle around observation \( C \) are at the same distance.
from it. If the data have been standardized, the circles become ellipses with the axes proportional in length to the variance of the variable. The Mahalanhois distance rotates the ellipses by an angle proportional to the correlation between the two variables.

Other definitions of distances can be created in special cases. An example of a special definition of distance is the mean angle distance defined by Sirois and Bottenheim (1995) in the case of air mass back-trajectories. The reader is referred to their paper for more details.

3. Clustering Techniques

The different types of clustering techniques can be grouped into two categories, namely, hierarchical techniques and partitioning techniques. We will discuss each of these in the rest of this section.

3.1 Hierarchical Techniques

In hierarchical techniques, one performs a succession of fusions or divisions of the clusters. An important characteristic of hierarchical techniques is that the allocation of an observation to a cluster is irrevocable; that is, once an object is allocated, it cannot be reallocated.

Two types of hierarchical clustering exist, namely, agglomerative methods and divisive methods. In the first, one starts with each object constituting a cluster and then groups them until every object is in one cluster. In divisive methods, one starts with one cluster and divides clusters until all the resulting clusters contain only one object. These techniques are illustrated in Figure 3.1.

3.1.1 Agglomerative Methods

In agglomerative methods, one starts with n clusters; in other words, each observation is a cluster. The first step is to merge the two closest clusters to form (n-1) clusters, then one merges the next two closest clusters to form (n-2) clusters, and so on. This process is continued until all the objects or observations are in the same cluster. This is illustrated in Figure 3.1a. The fusion of two clusters has to be made n-1 times.
Various methods differ, however, in the way they measure the distance between two clusters. (Note that if the two clusters contain one object each, all the distances are the same.)

Examples of the use of agglomerative methods in atmospheric chemistry can be found in Slanina et al. (1983), Moody and Galloway (1988), and Moody and Samson (1989).

3.1.1.1 Single Linkage or the Nearest-Neighbor Method

In this method, the distance between clusters is defined as the distance between their nearest members. This is illustrated in Figure 3.2 and can be written mathematically by:

\[ D_{ij} = \min_{m=1,\ldots,n_i, n=1,\ldots,n_j} (d_{mn}) \]  

(3.1)

where \( D_{ij} \) is the distance between clusters \( i \) and \( j \); \( n_i \) and \( n_j \) are the number of observations in clusters \( i \) and \( j \), respectively; and \( d_{mn} \) is the distance between the objects \( n \) and \( m \). Note that \( n \) and \( m \) index the observations in clusters \( i \) and \( j \) respectively.

This method is subject to chaining (chaining forms linear groups of objects). An example of chaining is shown in Figure 3.3.

3.1.1.2 Complete Linkage or the Furthest-Neighbor Method

The definition of distance in this method is the opposite of the one given in the preceding section. Here, the distance between two clusters is defined as the distance between the most distant pair of observations. This is illustrated in Figure 3.4 and can be written mathematically as:

\[ D_{ij} = \max_{m=1,\ldots,n_i, n=1,\ldots,n_j} (d_{mn}) \]  

(3.2)

This method tends to overcome the problem of chaining.
3.1.1.3 Centroid Method

In this method, the distance between two clusters is defined as the distance between each group’s centroid. This is illustrated in Figure 3.5. The distance between clusters i and j can be written as:

$$D_{ij} = \sum_{k=1}^{p} (\bar{x}_k^i - \bar{x}_k^j)^2$$  \hspace{1cm} (3.3)

where $\bar{x}_k^i$ is the kth coordinate of the centroid for cluster i. It is equal to:

$$\bar{x}_k^i = \frac{1}{n_i} \sum_{m=1}^{n_i} x_{mk}$$ \hspace{1cm} (3.4)

The summation is performed on the $n_i$ objects in cluster i.

3.1.1.4 Average Linkage

In this method, the distance between two clusters is defined as the average distance between all pairs of objects, one from one cluster and one from the other. This can be written as:

$$D_{ij} = \frac{1}{n_in_j} \sum_{i=1}^{n_i} \sum_{j=1}^{n_j} d_{ij}$$ \hspace{1cm} (3.5)

This method tries to minimize within-group variance and maximize between-group variance.

3.1.1.5 Ward’s Error Sum of Squares Method

This method was developed by Ward (1963). The technique estimates the loss of information resulting from grouping objects into clusters by summing the squared deviations of every observation from the mean of the cluster to which they belong. The assignment rule rests on the increase in the error sum of squares induced from combining every possible pair of clusters. At each step in the analysis, the union of every possible pair of clusters is tested and the two clusters whose fusion results in the minimum increase in the error sum of squares are combined.

If we denote by $x_{ijk}$ the observation of the ith of the p variables for the jth of the nk objects in the kth of the h clusters, the mean for variable i and cluster k is:

$$\bar{x}_{ik} = \frac{1}{n_k} \sum_{j=1}^{n_k} x_{ijk}$$ \hspace{1cm} (3.6)

The error sum of squares for cluster k is given by:

$$E_k = \sum_{i=1}^{p} \sum_{j=1}^{n_k} (x_{ijk} - \bar{x}_{ik})^2 = \sum_{i=1}^{p} \sum_{j=1}^{n_k} x_{ijk}^2 - n_k \sum_{i=1}^{p} \bar{x}_{ik}^2$$ \hspace{1cm} (3.7)

and the error sum of squares is:
Table 3.1 Coefficients for the Lance and Williams recurrence formula

<table>
<thead>
<tr>
<th>Method</th>
<th>( \alpha_i )</th>
<th>( \alpha_j )</th>
<th>( \beta )</th>
<th>( \gamma )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Single Linkage</td>
<td>1/2</td>
<td>1/2</td>
<td>0</td>
<td>-1/2</td>
</tr>
<tr>
<td>Complete Linkage</td>
<td>1/2</td>
<td>1/2</td>
<td>0</td>
<td>1/2</td>
</tr>
<tr>
<td>Centroid</td>
<td>( \frac{n_i}{n_i + n_j} )</td>
<td>( \frac{n_j}{n_i + n_j} )</td>
<td>(-\alpha_i \alpha_j )</td>
<td>0</td>
</tr>
<tr>
<td>Average Linkage</td>
<td>( \frac{n_i}{n_i + n_j} )</td>
<td>( \frac{n_j}{n_i + n_j} )</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Ward's method</td>
<td>( \frac{n_k + n_i}{n_k + n_i + n_j} )</td>
<td>( \frac{n_k + n_j}{n_k + n_i + n_j} )</td>
<td>(-n_k )</td>
<td>0</td>
</tr>
</tbody>
</table>

\[
E_{SS} = \frac{1}{k} \sum_{k=1}^{b} E_k \quad (3.8)
\]

Ward's method tends to join clusters with a small number of observations and is strongly biased toward producing clusters with roughly the same number of observations. The method is very sensitive to outliers.

3.1.1.6 Lance and Williams Flexible Method

Lance and Williams (1967) developed a flexible algorithm that can be used for all the techniques presented in the preceding sections. The distance measures between groups used by these methods satisfy a recurrence formula for the distance between a cluster \( k \) and a cluster \((ij)\) formed by the fusion of clusters \( i \) and \( j \). The recurrence formula of Lance and Williams is:

\[
D_{k(ij)} = \alpha_i D_{ki} + \alpha_j D_{kj} + \beta D_{ij} + \gamma |D_{ji} - D_{kj}| \quad (3.9)
\]

where the coefficients \( \alpha \), \( \beta \) and \( \gamma \) are given in Table 3.1 for the different methods.

The Lance and Williams algorithm is very useful because the same program can be used for all the different agglomerative techniques presented in this section. Only the coefficients in relation 3.9 have to be changed.

3.1.2 Divisive Methods

In divisive methods, one starts with all the observations in one cluster. The first step in the method is to divide the cluster into two groups or clusters. If there are \( n \) observations, there are \( 2^{n-1} - 1 \) different ways to form these two subsets. Even if \( n \) is not particularly large, the most powerful computer would have trouble in testing all these combinations. Thus, any divisive method must give an efficient algorithm for dividing the first cluster into two groups.

Once the initial split is made, objects are moved from one cluster to another, or finer subdivisions of the already formed clusters are made.
The various divisive methods differ according to how the initial split is effected and how already formed clusters are subdivided.

One of the simpler methods is the Splinter-Average Distance Method of MacNaughton-Smith et al. (1962). With this technique, one starts by splintering out that object having the largest distance from the other objects. That operation leaves two groups containing the splintered object and the other containing the remaining objects. Next, one computes two distances: (1) the average distance of each object in the main group to the object(s) in the splinter group and (2) the average distance of each object in the main group to the other observations in the group. If an object’s average distance to the object(s) in the splinter group is less than its average distance to the other objects in the main cluster, it is moved to the splinter group. This process is continued until the composition of the two clusters has stabilized, which is the case when each object’s average distance to objects in its own cluster is smaller than its average distance to objects in the other cluster. The process is then repeated, with each of the two clusters being splintered again. The process is continued until all clusters contained only one object.

Another example of the divisive method is Automatic Interaction Detection (AID). Sonquist and Morgan (1964) give details of this technique.

To the knowledge of the author, no divisive methods have been used in atmospheric chemistry.

3.2 Partitioning Methods

In partitioning methods, the allocation of an observation to a cluster is not irrevocable. Objects may be reallocated if the initial assignments were inaccurate.

Partitioning techniques partition the data by optimizing some formal and predefined criterion. The use of partitioning techniques assumes that the number of final clusters is known and specified in advance, although some methods allow the number to vary during the course of the analysis.

Partitioning techniques differ with respect to (1) how clusters are initialized, (2) how objects are allocated to clusters, and (3) how some or all of the objects already clustered are reallocated to other clusters.

An example of the use of partitioning cluster analysis can be found in Darling et al. (1992a, 1992b).

3.2.1 K-Means Clustering

Suppose that one has \( n \) observations of \( p \) variables, denoted as \( X_i \) (\( i = 1, 2, \ldots, n \); \( j = 1, 2, \ldots, p \)). We will denote by \( \bar{X}_{jk} \) the mean value for variable \( j \) in cluster \( k \) (\( k = 1, 2, \ldots, h \)). The process starts by generating an initial mean vector, called a seed, for each of the \( h \) clusters. Different techniques have been proposed to generate them. The easiest uses the first \( h \) observations as the seeds. A second technique is to select \( h \) observations randomly and use them as seeds. A third technique is to use the \( h \) observations that are mutually furthest apart. A final technique is to employ prior knowledge to generate the \( h \) seeds.

Once the \( h \) seeds have been selected, each observation is allocated to the closest cluster. New mean values for each cluster can be calculated using the observations in it. The
observations are then re-allocated to the closest cluster, and new mean vectors for each cluster are obtained. This process is continued until no observation can be transferred from one cluster to another.

This technique is very simple but it can take a long time to achieve convergence if the number of observations is large. In some cases there may be problems in achieving convergence. These usually occur when a few observations are alternatively shifted between two clusters because they are at the periphery of the two clusters.

3.2.2 Methods Based on Trace

Some partitioning algorithms are based upon minimizing (or maximizing) within-group (between-group) dispersion. If \( T \) denotes the total dispersion matrix, \( W \) the within-clusters dispersion matrix, and \( B \) the between-clusters dispersion matrix, then:

\[
T = W + B
\]

(3.10)

where, assuming \( h \) clusters, \( W = \sum_{k=1}^{h} W_k \). For any given set of data, \( T \) is fixed, so the clustering criteria can be defined in terms of \( W \) or \( B \). Several possibilities have been suggested, including:

1. Trace of \( W \). This criterion attempts to minimize the trace of the pooled within-cluster matrix of sums-of-squares and cross-products. Note that minimizing \( \text{tr}(W) \) is equivalent to maximizing \( \text{tr}(B) \), since \( \text{tr}(T) = \text{tr}(W) + \text{tr}(B) \).

2. Determinant of \( W \). Minimizing \( |W| \) is equivalent to maximizing \( |T|/|W| \). Friedman and Rubin (1967) suggest that this latter criterion be modified to \( \log(\max |T|/|W|) \).

3. Trace of \( BW^{-1} \). This criterion may also be expressed in terms of the eigenvalues, \( \lambda_1, \lambda_2, \ldots, \lambda_p \), of \( BW^{-1} \), since \( \text{tr}(BW^{-1}) = \sum \lambda_i \).

All of these techniques attempt to minimize or maximize some criterion. However, not every partition of the data is evaluated, and consequently less-than-optimal solutions may result. Most algorithms are designed to search for local optima of the criterion by employing hill-climbing or hill-descending methods, depending on whether the criterion is maximized or minimized. In either case, the existing partitions are rearranged, and only those new arrangements that yield an improvement in the criterion are kept. The problem of local optima can be avoided with the use of a dynamic programming algorithm developed by Jensen (1969), which guarantees a global optimum.

Note that the use of certain of the proposed criteria restricts the shape of the clusters formed. For example, use of the \( \text{tr}(W) \) criterion will mean that the uncovered clusters will be only spherical, even if natural clusters of other shapes are more appropriate. The \( |W| \) does not restrict the clusters to being spherical, but it does assume that all clusters have the same shape. For this reason, Scott and Symons (1971) suggest that \( \prod_{i=1}^{h} |W_i|^{n_i} \), where \( n_i \) is the number of objects in cluster \( i \), be used.

4. Graphical Methods
Table 4.1 Percentage of people agreeing with eleven statements about eight brands of cereals (from Chakrapani and Ehrenberg, 1981)

<table>
<thead>
<tr>
<th>Statement</th>
<th>Corn Flakes</th>
<th>Weet Abix</th>
<th>Rice Krispies</th>
<th>Shredded Wheat</th>
<th>Sugar Puffs</th>
<th>Special K</th>
<th>Frosties</th>
<th>All Bran</th>
</tr>
</thead>
<tbody>
<tr>
<td>come back to taste nice</td>
<td>65</td>
<td>31</td>
<td>10</td>
<td>10</td>
<td>5</td>
<td>6</td>
<td>6</td>
<td>7</td>
</tr>
<tr>
<td>popular with all the family</td>
<td>64</td>
<td>40</td>
<td>32</td>
<td>23</td>
<td>29</td>
<td>17</td>
<td>22</td>
<td>11</td>
</tr>
<tr>
<td>very easy to digest</td>
<td>59</td>
<td>30</td>
<td>20</td>
<td>13</td>
<td>15</td>
<td>7</td>
<td>13</td>
<td>5</td>
</tr>
<tr>
<td>nourishing</td>
<td>60</td>
<td>42</td>
<td>24</td>
<td>18</td>
<td>20</td>
<td>2</td>
<td>19</td>
<td>17</td>
</tr>
<tr>
<td>natural flavor reasonably priced</td>
<td>47</td>
<td>39</td>
<td>11</td>
<td>28</td>
<td>6</td>
<td>15</td>
<td>5</td>
<td>18</td>
</tr>
<tr>
<td>a lot of food value stays crispy in milk</td>
<td>60</td>
<td>37</td>
<td>9</td>
<td>12</td>
<td>5</td>
<td>6</td>
<td>5</td>
<td>6</td>
</tr>
<tr>
<td>helps to keep you fit fun for children to eat</td>
<td>27</td>
<td>38</td>
<td>9</td>
<td>26</td>
<td>9</td>
<td>17</td>
<td>7</td>
<td>17</td>
</tr>
<tr>
<td></td>
<td>42</td>
<td>6</td>
<td>23</td>
<td>18</td>
<td>13</td>
<td>8</td>
<td>11</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td>24</td>
<td>28</td>
<td>10</td>
<td>21</td>
<td>9</td>
<td>29</td>
<td>9</td>
<td>40</td>
</tr>
<tr>
<td></td>
<td>17</td>
<td>12</td>
<td>57</td>
<td>5</td>
<td>50</td>
<td>5</td>
<td>43</td>
<td>0</td>
</tr>
</tbody>
</table>

Another set of techniques to classify objects or observations uses the graphical methods. Many graphical techniques have been proposed in the past. The most frequently used are glyphs and metroglyphs, Fourier series, and Chernoff faces.

