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

(Helsinki, 14-18 September 1998)
NOTE

The designation employed and the presentation of material in this document/publication do not imply the expression of any opinion whatsoever on the part of the Secretariat of the World Meteorological Organization concerning the legal status of any country, territory, city or area or of its authorities, or concerning the delimitation of its frontiers or boundaries.

This report has been compiled from information furnished to the WMO Secretariat. It is not an official WMO publication and its distribution in this form does not imply endorsement by the Organization of the ideas expressed.
WMO/EMEP Workshop
on Advanced Statistical Methods and their
Application to Air Quality Data Sets

(Helsinki, 14-18 September 1998)
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. Tarrasson)
   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
Annex B AGENDA
Annex C Principal Component Analysis and other Dimensionality Reduction Techniques: a short overview
Annex D A brief overview of Cluster Analysis
Annex E A brief and brazed overview of Time Series Analysis
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 Atmospheric 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 Atmospheric 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 give 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 Synthezing Centres East and West (MSC-W, MSC-E) and the EMEP’s Chemical Coordinating Centre (CCC) on both modelling and monitoring.

![Figure 1](image)

2.2 Goals of the Workshop

Ms. Sonja Vidić, 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...
GAW 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. Vidic 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 \( \text{VOC} \) 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 on 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 co-operate 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 (SO\textsubscript{2}, SO\textsubscript{4}, NO/NO\textsubscript{2}, NO\textsubscript{3}, NH\textsubscript{3}, NH\textsubscript{4}) 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 SO$_2$, NO$_x$ or NH$_3$, 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 NO$_x$ 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.

Basicstatistical 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)
POPs 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/trajecory 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
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, PTL/AQ).

10
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 (eg. 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 database;

• 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 4565666
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.bg

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
Annex A, p 2

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
Geschwister 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
Uerlandstrasse 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 Environment 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: epoc@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
Øystein Hov
NILU
P.O. Box 100
N-2007 KJELLER
Norway
Tel: 47 63 898000
Fax: 47 63 898050
Email: oystein.hov@nilu.no

Dr M. Lesnjak
Hydrometeorological Institute
Environmental Protection Department
Vojkova 1/B
61000 LJUBLJANA
Slovenia
Tel: 386 61 327461
Fax: 386 61 133 1396
Email: melanija.lesnjak@rzs-hm.si

Kaare Kemp
DMU
P.O. Box 358
DK-4000 ROSKILDE
Denmark
Tel: 45 46 301164
Fax: 45 46 301214

Maria Lopes
Instituto de Meteorologia
Rua C – Aeroporto
1700 LISBOA
Portugal
Tel: 351 1 8483961
Fax: 351 1 8402370
Email: amelia.lopes@meteo.pt

Dr Peter Korkutis
Lithuanian Hydrometeorological Service
6 Rudmios Street
2600 VILNIUS
Lithuania

Mrs Iraida Lyulko
Latvian Hydrometeorological Agency
Environmental Pollution Observation Centre
165 Maskavas
RIGA LV-1019
Latvia
Tel: 371 7 113274
Fax: 371 7 145154
Email: epoc@meteo.lv

Antti Kulmala
Finnish Meteorological Institute
Air Quality Research
Sahaajankatu 20 E
00810 HELSINKI
Finland
Tel: 358 9 19291
Fax: 358 9 19295403
Email:antti.kulmala@fmi.fi

Mrs Laura Manea
National Institute of Meteorology and Hydrology
Sos. Bucuresti-Ploiesti 97
71552 BUCHAREST
Romania
Tel: 401 230 3116
Fax: 401 230 3143
Email: manea@meteo.inmh.ro
Ole Manscher  
National Environmental Research Institute  
P.O. Box 358  
DK-4000 ROSKILDE  
Denmark  
Tel: 45 4630 1164  
Fax: 45 4630 1214  
Email: Ohm@dmu.dk

Sergei Paramonov  
Institute of Global Climate and Ecology  
Glebovskaya Str. 20-B  
107258 MОСКОW  
Russian Federation  
Tel: 7 095 1605867  
Fax: 7 095 1600831

John Miller  
WMO Secretariat Environment Division  
7bis avenue de la Paix  
Case Postale No 2300  
CH-1211 GENEVA  
Switzerland  
Tel: 41 22 7308240  
Fax: 41 22 7400984  
Email: millerjm@wmo.ch

Kleanthis L. Philaniotis  
Director  
Meteorological Service  
CY-1418 NICOSIA  
Cyprus  
Tel: 357 2 902377  
Fax: 357 2 304149  
Email: roc-mete@cytanet.com.cy

Marta Mitosinkova  
Slovak Hydrometeorological Institute,  
Department of Atmospheric Chemistry  
Jesenlova 17  
SK-83315 BRATISLAVA  
Slovak Republic  
Tel: 421 7 276196  
Fax: 421 7 375670

Pila Ripodas  
Instituto Nacional Meteorologia  
Izana GAW Station  
C/San Sebastian 77  
38071 SANTA CRUZ DE TENERIFE  
Spain  
Tel: 34 922 211718  
Fax: 34 922 225248  
Email:pripodas@inm.es

Dr Grazyna Mitosek  
Institute of Environmental Protection  
Air Protection Division  
Ul. Kolektorska 4  
01-692 WARSAW  
Poland  
Tel: 48 22 33 8037  
Fax: 48 22 33 8507/6928

Dr Ott Roots  
Ministry of Environment  
Mustamae tee 33  
10616 TALLIN  
Estonia  
Tel: 372 6 565373  
Fax: 372 6 564071  
Email:roots@ic.envir.ee

Volker Mohnen  
Fraunhofer Institute for  
Atmospheric Environmental Research (IFU)  
Kreuzzeichbahnstrasse 19  
82467 GARMINSCHE-PARTENKIRCHEN  
Germany  
Tel: 49 88 2183263  
Fax: 49 88 2173573  
Email: mohnen@ifu.fhg.de

Tuija Ruoho-Airola  
Finnish Meteorological Institute  
Air Quality Research  
P.O. Box 503  
00101 HELSINKI  
Finland  
Tel: 358 9 19291  
Fax: 358 9 19293146  
Email: tuija.ruoho-airola@fmi.fi
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 Vole 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 235182

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

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

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
AGENDA

Monday, 14 September

14.00-14.30 OPENING OF THE MEETING

Welcome  Jatila
WMO J. M. Miller
EMEP A. Eliassen

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 Interpretative software that 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. Strols
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

09.00-10.00 Time series analysis - A. Sirois - or someone from NILU
Discussion
10.00-11.00 Spatial interpolation analysis - J. Schaug
Discussion
11.00-11.30 Coffee
11.30-12.30 Evaluation of kriged and modelled concentrations - Tarrason
Discussion
12.30-14.00 Lunch