To illustrate these techniques, we will use the data given in Table 4.1. These data are the percentage of people agreeing with eleven statements about eight brands of cereals in the United States.

Note that the usefulness of these techniques is greatly reduced when the number of observations becomes large.

4.1 Glyphs and Metroglyphs

This graphical method was developed by E. Anderson (1954, 1957, 1960). A glyph is a point (or circle of fixed radius) with rays emanating from it. The rays correspond to the characteristics or variables measured: the position of a ray labels the characteristic, while its length reflects its values. The angles between adjacent rays are equal, and the number of rays corresponds to the number of variables. This is illustrated in Figure 4.1. Usually, the star is completed by connecting the extremities of the rays to form a polygon.

The glyphs for the data in Table 4.1 are shown in Figure 4.2. One can see the connection between the Corn Flakes and the Weet Abix. The results for Frosties and Sugar Puffs are also quite similar. Both of these cereals are very popular with children.

![Figure 4.1 Illustration of glyphs.](image-url)
4.2 Fourier Series

This graphical technique was suggested by Andrews (1972). He proposed to transform a p-dimensional response vector (i.e., p observed variables) by Fourier series as:

\[ f_X(t) = \frac{X_1}{\sqrt{2}} + X_2 \sin t + X_3 \cos t + X_4 \sin 2t + X_5 \cos 2t + \cdots \]

(4.1)

over the range \(-\pi \leq t \leq \pi\).

Figure 4.3 presents the Fourier graph for the cereal attitude data of Table 4.1. On this plots, one can detect three groups or clusters. The first is formed by the Corn Flakes and the Weet Abix (continuous lines). A second cluster (dashed lines) contains Rice Krispies, Sugar Puffs and Frosties, the cereals that are popular with children. The last cluster (short-long dashed lines) contains the Shredded Wheat, the Special K, and the All Bran.

Andrews' representation has a number of useful properties that make it well suited for exploratory analysis. In particular:

1. The function \( f_X(\cdot) \) preserves means in the sense that if \( \bar{X} \) is the mean of a set of \( n \) multivariate observations \( X_i \), then \( f_X(t) = \frac{1}{n} \sum f_{X_i}(t) \).

Figure 4.3 Andrews Fourier plot for data in Table 4.1.
2. The function $f_x(.)$ preserves distances. The implication of this property is that close points in a Euclidean sense will appear as close functions, and distant points as distant functions. This may be very useful for detecting outliers.

3. For a given $t_0$, $f_x(t_0)$ is proportional to the length of the projection of the vector $(X_1, X_2, ..., X_p)$ on the vector $f_x(t_0) = (1/\sqrt{2}, \sin t_0, \cos t_0, \sin 2t_0, \cos 2t_0, ...)$.

4. The function $f_x(.)$ preserves variances.

4.3 Chernoff Faces

This graphical method was introduced by Chernoff (1973). In its original form, up to 18 dimensions in a response vector were allowed. Each dimension becomes associated with one of 18 facial features.

The Chernoff faces for the cereal attitude data of Table 4.1 are shown in Figure 4.4. The association between the statements in Table 4.1 and the face characteristics are given in Table 4.2.

The connection between the Corn Flakes and the Weet Abix is easy to see. One can also see how similar the results are for the Rice Krispies, the Sugar Puffs and the Frosties. Additional similarity is apparent in the results for Shredded Wheat, the Special K, and the All Bran.

5. Application to Atmospheric Chemistry: An Example

To illustrate the use of cluster analysis in atmospheric chemistry, we will present the results of different clustering techniques applied to one year (1980) of 925 mb three-day back-trajectories from Chalk River, Ontario. Note that there is four back-trajectories for each day. Figure 5.1 presents all the three-day back trajectories for the year. It is

Table 4.2 Association between statement and face characteristic for Chernoff faces in Figure 4.4.

<table>
<thead>
<tr>
<th>Statement</th>
<th>Face Characteristic</th>
</tr>
</thead>
<tbody>
<tr>
<td>come back to</td>
<td>Area of face</td>
</tr>
<tr>
<td>taste nice</td>
<td>shape of face</td>
</tr>
<tr>
<td>popular with all the family</td>
<td>length of nose</td>
</tr>
<tr>
<td>very easy to digest</td>
<td>location of mouth</td>
</tr>
<tr>
<td>nourishing</td>
<td>curve of smile</td>
</tr>
<tr>
<td>natural flavor</td>
<td>width of mouth</td>
</tr>
<tr>
<td>reasonably priced</td>
<td>location of eyes</td>
</tr>
<tr>
<td>a lot of food value</td>
<td>separation of eyes</td>
</tr>
<tr>
<td>stays crispy in milk</td>
<td>angle of eyes</td>
</tr>
<tr>
<td>helps to keep you fit</td>
<td>shape of eyes</td>
</tr>
<tr>
<td>fun for children to eat</td>
<td>width of eyes</td>
</tr>
</tbody>
</table>

Figure 4.4 Four Chernoff faces for the cereal attitude data in Table 4.1.
easily seen that the back-trajectories cover much of the area west of the site and only a smaller area east of the site.

The first cluster analysis was made using an agglomerative hierarchical method (Ward's error sum of squares technique). The merging of clusters was stopped when the back-trajectories were grouped into six clusters. The Euclidean metric was used to measure distances between the back-trajectories. A stereographic projection true at 60° North was used to project the back-trajectories before the analysis was made.

Figure 5.2 presents the positions of the mean back-trajectories for each cluster. The mean positions were drawn for every 6 hours for 3 days. The ellipse around each mean value for time $t_i$ would contain 68% of the corresponding (in time) positions of the back-trajectories within the cluster, if these positions followed a bi-normal distribution.

Figure 5.1 925mb three-day back-trajectories for Chalk River, Ontario, in 1980.

Figure 5.2 Mean back-trajectories for the six clusters obtained using the Ward's technique and Euclidean distances.
From these results, one sees that the clusters differ by either direction or speed of displacement of the air. For example, the air in clusters 2 and 4 came from about the same direction, but the air came from a greater distance in cluster 2 than in cluster 4. The same is
Figure 5.5 Three-day back-trajectories from Chalk River, Ontario, for the six clusters obtained using the K-Means technique and Euclidean distances.

true for clusters 1 and 6. The three-day back-trajectories for the six clusters are shown in Figure 5.3.

The next step in the analysis was to redo it using the K-mean method with

Figure 5.6 Mean back-trajectories for the six clusters obtained using the K-Means technique and angle distances.
Euclidean distances. As in the preceding case, the back-trajectories were projected before the analysis.

The results of the clustering are presented in Figures 5.4 and 5.5. Some of the clusters are similar to the ones obtained using the hierarchical technique. For example, Clusters 2 and 6 of Figure 5.5 are similar to clusters 1 and 4 of Figure 5.3 respectively. Other clusters are combinations of portions of clusters in Figures 5.2 and 5.4. For example, cluster 3 in Figures 5.4 and 5.5 is a mixture of part of cluster 2 and all of cluster 5 in Figures 5.2 and 5.3.

The preceding example illustrates the fact that two different methods produce different results even if the same distance metric is used. In the same way, the use of a different distance metric produces different results even if the same technique is used. This is illustrated in Figures 5.6 and 5.7. These figures present the result of clustering using the K-means method with the angle distance metric defined in Sirois and Bottenheim (1995). Using the angle between the back-trajectories as a distance metric eliminates the effect of the velocity of the air. This effect, as we have seen in Figures 5.2 and 5.3, results in the generation of clusters having the same direction but different air velocities. The angle metric can correct for this effect.

In Figures 5.6 and 5.7, the clusters point in different directions (in the mean). Some of the clusters are similar to the ones obtained by the other techniques. Some are somewhat different.

One important characteristic of all the results is that the back-trajectories would overlap at the edges of the clusters if the clusters were plotted together. This is a characteristic of the back-trajectory clusters.
6. Final Remarks

Lack of time and space have made it impossible to mention many subjects related to cluster analysis in this overview. The most important of these omitted subjects deals with methods for selecting the number of clusters that one should retain. However, the reader can gain fuller understanding of cluster analysis by consulting Anderberg (1973), Hartigan (1975), Späth (1980) and Everitt (1980). The first three of these books include Fortran programs to implement most of the techniques presented here. In addition, a short introduction to cluster analysis can be found in Dillon and Goldstein (1984).

Most statistical software packages include functions or programs for the majority of the techniques presented in this overview.

Finally, in cluster analysis, as in most other statistical fields, there is no technique or method that can be used blindly for all data sets. In the present instance, it is the usefulness of the clusters that determines whether a technique was appropriate. As in all statistical analyses, the user of cluster analysis must apply his or her judgment and not accept results without due consideration.
7. References


A Brief and Biased Overview of Time-Series Analysis
or
How to Find that Evasive Trend

Alain Sirois

Atmospheric Environment Service
Environment Canada
2121, Voie de Service Nord, Suite 500
Route Trans-canadienne,
Dorval, Quebec, Canada, H9P 1J3
Notice

This report is the background document for the presentation of the same name given at the WMO-EMEP Workshop on Advanced Statistical Methods and their Application to Air Quality Data Sets, Finnish Meteorological Institute, Helsinki, Finland, 14-18 September, 1998. The author could be reached at the following address: Dr. Alain Sirois (ARQI), Service de l'Environnement Atmosphérique, 2121, Voie de Service Nord, Suite 500, Route Trans-canadienne, Dorval, Quebec Canada, H9P 1J3. Tel: (514) 421-4674; Fax: (514) 421-2106; E-mail: Alain.Sirois@ec.gc.ca
Preface

The bias in the title of this presentation is a bias in favour of utility. In that spirit, the author has selected the techniques to be discussed here on the basis of two criteria:

- the technique has been used before by somebody;
- the technique might possibly be useful in the future.

In making selections, the author's agreement with any particular technique was not a consideration as long as that technique was mathematically correct.

Another factor that may have affected the selection process is the simple fact that the author does not know everything. Therefore, some techniques could have been omitted, not because they are incorrect or uninteresting but because they were unknown to the author. Unfortunately, the time when someone could everything know that could possibly be known about a subject is now entirely gone.

A final factor has been the need for brevity in both the presentation and the written paper. As a result, the author has been able to consider only a limited number of techniques.

The author has tried in this overview not only to give a catalogue of both past and present techniques for time series analysis but also to present an overall framework for the subject. Time series analysis involves more than just applying statistical techniques blankly. It should be extended to a complete analysis of the data. Therefore, the author will also try in this overview to indicate how such an analysis, in his opinion, should be done. The views expressed in this overview reflect only the personal and biased opinions of the author, and he is the first to admit any shortcomings in his approach. However, the author feels that his methods are at least comprehensive and provide a useful guide to time series analysis.

The overview consists first of a description of techniques that may help in exploring the main characteristics of a given time series. Then, different techniques to test for the existence of or to extract the long-term trends and/or other temporal variations present in the time series are discussed. The last, technical, part of the overview examines techniques that permit the verification of the assumptions made when applying any of the tests or other techniques. This last point is a very important aspect of time series analysis.

An example illustrating the use of some of the techniques presented here is given at the end. Note, however, that many other examples of particular techniques or types of plots will be given in the sections dealing with those topics. The purpose of the final example is to provide an integrated illustration of the author's approach to time series analysis.

Time for the preparation of this overview was limited. Although the author has tried as much as possible to make sure that the equations, examples, and descriptions are correct, errors could easily have crept in, especially in unifying the notations used by the different authors consulted. The author would therefore recommend that, before using the equations presented here, the reader verify in the references cited.
## Contents

Preface iii

Contents v

1. Introduction 1
   2. Exploring the Time Series 3
      2.1 Introduction 3
      2.2 Time-series Plots 4
      2.3 Missing Value Statistics 6
      2.4 Below-detection-limit Data 7
      2.5 Normal Probability Plots or Other Q-Q Plots 8
      2.6 Scatter Plots 10
      2.7 Spectral Analysis 11
      2.8 Smoothers and First Estimations 17
      2.9 Transforming the Data? 19
      2.10 Aggregating the Data into Longer Periods than the Original One? 21

3. Formulation of a General Time-series Model 23

4. Techniques of Time-series Analysis 25
   4.1 Introduction 25
   4.2 Nonparametric Techniques Based on Rank 25
      4.2.1 Introduction 25
      4.2.2 Testing for a Long-term Trend 26
         4.2.2.1 Mann-Kendall Test 26
         4.2.2.2 Farrell Test 27
         4.2.2.3 Seasonal Kendall Test 29
         4.2.2.4 Hirsch-Slack Test 31
         4.2.2.5 Dietz-Killen Test 32
         4.2.2.6 van Belle-Hughes’ Test 33
      4.2.3 Slope Estimation 35
         4.2.3.1 Sen’s Slope Estimator 35
         4.2.3.2 Seasonal Kendall’s Slope Estimator (Gilbert, 1987) 36
      4.2.4 Discussion 37
   4.3 Smoothers 37
      4.3.1 Introduction 37
      4.3.2 Running Mean Smoother 38
      4.3.3 Running Median Smoother 41
      4.3.4 Robust Locally Weighted Regression (Loess) Smoother 41
      4.3.5 Kernel Smoothers 44
      4.3.6 Smoothing Splines 50
      4.3.7 Periodic Smoothers 53
      4.3.8 Selecting the Smoothing Parameters 54
         4.3.8.1 Cross-validation 54
         4.3.8.2 Risk Estimation 55
      4.3.9 Discussion 56
1. Introduction

In atmospheric chemistry, an ever-increasing amount of data has been collected and archived in the past 20 to 30 years. Some of these data come from short-term field studies lasting from a week to a few months. Others, collected in the context of network monitoring over many years or decades, can be used to construct long time series. It is in this latter context that the question of temporal variation in time series arises.

The study of the temporal variation of time-series data is not only a matter of checking to see if a long-term trend is present and, if so, estimating it. There are many other characteristics of a time series that have to be analyzed. In addition to long-term variations, one has to consider long-term cycles, seasonal cycles, and others. The study of all these variations can add something to our knowledge of the data. It is therefore important to isolate and examine all these temporal variations in the time series under study. Another characteristic of time series data in atmospheric chemistry is that the data are not independent of one another. Consecutive samples are usually highly correlated. Some of the correlation can be explained by the presence of the long-term trend, long-term cycles, seasonal cycles, and other explainable variations. However, there usually remains a part of the between-samples correlation that cannot be explained by these factors. It is therefore important to study this remaining part also.

In statistical data analysis in general and in time series analysis in particular it is important to keep in mind two rules:

1. Know your data;
2. Know the statistical techniques you want to use.

The first of these rules implies a general knowledge of how the data were collected, their uncertainties, the presence of missing data, and so on. In addition, knowing the time series means that we have some idea of the kind of temporal variations that can be found in it. It is important to have some general idea of the types of temporal variations that are present in the time series before subjecting the data to any statistical analysis.

The second rule is related to the fact that there are hypotheses are made in all statistical techniques. Therefore, it is important to know these hypotheses and also to know as much as possible about the consequences of breaking them. These assumptions or hypotheses may be quite general in some cases and quite specific in others. For example, one statistical technique may require only that the data be from a continuous distribution. Other techniques, however, may require that the data be from a normal distribution.