14.00-15.30 DISCUSSION AND NATIONAL PRESENTATIONS

Question period concerning lectures, countries' experience and short communications - conclusions
15.30-16.00 Coffee
16.00-17.30 Visit to FMI, Air Quality Research
20.00- Workshop dinner hosted by the Ministry of the Environment

Thursday, 17 September

09.00-12.30 LECTURES

09.00-10.00 POPs and heavy metal data and analysis at MSC-E - A. Ryaboshapko
Discussion
10.00-11.00 HY split trajectory model and clustering – R. Draxler
Discussion
11.00-11.30 Coffee
11.30-12.30 Tropospheric ozone data analysis - Ø Hov
Discussion
12.30-14.30 Lunch
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 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

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
3.2 Technical Details 38
3.3 Example of Absolute Principal Component Analysis 39
3.4 Remarks 40
3.5 Computer Software 41
3.6 References 41

4. Target Principal Component Analysis (TPCA) 43
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 Communality 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
7.2.2 Algorithm for PMF 80
7.2.3 Error Estimation 81
7.2.4 Missing Data 82
7.2.5 Robust Analysis 82
7.2.6 Lognormal Data 82
7.2.7 Normalization of Results 82
7.2.8 Uniqueness of Solution and Rotation 83
7.3 Example of Positive Matrix Factorization 83
7.4 Remarks 84
7.5 Computer Software 85
7.6 References 85

8. Rotation 87
8.1 Introduction 87
8.2 Orthogonal Rotations 87
  8.2.1 The Quartimax Method 88
  8.2.2 The Varimax Method 88
  8.2.3 The Equamax Method 89
  8.2.4 The McCanmon Method 89
8.3 Oblique Rotations 89
  8.3.1 The Oblimax Method 90
  8.3.2 The Quartimin Method 90
  8.3.3 The Covarimin Method 90
  8.3.4 The Biquartimin Method 90
  8.3.5 The Oblimix Method 91
  8.3.6 The Binormamin Method 91
  8.3.7 The Promax Method 91
  8.3.8 The Crawford-Ferguson Method 92
8.4 Computer Software 92
8.5 References 92

9. Nonparametric Confidence Intervals 93
9.1 Introduction 93
9.2 The Jackknife Technique 93
9.3 The Bootstrap Technique 95
9.4 Realignment of Loadings Matrices 95
9.5 Examples 96
9.6 Computer Software 97
9.7 References 97

10. Examples of the Use of Dimensionality Reduction Techniques 99
  10.1 Arctic Lower Tropospheric Aerosol at Alert, Canada 99
  10.2 Sulphate Concentration in precipitation in Eastern North America 107

11. Short Guideline to the Use of Dimensionality Reduction Techniques 117
  11.1 Research Goals 117
  11.2 Data Selection 117
  11.3 Data Preprocessing 118
  11.4 Selection of the Method to Use 118
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.

![Diagram](image)

**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\(^-\), Mg\(^{2+}\), Ca\(^{2+}\) and also some contribution to SO\(_2\)\(^2-\) ionic concentration; and b) an anthropogenic factor which may influence SO\(_2\)\(^2-\), NH\(_4\)\(^+\), and NO\(_3\)\(^-\) 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:

\[ \mathbf{x} = \mu + \Lambda \mathbf{f} + \mathbf{e} \]  \hspace{1cm} (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 \textit{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 \( X_1, X_2, \ldots, X_p \); the first component \((P_1)\) is written as:

\[
P_1 = a_{(1)1}X_1 + a_{(1)2}X_2 + \ldots + a_{(1)p}X_p
\]

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

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.

\[ P = A^T(X - \mu) \]  

(2.2)

where \( P, A = (a_{1}, a_{2}, \ldots, a_{p}) \); \( a_{0} = (a_{01}, a_{02}, \ldots, a_{0p})^T \); \( X = (X_{1}, X_{2}, \ldots, X_{p})^T \) and \( \mu = (\mu_{1}, \mu_{2}, \ldots, \mu_{p})^T \) are px1, (p)x(p), (p)x1 and (p)x1 matrices respectively. \( \mu \) is the matrix of variable means. As \( A \) is an orthogonal matrix, we have \( A^{-1} = A^T \). Therefore, we can write:

\[ X = \mu + AP \]  

(2.3)
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 is 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 $X = [X_1, X_2, \ldots, X_p]$ 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:

$$\text{var}(P_1) = \text{var}(a_{(1)}^T X)$$

$$= a_{(1)}^T \Sigma a_{(1)}$$

(2.4)

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:

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

(2.5)

We write therefore:

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

(2.6)

which can be written as:

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

(2.7)

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:

$$\det(\Sigma - \lambda I) = 0$$

(2.8)

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:

$$\text{var}(a_{(1)}^T X) = a_{(1)}^T \Sigma a_{(1)}$$

$$= a_{(1)}^T \lambda I a_{(1)}$$

$$= \lambda$$

(2.9)

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 $P_2 = a_{(2)}^T X$ such that $a_{(2)}^T a^2 = 1$ and $P_2$ is uncorrelated to $P_1$. We have that:
\[
\text{cov}(P_2, P_1) = \text{cov}(a_{(2)}^T X, a_{(1)}^T X)
\]

\[
= E[a_{(2)}^T (X - \mu)(X - \mu)^T a_{(1)}]
\]

\[
= a_{(2)}^T \Sigma a_{(1)}
\] (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:

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

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

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

If we multiply relation 2.12 by \( a_{(2)}^T \), we obtain:

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

but from relation 2.10, we also require that \( a_{(1)}^T \Sigma a_{(2)} \) equals zero. Therefore, \( \delta \) must also be zero and relation 2.12 becomes:

\[
(\Sigma - \lambda \mathbf{I}) a_2 = 0
\] (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:

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

Using the fact that:

\[
\text{var}(\mathbf{P}) = \Lambda^T \Sigma \Lambda
\] (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:

$$\text{trace}(\Lambda) = \text{trace}(A^T \Sigma A)$$

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

$$= \text{trace}(\Sigma)$$

$$= \sum_{i=1}^{p} \text{var}(X_i)$$  \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 $i$th principal component accounts for a proportion $\frac{\lambda_i}{\sum_{j=1}^{p} \lambda_j}$ 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 $(n \times p)$ data matrix as:

$$X_{(n \times p)} = \begin{pmatrix}
X_{11} & X_{12} & \cdots & X_{1p} \\
X_{21} & X_{22} & \cdots & X_{2p} \\
\vdots & \vdots & \ddots & \vdots \\
X_{n1} & X_{n2} & \cdots & X_{np}
\end{pmatrix}$$  \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:
\[ M = \begin{pmatrix}
X_1 & X_2 & \ldots & X_p \\
X_1 & X_2 & \ldots & X_p \\
\vdots & \vdots & & \vdots \\
X_1 & X_2 & \ldots & X_p
\end{pmatrix} \]  

(2.23)

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:

\[ \det(S - \ell I) = 0 \]  

(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:

\[ (S - \ell_{(i)} I)a_{(i)} = 0 \]  