Once one knows the characteristics of the data set that one wants to study and the hypotheses made by the statistical techniques that one wants to use, it is possible to estimate whether these hypotheses apply to this particular set of data. If such an analysis is not made, one risks being the victim of the old saying, “Garbage in → Garbage out”, without even knowing it. Any statistical technique, and time series analysis techniques in particular, should not be used as a black box where we put data through the In door and expect to get perfect results from the Out door. Such an approach produces generally random and sometimes embarrassing results and is not to be recommended.

We will also not recommend rejecting the results of a statistical analysis because the data do not completely respect the assumptions made by the techniques. The effects on the results of a statistical technique when data do not conform to its assumptions depend first on
the technique and secondly on the importance of the nonconformity. For example, the Student t-test is somewhat tolerant of data that may deviate from the normal distribution, but the F-test used in the analysis of variance technique is not nearly as tolerant. To take another example, a test that assumes that the data are independent may tolerate small correlations between the data but not medium or large ones. In such a case, it is important to know what is small for that test and what the correlation between the data is before being able to decide if the results of the test can be used. In summary, the user of statistical techniques has to judge the validity of the study. The only way, he or she can do that is to know the data and to understand the techniques that he or she is using.

This overview has been written to help the reader accomplish these objectives. It is based on a general framework developed by the author over the past 15 years, during which he has spent much of his time studying time series. This framework can be summarized by the following five procedural steps:

1. Explore the data.
2. Formulate a general conceptual time series model.
3. Select an appropriate statistical technique(s).
4. Applying the technique(s) to the data.
5. Verify the hypothesis made by the statistical technique(s).

We will discuss and describe these steps in the remaining sections of this overview. Note that steps 4 and 5 are related and will be discussed together in Section 4.

The data that constitute a time series may have been collected in many different ways. They may represent samples of the same length of time or not. They may have been collected at a regular frequency or not. In the present overview, we will only consider data that represent, in theory, sampling of the same length, although a small fraction of the data may differ slightly from that ideal. The data may or may not be collected with equal frequency. Examples of the type of data set that are considered here are: (1) hourly O₃ measurements, (2) 24-hours filter pack measurements of ambient air, and (3) daily sampling of precipitation. This last is an example of equal length samples that are not collected at a regular frequency.
2. Exploring the Time Series

2.1 Introduction

The first question that the reader might ask at this stage is: "What should I know about my data?" There is no universal answer to that question, but in the context of time series analysis any answer should address the following points:

- Are there any missing data? If yes, what fraction of the data is missing? Is there a pattern to the missing data?
- Are all or part of the data missing because they were collected irregularly? Precipitation data are an example of irregularly collected data.
- Are there any outliers? If yes, are they many or few? Are they very different from the other data?
- Are there any below-detection-limit values? If yes, how should we treat them?
- What seems to be the distribution of the data? Is it normal or log-normal? Is there any transformation that will help in normalizing the data?
- Is there heteroscedasticity in the data (i.e., does the variance of the data changes with time)? If yes, could the data be transformed to stabilize the variance? Usually the same transformation will help in normalizing the data and stabilizing the variance.
- Is there a long-term trend in the time series? If yes, is it decreasing monotonically, increasing monotonically, or does it exhibit some other pattern?
- Are there other variations (e.g., long-term cycles) with long periods?
- If the time series extends over more than a year, are there some seasonal cycles? If yes, are they characterized by a simple sine or cosine function or by more complex ones?
- If the data are sampled at periods shorter than a day, are there some diurnal cycles?
- Should I take into account the relationship of the studied variables with each other before looking for temporal variations? An example of such a relationship is that between ionic concentration in precipitation and the amount of precipitation.

This list of questions is not exhaustive, and the reader could easily add several other pertinent questions to it.

We will try in the remainder of this section to present and discuss some of the techniques that may help to answer these questions. The techniques presented here are ones that the author has found useful in the past to get a better understanding or "feeling" about the time series that he was trying to analyze. They are presented here in the hope that they may also be useful to the reader.
2.2 Time-series plots

This is the most basic type of plot. An example of plotting the data against time is presented in Figure 2.1. In it, the SO₂ ambient concentration at Chalk River, Ontario, is shown between 1979 and 1994 without any transformation of the data (see Appendix I for a map showing the location of the CAPMoN sites). Although interesting, this plot is not very informative. All that can be derived from this plot is that: (1) there are some missing data; (2) there seems to be some seasonal variation; and (3) the data are mainly concentrated between 0 and about 15 μg m⁻³ with some very high values present. From this last characteristic, one can
arrive at the conclusion that the data are not normally distributed. A more informative way to plot these data is to use a logarithmic scale for the ordinate axis.

The result is illustrated in Figure 2.2. One gets a better idea of the data in this figure than in the preceding. One can see first that most of the data in the second half of 1988 are missing. This is the only long period with missing data. Secondly, there are seasonal cycles in the data. An idea of these cycles can be had by creating a plot like the one presented in Figure 2.3. In this plot, all the years have been merged into one, and well-defined seasonal cycles are evident.

A long-term trend is also evident in Figure 2.2. SO₂ ambient air concentration has decreased during the period. This decrease seems to be

Figure 2.3 Seasonal variation of ambient SO₂ concentration in air at Chalk River, Ontario, using a logarithmic axis.

Figure 2.4 Box-plots for ambient SO₂ concentrations in air at Chalk River, Ontario, using a logarithmic axis and grouped by year. In each box, the central bar is the median and the lower and upper limits are the first and third quartiles respectively. The lines extending vertically from the box indicate the spread of the distribution, with the length being 1.5 times the difference between the first and third quartiles. Observations falling beyond the limits of these lines are considered to be outliers and are indicated by dots.
monotonic and more important between 1979 and 1985 than after. This example illustrates that it is not enough to show just a time series of the data. One must also respect the distribution of the data if one wants to get useful information from them.

Another type of graph that may be useful in detecting the presence of a long-term trend is a box-plot, as shown in Figure 2.4. In this figure, the data are grouped by year and a long-term trend is evident. The corresponding figure for seasonal variation is presented in Figure 2.5. These plots are very useful for detecting a specific type of temporal variation. Similar plots could be done also for diurnal data, such as hourly O$_3$ mixing ratios.

2.3 Missing Value Statistics

As mentioned in the introduction, it is important before doing any time series analysis to check first for missing values, and, if they are present, to consider their number and their distribution across the time series.

Missing values in a time series may result from many different causes. On the one hand, they may be due to the fact that the data are not sampled regularly. This would be the case, for example, for ionic concentration in precipitation. Precipitation does not occur every day, and a period of a few days with precipitation can be followed by a few days without it. These dry days will appear in the data bank as a data gap. On the other hand, a short data gap may be due to a broken instrument, for example. In the case of monitoring that extends over many years, much longer data gaps may occur for a variety of reasons.

An interesting way to summarize information about missing values is illustrated in Figures 2.6 and 2.7. For these plots, the length of each missing period was calculated. The plots present the number of such periods in the time series. Figure 2.6 is for ambient SO$_2$ concentrations in air at Chalk River, Ontario. Figure 2.7 is for the ionic concentrations of SO$_4^{2-}$ in precipitation at the same site.
One notices first the large difference in the number of missing values between the two time series. They represent less than 2.5 % of the data in the case of ambient \( \text{SO}_2 \) air concentrations but about 20% of the data in the case of ionic concentrations of \( \text{SO}_4^{2-} \) in precipitation. The shapes of the histograms are also quite different. For the \( \text{SO}_2 \) concentration in air, most of the missing periods are one day long and the number of cases decreases rapidly for longer periods. For \( \text{SO}_4^{2-} \) concentration in precipitation, the decrease in the number of cases is more gradual.

Plots like Figures 2.6 and 2.7 help in getting a better understanding of the pattern of the missing data. This also helps in estimating their impacts on the results.

2.4 Below-detection-limit Data

In some data sets, the number of below-detection-limit values may cause some problems. In the techniques presented in this overview, there is no method for taking below-detection-limit data into consideration. If the detection limit is not known, these items are usually treated as missing values, although it might be more correct to replace them by a

![Figure 2.7](image_url) Number of periods of missing data of different length for \( \text{SO}_4^{2-} \) concentrations in precipitation at Chalk River, Ontario.

![Figure 2.8](image_url) Kernel smoother estimate of the long-term trend for weekly mean \( \text{H}^+ \) concentrations at Alert, Canada.
small value. (Note that the use of 0 may create problems later in the analysis. The author personally uses the value corresponding to the number of digits reported; for example, he will use a value of 0.01 for data reported with two decimals after the decimal point.) If the detection limit is reported, one usually uses 1/2 or 2/3 of the detection limit. If the former fraction is used, it is assumed that data that are lower than the detection limit follow a rectangular distribution. If the latter fraction is used, they are assumed to follow a triangular distribution, with the hypotenuse going from zero to the detection limit. The author prefers using the second value.

The impact of below-detection-limit data on the results of any time series depends on the number and temporal distribution of such data. If they are equally spread across the time series, their main impact will be to shift any long-term trend downwards, although they will usually have no impact on the shape of the trend. If, in addition, they occur mainly in one season, they may influence the seasonal cycle.

Examples of the impact of a large number of below-detection-limit data on the long-term trend and on the seasonal cycle are given in Figures 2.8 and 2.9. The kernel smoother (see Section 4.3.5) was used to estimate the long-term trend and the seasonal cycle. The data are weekly high volume samples collected at Alert, Canada (see Sirois and Barrie, 1999). As can be seen in the two figures, the below-detection-limit values are well distributed across the years but occur mainly in summer. In Figure 2.8, although the shape of the long-term trend seems to be correct, the curve looks as if it was shifted downward. Its mean should be about 8-9 ngm$^{-3}$ and not about 2 ngm$^{-3}$. The impact of the presence of the below-detection-limit data is greatest in the estimation of the seasonal cycle (Figure 2.9) because most of the below-detection-limit data occur during the summer months, thus influencing mainly the estimated seasonal cycle during those months.

When the number of below-detection-limit values is large, it is important to estimate their impact on the results of the time series analysis. This can be done by redoing this analysis with different values for the below-detection-limit data. This technique usually gives some idea of the impact of the below-detection-limit data.

2.5 Normal Probability Plots or Other Q-Q Plots

Normal probability plots have been used for a long time to verify the possible normality of the distribution of data. Normal probability plots are only one example of a class of graphs called Q-Q plots (see Wilk and Gnanadesikan, 1968; Michael, 1983). The name indicates that the quartiles of one distribution is plotted against the quartiles of another. In

![Figure 2.9 Periodic kernel smoother estimate of the seasonal cycle for weekly mean H$^+$ concentrations at Alert, Canada.](image)
normal probability plots, the quartiles of the observed data are drawn against the quartiles of the normal distribution. That concept may be extended to other types of distributions.

In the context of time series analysis, these types of graphs can be used first to see if the data should be transformed before the analysis and also to test, in the case of some techniques, whether one or more of their hypotheses are valid for the data used.

An example of a normal probability plot is given in Figure 2.10. To create this graph, the \( n \) concentration observations, \( e_1, e_2, e_3, \ldots, e_n \) are first ordered in increasing order to produce the order statistics, \( C_{[1]}, C_{[2]}, C_{[3]}, \ldots, C_{[n]} \). Then the \( C_{[i]} \) (ordinate) are plotted against \( N^{-1}((i-1)/2) \) (abscissa) where \( N^{-1}(\cdot) \) is the quartile of the normal distribution. In Figure 2.10, although the quartiles are used, the percentile values are indicated to simplify the interpretation. If the data were normally distributed, they would form a straight line, which is not the case for the present data set.

A better fit to a straight line can be obtained if a log-normal distribution is assumed. One should note that for this example (Figure 2.11) the fit is not yet perfect but is better than if a normal distribution is assumed.

This type of graphs can easily be extended to other distributions, such as Gamma or Weibull distributions.

These plots may also point to the necessity of transforming the data before plotting or analyzing them. They may also help in determining the type of transformation to use. This is illustrated in Figure 2.12. The data, after being transformed by means of a power transformation, were plotted on normal probability paper. One sees that, except for

---

**Figure 2.10** Normal probability plot for the ambient SO\(_2\) concentrations in air at Chalk River, Ontario.

**Figure 2.11** Log-normal probability plot for the ambient SO\(_2\) concentrations in air at Chalk River, Ontario.
for the lowest 0.5%, the data come quite close to falling along a straight line if they are raised either to the one fourth or one fifth power.

One should note that it may be best to use these transformations before applying the model, because part of the non-normality in the data may be due to the presence of a long-term trends, a seasonal cycles, or other such variations.

2.6 Scatter Plots

We have mentioned earlier that the time series that we want to analyze may depend on another time series. An example in atmospheric chemistry is the correlation between the ionic concentration in precipitation and the precipitation amount. As the precipitation amounts may have its own temporal variations that may be due to other factors than the ones we want to detect in the ionic concentrations, it is important to take that dependence into consideration. How we take that information into account depends on the technique we employ to analyze the time series.

Figure 2.13 Scattergram of NO₃⁻ concentrations in precipitation versus precipitation amount at Chalk River, Ontario between 1979 and 1995.
The dependence of the studied time series on others can be checked by calculating the regression coefficients or by plotting scattergrams. This last technique is illustrated in the case of NO$_3^-$ concentration in precipitation at Chalk River, Ontario (Figure 2.13). This figure shows that the concentrations decrease as the precipitation amount increases. This should be taken into account when studying the temporal variation of NO$_3^-$ concentration in precipitation.

2.7 Spectral Analysis

Until now we have only looked at the general characteristics of the time series, discussing, for example, how to deal with missing data or below-detection-limit data. Before
testing for or estimating some of the temporal variations present in the data, we would like to have some idea of which types of temporal variation or which characteristics of the time series we should take into account. In other words, we would like to know what type of temporal variations are present in the time series. A very powerful tool for answering that question is spectral analysis.

Programs that estimate the power spectrum of a time series can be found in any statistical computer package. However, these programs usually assume that there are no missing data in the time series, which is not often the case in atmospheric chemistry. Two techniques exist to solve that problem. The first technique is based on the technique developed by Dunsmuir (1984) to estimate the auto-correlation function in the presence of missing data. The theory is described in Sirois et al. (1995), and the necessary equations are given in Appendix II. In the second technique, the power is calculated directly. This technique was developed by Lomb (1976) and based on earlier work by Barning (1963) and Vanicek (1971). The equations for the power spectrum are given in Appendix III. This second technique may be more appropriate in the case of ionic concentrations, in precipitation, for example.

The basic relation behind the interpretation of the power spectrum is Parseval’s relation (see Priestley, 1981) which can be written as:

\[ \sigma^2 = 2\pi \int_0^{0.5} f(\omega) \, d\omega \]  

(2.1)

where \( \sigma^2 \) is the variance of the data, \( f(\omega) \) is the power spectrum, and 0.5 is the Nyquist limit. Parseval’s relation indicates that the power spectrum is, in fact, the decomposition of the total variance into frequency dependent components.

An example of a power spectrum is given in Figure 2.14 for ambient \( \text{SO}_4^{2-} \) air concentrations at Kejimkujik, Nova Scotia, from 1979 to 1994. Note that the data were transformed using a logarithmic transformation before the power spectrum was calculated. Although relation 2.1 can be used directly, this plot is difficult to interpret. It is usually more
informative to plot the logarithm of the power spectrum against the logarithm of the frequency. This is illustrated in Figure 2.15. This latter graph is easier to interpret than the preceding one. The reader should remember that Parseval’s relation cannot be applied directly in this case. In this plot, the frequency of the seasonal cycle is about 0.0027, that of the first harmonic is about 0.0055, and that of the third harmonic, about 0.0082.

Instead of trying to deduce all the types of temporal variations that can be present in the time series from Figure 2.15, we will try to show the impact on the power spectrum of adding different types of temporal variations, starting with random noise and adding autocorrelation, a seasonal cycle, a long-term cycle, and a long-term trend in that order. To do that we will create an artificial 15-year time series of daily data.
Figure 2.18 Power spectra for time series with seasonal cycles and auto-regressive model of order 1 with $a_1 = 0.5$. Plots (a) and (c) incorporate one cycle. Plots (b) and (d) incorporate one cycle and one harmonic.

Figure 2.16 presents the power spectrum of normally distributed random errors, or white noise. In this figure, one sees that the power spectrum of random noise consists of random variation around a straight line. If one compares Figure 2.16 to Figures 2.14 and 2.15, one sees that the ambient $\text{SO}_4^{2-}$ air concentrations at Kejimkujik do not constitute a random time series.

It is well known in atmospheric chemistry that consecutive data are usually not independent. This process can be mathematically described by the following auto-regressive model:
where $a_i$ are coefficients, $K$ is the order of the auto-regressive model, and $\varepsilon_t$ is a random error. The power spectrum of a time series that follows such a model depends on the order of the auto-regressive model and on the values of the coefficients $a_i$. In atmospheric chemistry, auto-regressive models of order 1 to 3 are usually found. Typically, the first auto-regressive coefficient is much larger than the other ones. Two examples of power spectra for an auto-regressive model of order 1 are shown in Figure 2.17.

In Figures 2.17a and 2.17c, $a_1$ is equal at 0.25, and in Figures 2.17b and 2.17d it is equal to 0.50. If one compares these figures to Figure 2.16, one can see that the presence of the auto-regressive model has modified the power spectrum and that the modification is more important when the auto-regressive coefficient is largest. The modifications result in a transfer of the variance or energy from the highest frequencies to the lowest. Note that the effect will be opposite if the auto-regressive coefficient is negative. When the order of the auto-regressive model is greater than 1, its effects on the power spectrum are more complex. However if the first coefficient is positive and much larger than the others, the effects will be similar to the results shown in Figure 2.17. From a comparison of Figure 2.17 with Figure 2.15, one can conclude that auto-correlations are present in the data. The first coefficient of the auto-regressive model will be positive and much larger than any other coefficients present.

The next step is to add a seasonal cycle. Such a cycle can usually be described by a relation such as:

\[ C_t = \sum_{i=1}^{K} a_i C_{t-i} + \varepsilon_t \]  (2.2)

Figure 2.19: Smoothed power spectra for an artificial time series containing a decreasing linear trend, seasonal cycles, an one-order auto-regression model, and normally distributed white noise. The plots correspond to different values for the slope of the linear trend.
Figure 2.20 Artificial time series corresponding to the power spectrum of Figure 2.19a.

\[ \sum_{m=1}^{M_s} \left( C_{sm} \sin \left( \frac{2 \pi m t}{365.25} \right) + C_{sm} \cos \left( \frac{2 \pi m t}{365.25} \right) \right) \]  

(2.3)

where \( M_s \) is the number of waves. If \( M_s \) is greater than 1, harmonics are present.

Figure 2.18 shows the power spectrum when seasonal cycles are added to the auto-regressive model of order 1 with \( a_1 = 0.50 \). In Figures 2.18a and 2.18c, a simple cosine wave was used to describe the seasonal cycle. In the remaining plots of the figure, the seasonal cycle was described by a cosine wave plus a sine wave with a 6-months period and half the amplitude. Figure 2.18 shows that it is more and more difficult to interpret the power spectrum if one does not plot the logarithm of the smoothed power against the logarithm of the frequency. The presence of the seasonal cycles is very clear in these plots. If now we return to Figure 2.15, it is easy for the reader to detect the presence of a seasonal cycle with one main cycle and two harmonics.

To complete our study, we will now add a decreasing linear trend to the log-transformed data. The power spectra for two linear decreasing trends are given in Figure 2.19 for two values of the slope. The slope in Figure 2.19b is twice that in Figure 2.19a. The main effect of adding the trend is to increase the energy or variance at frequencies lower than about 0.001 cycles day\(^{-1}\).

A comparison of the power spectra in Figure 2.19 with the spectrum in Figure 2.15 shows that all the spectra are very similar. Therefore, using a simple model (linear long-term trend, plus a seasonal cycle, plus a one-order auto-regressive model, plus normally distributed white noise), we have reproduced the general lines of the power spectrum of the observed data.

In practice, we have to do the inverse of what we have done here, namely, start from the power spectrum and deduce the main types of temporal variations that contribute to the
variance observed in the data. This may seem a difficult proposition for a beginner, but with time one becomes adept at interpreting these spectra. The author hopes that the short demonstration presented here may help the reader in acquiring a better understanding of power spectra.

Although the power spectra produced in this section appear realistic, readers may ask whether the artificial time series from which they were realistically simulates observed ambient air concentrations. The time series corresponding to the power spectrum of Figure 2.19a is shown in Figure 2.20. The line is the kernel smoother estimation (see Section 4.3.5) of the long-term trend. One notices first that the variability may be somewhat too large when compared with most 24-hour ambient air sampling data. This is due to the fact that too large a value was selected for the variance of the white noise. This does not affect the fact that the time series appears to be as complex as an observed one, although it is only the sum of a few different types of temporal variations.

We have shown in this section that spectral analysis can be effective in detecting the presence of different types of temporal variation. However, it does not help us in getting an estimate of the shape or form of these temporal variations. We will address that problem in the next section.

2.8 Smoothers and First Estimations

Although spectral analysis can help in determining which types of temporal variations, if any, are present in a time series, it cannot help in determining their characteristics. One tool that can give a first idea of those characteristics is smoothing. Different smoothers are available (see Section 4.3). As an example, we will use the kernel smoother (Section 4.3.5; Sirois, 1993, 1997) and the daily ambient SO\textsubscript{2} concentration at the Experimental Lakes Area in Ontario.

The power spectrum is shown in Figure 2.21. We can deduce from it that a strong long-term trend is present. A seasonal cycle with a first harmonic at least is present. There are also some indications of auto-correlations between the data.

Figure 2.22 presents the kernel smoother estimation of the long-term trend. One sees that the concentrations decrease mainly between 1979 and 1985. They may also increase again slightly between 1987 and 1991. The SO\textsubscript{2} concentrations appear to have decreased after 1991. It is interesting to note that even the missing data for 1983 do not prevent the kernel smoother from giving a good approximation of the long-term trend. In summary, the SO\textsubscript{2}
concentrations at the Experimental Lakes Area have decreased almost monotonically since 1979 but not linearly.

After eliminating the long-term trend, one can estimate the seasonal cycle. Figure 2.23 presents the periodic kernel smoother estimate of the seasonal cycle. One can see that the seasonal cycle is more complex than a simple sinusoidal wave.

After eliminating the estimated seasonal cycle, we would like to test for autocorrelation. One easy visual way to do this is to plot the scattergram of $C_t$ versus $C_{t-1}$; $C_t$ versus $C_{t-2}$, etc. Some of these graphs are shown in Figure 2.24.

One sees a net relationship between the concentrations at time $t$ and time $t-1$. The relationship becomes less clear in the case of the concentrations at time $t$ and time $t-2$, and at time $t$ and time $t-3$. One should therefore try to fit the auto-regressive model given in relation 2.2. There are many ways to fit such a model. However, because there are missing data in the time series, one has to use a technique that will allow the presence of missing data. The easiest way is to use the regression technique (see Chatfield, 1984; Botteheim and Sirois, 1996). In that technique, one regresses the concentration at time $t$ against the concentration at time $t-1$, $t-2$, etc. For

---

**Figure 2.22** Kernel smoother estimate of long-term trend for daily ambient SO₂ concentrations at the Experimental Lakes Area, Ontario.

**Figure 2.23** Periodic kernel smoother estimate of seasonal cycles for daily ambient SO₂ concentrations at the Experimental Lakes Area, Ontario.
the present data, the results indicate that an auto-regressive model of order 2 is applicable, with $a_1 = 0.43$ and $a_2 = 0.06$. One should keep in mind that, as in the other results presented in this section, the order of the auto-regressive model and the coefficients obtained are only first approximations.

The results presented in this section are very useful for the remainder of the study. However, they cannot be taken as definitive because (1) some of the hypotheses assumed for kernel smoothing are not usually fulfilled, and (2) there is no way to check the statistical significance of the results. We will discuss these points later in Section 4.4.

Those results are also very useful for the next step in the statistical study of temporal variations, namely, the formulation of a conceptual general time series model to describe the time series one wants to study. Before discussing that step, however, we will discuss two other important subjects: (1) the possible transformation of the data and (2) their aggregation to form longer sampling periods.

### 2.9 Transforming the Data?

Should I transform the data? This question is sometimes very difficult to answer. The answer is usually related to the technique that one wants to use to study the data. In the case of nonparametric techniques (Section 4.2), transformation is not usually necessary. However, it may be necessary to transform the data, not to get them to normality, but to make their variance independent of time, since one of the usual assumptions of nonparametric techniques is that the variance is independent of time. Therefore, most of the time the data collected in atmospheric chemistry will have to be transformed because they are log-normally distributed, which means that the

![Figure 2.24 Scattergrams of de-trend and deseasonal SO$_2$ concentrations at time $t$ and $t-1$ (a), $t-2$ (b), and $t-3$ (c) for daily ambient SO$_2$ concentrations at the Experimental Lakes Area, Ontario.](image-url)
variance is linearly related to the concentration (see Sirois, 1991). A log-transformation of the data eliminates that relationship. The same is true for techniques such as smoothing (Section 4.3). The data may have to be transformed to get them close to normality if techniques like the anova techniques (Section 4.5) and the maximum likelihood technique (Section 4.6) are used. Finally, it may be necessary in some cases to transform the data to make their distribution symmetrical.

In most cases, a simple power transformation will be enough to transform the data to normality and/or stabilize the variance, and/or get a symmetrical distribution. Such a transformation can be expressed as:

$$T_p(x) = \begin{cases} x^p & (p > 0) \\ \log(x) & (p = 0) \\ -x^p & (p < 0) \end{cases}$$

The most frequently used values are $p = 0$ and $p = 1/2$. Power transformations are discussed in Emerson and Stoto (1983) and Emerson (1983).

In the case of logarithmic transformations, it is sometimes necessary to use a shifting parameter. The transformation then becomes:

$$T(x) = \log(x + \delta)$$

Techniques to estimate $\delta$ are given in Aitchison and Brown (1957), Crow and Shimizu (1988), and Wheeler (1980).

Another system of transformation to normality is the Johnson System (Johnson, 1949). This system consists of three families of distributions, which are generated by a transformation of the form:

$$T(x) = \gamma + \eta k_i(x; \lambda, \epsilon)$$

where $T(x)$ is a standard normal variable, and the $k_i$ are selected to cover a wide range of shapes. The three families suggested by Johnson (1949) are:

1) the $S_U$ distribution with $k_1(x; \lambda, \epsilon) = \sinh^{-1}\left(\frac{x - \epsilon}{\lambda}\right)$

2) the $S_B$ distribution with $k_2(x; \lambda, \epsilon) = \ln\left(\frac{x - \epsilon}{\lambda + \epsilon - x}\right)$

3) the $S_L$ distribution with $k_3(x; \lambda, \epsilon) = \ln\left(\frac{x - \epsilon}{\lambda}\right)$

The $S_3$ distribution is, in essence, a three-parameter log-normal distribution, since the parameter $\lambda$ can be eliminated by letting $\gamma^* = \gamma - \eta \ln \lambda$, so that $T(x) = \gamma^* + \eta \ln(\lambda - \epsilon)$. The $S_B$ is a distribution bounded on $(\epsilon, \epsilon + \lambda)$, and the $S_U$ is an unbounded distribution. The first step in using Johnson's system of transformation is to select which of the three families to use. Then, the parameters of the family can be estimated. More details about Johnson's system of transformations can be found in Johnson (1949, 1965), Draper (1952), Bukac (1972), Mage (1980), Shifker and Shapiro (1980), and Wheeler (1980). Fortran programs to fit Johnson's system can be found in Hill et al. (1976) and Hill (1976).
In summary, the important thing to remember is that one transforms the data not necessarily to normalize them but also to stabilize the variance or render the distribution symmetrical. Therefore, one sometimes has to transform the data, even if one wants to use nonparametric techniques to analyze the temporal variation in the data.

2.10 Aggregating the Data into Longer Periods than the Original One?

In the past, researchers often aggregated data to a longer period before analyzing them. This was sometimes done because one can usually get better normal approximations with longer sampling periods. Another reason was that one could reduce or eliminate the effects of missing data. Frequently, one would calculate aggregated data to eliminate autocorrelation and therefore simplify the analysis. However, when data gaps were present, this became a very complex problem because: (1) the aggregation technique could introduce a bias into the aggregated data estimate, and (2) the increase in the variance of the aggregated estimate would usually depend on parameters like the percentage of data missing. Therefore, the variance of the data would not be constant in time except if the number of data items missing was also constant in time.

As an example, Sirois (1990) has shown that, when data are missing, estimates of the monthly, seasonal, or annual precipitation-weighted-mean concentration of ions in precipitation are biased, and that the magnitude of the bias depends on the percentage of data missing and the percentage of the total precipitation associated with missing data. He has also shown that the variance of the estimation depends on those two parameters as well.

In summary, although it may seem that aggregating data to longer sampling periods will solve some of the difficulties usually associated with the analysis of temporal variation in atmospheric chemistry, the procedure introduces characteristics into the data (i.e., biases and variability that depend on parameters that vary with time) that invalidate most of the techniques presented in the present overview. Therefore, the author advises against aggregating the data, except if the amount of missing data is very small.
3. Formulation of a General Time-series Model

This step in the temporal analysis of a time series is a very important one for many reasons. First, this model embodies our knowledge of the characteristics of the time series. Secondly, this model also delimits the techniques that can be used to analyze the time series. In practice, the conceptual model and the technique to be used are selected at the same time, because behind all the techniques presented in the next section there is some conceptual model of the time series.

This conceptual model may be very simple or very complex; very general or very specific. An example of a very simple conceptual model that would apply to most of the smoothing techniques presented in Section 4.3 is the following:

\[ C_i = f(t_i) + \varepsilon_i \]  

where \( f(t) \) is a continuous function and the \( \varepsilon \) values are random noises from the same distribution with mean zero and constant variance. In most cases also, the existence of the first and second derivatives of \( f(t) \) is assumed.

Model 3.1 also applies in the case of nonparametric techniques if we add the condition that \( f(t) \) should be monotonic. This can be written as:

\[ f(t_i) \geq f(t_j) \quad \text{if} \quad t_i > t_j \]  

for a monotonically increasing function and

\[ f(t_i) \leq f(t_j) \quad \text{if} \quad t_i > t_j \]  

for a monotonically decreasing function.

For some of the techniques like the anova technique, model 3.1 is valid but the distribution of \( \varepsilon \) is assumed to be normal.