(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:

\[ A = \begin{pmatrix}
a_{(1)1} & a_{(2)1} & \ldots & a_{(p)1} \\
a_{(1)2} & a_{(2)2} & \ldots & a_{(p)2} \\
\vdots & \vdots & & \vdots \\
a_{(1)p} & a_{(2)p} & \ldots & a_{(p)p}
\end{pmatrix} \]  

(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 \( \text{(SVD)} \), 
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:

\[
X - M = ULV^T
\]  
(2.29)

where :

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

\[
A = V
\]  
(2.30)

and

\[
P = UL
\]

The eigenvalue \( \ell_{ij} \) is equal to \( l_{ii}/(n-1) \) (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:

\[
X^* = (X - M)D^{-1/2}
\]  
(2.31)

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

$\textbf{R} = \frac{1}{n-1} \textbf{X}^* \textbf{X}^*$

$= \frac{1}{n-1} ((\textbf{X} - \textbf{M})\textbf{D}^{-\frac{1}{2}})^T (\textbf{X} - \textbf{M})\textbf{D}^{-\frac{1}{2}}$

$= \frac{1}{n-1} \textbf{D}^{-\frac{1}{2}} (\textbf{X} - \textbf{M})^T (\textbf{X} - \textbf{M})\textbf{D}^{-\frac{1}{2}}$

$= \textbf{D}^{-\frac{1}{2}} \textbf{S} \textbf{D}^{-\frac{1}{2}}$  \hspace{1cm} (2.33)

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

$\text{det}(\textbf{R} - \ell \textbf{I}) = 0$  \hspace{1cm} (2.34)

and

$(\textbf{R} - \ell (i) \textbf{I}) \textbf{a}_{(i)} = 0$  \hspace{1cm} (2.35)

The scores matrix $\textbf{P}$ is now equal to:

$\textbf{P} = \textbf{X}^* \textbf{A} = (\textbf{X} - \textbf{M})\textbf{D}^{-\frac{1}{2}} \textbf{A}$  \hspace{1cm} (2.36)

and relation 2.24 becomes:

$\textbf{X} = \textbf{M} + \textbf{P} \textbf{A}^T \textbf{D}^{\frac{1}{2}}$  \hspace{1cm} (2.37)

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

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

$\sum_{i=1}^{p} \ell_{(i)} = p$  \hspace{1cm} (2.38)

It is important to note that the principal components obtained using the correlation matrix $\textbf{R}$ are different from those obtained using the variance-covariance matrix $\textbf{S}$. 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_j by the mean value, \bar{x}_j, of the observations for which the value of x_j 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 distributions, 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:

\[
\frac{\ell^{(i)}_1 + \sqrt{2/ \frac{n}{n-1}} \ Z_{\alpha/2}}{1 + \sqrt{2/ \frac{n}{n-1}} \ Z_{\alpha/2}} \leq \lambda_i \leq \frac{\ell^{(i)}_n - \sqrt{2/ \frac{n}{n-1}} \ Z_{\alpha/2}}{1 - \sqrt{2/ \frac{n}{n-1}} \ Z_{\alpha/2}}
\]  

(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):

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

(2.40)

or for \( \lambda_i \) :

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

(2.41)
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:

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

(2.42)

where \(q = p - r\) and

\[
-\ell_q = \frac{1}{q} \sum_{i=r+1}^{p} \ell_{(i)}
\]  

(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:

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

(2.44)

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

(2.45)

and

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

(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, \(\alpha_{ij}\), of the loadings matrix is:

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

(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

16
Here we want to test:

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

\[ H_1: \text{not all equal} \]  

(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

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

(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

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

(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:

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

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

(2.51)

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

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

(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:

\[
\chi^2 = \left[ n - r - \frac{1}{6q} (2q^2 + q + 2) + \sum_{i=1}^{q} \frac{(\bar{\ell}_q)^2}{(\ell_{(i)} - \bar{\ell}_q)^2} \right] \ln \left( \frac{1}{q} \sum_{i=1}^{q} \ell_{(i)} \right) - q \ln \left( \frac{1}{q} \sum_{i=1}^{q} \ell_{(i)} \right)
\]

(2.53)

where:

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

(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

$$\Theta = \frac{\sum_{i=r+1}^{p} \lambda_i}{\sum_{i=1}^{p} \lambda_i}$$  \hspace{1cm} (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:

$$0 \leq \Theta \leq \frac{\sum_{i=r+1}^{p} \ell_{(i)}}{\sum_{i=1}^{p} \ell_{(i)}} + Z_{2\alpha} \sqrt{n\left(\sum_{i=1}^{l} \ell_{(i)} \right)^2 \left(\sum_{i=r+1}^{p} \ell_{(i)} \right)^2}$$  \hspace{1cm} (2.56)

where $Z_{2\alpha}$ is the upper 100(1-2\(\alpha\))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:

$$m \hat{X} = M + m P m A^T$$  \hspace{1cm} (2.57)

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

$$m E = X - m \hat{X}$$  \hspace{1cm} (2.58)

Therefore, $X$ can be written as:

$$X = M + m P m A^T + m E$$  \hspace{1cm} (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, $m E$ 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:

\[
\hat{X} = \hat{X}^* = mP_mA^T
\]

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

\[
\hat{X} = M + mP_mA^TD^{1/2}
\]

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:

\[
h_i^2 = \frac{\sum_{j=1}^{m} \ell_{ij}^2}{S_{ii}}
\]

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

\[
h_i^2 = \sum_{j=1}^{m} \ell_{ij}^2
\]

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>0.750</td>
<td>0.250</td>
<td>0.250</td>
<td>0.111</td>
<td>0.111</td>
<td>0.111</td>
<td>0.111</td>
<td>0.111</td>
<td>0.111</td>
</tr>
<tr>
<td>3</td>
<td>0.611</td>
<td>0.611</td>
<td>0.278</td>
<td>0.278</td>
<td>0.146</td>
<td>0.146</td>
<td>0.063</td>
<td>0.063</td>
<td>0.063</td>
<td>0.063</td>
</tr>
<tr>
<td>4</td>
<td>0.521</td>
<td>0.521</td>
<td>0.271</td>
<td>0.271</td>
<td>0.157</td>
<td>0.157</td>
<td>0.090</td>
<td>0.090</td>
<td>0.090</td>
<td>0.090</td>
</tr>
<tr>
<td>5</td>
<td>0.457</td>
<td>0.457</td>
<td>0.257</td>
<td>0.257</td>
<td>0.157</td>
<td>0.157</td>
<td>0.090</td>
<td>0.090</td>
<td>0.090</td>
<td>0.090</td>
</tr>
<tr>
<td>6</td>
<td>0.408</td>
<td>0.408</td>
<td>0.242</td>
<td>0.242</td>
<td>0.158</td>
<td>0.158</td>
<td>0.103</td>
<td>0.103</td>
<td>0.103</td>
<td>0.103</td>
</tr>
<tr>
<td>7</td>
<td>0.370</td>
<td>0.370</td>
<td>0.228</td>
<td>0.228</td>
<td>0.156</td>
<td>0.156</td>
<td>0.109</td>
<td>0.109</td>
<td>0.109</td>
<td>0.109</td>
</tr>
<tr>
<td>8</td>
<td>0.340</td>
<td>0.340</td>
<td>0.215</td>
<td>0.215</td>
<td>0.152</td>
<td>0.152</td>
<td>0.111</td>
<td>0.111</td>
<td>0.111</td>
<td>0.111</td>
</tr>
<tr>
<td>9</td>
<td>0.314</td>
<td>0.314</td>
<td>0.203</td>
<td>0.203</td>
<td>0.148</td>
<td>0.148</td>
<td>0.111</td>
<td>0.111</td>
<td>0.111</td>
<td>0.111</td>
</tr>
<tr>
<td>10</td>
<td>0.293</td>
<td>0.293</td>
<td>0.193</td>
<td>0.193</td>
<td>0.143</td>
<td>0.143</td>
<td>0.110</td>
<td>0.110</td>
<td>0.110</td>
<td>0.110</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:

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

(2.64)

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, $h_i^2$, 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 than the mean of the eigenvalues, $\bar{\ell} = \sum_{i=1}^{p} \ell_{(i)}$. 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 kth longest segment is:

\[ \ell^*_k = \frac{1}{p} \sum_{j=k}^{\infty} \frac{1}{j} \]  

(2.65)

One way of deciding whether the proportion of variance accounted for by the kth 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 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 \( \log(\ell_{ij}) \), rather than \( \ell_{ij} \) 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 even 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.

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 $\hat{X}$ 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:

$$v_m = \left( \frac{\sum_{i=1}^{n} \sum_{j=1}^{p} (x_{ij}^* - \bar{x}_j^*)^2}{\sum_{i=1}^{n} \sum_{j=1}^{p} (x_{ij}^* - \bar{x}^*)^2} \right)^{1/2}$$ (2.66)

where $x_{ij}^*$ and $\bar{x}_j^*$ are the elements of the standardized matrices $X^*$ and $\hat{X}^*$ 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., \( \hat{X}_m \)). Then the predicted sum of squares

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

(2.67)

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

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

(2.68)

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 \( \hat{x}_{ij} \) 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:

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

(2.69)

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

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

(2.70)

A similar relation exists for the principal components extracted from the correlation matrix. Let
\[ R' = D^{-1/2} E^T R E D^{-1/2} \]  
(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

\[ f_r = \sum_{i=1}^{p} \sum_{j=1}^{p} \frac{(r_{ij}')^2}{p(p-1)} \]  
(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 \( TT^T = T^T T = I \). Therefore, relation 2.57 becomes:

\[ \hat{X}_m = M + m P T T^T m A^T \]
\[ = M + m P T (m A T)^T \]
\[ = M + m P^R (m A^R)^T \]  
(2.73)
where \( P^R = P^T \) and \( A^R = A^T \). Therefore, the orthogonal rotation \( T \) does not change the values of \( \hat{X} \) 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, \( A^R \) is still orthogonal and \( (A^R)^T A^R = I \). However, the scores matrix, \( P^R \), is no longer orthogonal; that is, \( (P^R)^T P^R \) is not a diagonal matrix. To get an orthogonal matrix for \( P^R \), we have to replace the loadings matrix, \( A \), by \( A\Lambda^{1/2} \) 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 \( P^R \) by \( P^R\Lambda^{-1/2} \). However, in that case the new rotated loadings matrix \( A^R \) is not orthogonal anymore. In summary, one cannot have both \( A^R \) and \( P^R \) 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 \( P^R = P^T \) and \( A^R = A^T \).

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

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

(ii) \( \hat{X}^T \hat{X} = (A^R)^T A^R (A^R)^T \).

(iii) \( (A^R)^T = (A^R)^T (A^R)^{-1} (A^R)^T \).

(iv) \( \hat{X} = A^R (A^R)^T \).

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 nothing 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 (λ_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 (λ_1 - λ_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>Loadings</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Total population</td>
<td>0.7812</td>
<td>0.0709</td>
<td>0.0037</td>
<td>-0.5417</td>
<td>-0.3020</td>
</tr>
<tr>
<td>Median school years</td>
<td>0.3056</td>
<td>0.7639</td>
<td>-0.1618</td>
<td>0.5448</td>
<td>-0.0093</td>
</tr>
<tr>
<td>Total employment</td>
<td>0.3344</td>
<td>-0.0829</td>
<td>0.0148</td>
<td>-0.0510</td>
<td>0.9373</td>
</tr>
<tr>
<td>Health services</td>
<td>0.4260</td>
<td>-0.5795</td>
<td>0.2205</td>
<td>0.6360</td>
<td>-0.1721</td>
</tr>
<tr>
<td>Median home value</td>
<td>-0.0543</td>
<td>0.2624</td>
<td>0.9617</td>
<td>-0.0513</td>
<td>0.0245</td>
</tr>
</tbody>
</table>

\[
R = \begin{pmatrix}
1.00 & 0.610 & 0.971 & 0.740 & -0.172 \\
0.610 & 1.000 & 0.494 & 0.095 & 0.186 \\
0.971 & 0.494 & 1.000 & 0.848 & -0.249 \\
0.740 & 0.095 & 0.848 & 1.000 & -0.358 \\
-0.172 & 0.186 & -0.249 & -0.358 & 1.000
\end{pmatrix}
\]

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></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>3.0289</td>
<td>1.2911</td>
<td>0.5725</td>
<td>0.0954</td>
<td>0.0121</td>
</tr>
<tr>
<td>Difference ($\ell_{k-1} - \ell_k$)</td>
<td>1.7378</td>
<td>0.7186</td>
<td>0.4771</td>
<td>0.0833</td>
<td></td>
</tr>
<tr>
<td>Proportion</td>
<td>0.6058</td>
<td>0.2582</td>
<td>0.1145</td>
<td>0.0191</td>
<td>0.0024</td>
</tr>
<tr>
<td>Cumulative Proportion</td>
<td>0.6058</td>
<td>0.8640</td>
<td>0.9785</td>
<td>0.9976</td>
<td>1.0000</td>
</tr>
</tbody>
</table>

<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 home value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Loadings</td>
<td>0.5584</td>
<td>0.1314</td>
<td>0.0079</td>
<td>-0.5506</td>
<td>-0.6065</td>
</tr>
<tr>
<td></td>
<td>0.3133</td>
<td>0.6289</td>
<td>-0.5490</td>
<td>0.4526</td>
<td>0.0067</td>
</tr>
<tr>
<td></td>
<td>0.5683</td>
<td>0.0043</td>
<td>0.1173</td>
<td>-0.2681</td>
<td>0.7690</td>
</tr>
<tr>
<td></td>
<td>0.4866</td>
<td>-0.3096</td>
<td>0.4549</td>
<td>0.6480</td>
<td>-0.2013</td>
</tr>
<tr>
<td></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, $h_i^2$, for the five observed variables with the number of components retained for the two types of principal components.