An example of a more complex conceptual model is one used by Sirois (1997). It can be written as:

\[
\log(C(t_i)) = C_0 + f^T(t_i) + f^C(t_i) + f^S(t_i) + f^D(t_i)
\]  

where \( C_0 \) is a constant; \( f^T, f^C, f^S \) are general functions that describe the long-term trend, the long-term cycles, and the seasonal cycles respectively; and \( f^D \) is an auto-regressive model written as:

\[
f^D(t_i) = a_1f^D(t_{i-1}) + a_2f^D(t_{i-2}) + \ldots + a_Nf^D(t_{i-N}) + \varepsilon_i
\]

where \( N \) is the order of the auto-regressive model; \( \varepsilon_i \) is the random noise component drawn from a normal distribution with mean zero and constant variance \( \sigma^2 \). This type of model is useful for techniques like those discussed in Section 4.6.

The important notion to keep in mind is that one should have a conceptual model and that it should be consistent with the technique used. It should also be consistent with what it is known about the data and especially with the results obtained in the preceding section. If the conceptual model do not describe the time series properly, there is a good chance that the method one wants to use does not apply to the data. One should always keep in mind that the results of any statistical analysis are valid only if the assumptions related to that technique are fulfilled by the data. As mentioned earlier, small deviations from these conditions may not
always invalidate the results, but they will increase the uncertainties, depending on the importance of the deviations.

Formulating the conceptual model also helps in the selection of the correct statistical techniques to use. One usually starts by extracting the main characteristics of the time series, employing the techniques described in the preceding section. A conceptual model is then constructed, and finally one or more of the statistical techniques compatible with the conceptual model is selected. *The reader should always remember that the data should drive the construction of the conceptual model and not the wishes of the researcher. We should always bear in mind that we do not control nature yet.*
4. Techniques of Time-series Analysis

4.1 Introduction

Many techniques have been developed during the past century or so to analyze the temporal variations of series of numbers or time series. These techniques range from the simple fitting of a straight line by least squares to a complex ARIMA (Auto-Regressive Integrated Moving-Average) model. In each of these techniques some hypothesis is made about the nature of the data and the characteristics of their temporal variations. These hypothesis can usually be summarized into a conceptual model of the data to which the technique applies. As we have seen in the preceding section, some of these models may be simple and others more complex. In either cases, the observed data are the important factor that determines the correct selection of the conceptual model and thus of the technique used to analyze them.

In this section, we will describe some of the techniques that have been used in atmospheric chemistry to analyze the temporal variation of different observed pollutants. For each of these techniques, we will not only describe how to apply them but will also explain the hypotheses about the data and the conceptual models that are implied when using them. We will also point out their limitations, if any.

The techniques have been grouped according to their main characteristics. We will first describe nonparametric techniques based on rank. We will then discuss smoothers that are nonparametric but that are associated with a somewhat different conceptual model. Next, we will mention the use of spectral analysis to test and estimate some of the temporal variations. We will finish by examining techniques that require the construction of more hypotheses but that are also more general in their application.

The present overview, although not exhaustive, should give a good idea of the techniques available for the analysis of temporal variation of observed data.

4.2 Nonparametric techniques based on rank

4.2.1 Introduction

The techniques presented in this section use basically the same conceptual model, which can be written as:

\[ C_i = f(t_i) + \varepsilon_i \]  

(4.1)

It is assumed that \( f(t) \) is a continuous monotonic increasing or decreasing function. The \( \varepsilon_i \) values are assumed to be from the same distribution with zero mean. It is therefore assumed that the dispersion (i.e., variance) of the distribution is constant in time.

In some of the techniques, the data are grouped to form a two-way or even three-way layout for the regional version of the tests. Usually, one of the layout indices will be the year; another will be either the week, month, or season; and in the cases of regional tests, the third will be the site. In these techniques, it is usually assumed that there is one observation for each layout cell, although it is possible in some cases to extend the technique to include more
than one observation per cell. We will indicate those cases and the changes that must be made to the equations.

Two types of analysis can be carried out using nonparametric techniques. First, one may test for the presence of a monotonic increasing or decreasing long-term trend, and secondly one may estimate the slope of a linear trend. We will address these two applications separately.

4.2.2 Testing for a Long-term Trend

Many nonparametric statistical tests exist. Since we cannot report all of them, we decided to limit ourselves to those that are most important or useful in atmospheric chemistry. Six tests are described here. Except for the first, all try to take the presence of seasonality into account, and assume monotonic trends for each season. Unless otherwise mentioned, the trends are assumed to be homogeneous across the seasons. The random errors are assumed to be independent, except in the case of the Hirsch-Slack test.

4.2.2.1 Mann-Kendall Test

Let be a sequence of observations, \( x_1, x_2, \ldots, x_n \), ordered by time. One wants to test the following hypothesis:

\[ H_0: \text{the observations are randomly ordered, i.e., } x_1, x_2, \ldots, x_n \text{ are} \]
\[ \text{independent identically distributed (or i.i.d.) samples,} \]

against the hypothesis:

\[ H_1: \text{there is a monotone trend over time, i.e., } F_{x_i}(x) \geq (\text{or} \leq) F_{x_j}(x) \text{ for all } i < j \text{ with at least one strict inequality,} \]

where \( F_{x_i}(x) \) is the cumulative distribution function of the random vector \( x_i \). Using conceptual model 4.1, the test can be described as \( H_0: f(t) = \text{constant against } H_1: f(t) \text{ a monotonic increasing or decreasing function.} \)

Mann (1945) described a nonparametric techniques for applying such a test. Mann’s test is in fact a particular application of Kendall’s test for correlation (Kendall, 1938, 1975), commonly known as Kendall’s tau. The test statistic \( S \) is defined as

\[
S = \sum_{k=1}^{n-1} \sum_{j=k+1}^{n} \text{sgn}(x_j - x_k)
\]

where

\[
\text{sgn(\theta)} = \begin{cases} 
1 & \text{if } \theta > 0 \\
0 & \text{if } \theta = 0 \\
-1 & \text{if } \theta < 0 
\end{cases}
\]  

Note that the statistic \( T \) used by Mann (1945) is related to the statistic \( S \) used by Kendall (1938, 1975) by \( S = 2T - n(n-1)/2 \). Mann showed that under \( H_0 \) the distribution of \( T \) and
Barrie (1999). Note that the S-Plus® function “arima.mle” which implements the method of Kohn and Ansley (1986) was used in these examples.

To illustrate this method, we will return to the example presented in Section 4.6.2. The model described by relation 4.167 was fitted to the data, but this time, the residuals were assumed to follow an auto-regressive model of order 1. The new model explained 33.2% of the total variance. The relative contribution of the relationship with the precipitation amount, of the long-term trend, and of the seasonal cycle did not change noticeably. All of the increase in the percentage of the total variance explained by the model is due to the inclusion of the auto-regressive model. The auto-regression coefficient is equal to 0.33. The long-term trend is shown in Figure 4.30 and the seasonal cycles in Figure 4.31. As we can easily see, there are no visible differences in this example between the results with the auto-regressive model and those without. Note that the auto-correlation was not very large. The coefficients in relation 4.167 have varied by less than 11%. Note also that the most important changes occurred in the seasonal cycles. We can therefore conclude that the presence of auto-correlations in the residuals did not strongly influence the estimations of the parameters.

4.6.3 Discussion

The techniques presented in this section are the most versatile and complete of all the techniques presented in this overview. They are versatile because any polynomial or trigonometric function can be used to describe temporal variations on any time scale. They are complete because the different components contributing to the total variation can be separated and modeled in the same general model. The only extra limiting assumption, compared with other techniques, is that the residuals should be normally distributed, and this is not as limiting as it seems when the number of data items is large. The other assumptions (i.e., independence of the residuals and constant variance) are common to all the techniques presented in this overview.
The only important drawbacks to the use of the techniques presented here are their complexity if serial correlations are present and the fact that the search for the best model can be labor intensive. In the case of this last restriction, the use of smoothers and spectral analysis can usually help accelerate the process of model selection. In the case of the first drawback, the use of statistical computer packages such as S-Plus® that includes programs that implement some or most of the techniques presented here is usually helpful.

One important consideration to keep in mind is that the presence of small serial correlations in the data will not usually have much of an influence on the estimation of the long-term trend or seasonal cycles if the number of observations is large. However, their presence will cause temporal variations that are not very strong to be considered not statistically significant. For example, if only a very weak long-term trend is present in the data, it may be rejected because of the presence of serial correlations. If these are taken into account, the long-term trend will then be judged significant. If the long-term trend is a strong one, however, the presence of serial correlations will only have a slight effect on the estimation of its shape. In conclusion, it is not always necessary to use models that include serial correlations but it is usually the cautious thing to do.

As we mentioned in the introduction of this section, it has been customary in the past to assume a priori that the long-term trend follows a straight line. That, in the opinion of the author, is a practice that should be avoided. Unless there are good physical reasons to assume any particular shape for the long-term trend, the reader should let the data tell him what the real shape of the long-term trend is. In all his work on time series analysis in atmospheric chemistry, the author has found only a few cases for which the straight line model was the appropriate one to describe the long-term variation. It is therefore recommended that other techniques such as smoothers be used to test the possible shape of the long-term trend before accepting any a priori conclusions. This remark also applies to any of the other temporal variations present in the data.

In this section, we have described techniques for including auto-regressive processes in the temporal model. These techniques can be easily extended to include MA (moving average) and ARMA (auto-regressive moving-average) models for the residuals. The reader is referred to Chatfield (1984) or Brockwell and Davis (1987) for details on these types of models. Descriptions of the techniques to use in these cases are given in Seber and Wild (1989) and Kohn and Ansley (1985,1986).
5. Verification of the Hypothesis behind the Techniques Used

5.1 Introduction

In Section 4, we presented a number of techniques for estimating and/or testing the different temporal variations that can be present in a time series. In all of these techniques some hypotheses were made, either about the data themselves or about the residuals after the elimination of some estimated model. It is therefore important after fitting a model or testing one for statistical significance to verify whether these hypotheses have been fulfilled.

The first thing to check is the validity of the assumed model. For example, if one assumes a straight line model for the long-term trend, the residuals will not show any long-term trend if the model is correct. It is therefore important to verify that such is the case. This applies to all the components of the temporal variations that are assumed to be present in the time series.

The most common hypothesis made in the techniques presented in Section 4 is that the residuals are independent. It is therefore necessary to test that hypothesis. We will present different methods here that can be used to test for independence. It is also usually assumed that the variance of the distribution of the residuals is constant with time. We will therefore discuss how to determine if that hypothesis has been fulfilled.

In some of the techniques that we have discussed, it is assumed that the data or the residuals are normally distributed. Many techniques exist to test whether this is the case. We will summarize some of these tests here.

It is important not only to assess whether the hypotheses are fulfilled but also to determine the importance of the deviations from them when they are not fulfilled. As we have mentioned often in Section 4, some of the techniques are still approximately valid even if some of the hypotheses are not respected, as long as the deviation from them is not too large. It is therefore necessary to have some idea of the importance of any possible deviation.

Although verification of the hypotheses may seems unnecessary to some readers, this procedure is a very important part of any time series analysis. Not doing it is like walking blindfolded into an unknown place. We thus recommend that the reader always verifies the validity of the model and/or techniques used by employing the techniques presented here or others that are appropriate.

5.2 Validity of the Model

In some of the techniques presented earlier, the shape of the long-term trend, seasonal cycles, or other temporal variations has to be specified before the model is fitted. We have mentioned that it has been popular in the past to assume a straight line for the long-term trend and a simple trigonometric function for the seasonal cycle. The most current model to describe the temporal variation in the observed ionic concentrations is given by relation 4.145 or other similar models. If that model is the right one, the mean of the residuals (i.e., the $e_i$'s in relation 4.145) should be constant and equal to zero. Therefore, they should not present any indication of long-term trends, or seasonal cycles, or other temporal variations. One easy way to verify that this condition has been met is to plot the residual time series and use one of the smoothing techniques presented in Section 4.3 to estimate the temporal variations, if there are any.
To illustrate the technique, we have fitted the model given by relation 4.145 to the $\text{SO}_4^{2-}$ ionic concentrations in precipitation at Chalk River. We have seen in Section 4.6.2 that a good model is the one given in relation 4.167, with the $\varepsilon_i$'s following an auto-regressive model of order one.

The long-term trend in relation 4.167 is a third order polynomial. Thus, the residuals should present a long-term trend if a straight line is used to model it. This is illustrated in Figure 5.1, which presents the residuals and a kernel smoother estimate of the long-term trend of the residuals. As we can see, a noticeable long-term trend is still present in the data. The model underestimates the concentrations before 1983, overestimates them between 1984 and 1990, and is close to them after that date. This is not the case if a third order polynomial is used to describe the long-term trend as illustrated in Figure 5.3. Unlike Figure 5.1, the estimated long-term trend does not show any large variations and can therefore be assumed to be constant and equal to zero.

We will now consider the seasonal cycle. Figure 5.2 presents the seasonal variations of the residuals after eliminating the model in relation 4.145. The solid line presents the estimated seasonal cycle using a periodic kernel smoother.
smoother. It is easy to see that the simple trigonometric functions used by the model do not completely describe the seasonal cycle. If one now uses the seasonal cycle given in relation 4.167, one sees that the residuals do not present any seasonal cycle (Figure 5.4).

This example illustrates the importance of checking the validity of the assumed model. The reader should keep in mind that temporal variations that are not explained will be translated into serial correlations in the residuals that may invalidate any test of significance that is made. It is therefore essential that the model used to describe the temporal variations present in the data should be as comprehensive as possible.

5.3 Independence of the Residuals

As mentioned earlier one of the most universal assumptions made in time series analysis techniques is that the data or the residuals are independent. It is therefore very important to be able to test whether the residuals are independent. Many techniques have been developed in the past to do this, and we will review some of them here.

5.3.1 Testing for Independence

Figure 5.3 Residuals after eliminating temporal model described by relation 4.167 from the observed SO$_4^{2-}$ ionic concentrations in precipitation at Chalk River. The solid line is the estimated long-term trend using a kernel smoother.

Figure 5.4 Residuals after eliminating temporal model described by relation 4.167 from the observed SO$_4^{2-}$ ionic concentrations in precipitation at Chalk River. The solid line shows the estimated seasonal cycle using a kernel smoother.
We will discuss here two possible tests for determining whether the residuals are independent. For these tests it is assumed that no long-term trend, seasonal cycles, or other temporal variations are present in the data. The first test is nonparametric, and the second assumes that the residuals are normally distributed. In what follows \( e_i \) \( (i = 1, 2, \ldots, n) \) will denote the residuals (i.e., \( e_i = y_i - (X \hat{b}_i) \)).

5.3.1.1 Turning Points Test

The turning points test (also known as the peaks and troughs test) is one of the easiest tests of randomness to apply (see Kendall, 1976; Kendall et al., 1983). The idea behind this test is to count the number of peaks or troughs that the time series exhibits. A peak is defined as a value greater than its two neighbouring values and a trough as a value less than its two neighbouring values.

Define \( X_i \) \( (i = 1, 2, \ldots, n-2) \) as:

\[
X_i = \begin{cases} 
1 & \text{if } e_i < e_{i+1} > e_{i+2} \text{ or } e_i > e_{i+1} < e_{i+2} \\
0 & \text{otherwise}
\end{cases}
\]  

The number of turning points (i.e., the sum of peaks and troughs) \( p \) in the time series is:

\[
p = \sum_{i=1}^{n-2} X_i
\]  

and we have:

\[
E(p) = \frac{2}{3} (n - 2)
\]  

and

\[
\text{var}(p) = \frac{16n - 29}{90}
\]  

If \( n \) is large, the variable \( P \) is equal to:

\[
P = \frac{p - E(p)}{\sqrt{\text{var}(p)}}
\]  

and follows a normal distribution with mean 0 and variance 1. If \( P \) is statistically significantly different from zero, the residuals are not independent. A small value for \( p \) (i.e., \( P \) negative) indicates that the successive values are positively correlated. A large value for \( p \) (i.e., \( P \) positive) indicates rapid non-random oscillations in the time series. Note that no missing values are allowed. It is, however, possible to calculate \( P_i \) independently for section \( i \) of the time series without missing values. Assuming that these series are independent and that the length of each series is not too small, the variable \( P^T = \sum_{i=1}^{n_i} P_i \) is normally distributed with mean zero and standard deviation equal to \( 1/\sqrt{n_i} \) where \( n_i \) is the number of series.