<table>
<thead>
<tr>
<th></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 $\hat{\lambda}$ 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 Joliffe (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 $o_{\lambda}^{1/2}$ 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>Eigenvalue ($\ell_k$)</td>
<td>6.9311</td>
<td>1.7851</td>
</tr>
<tr>
<td>Difference ($\ell_{k-1} - \ell_k$)</td>
<td>5.1460</td>
<td>1.3955</td>
</tr>
<tr>
<td>Proportion</td>
<td>0.7413</td>
<td>0.1909</td>
</tr>
<tr>
<td>Cumulative Proportion</td>
<td>0.7413</td>
<td>0.9323</td>
</tr>
<tr>
<td>Loadings</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total population</td>
<td>2.0567</td>
<td>0.0947</td>
</tr>
<tr>
<td>Median school years</td>
<td>0.8047</td>
<td>1.0206</td>
</tr>
<tr>
<td>Total employment</td>
<td>0.8805</td>
<td>-0.1108</td>
</tr>
<tr>
<td>Health services employment</td>
<td>1.1215</td>
<td>-0.7742</td>
</tr>
<tr>
<td>Median home value</td>
<td>-0.1431</td>
<td>0.3505</td>
</tr>
</tbody>
</table>

Component 1 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>Eigenvalue ($\ell_k$)</td>
<td>3.0289</td>
<td>1.2911</td>
</tr>
<tr>
<td>Difference ($\ell_{k-1} - \ell_k$)</td>
<td>1.7378</td>
<td>0.7186</td>
</tr>
<tr>
<td>Proportion</td>
<td>0.6058</td>
<td>0.2582</td>
</tr>
<tr>
<td>Cumulative Proportion</td>
<td>0.6058</td>
<td>0.8640</td>
</tr>
<tr>
<td>Loadings</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total population</td>
<td>0.9718</td>
<td>0.1493</td>
</tr>
<tr>
<td>Median school years</td>
<td>0.5452</td>
<td>0.7148</td>
</tr>
<tr>
<td>Total employment</td>
<td>0.9890</td>
<td>0.0048</td>
</tr>
<tr>
<td>Health services employment</td>
<td>0.8469</td>
<td>-0.3517</td>
</tr>
<tr>
<td>Median home value</td>
<td>-0.3033</td>
<td>0.7965</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.
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.

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 2.5 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 (n x p) matrices. There are many different ways to do that. Examples are the two-stage principal components method (see Bourroche and Dussaux, 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 Flitnoff (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:

\[ C = AS \]  \hspace{1cm} (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 case 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 $m P^R$ and $m A^R$ in relation 2.73.

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

(3.2)

and for the principal component from a correlation matrix:

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

(3.3)

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

The \((1 \times m)\) scores matrix, \( m \mathbf{P}_0^R \), of the *absolute zeros* of the observed variables can be obtained by replacing \( m \hat{\mathbf{X}}^R \) and \( \mathbf{M} \) in relations 3.2 or 3.3 by the following two \((1 \times p)\) matrices:

\[ m \hat{\mathbf{X}}^R = \begin{pmatrix} 0 & 0 & \cdots & 0 \end{pmatrix} \]  

(3.4)

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

(3.5)

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

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

(3.6)

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

\[ m [\mathbf{AP}]^R = m \mathbf{P}^R - m \mathbf{\Pi}^R \]  

(3.7)

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

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

(3.8)

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

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

(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:

\[ C_k = \alpha_0 + \sum_{j=1}^{m} \beta_j \_{m}[\text{AP}]_j^R \] (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 \( \_{m}[\text{AP}]_j^R \) by least squares, one will obtain:

\[ \alpha_0 = [\text{M}+_{k} \Pi^R (_{m}A^R)^T]_k \] (3.11)

and

\[ \beta_j = \left( _{m}A^R \right)^T \] (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 \( \_{m}[\text{AP}]_j^R \), which is not the

Figure 3.2 Scattergrams of absolute scores for the first three principal components of example in Section 2.3 extracted from the correlation matrix and after a varimax rotation.
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, \( m_{jk} \), using the relation:

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

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

5. Regress the concentrations, \( C_k \), on \( S_j = (\xi_j m_{jk}) \).

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 estimate these 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.
Figure 4.1 Illustration of principal components in 2 dimension: (a) observed variables; (b) principal components using non-centered correlation matrix.

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 *a priori* loadings matrix of the factor.
7. Scale the results if the total particulate masses are available.

4.2 Technical Details

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 *absolute zero* is:

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

(4.1)

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

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

(4.2)

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

$$ v_{ii} = \left( \frac{1}{p-1} \sum_{j=1}^{p} x_{ij}^2 \right)^{1/2} $$

(4.3)

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

$$ A_Q = X^T P_Q $$

(4.4)

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 $m$ ($< p$) components are retained.

If the correlation matrix is used, relation 4.4 becomes:

$$ A_Q = (V_Q^{-1} X)^T P_Q $$

(4.5)

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

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

(4.6)

if $S_Q$ is used, and to

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

(4.7)

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:
\[
L_{(m \times w)} = \begin{bmatrix}
\sum_{i=1}^{p} a_{1i}^2 & 0 & \cdots & 0 \\
0 & \sum_{i=1}^{p} a_{2i}^2 & \cdots & 0 \\
\vdots & \vdots & \ddots & \vdots \\
0 & 0 & \cdots & \sum_{i=1}^{p} a_{wi}^2
\end{bmatrix}
\]  

(4.8)

where the \(a_{ij}\) are the elements of matrix \(m \times A_Q\).

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:

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

(4.9)

and

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

(4.10)

respectively, where \(V_R\) is a (p \(\times p\)) diagonal matrix with elements given by:

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

(4.11)

If \(S_R\) is used, relation 2.22 for the scores matrix, \(P\), becomes:

\[
P_R = X A_R
\]  

(4.12)

where \(A_R\) is the standardized loadings matrix. If \(R_R\) is used, one gets:

\[
P_R = (X V_R^{-1}) A_R
\]  

(4.13)

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

\[
m \hat{X}_R = m P_R (m A_R)^T
\]  

(4.14)

and

\[
m \hat{X}_R = m P_R (m A_R)^T V_R
\]  

(4.15)

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:

\[
X = TD^T S^T V_R^* = V_Q^* TDS^T
\]  

(4.16)
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 $X^TX$ matrix, and $D$ is the $(p \times n)$ diagonal matrix containing the square roots of the first $n$ eigenvectors obtained from the $X^TX$ 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, $A$, 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, $A$, with the input test vector, $b$, by least squares maximization of the overlap between a rotated axis of $A$ and the test vectors.