5.3.1.2 Durbin-Watson Test

In this test, one calculates the statistic,
An exact test procedure is not available, but Durbin and Watson (1950, 1951) have obtained lower and upper bounds $d_1$ and $d_u$ such that a value of $D$ outside these bounds leads to a definite decision. The decision rule for a test at level $\alpha$, is as follows:

1. **One-sided test against alternatives, the auto-correlation coefficient, $\rho$, greater than 0:**
   - If $D > d_u$, conclude that $\rho$ is not statistically different from zero.
   - If $D < d_1$, conclude that $\rho > 0$.
   - If $d_1 \leq D \leq d_u$, the test is inconclusive.

2. **One-sided test against alternatives, the auto-correlation coefficient, $\rho$, less than 0:**
   - Repeat (1) using $(4 - D)$ in place of $D$.

3. **Two-sided test against alternatives, the auto-correlation coefficient, $\rho$, not equal to 0:**
   - If $D < d_u$ or $4 - D < d_1$, conclude that $D$ is significant and reject the hypothesis of no auto-correlation at level $2\alpha$.
   - If $D > d_u$, conclude that $D$ is not significant and do not reject the hypothesis of no auto-correlation at level $2\alpha$.
   - If $D > d_u$ and $4 - D > d_u$, conclude that $D$ is not significant and do not reject the hypothesis of no auto-correlation at level $2\alpha$.

   Otherwise, the test is said to be inconclusive.

It has been discovered that, in many situations, treating the test as though $d_1$ did not exist and $d_u$ were the appropriate single critical value provides a very good approximation to the truth. A simplified approximate test procedure is as follows:

1. **Simplified one-sided test against alternatives $\rho > 0$:**
   - If $D < d_u$, conclude that $\rho > 0$.
   - Otherwise, do not reject that $\rho = 0$.

2. **Simplified one-sided test against alternatives, $\rho < 0$:**
   - If $4 - D < d_u$, conclude that $\rho > 0$.
   - Otherwise, do not reject that $\rho = 0$.

3. **Two-sided test against alternatives, $\rho \neq 0$:**
   - If $D < d_u$ or $4 - D < d_u$, conclude that $\rho \neq 0$ at level $2\alpha$. 

\[ D = \frac{\sum_{i=2}^{n} (e_i - e_{i-1})^2}{\sum_{i=1}^{n} e_i^2} \] (5.6)
Otherwise, do not reject that $\rho = 0$.

(5.8)

Tables of $d_u$ and $d_v$ for different numbers of $n$, for two levels of significance (0.05 and 0.01), and for various numbers of independent variables in the temporal model can be found in Neter and Wasserman (1974), Kendall (1976), or Draper and Smith (1981).

The present test assumes that there are no missing values. A variant of the test that takes the missing data into account was developed by Savin and White (1978). However, Bhargava (1989) has shown that there is no advantage in employing such a test. Using the usual Durbin-Watson test and ignoring the gaps in the data is just as effective. Savin and White (1978) have proposed other possible tests of independence for the residuals.

5.3.2 Plots of Auto-correlation and Integrated Periodogram

Many different types of graphics can be drawn to test the independence of the residuals visually or statistically. One of the simpler methods involves drawing scattergrams of $e_i$ versus $e_{i-1}$. We used these types of plots in Section 2.8 when trying to familiarize ourselves with the time series. This method is illustrated in Figure 5.5 for the residuals obtained by subtracting fitted models 4.145 or 4.167 from the observed $\text{SO}_4^{2-}$ ionic concentrations in precipitation at Chalk River. The straight lines represent least squares fits to the data. One can easily see that there are serial correlations in the residuals if model 4.145 is used but that this is not the case if model 4.167 (with a first order auto-regressive model to describe the error term) is used.

A second technique that can be used is the spectral analysis technique described in Section 4.4.2. We mentioned in that section that if the data are independent, their power spectrum should be a random variations around a straight line. A technique to test for

![Figure 5.5 Scattergrams of $e_i$ versus $e_{i-1}$ for the residuals after fitting model 4.145(a) or 4.167(b) (with a first order auto-regressive model to model the error term) to the observed $\text{SO}_4^{2-}$ ionic concentrations in precipitation at Chalk River. The straight lines represent least squares fits to the data.](image-url)
randomness was also presented in that section. The smoothed power spectra for the residuals after fitting model 4.145 or 4.167 (with a first order auto-regressive model to describe the error term) to the observed SO$_4^{2-}$ ionic concentrations in precipitation at Chalk River are shown in Figures 5.6a and 5.6b respectively. It is clear from the power spectrum that some part of the long-term trend and seasonal cycle remains if the model described by relation 4.145 is used. The presence of auto-correlations is also quite evident. However, if we use the model described by relation 4.167 with the error term following an auto-regressive model of order one, the smoothed power spectrum indicates that the residuals can be considered to be white noise or independent.

A simpler technique than using the smoothed power spectrum plots the auto-regression function between lag 1 and $N$ where $N$ is some predefined number. If there are no missing data in the time series, the auto-covariance coefficient at lag $k$ is given by (Chatfield, 1984):

$$c_k = \frac{1}{n} \sum_{i=1}^{n-k} (e_i - \bar{e})(e_{i+k} - \bar{e})$$  \hspace{1cm} (5.9)

where $\bar{e}$ is the mean value of the residuals and therefore equal to zero. The auto-correlation coefficient of lag $k$ is equal to:

$$r_k = \frac{c_k}{c_0}$$  \hspace{1cm} (5.10)

If there are missing values, the technique developed by Dunmuir (1984) and presented in Appendix II can be used to estimate the auto-covariance function.

The auto-correlation coefficients have to be compared with the limits $\pm 1.96/\sqrt{n}$ (Chatfield, 1984).

Examples of the auto-correlation function for the residuals after fitting model 4.145 or 4.167 (with a first order auto-regressive model to model the error term) to the observed SO$_4^{2-}$ ionic concentrations in precipitation at Chalk River are presented in Figures 5.7a and 5.7b respectively. They were drawn up to a lag of 100. The auto-correlation function indicates the presence of a net auto-correlation in the case of the residuals of model 4.145 but not in the
case of the second model. This is consistent with the results obtained using the other techniques.

A final technique, which is related to the power spectrum technique, is the integrated or cumulative periodogram. It is defined as:

$$CP(\omega) = \frac{\int_0^\omega f(\lambda) d\lambda}{\int_0^{2\pi} f(\lambda) d\lambda} \quad 0 \leq \omega \leq 0.5 \quad (5.11)$$

where $f(\omega)$ is the power spectrum. For the integrated or cumulative periodogram plot, one draws $CP(\omega)$ versus $\omega$. Notice that, by definition, $CP(\omega)$ varies between zero for $\omega = 0$ and 1 for $\omega = 0.5$. We have seen in Section 4.4.1 that the power spectrum of a random or white noise is a constant equal to $\sigma^2 / \pi$ (relation 4.98b). Therefore, its cumulative periodogram is equal to:

$$CP(\omega) = 2\omega \quad (5.12)$$

The cumulative periodogram of random noise is therefore a straight line crossing the graph. As was mentioned in Section 4.4.2, the cumulative periodogram for a sample from a random noise would oscillate around that line. Confidence intervals for those oscillations can be obtained by using the Kolmogorov-Smirnov test, if the random noise is assumed to have been sampled from a normal distribution. The confidence band is defined by the following two lines (Brockwell and Davis, 1987):

$$\omega \pm \frac{k_a}{\sqrt{n/2 - 1}} \quad (5.13)$$

where $k_{0.05} = 1.36$ and $k_{0.01} = 1.63$.

To illustrate the use of the integrated or cumulative periodogram, we will use the same data used to illustrate the other techniques presented in this section. Figure 5.8 presents the cumulative periodogram for the two sets of residuals. Relation 4.95 was used to estimate...
the power spectrum. It is quite clear that the residuals of model 4.145 are not independent. As for the other set of residuals, although there is some deviations from independence, they are small and should not influence the validity of the model much.

We have reviewed four graphical techniques in this section that can be used to verify the independence of the residuals. In the example they all give the same results. Of the four techniques, the scattergram and plotting the auto-correlation function are the simplest to use. The smoothed power spectrum is more informative because it can indicate which types of temporal variation are present in the residuals. The cumulative periodogram can also be applied in this way, but it is more difficult to use.

5.4 Normality of the Residuals

In some of the techniques presented in Section 4, it was assumed that the residuals were normally distributed. It is therefore important when using these techniques to be able to assess whether that hypothesis has been fulfilled.

As many tests for normality exist, we will concentrate on the most frequently used, namely, the $W$ test of Shapiro and Wilk (1965), D’Agistino’s test (D’Agostino, 1971), the range test of David et al. (1954), and finally the Kolmogorov-Smirnov goodness of fit test (Kolmogorov, 1941; Smirnov, 1948). Other tests can be found in Sachs (1984) and Stephens (1970, 1974). The power of many of these tests is compared in this last paper.

Although it is useful to test for non-normality, it is sometimes more useful to know the importance of the deviation from normality. Normal and log-normal probability paper (or plots) are very useful in these cases. We will therefore described the uses of these tests here.

Figure 5.8: Integrated periodogram for the residuals after fitting model 4.145(a) or 4.167(b) (with a first order auto-regressive model to model the error term) to the observed $SO_4^{2-}$ ionic concentrations in precipitation at Chalk River. The straight lines represent least squares fits to the data.
5.4.1 Testing for Normality

5.4.1.1 The W Test

This test was developed by Shapiro and Wilk (1965) to test the hypothesis:

\[ H_0 : \text{The population has a normal distribution} \tag{5.14} \]

against the hypothesis:

\[ H_1 : \text{The population does not have a normal distribution} \tag{5.15} \]

It can therefore be used to test whether the residuals are from a normal distribution. Note that it can only be used for \( n \leq 50 \). The test is as follows:

1. Calculate the parameter \( d \) given by:

\[
d = \frac{\sum_{i=1}^{n} (e_i - \bar{e})^2}{\sum_{i=1}^{n} e_i^2 - \frac{1}{n} \left( \sum_{i=1}^{n} e_i \right)^2} \tag{5.16}
\]

2. Order the residuals in increasing order to obtain the order statistics \( e_{[1]} \leq e_{[2]} \leq \ldots \leq e_{[n]} \).

3. Compute \( k \) as:

\[
k = \begin{cases} 
\frac{n}{2} & \text{if } n \text{ is even} \\
\frac{n-1}{2} & \text{if } n \text{ is odd}
\end{cases} \tag{5.17}
\]

4. Turn to Table 5 in Shapiro and Wilk (1965) or Table A6 in Gilbert (1987) to get the coefficients \( a_1, a_2, \ldots, a_k \).

5. Compute \( W \), given by:

\[
W = \frac{1}{d} \left[ \sum_{i=1}^{k} a_i (e_{[n-(i-1)]} - e_{[i]}) \right]^2 \tag{5.18}
\]

6. Reject \( H_0 \) at the \( \alpha \) significance level if \( W \) is less than the quantile given in Table 6 of Shapiro and Wilk (1965) or Table A7 of Gilbert (1987).

This test can be used to test for log-normality when the data are all greater than zero by first using a logarithmic transformation to transform them.

5.4.1.2 D'Agostino’s Test

This test can be applied only if the number of samples is greater than or equal to 50. Thus, it may serve as a complement to the \( W \) test. D’Agostino’s test can be summarized as follows:

1. Order the residuals in increasing order to obtain the order statistics \( e_{[1]} \leq e_{[2]} \leq \ldots \leq e_{[n]} \).
2. Compute the statistic $D$ as:

$$D = \frac{\sum_{i=1}^{n} \left[ i - \frac{1}{2} (n+1) \right] e_i}{n^2 s}$$

where

$$s = \left[ \frac{1}{n} \sum_{i=1}^{n} (e_i - \overline{e})^2 \right]^{1/2}$$

3. Calculate the statistic $Y$ given by:

$$Y = \frac{D - 0.28209479}{0.02998598 / \sqrt{n}}$$

Note that one should aim for five-place numerical accuracy in computing $D$.

4. Reject at the $\alpha$ significance level the null hypothesis (relation 5.14) if $Y$ is less than the $\alpha/2$ quantile or greater than the $1 - \alpha/2$ quantile of the distribution of $Y$. These quantiles are given in D'Agostino (1971) and in Table A8 of Gilbert (1987).

5.4.1.3 Range Test of David et al. (1954)

A very rapid and simple method to test for normality is as follows.

1. Calculate the ratio:

$$\frac{\text{range}}{\text{standard deviation}} = \frac{R}{s}$$

2. If the quotient does not lies between the tabulated critical values given in Table 3 of Pearson and Stephens (1964) or Table 72 of Sachs (1984), then the hypothesis of normality is rejected at the respective significance level.

It is quite apparent that this test is easy to use. Extensive tables for this procedure can be found in Pearson and Stephens (1964).

5.4.1.4 Kolmogorov-Smirnov Test

The Kolmogorov-Smirnov test can be described by the following steps:

1. Estimate the mean and standard deviation:

$$\overline{e} = \frac{1}{n} \sum_{i=1}^{n} e_i$$

and

$$s = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (e_i - \overline{e})^2}$$
Note that in the case of the residuals $\bar{e} = 0$ by definition.

2. Order the residuals in increasing order to obtain the order statistics $e_{(1)} \leq e_{(2)} \leq \ldots \leq e_{(n)}$.

3. Calculate the variables $z_i$ equal to

$$z_i = \frac{e_{(i)} - \bar{e}}{s}$$

(5.25)

4. Evaluate the following statistics:

$$D^+ = \max \{|i/n - z_i|\}$$

(5.26)

$$D^- = \max |z_i - (i - 1)/n|$$

(5.27)

$$D = \max (D^+, D^-)$$

(5.28)

5. Compare the value of $D$ with the critical values for the test (see Lilliefors, 1967, for tables). If $D > D_\alpha$, where $D_\alpha$ is the critical value for a test at significance level $\alpha$, the hypothesis that the sample is from a normal distribution can be rejected.

Note that $D_\alpha$ can be approximated by the relation (Stephens, 1974):

$$D_\alpha \approx \frac{K_\alpha}{\sqrt{n - 0.01 + 0.85/\sqrt{n}}}$$

(5.29)

where the $K_\alpha$ are equal to:

<table>
<thead>
<tr>
<th>$\alpha$</th>
<th>0.15</th>
<th>0.10</th>
<th>0.05</th>
<th>0.025</th>
<th>0.01</th>
</tr>
</thead>
<tbody>
<tr>
<td>$K_\alpha$</td>
<td>0.775</td>
<td>0.819</td>
<td>0.895</td>
<td>0.955</td>
<td>1.035</td>
</tr>
</tbody>
</table>

The approximation becomes better as $n$ increases.

5.4.2 Normal Probability Plots

We have already mentioned normal probability plots and described some of their uses in Section 2.5. We have seen that they can be used to estimate the importance of any deviation from normality.

The normal probability plot is constructed by first ordering the observations so that $e_{(1)} \leq e_{(2)} \leq \ldots \leq e_{(n)}$ and then plotting the $e_{(i)}$ against $\Phi^{-1}\left(\frac{i - \frac{1}{2}}{n}\right)$, where $\Phi$ is the cumulative distribution function for the standard normal distribution defined as:

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{t^2}{2}} dt$$

(5.30)
If the data are normally distributed, one will obtain a straight line. An $\alpha$ confidence band around that theoretical straight line, based on the Kolmogorov-Smirnov test, is given by the following two lines (see Michael, 1983):

$$y = \bar{e} + s\Phi^{-1}\left\{\Phi(x) \pm \left(\frac{D_n}{\sqrt{n}}\right)\right\}$$

(5.31)

where $\bar{e}$ and $s$ are defined by relations 5.23 and 5.24 and relation 5.29 can be used to estimate $D_n$. If the observed cumulative distribution gets out of this band, one can say that, with a level of confidence $\alpha$ the hypothesis that the sample is from a normal distribution can be rejected using the Kolmogorov-Smirnov test.

An example of such a plot is presented in Figure 5.9 for the residuals after fitting model 4.167 (with a first order auto-regressive model to model the error term) to the observed $\text{SO}_4^{2-}$ ionic concentrations in precipitation at Chalk River. One sees first that, using the Kolmogorov-Smirnov test, the hypothesis of normality can be rejected at both significance levels. However, one also sees that the deviation from normality is not extreme and therefore should not have much influence on the results of the analysis, as the number of observations available is large.

This simple example illustrates how useful these types of graphics can be to verify the normality of the distribution of sampled data.

5.5 Homoscedasticity of the Residuals

One of the basic hypotheses of all the techniques presented in Section 4 is that the variance of the residuals remains constant in time. It is therefore important to check if this hypothesis is fulfilled by the data studied.

The easiest way to check this hypothesis is to determine visually whether the spread in the residuals changes with time. If we consider the residuals in Figure 5.10, one notices that, apart from a few low values, the residuals seem to spread out equally around zero during all periods.

A better way to study the variation of the spread of the residuals with time is to consider the variations of the first and third quartiles with time. The difference between the two gives the variation of the range with time. This may help in getting a better idea of the variance changes with time. The variation of the quartiles may be estimated using a running quartile smoother. This estimator is similar to the running median smoother discussed in Section 4.3.3, except...
that the first or the third quartiles is estimated instead of the median. Figure 5.11 presents examples of the estimated running first and third quartiles for the residuals of Figure 5.10. These lines suggest that the variance may have increased slightly with time. The effect is very small, however, and can be neglected.

The running quartile smoother can also be used to verify that the distribution of the residuals is symmetric around zero. This is evidently true in the present example.

If we assume that the distribution of the residuals remains symmetric around zero, one
Figure 5.12 Absolute value of the residuals after fitting 4.167 (with a first order autoregressive model to model the error term) to the observed SO$_2$-ionic concentrations in precipitation at Chalk River. The line shows the running median smoother estimate of the variations of the median with time

can consider the changes in time of the distribution of the absolute values of the residuals. The running median smoother or any other smoother can be used to see if there is any indication that the variance has changed with time. This is illustrated for the residuals of Figure 5.10 in Figure 5.12. The running median smoother was used for this example. One sees that, as mentioned earlier, there is some indication of a small increase in the variance with time. However the effect is small and should not influence unduly the results of the analysis.

One way to test the statistical significance of this variation is to calculate either Spearman’s correlation coefficient or Kendall’s $\tau$ correlation coefficient between the absolute value of the residuals and time (see Hollander and Wolfe, 1973, for the definition of these correlation coefficients). Both correlation coefficients are nonparametric and based on rank. For the example studied in this section, Spearman’s and Kendall’s correlation coefficients are equal to 0.12 and 0.08 respectively, and both are significantly different from zero. This confirms the fact that a small increase in the variance is present in the residuals.

5.6 Discussion

In this section we have presented some of the techniques that can be used to verify the fulfillment of the hypothesis underlying various methods for analyzing the temporal variations of a time series. The reader should be aware that the list of techniques presented here is far from comprehensive and that many others exist.

The main goal of the author in this section was to illustrate the importance of testing the fulfillment of these hypotheses and indeed of knowing which hypotheses underlie the techniques one wants to use. Even if one accepts that some or all of these hypotheses will not
be fulfilled by the data at hand, it is still very important to have some idea of the effects of not conforming to them. These techniques can therefore be used to estimate if the techniques or models used are acceptable for a particular time series.
6. A Case Study: Particulate \( \text{SO}_4^{2-} \) Concentrations in Ambient Air at ELA, Ontario.

Although examples have been given of most of the techniques presented in this overview, it may be useful to present a complete example of the different steps involved in the study of a time series. We have selected as an example the variations of particulate \( \text{SO}_4^{2-} \) daily concentration in ambient air at the Experimental Lakes Area (ELA), Ontario. Because of lack of space and preparation time, we will not use all the techniques presented in this overview but rather a selection of the most important ones. Also, we will not try to interpret the results here in any detail.

The time series is presented in Figure 6.1. The first thing that one notices is that although some long-term trend may be present in the data no clear increase or decrease occurred during the sampling period. There is also some indications that a seasonal cycle is present in the time series. There is also evidence of missing data especially, in 1989. As Figure 6.2 illustrates, most of the missing periods are smaller than 3 days although there are a few very long ones.

The possible presence of a non-monotonic long-term trend is indicated by the annual box-plots presented in Figure 6.3. The presence of a marked seasonal cycle is confirmed by the monthly box-plots shown in Figure 6.4. The presence of both variations in the time series is confirmed by the smoothed power spectrum presented in Figure 6.5. The spectrum also shows that auto-correlations are present in the time series.

From Figure 6.6, we can conclude that a logarithmic transformation should be applied to the data before the analysis. One should, however, note that the distribution seems to deviate from the normal distribution (i.e., the observations do not follow a straight line). Part of that deviation could be due to the presence of the long-term trend, seasonal cycles, and

![Figure 6.1 Temporal variations of particulate \( \text{SO}_4^{2-} \) daily concentrations in ambient air at the Experimental Lakes Area, Ontario.](image-url)
possibly of other temporal variations. We will therefore assume that a logarithmic transformation will be appropriate for now.

A first approximation of the long-term trend was obtained using a Nadaraya-Watson kernel estimator. The result of the analysis is presented in Figure 6.7. Broadly speaking, the trend here is similar to the long-term trend obtained using the annual box-plots in Figure 6.3. After a decrease in the beginning of the 1980s, the concentration increased up to about the end of 1989 and the beginning of 1990 and decreased slowly thereafter.
Figure 6.4: Monthly box-plots for the data presented in Figure 6.1. The squares are the monthly means.

After elimination of the long-term trend, a periodic smoother was used to estimate the seasonal cycle. The resulting seasonal cycle is shown in Figure 6.8. The result is very similar to the one obtained using the monthly box-plot of Figure 6.4. The concentration reaches a maximum in March and a minimum in July and September.

After elimination of both the long-term trend and seasonal cycle estimates, scattergrams of the concentrations at time $t$ versus the concentrations at times $t-1$, $t-2$, and $t-3$ were constructed. They are shown in Figure 6.9. One can detect a net relationship between the concentration at time $t$ and the one at time $t-1$. The other two graphs are not as clear as this one. If one uses the regression technique described in Section 2.8, one obtains the following auto-regressive model:

$$e_t = 0.60e_{t-1} - 0.12e_{t-2} + 0.04e_{t-3} + \eta_t$$

(6.1)

If one puts all these results together, we arrive at the following general temporal model to describe the time series:

1. A long-term trend that is neither a straight line nor monotonic.
2. A seasonal cycle that seems to be more complex than a simple sine/cosine wave.
3. An auto-regressive model of order two or three.
4. The smoothed power spectrum may indicate the presence of a long-term cycle with a period of about 3 years.

Figure 6.5: Smoothed power spectrum for the data presented in Figure 6.1.

Figure 6.6: Log-normal probability plot for the data of Figure 6.1.
Table 6.1 Results of the seasonal Kendall test.

<table>
<thead>
<tr>
<th>Month</th>
<th>S</th>
<th>var(S)</th>
<th>Pvalue</th>
</tr>
</thead>
<tbody>
<tr>
<td>January</td>
<td>9655</td>
<td>11895508</td>
<td>0.0051</td>
</tr>
<tr>
<td>February</td>
<td>981</td>
<td>9522241</td>
<td>0.7506</td>
</tr>
<tr>
<td>March</td>
<td>6385</td>
<td>12584630</td>
<td>0.0719</td>
</tr>
<tr>
<td>April</td>
<td>9084</td>
<td>10734322</td>
<td>0.0056</td>
</tr>
<tr>
<td>May</td>
<td>5211</td>
<td>11820522</td>
<td>0.1296</td>
</tr>
<tr>
<td>June</td>
<td>11022</td>
<td>9073361</td>
<td>0.0003</td>
</tr>
<tr>
<td>July</td>
<td>-4264</td>
<td>6578558</td>
<td>0.0964</td>
</tr>
<tr>
<td>August</td>
<td>1162</td>
<td>7313990</td>
<td>0.6674</td>
</tr>
<tr>
<td>September</td>
<td>8855</td>
<td>9200036</td>
<td>0.0035</td>
</tr>
<tr>
<td>October</td>
<td>6645</td>
<td>12352066</td>
<td>0.0587</td>
</tr>
<tr>
<td>November</td>
<td>5576</td>
<td>10113921</td>
<td>0.0796</td>
</tr>
<tr>
<td>December</td>
<td>2106</td>
<td>10113714</td>
<td>0.5078</td>
</tr>
<tr>
<td>Total</td>
<td>62418</td>
<td>121302864</td>
<td>0.0000</td>
</tr>
</tbody>
</table>

5. A logarithmic transformation is necessary because the data seem to be close to a log-normal distribution.

From these results, we can conclude that the nonparametric techniques (Section 4.2) are not applicable because the long-term trend is not monotonic. If, however, the seasonal Kendall test (Section 4.2.2.3) is used, we obtain the results presented in Table 6.1. Although the test indicates a long-term trend at a 99.99% confidence level, the validity of the test is strongly in doubt, as the results differ widely from month to month. In fact, the test indicates that the trend may be significant for only three of the months. That problem would also be present if one used the seasonal Kendall's slope estimator (Section 4.2.3.2).

![Figure 6.7 Nadaraya-Watson kernel estimate of the long-term trend present in time series of Figure 6.1.](image)
The long-term trend obtained using the Nadaraya-Watson kernel is shown in Figure 6.7. A comparison of this estimate with the running mean, running median, loess, spline, and super smoother estimates is shown in Figure 6.10. The same span of 3 years was used for all the estimators. Except for the running median and loess estimators (dashed lines), all the estimators give about the same results. The shape of the running median is similar to the others but shifted slightly to larger values in comparison to the results of the other estimators. The loess estimate oscillates around the others, although a result similar to the others could be obtained if a span of about 4 years were used. In summary, all the smoothers, with the exception of the running median, produce very similar long-term trend estimates if the right span is used. In the case of the running median estimator, the mean long-term trend is slightly higher than that for the other smoothers. This may result from a slight asymmetry in the distribution of the observations, even after logarithmic transformation.

After elimination of the long-term trend using the Nadaraya-Watson estimate of Figure 6.7, one can look at the seasonal cycle. The periodic kernel estimation is shown in Figure 6.8. Results from three other smoothers, namely the running mean, running median,
Figure 6.10: Estimated long-term trend using the kernel, running mean, running median, loess, spline and super smoother techniques. A span of three years was used for all smoothers.

and the loess smoothers, are compared to the result of the kernel smoother in Figure 6.11. The estimates are quite close to one another. The one showing the biggest difference is the loess smoother (dashed line). As with for the long-term trend, the fit between the loess estimates and the others can be improved by slightly increasing the span.

As there are many data items missing, it is difficult to use a filter technique. Spectral analysis can still be used, however, to test the statistical significance of the different components of the temporal variations.

Figure 6.12 compares the smoothed power spectrum for the observations with the theoretical spectrum for white noise (i.e., random errors) and for a first order autoregressive process with an autocorrelation coefficient equal to 0.4. From this figure it is clear that an autoregressive process is present in the data and should be taken into account.

Figure 6.13a shows what happens when we eliminate from the data the long-term trend estimated with the kernel smoother. One can see that

Figure 6.11: Seasonal cycle estimates using the periodic kernel, running mean, running median, and loess periodic smoothers.
a) White Noise

![White Noise Power Spectrum](image1)

b) First order auto-correlation coefficient = 0.4

![First Order Auto-correlation Coefficient](image2)

a) Without Long-term trend

![Without Long-term trend Power Spectrum](image3)

b) Without long-term trend and seasonal cycles

![Without Long-term and Seasonal Cycles](image4)

**Figure 6.12** Smoothed power spectra.

**Figure 6.13** Smoothed power spectra.

the spectrum associated with the long-term trend is no longer significant. The reader should note the significant long-term wave with a period of about 3 years.

The power spectrum after the periodic kernel estimates of the seasonal cycle have been eliminated from the data is shown in Figure 6.13b. One sees that most of the temporal variations present in the data have been explained, except perhaps for a long-term cycle with a period of 3 years. Note that the amplitude of this long-term cycle is small.

The monthly medians of the data are given in Table 6.2, and their temporal variations are shown in Figure 6.14. Unlike the example in Section 4.5.3, the present example shows no
Table 6.2 Monthly median SO$_4^{2-}$ ambient ionic concentrations in air at Experimental Lake Area, Ontario. Numbers in brackets have been interpolated.