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

$$\varepsilon = Ar - b \quad (4.17)$$

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

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

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

$$= r^TA^TAr - r^TA^Tb - b^TAr + b^Tb \quad (4.18)$$

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

$$\varepsilon^2 = r^TA^TAr - 2r^TA^Tb + b^Tb \quad (4.19)$$

Taking the derivative and setting it equal to 0 gives:

$$\frac{\partial \varepsilon^2}{\partial r} = 0 = 2A^TAr - 2A^Tb \quad (4.20)$$

Thus, one has:

$$r = (A^TA)^{-1}A^Tb \quad (4.21)$$

It has been found that using weighted least squares greatly enhances the ability of the analysis to resolve sources with similar concentration profiles (Rosco and Hopke, 1981b). The weighted target transformation rotation is given by:
\[ r = (A^T W A)^{-1} A^T W b \]  \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

\[ b' = A r \]  \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 \( 10^4 \), 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 \( m A^R \) and \( R_T \) matrices such that:

\[ m A^R = m A R_T. \]  \hspace{1cm} (4.24)

An important characteristic of the loadings matrix, \( m A^R \), that is thus obtained is that it no longer has to be orthogonal, as the rotation, \( R_T \), 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:

\[ m P^R = m P (R_T^T)^{-1} \]  \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 \( m P^R \). They used the final matrix of source profiles, \( m A^R \), and the weighting matrix, \( W \), used in relation 4.22, to calculate least squares fits to the columns of the observed data matrix \( X \). \( m P^R \) is equal to:

\[ m P^R = (m A^R)^T W m A^R)^{-1} (m A^R)^T WX \]  \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 \( (m P^R)_j \) 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></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>4.6162</td>
<td>0.3043</td>
<td>0.0572</td>
<td>0.0203</td>
<td>0.0021</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>
<td></td>
</tr>
<tr>
<td>Proportion</td>
<td>0.9232</td>
<td>0.0609</td>
<td>0.0114</td>
<td>0.0041</td>
<td>0.0004</td>
</tr>
<tr>
<td>Cumulative Proportion</td>
<td>0.9232</td>
<td>0.9841</td>
<td>0.9955</td>
<td>0.9996</td>
<td>1.0000</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Loadings</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Total population</td>
<td>0.4569</td>
<td>-0.2178</td>
<td>-0.5915</td>
<td>-0.2507</td>
<td>-0.5754</td>
</tr>
<tr>
<td>Median school years</td>
<td>0.4524</td>
<td>0.3714</td>
<td>-0.0482</td>
<td>0.8505</td>
<td>-0.0827</td>
</tr>
<tr>
<td>Total employment</td>
<td>0.4592</td>
<td>-0.2588</td>
<td>-0.2916</td>
<td>-0.0745</td>
<td>0.7947</td>
</tr>
<tr>
<td>Health services</td>
<td>0.4371</td>
<td>-0.5449</td>
<td>0.6938</td>
<td>0.0294</td>
<td>-0.1743</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 Procrustes rotation. This name come 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:

$$M_j = \sum_{k=1}^{m} s_k (m^R P_k)_j.$$  \hfill (4.27)

or

$$M_j = s_0 + \sum_{k=1}^{m} s_k (m^R P_k)_j.$$  \hfill (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.
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>

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:

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

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 4.3 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>NO₂⁻</th>
<th>SO₄²⁻</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nov. 3</td>
<td>0.60</td>
<td>4.80</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>3.00</td>
<td>2.80</td>
<td>0.10</td>
<td>0.50</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>

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></th>
<th>Component</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
</tr>
<tr>
<td>Eigenvalues($\eta_k$)</td>
<td>7.557</td>
</tr>
<tr>
<td>Difference($\eta_{k-1}$-$\eta_k$)</td>
<td>6.014</td>
</tr>
<tr>
<td>Proportion</td>
<td>0.687</td>
</tr>
<tr>
<td>Cumulative Proportion</td>
<td>0.687</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Loadings</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pb</td>
<td>0.343</td>
</tr>
<tr>
<td>Fe</td>
<td>0.319</td>
</tr>
<tr>
<td>Ca</td>
<td>0.339</td>
</tr>
<tr>
<td>K</td>
<td>0.190</td>
</tr>
<tr>
<td>Si</td>
<td>0.323</td>
</tr>
<tr>
<td>Al</td>
<td>0.326</td>
</tr>
<tr>
<td>Na$^+$</td>
<td>0.287</td>
</tr>
<tr>
<td>NH$_4^+$</td>
<td>0.306</td>
</tr>
<tr>
<td>Cl$^-$</td>
<td>0.294</td>
</tr>
<tr>
<td>NO$_3^-$</td>
<td>0.329</td>
</tr>
<tr>
<td>SO$_4^{2-}$</td>
<td>0.219</td>
</tr>
</tbody>
</table>

\[
\begin{bmatrix}
1 & 0.358 & 1 \\
0.153 & 0.204 & 1 \\
0.182 & 0.260 & 0.215 & 1
\end{bmatrix}
\]

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., \( Y^T Y \) is diagonal). If matrix \( B \) is scaled such that \( B^T RB = I \), where \( R \) is the correlation matrix = \( X^T X \), then \( Y^T Y = B^T X^T XB = B^T RB = I \).

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

\[ R_{xy} = X^T Y = X^T XB = RB \]  
(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 \( r_{xy} \)) and are such that the relationship \( Y^T Y = I \) 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):

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

(6.1)

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 common factors; and \( \mathbf{e} \) is a \( p \)-vector of unobservable 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, \( \Delta \), 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, \( \Phi \), 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 factor pattern, and the loadings matrix, \( \Lambda \), can be referred to as the 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 factor structure matrix, or more briefly, a 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 \( j \)th common factors and the \( j \)th 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):

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

(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:
\[ S = \Lambda \Lambda^T + \Delta \]  \hspace{1cm} (6.4)

where \( S \) is the \((p \times p)\) variance-covariance matrix (relation 2.25). Fitting the model involves finding the \( \Lambda \) 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., \( T^T T = T T^T = I \)). We have then:

\[ (\Lambda T)(\Lambda T)^T = \Lambda T T^T \Lambda^T = \Lambda \Lambda^T. \]  \hspace{1cm} (6.5)

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

(6.6)

where

\(x_j\) = jth observed variable

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

\(e_i\) = ith unique or specific factor

\(f_j\) = jth common factor

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

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

(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:

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

(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:

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

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

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

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

(6.10)

where \(\Phi = \mathbf{f} \mathbf{f}^T\) 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:

\[
\Sigma = \Lambda \Lambda^T + \Delta.
\]

(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:

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

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

$$h_i^2 = \sum_{j=1}^{m} \lambda_{ij}^2$$

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:

$$\text{total variance} = tr(\Sigma) = \sum_{i=1}^{p} \sigma_{ii}$$

$$= \sum_{i=1}^{p} (h_i^2 + \delta_i^2)$$

$$= \sum_{i=1}^{p} \sum_{j=1}^{m} \lambda_{ij}^2 + \sum_{j=1}^{p} \delta_j^2$$

$$= V + \delta^2$$

where $V = \sum_{i=1}^{p} \sum_{j=1}^{m} \lambda_{ij}^2$ is the total communality and $\delta^2 = \sum_{j=1}^{p} \delta_j^2$.