<table>
<thead>
<tr>
<th>Year</th>
<th>J</th>
<th>F</th>
<th>M</th>
<th>A</th>
<th>M</th>
<th>J</th>
<th>A</th>
<th>S</th>
<th>O</th>
<th>N</th>
<th>D</th>
</tr>
</thead>
<tbody>
<tr>
<td>1979</td>
<td>1.11</td>
<td>1.40</td>
<td>2.51</td>
<td>1.77</td>
<td>0.70</td>
<td>0.54</td>
<td>1.08</td>
<td>0.82</td>
<td>0.55</td>
<td>0.65</td>
<td>0.81</td>
</tr>
<tr>
<td>1980</td>
<td>0.85</td>
<td>1.30</td>
<td>1.09</td>
<td>0.80</td>
<td>1.02</td>
<td>0.83</td>
<td>(1.00)</td>
<td>0.83</td>
<td>0.78</td>
<td>0.54</td>
<td>0.96</td>
</tr>
<tr>
<td>1981</td>
<td>1.57</td>
<td>1.45</td>
<td>0.95</td>
<td>1.29</td>
<td>1.12</td>
<td>0.60</td>
<td>0.83</td>
<td>0.97</td>
<td>0.43</td>
<td>0.77</td>
<td>0.79</td>
</tr>
<tr>
<td>1982</td>
<td>0.95</td>
<td>1.23</td>
<td>1.94</td>
<td>1.69</td>
<td>1.40</td>
<td>0.32</td>
<td>0.47</td>
<td>0.33</td>
<td>0.45</td>
<td>0.57</td>
<td>0.92</td>
</tr>
<tr>
<td>1983</td>
<td>0.85</td>
<td>2.07</td>
<td>2.61</td>
<td>0.99</td>
<td>0.68</td>
<td>0.76</td>
<td>0.98</td>
<td>1.17</td>
<td>0.67</td>
<td>0.82</td>
<td>0.40</td>
</tr>
<tr>
<td>1984</td>
<td>1.76</td>
<td>2.18</td>
<td>2.53</td>
<td>1.71</td>
<td>1.41</td>
<td>1.19</td>
<td>0.70</td>
<td>1.16</td>
<td>0.57</td>
<td>1.41</td>
<td>1.43</td>
</tr>
<tr>
<td>1985</td>
<td>0.97</td>
<td>1.22</td>
<td>0.93</td>
<td>1.21</td>
<td>1.07</td>
<td>0.92</td>
<td>0.62</td>
<td>0.66</td>
<td>0.36</td>
<td>0.70</td>
<td>1.10</td>
</tr>
<tr>
<td>1986</td>
<td>0.99</td>
<td>1.40</td>
<td>1.49</td>
<td>0.89</td>
<td>1.51</td>
<td>0.67</td>
<td>0.61</td>
<td>0.53</td>
<td>0.47</td>
<td>1.04</td>
<td>1.00</td>
</tr>
<tr>
<td>1987</td>
<td>1.12</td>
<td>1.87</td>
<td>2.03</td>
<td>1.41</td>
<td>1.04</td>
<td>0.97</td>
<td>0.55</td>
<td>0.83</td>
<td>0.75</td>
<td>0.77</td>
<td>0.93</td>
</tr>
<tr>
<td>1988</td>
<td>1.30</td>
<td>1.14</td>
<td>1.53</td>
<td>1.45</td>
<td>1.32</td>
<td>0.94</td>
<td>0.66</td>
<td>0.94</td>
<td>0.69</td>
<td>0.73</td>
<td>1.58</td>
</tr>
<tr>
<td>1989</td>
<td>1.67</td>
<td>1.83</td>
<td>2.30</td>
<td>1.86</td>
<td>1.25</td>
<td>0.82</td>
<td>(0.70)</td>
<td>0.61</td>
<td>0.85</td>
<td>0.94</td>
<td>0.99</td>
</tr>
<tr>
<td>1990</td>
<td>1.64</td>
<td>1.62</td>
<td>1.84</td>
<td>1.94</td>
<td>1.77</td>
<td>0.89</td>
<td>0.74</td>
<td>1.01</td>
<td>0.47</td>
<td>0.85</td>
<td>0.93</td>
</tr>
<tr>
<td>1991</td>
<td>1.60</td>
<td>1.13</td>
<td>1.64</td>
<td>1.13</td>
<td>0.96</td>
<td>0.79</td>
<td>0.62</td>
<td>0.75</td>
<td>0.76</td>
<td>0.78</td>
<td>0.84</td>
</tr>
<tr>
<td>1992</td>
<td>1.70</td>
<td>1.77</td>
<td>1.95</td>
<td>1.70</td>
<td>1.00</td>
<td>0.70</td>
<td>0.80</td>
<td>0.95</td>
<td>1.15</td>
<td>0.93</td>
<td>1.11</td>
</tr>
<tr>
<td>1993</td>
<td>1.01</td>
<td>1.13</td>
<td>2.09</td>
<td>1.95</td>
<td>1.24</td>
<td>0.96</td>
<td>0.68</td>
<td>1.07</td>
<td>0.55</td>
<td>0.69</td>
<td>1.32</td>
</tr>
<tr>
<td>1994</td>
<td>0.81</td>
<td>1.63</td>
<td>1.77</td>
<td>1.44</td>
<td>0.84</td>
<td>1.08</td>
<td>0.58</td>
<td>0.44</td>
<td>1.18</td>
<td>1.01</td>
<td>1.16</td>
</tr>
</tbody>
</table>

net decrease. However, there seems to be a minimum in the middle of the 1980s and a maximum at the beginning of the 1990s. The long-term trend and the seasonal cycles estimates using the median polish and anova techniques are shown in Figures 6.15 and 6.16. The anova technique indicates that both the yearly and monthly variations are statistically significant. A simple additive model was used.

The seasonal cycles are similar to those obtained with the smoothers (Figures 6.15 and 6.14 Temporal variations of monthly median SO$_4^{2-}$ ambient concentrations in air at the Experimental Lakes Area, Ontario.)
6.11. It is interesting to note that the long-term trends obtained here are similar to those obtained with the loess smoother. They differ from the results of the other smoothers because of the presence of a local maximum in 1984. Note that the other smoothers will also show this maximum if they use a span of less than three years. One should finally note that the results of the median polish and anova techniques are very similar. The values of both the long-term trend and the seasonal cycle are smaller in the former. This is due, in part, to the fact that the estimate of the overall mean (which was added to the yearly or monthly effects to produced these two graphs) is smaller for the median polish method.

Figure 6.15 Estimated long-term trend using anova (continuous line) and median-polish (dashed line). The overall mean was added to the yearly effects.

Figure 6.16 Estimated seasonal cycle using anova (continuous line) and median-polish (dashed line). The overall mean was added to the monthly effects.

Figure 6.17 Estimated long-term trend using the maximum likelihood technique of Section 4.6.3.2
Table 6.3 Percentile contribution of model components to the total variance observed in the data.

<table>
<thead>
<tr>
<th>Component</th>
<th>Percentile Contribution to Total Variance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Long-term Trend</td>
<td>0.79</td>
</tr>
<tr>
<td>Long-term cycle (period = 3 years)</td>
<td>0.75</td>
</tr>
<tr>
<td>Seasonal Cycle</td>
<td>12.27</td>
</tr>
<tr>
<td>Auto-regressive model of order 2</td>
<td>19.10</td>
</tr>
<tr>
<td>$a_1 = 0.487$</td>
<td></td>
</tr>
<tr>
<td>$a_2 = -0.044$</td>
<td></td>
</tr>
</tbody>
</table>

The last technique to be illustrated fits a comprehensive model (see relation 4.147) which includes a long-term trend, long-term cycles, a seasonal cycle, and an auto-regressive model. The maximum likelihood technique described in Section 4.6.3.2 was used.

The model fitted to the data explains about 33% of the total variance observed. The components of the model fitted to the observations are listed in Table 6.3, with the fraction of the total variance they explain. The long-term trend and the seasonal cycle are shown in Figures 6.17 and 6.18. Although statistically significant, the long-term trend and the long-term-cycle represent less than 1% of the total variance. The presence of this long-term cycle confirms our earlier analysis of the power spectrum. The highest contributors to the total variance are the seasonal cycle and the second order auto-regressive component, which contributed 12 and 19% respectively. Although statistically significant, the second component of the auto-regression model is very small.

It is interesting to compare this long-term trend with those obtained using the smoothers. Such a comparison can be seen in Figure 6.19. One can see that the differences

Figure 6.18 Estimated seasonal cycle using the maximum likelihood technique of Section 4.6.3.2
between the various estimates of the long-term trend are small, and that, overall, they give the same picture of the long-term trend, a small decrease in the first few years followed by an increase until the beginning of the 1990s, followed again by a decrease.

The correspondence between the maximum likelihood estimate and the results obtained using the smoothers holds for the
seasonal cycle. This is illustrated in Figure 6.20. One sees that, overall, the fit between the different estimators is quite good.

To illustrate the different methods presented in Section 5 for verifying the validity of models and techniques, we will employ some of these methods to analyze the residuals obtained when using the maximum likelihood technique to fit a comprehensive model.

The smoothed power spectrum of the residuals (Figure 6.21) indicates that the hypothesis that they are random cannot be rejected. We can arrive at the same conclusion by considering the auto-correlation coefficient up to lag 100 as shown in Figure 6.22.

The distribution of the residuals seems to deviate noticeably from the normal

---

**Figure 6.22** Auto-correlation coefficients for the residuals. The dashed lines indicate the 95% confidence band for a random process.

**Figure 6.23** Log-normal probability plot for the residuals.
distribution, as Figure 6.23 shows. Although the deviation of the distribution is important, this should not influence the results drastically as a large number of observations are available. One should also note that the variance of the distribution may have changed slightly with time, as Figure 6.24 shows. The variation is small, however, as the differences between the third and the first quartiles change only slightly with time.

In summary, all of these analyses have shown that the ambient SO$_4^{2-}$ air concentrations at the Experimental Lakes Area have varied between 1979 and 1994, but with no net decrease or increase. The long-term trend cannot be approximated by a straight line and it is not monotonic. Thus, the nonparametric techniques of Section 4.2 should not be applied in this case.

A long-term cycle with a period equal to about 3 years is also present in the data. The amplitude of the cycle is small, and it contributes only a small fraction of the total variance. However, this contribution is about the same order of magnitude as the long-term trend.

All the analyses presented in this section indicate that the seasonal cycle is a complex and cannot be described by a simple sine/cosine wave. Therefore, a more complex description using a combination of trigonometric functions has to be used.

The presence of auto-correlations in the observations is well established. Although an auto-regressive model of order 2 was used, a model of order 1 should be enough to describe the auto-correlations.
7. Final Remarks

In this overview of time series analysis, the author has tried to give the reader more than just a catalogue of techniques. He has tried, in particular, to present a framework that can be used in the study of the temporal variations of time series. This is not the only possible framework, but it has the advantage of reducing the number of a priori hypotheses as much as possible. The main driving force behind this framework is the belief that the data should guide the analysis and the choice of conceptual model as much as possible.

The main points of this framework are the following:

1. The study of the temporal variations in a series of numbers should not be done using statistical techniques as black-boxes.
2. Users should have a good knowledge of their data. To acquire that knowledge, they have to explore the data first, giving special attention to its distribution and temporal characteristics.
3. They must also have a good knowledge of the technique(s) they want to use.
4. They must check that the technique selected can be applied to the data. It very often happens that one or many of the assumptions that are made when using a technique are not fulfilled by the data.
5. It is generally good practice to use more than one technique to explore and analyze the data.
6. Finally, one should test or verify the hypotheses made after using a particular technique.

Naturally, the use of the complete framework is not possible in some circumstances; such as when we have to analyze a large number of time series. In those situations, the use of a combination of techniques can be very useful. For example, spectral analysis and smoothing techniques can easily be used to analyze a large number of time series. The calculation and drawing of a number of smoothed power spectra and smoother estimates is not such a difficult and/or lengthy task. These techniques can be employed to isolate the time series with the more interesting characteristics which can be studied later with more demanding techniques such as the fitting of a general model by maximum likelihood.

Throughout this overview, the author has tried to present the advantages and disadvantages of the different techniques discussed. As a final review, it may be useful to give some general comments on the different types of techniques.

Nonparametric tests and slope estimators may seem to have a net advantage in that they do not assume any specific form for the distribution of the error term. However, they in fact present many disadvantages that outweigh this one advantage. First, the long-term trends are assumed to be monotonic or even to be a straight line in the case of the estimators. This is a very stringent hypothesis. Therefore, the author does not recommend the use of these techniques, except in special circumstances.

The smoothing techniques are very useful for exploring the data, and in most cases when the quantity of data is large will give very good estimates for some temporal variations, such as the long-term trend or seasonal cycle, even if some of the hypotheses (such as the independence of the residuals) are not fulfilled. We have seen an example of that in the case study presented in Section 6.
Spectral analysis is another very useful and general technique for exploring temporal variations in a time series. This technique can also be used to determine the statistical significance of those variations. On the other hand, the use of filters is quite limited because the filters have difficulties in taking missing data into account. If the number of missing data items is small, filters can still be employed, but their use becomes problematic if large amounts of data are missing.

As the reader may already have discerned, the author's preferred technique is the fitting of a very general model to the data by maximum likelihood (Section 4.6). The model used, however, should not be restricted to long-term trends that are straight lines. The more general the model, the better it can fit the data without doing injury to them.

The author would like to conclude by pointing out that the job of data analysis is not to fit the data to an *a priori* concept of what they should be but to extract what they have to say.
The Canadian Air and Precipitation Monitoring Network as it was in November 1997.
Appendix II: Estimation of Smoothed Power Spectrum When Observations Are Missing

The smoothed power spectrum is obtained from the auto-covariance function using the relation (Chatfield, 1984):

\[ f(\omega) = \frac{1}{\pi} \left\{ \lambda_0 c_0^a + 2 \sum_{k=1}^{M} \lambda_k c_k^a \cos(2\pi \omega k) \right\} \]  

(A1)

where \( \omega (0 \leq \omega \leq 0.5) \) is the frequency in cycles per hour, day, or other period, depending on the data observation frequency; \( c_k^a \) is the auto-covariance coefficient of lag \( k \); \( M (< N) \) is the truncation point that has to be chosen (the lower the value of \( M \), the smoother the spectrum); \( \lambda_k \) is the smoothing window; and \( N \) is the number of hourly, daily, or other periods between the first and last data items. One of the possible windows is the Parzen window (Chatfield, 1984) given by:

\[
\lambda_k = \begin{cases} 
1 - 6 \left( \frac{k}{M} \right)^2 + 6 \left( \frac{k}{M} \right)^3, & 0 \leq k \leq M / 2 \\
2 \left( 1 - \frac{k}{M} \right)^3, & M / 2 \leq k \leq M 
\end{cases} \]  

(A2)

Other possible windows are described by Bloomfield (1976) and Chatfield (1984).

Because of the possibility of missing data, a special technique has to be used to estimate the auto-covariance function, \( c_k^a \). The technique developed by Dunsmuir (1984) is used. If \( X_1, X_2, \ldots, X_N \) is a sample in which some of the \( X_i \) can be missing, one first defines the indicator variable \( a_i \) as:

\[
a_i = \begin{cases} 
0 & \text{if data } i \text{ is missing} \\
1 & \text{otherwise} 
\end{cases}
\]  

(A3)

i=1, 2, ..., N, where \( N \) is the total number of possible data points; then a new variable defined as

\[
Y_i = a_i X_i \]  

(A4)

is calculated. To estimate the auto-covariance function, one first calculates

\[
\bar{a} = \frac{1}{N} \sum_{n=1}^{N} a_n 
\]  

(A5)

\[
\bar{Y} = \frac{1}{N} \sum_{n=1}^{N} Y_n 
\]  

(A6)

\[
\bar{\mu} = \frac{\bar{Y}}{\bar{a}} 
\]  

(A7)

and then defines

\[
c_k = \frac{1}{N} \sum_{n=1}^{N} a_n a_{n+k} 
\]  

(A8)
Finally the auto-covariance $c_k^a$ is equal to

$$c_k^a = c_k^r / c_k \quad \text{if} \quad c_k \neq 0$$
Appendix III : Lomb's Technique for Estimating Power Spectrum with Missing Data

Suppose that there are $N$ data points $c_i = c(t_i)$, $i = 1, 2, ..., N$. Then, the mean and variance of the data are found in the usual way:

$$\bar{c} = \frac{1}{N} \sum_{i=1}^{N} c_i \quad \text{(A11)}$$

$$s^2 = \frac{1}{N-1} \sum_{i=1}^{N} (c_i - \bar{c})^2 \quad \text{(A12)}$$

Now, the Lomb normalized periodogram is defined by:

$$P_N(\omega) = \frac{1}{2s^2} \left\{ \frac{\sum_j (c_j - \bar{c}) \cos \omega(t_j - \tau)}{\sum_j \cos^2 \omega(t_j - \tau)} \right\}^2 + \frac{\sum_j (c_j - \bar{c}) \sin \omega(t_j - \tau)}{\sum_j \sin^2 \omega(t_j - \tau)} \quad \text{(A13)}$$

where $\tau$ is defined by:

$$\tan(2\omega\tau) = \frac{\sum_j \sin 2\omega t_j}{\sum_j \cos 2\omega t_j} \quad \text{(A14)}$$

The constant $\tau$ is a kind of offset that makes $P_N(\omega)$ completely independent of shifting all the $t_i$'s by any constant. See Lomb (1976) and Scargle (1982) for more details. Computer programs in Fortran and C can be found in Press et al. (1992 a and b).
REFERENCES