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:

$$\Gamma = \Lambda \Lambda^T + \Delta$$

or

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

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

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

The $(p \times p)$ matrix $\Gamma^*\lambda$, 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 \( \lambda_i = 1 \) 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:

\[
\Sigma = ULU^T
\]  
\[
\text{(6.17)}
\]

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

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

\[
\text{(6.18)}
\]

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

\[
\Sigma = UL^{1/2}L^{1/2}U^T
\]

\[
= UL^{1/2}(UL^{1/2})^T
\]

\[
= UL^{1/2}(UL^{1/2})^T + 0
\]

\[
= L^*(L^*)^T + 0
\]

\[
\text{(6.19)}
\]

\[
\text{(6.20)}
\]

\[
\text{(6.21)}
\]
where \( \mathbf{L}^* = \mathbf{UL}^{1/2} \). Relation 6.21 is identical to 6.15 if one writes \( \Lambda = \mathbf{L}^* \) and \( \Delta = \mathbf{0} \). 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 \( \mathbf{e} \) 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:

\[
\Sigma = \mathbf{mL}^*(\mathbf{mL}^*)^T + \Delta
\]

where \( \Delta \) is a diagonal matrix with elements \( \delta^2_i = \sigma^2_i - \sum_{j=1}^{m} \ell^2_{ij} \) for \( i = 1, 2, \ldots, p \) and \( \ell_{ij} \) are the elements of \( \mathbf{mL}^* \). In this method, \( \hat{\Lambda} \) is estimated by \( \mathbf{mL}^* \); \( \hat{\Delta} \) by the diagonal matrix with elements equal to \( \delta^2_i = \sigma^2_i - \sum_{j=1}^{m} \ell^2_{ij} \).

The communalities are estimated as:

\[
\hat{h}_i^2 = \sum_{j=1}^{m} \ell^2_{ij}.
\]

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, \( \Gamma^* \), \( \mathbf{R}^* \). 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, \( \mathbf{R}^* \):

\[
\mathbf{R}^* = \begin{bmatrix}
\hat{h}_1^2 & r_{12} & \cdots & r_{1p} \\
r_{12} & \hat{h}_2^2 & \cdots & r_{2p} \\
\vdots & \vdots & \ddots & \vdots \\
r_{1p} & r_{2p} & \cdots & \hat{h}_p^2
\end{bmatrix}
\]

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

62
\[ \mathbf{R}^* \equiv \mathbf{L}_r \mathbf{L}_r^T \]  

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

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

4. \( \hat{h}_i^2 \) = the square of the multiple correlation coefficient of variable \( i \) with the other variables = \( 1 - \frac{1}{p_{ii}} \), where the \( p_{ii} \) are the elements of \( \mathbf{R}^* \).

63
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 $\mathbf{0}$ 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, $\Theta$, of $\Lambda$, $\Phi$, and $\Delta$ is equal to:

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

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

(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 $\Theta$. 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:

\[
\frac{\partial F}{\partial \Lambda} = 2 \Sigma^{-1} (\Sigma - S) \Sigma^{-1} \Lambda
\]  
(6.28)

and

\[
\frac{\partial F}{\partial \Delta} = \text{diag}(\Sigma^{-1} (\Sigma - S) \Sigma^{-1})
\]  
(6.29)

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 \( 1/2(p+1) \) distinct elements in \( \Sigma \). There are, in total, \( pm \) parameters in \( \Lambda \), \( 1/2m(m+1) \) parameters in \( \Phi \), and \( \Delta \) 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 others words, without some restrictions on the system there is a basic indeterminacy. Specifically, if \( T \) is any non-singular \((m \times m)\) matrix, then:

\[
F(\Lambda T^{-1}, T \Phi T^T, \Delta) = F(\Lambda, \Phi, \Delta)
\]  
(6.30)

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:

\[
\Phi = I
\]  
(6.31)

and

\[
\Lambda^T \Delta^{-1} \Lambda = \text{diag}(\Omega)
\]  
(6.32)

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 \( \Lambda^T \Delta^{-1} \Lambda = \Omega \).

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 \( \lambda^2_i / \delta^2_i \), where \( \lambda_{ij} \) 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 \( \sum_i \lambda^2_{ir} / \delta^2_i \). This is the \( r \)th diagonal element of \( \Lambda^T \Delta^{-1} \Lambda \). 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 $\Delta$ is found. The basic equation to solve is:

$$ S \Delta^{-1} \Lambda = \Lambda (I + \Omega) \quad (6.33) $$

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 $\partial F / \partial \Delta$ be found when evaluated at $\Lambda = \Lambda_0$. The solution can be expressed as:

$$ \text{diag}(\Lambda_0^T \Delta_0^T + \Delta - S) = 0 \quad (6.34) $$

or

$$ \Delta = \text{diag}(S - \Lambda_0 \Lambda_0^T) \quad (6.35) $$

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, $\Delta$, are known for the factor model (relation 6.1 or 6.6); i.e., the factor model has been solved:

$$ x - \mu = \Lambda f + e \quad (6.36) $$

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

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

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

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 \( \Lambda^T \Lambda = \Omega \), where \( \Omega \) is a diagonal matrix. In that case, relation 6.37 can be written as:

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

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

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

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, \( \hat{\mathbf{f}}_j' \), are related to \( \hat{\mathbf{f}}_j \) by \( \hat{\mathbf{f}}_j' = \mathbf{T}^T \hat{\mathbf{f}}_j \), \( j = 1, 2, \ldots, n \).

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 \( \delta_i^2 \) 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 \( \mathbf{f} \) and the specific factors \( \mathbf{e} \) are jointly normally distributed with means and covariances given in Section 6.2.1, the linear combination \( \mathbf{x} - \mu = \Lambda \mathbf{f} + \mathbf{e} \) has p-dimensional normal distribution with mean zero and covariance \( \Lambda \Lambda^T + \Delta \). Moreover, the joint distribution of \( (\mathbf{X} - \mu) \) and \( \mathbf{f} \) is a \( (m+p) \times p \)-dimensional normal distribution with mean \( \mathbf{0} \) (i.e., a \( (m+p) \times p \) matrix of zeros) and covariance matrix:

\[
\Sigma^* = \begin{bmatrix}
\Sigma = \Lambda \Lambda^T + \Delta & \Lambda \\
\Lambda & \mathbf{I}
\end{bmatrix}
\]

where \( \Sigma \) is a \( (p \times p) \) matrix, \( \Lambda \) is a \( (p \times m) \) matrix, \( \mathbf{I} \) is a \( (m \times m) \) identity matrix.
The conditional distribution of \( f \mid x \) is multivariate normal (see Johnson and Wichern, 1982) with:

\[
mean = E(f \mid x) = \Lambda^T \Delta^{-1} (x - \mu) = \Lambda^T (\Lambda \Lambda^T + \Delta)^{-1} (x - \mu)
\]

where \( E(\cdot) \) is the expected value of \( \cdot \); and

\[
covariance = \text{Cov}(f \mid x) = I - \Lambda^T \Delta^{-1} \Lambda = I - \Lambda^T (\Lambda \Lambda^T + \Delta)^{-1} \Lambda.
\]

The quantities \( \Lambda^T (\Lambda \Lambda^T + \Delta)^{-1} \) 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 \( \hat{\Delta} \) as the true values, the scores vector for the jth sample is given by:

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

\( j = 1, 2, \ldots, n \).

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

\[
\hat{\Lambda} (\hat{\Lambda} \hat{\Lambda}^T + \hat{\Delta})^{-1} = (I + \hat{\Lambda}^T \hat{\Delta}^{-1} \hat{\Lambda})^{-1} \hat{\Lambda}^T \hat{\Delta}^{-1}.
\]

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:

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

Therefore,

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

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

\[
\hat{f}_j^{LS} = (\hat{\Lambda} \hat{\Lambda}^T + \hat{\Delta}^{-1} \hat{\Lambda} \hat{\Lambda}^T \hat{\Delta}^{-1}) \hat{f}_j^R = (I + \hat{\Lambda} \hat{\Lambda}^T \hat{\Delta}^{-1} \hat{\Lambda}) \hat{f}_j^R
\]

For the maximum likelihood estimates, \( (\hat{\Lambda} \hat{\Lambda}^T \hat{\Delta}^{-1})^{-1} = \hat{\Omega}^{-1} \) 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 \( \hat{\Sigma} = \hat{\Lambda} \hat{\Lambda}^T + \hat{\Delta} \). The factor scores are then given by:

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

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

(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 \( \hat{f}_j' = T^T \hat{f}_j \).

### 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, \( \Gamma = \Lambda \Lambda^T + \Delta \), and testing the adequacy of the \( m \) common factor model is equivalent to testing:

\[
H_0: \quad \Sigma = \Lambda \Lambda^T + \Delta
\]

(6.50)

versus

\[
H_1: \quad \Sigma \text{ any other positive definite matrix}.
\]

(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:

\[
|\hat{S}_n|^{-n/2} e^{-np/2}
\]

(6.52)

where \( \hat{S} = S_n = [(n-1)/n]S \) 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:

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

69
\[ = |\hat{\Lambda} \hat{\Lambda}^T + \hat{\Delta}|^{-n/2} e^{-\frac{1}{2} \text{trace}((\hat{\Lambda} \hat{\Lambda}^T + \hat{\Delta})^{-1}S_n)} . \]  

Equation (6.53)

The likelihood ratio statistic for testing \( H_0 \) is:

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

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

Equation (6.54)

with

\[ \nu - \nu_0 = \frac{1}{2} p(p+1) - [p(m+1) - \frac{1}{2} m(m-1)] = \frac{1}{2} [(p-m)^2 - p-m] \]  

Equation (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:

\[ U = n \ln \left( \frac{|\hat{\Lambda} \hat{\Lambda}^T + \hat{\Delta}|}{|S_n|} \right) . \]  

Equation (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:

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

Equation (6.57)

provided \( n \) and \( n-p \) are large. Since the number of degrees of freedom \( \nu = (p-m)^2-p-m \), must be positive, it follows that

\[ m < \frac{\nu}{2} (2p+1 - \sqrt{8p+1}) \]  

Equation (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 Maxell (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

![Pattern and Structure Loadings Diagram](image-url)
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>Specific variances</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Estimated Factor loadings</td>
<td></td>
<td>$\hat{\delta}_j^2$</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$F_1$ $F_2$ $F_3$ $F_4$</td>
<td>$F_1$ $F_2$ $F_3$ $F_4$</td>
<td></td>
<td>$\hat{\delta}_j^2$</td>
</tr>
<tr>
<td>1. 100-m run</td>
<td>0.695 0.239 -0.484 0.103</td>
<td>-0.090 0.341 0.830 -0.169</td>
<td>0.21</td>
<td>0.16</td>
</tr>
<tr>
<td>2. Long jump</td>
<td>0.750 0.149 -0.097 -0.184</td>
<td>0.065 0.433 0.595 0.275</td>
<td>0.37</td>
<td>0.38</td>
</tr>
<tr>
<td>3. Shot put</td>
<td>0.704 -0.544 0.058 0.181</td>
<td>-0.139 0.990 0.000 0.000</td>
<td>0.17</td>
<td>0.00</td>
</tr>
<tr>
<td>4. High jump</td>
<td>0.614 0.065 0.176 -0.322</td>
<td>0.156 0.406 0.336 0.445</td>
<td>0.48</td>
<td>0.50</td>
</tr>
<tr>
<td>5. 400-m run</td>
<td>0.629 0.565 -0.021 0.334</td>
<td>0.376 0.245 0.671 -0.137</td>
<td>0.17</td>
<td>0.33</td>
</tr>
<tr>
<td>6. 100-m hurdles</td>
<td>0.617 0.018 -0.043 -0.256</td>
<td>-0.021 0.361 0.425 0.388</td>
<td>0.55</td>
<td>0.54</td>
</tr>
<tr>
<td>7. Discuss</td>
<td>0.595 -0.474 0.106 0.224</td>
<td>-0.063 0.728 0.030 0.019</td>
<td>0.36</td>
<td>0.46</td>
</tr>
<tr>
<td>8. Pole vault</td>
<td>0.459 0.036 0.241 -0.213</td>
<td>0.155 0.264 0.229 0.394</td>
<td>0.68</td>
<td>0.70</td>
</tr>
<tr>
<td>9. Javelin</td>
<td>0.362 -0.258 0.145 0.077</td>
<td>-0.026 0.441 -0.010 0.098</td>
<td>0.78</td>
<td>0.80</td>
</tr>
<tr>
<td>10. 1500-m run</td>
<td>0.133 0.438 0.553 0.164</td>
<td>0.998 0.059 0.000 0.000</td>
<td>0.46</td>
<td>0.00</td>
</tr>
<tr>
<td>Cumulative proportion of total variance explained</td>
<td>0.34 0.36 0.53 0.58</td>
<td>0.12 0.37 0.55 0.61</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

\[
\begin{bmatrix}
100 \text{ m run} & Long jump & Shot put & High Jump & 400 \text{ m run} & 110 \text{ m hurdles} & dis - cuss & Pole vault & Jeve - lin & 1500 \text{ m run} \\
1.0 & 0.59 & 0.35 & 0.34 & 0.63 & 0.40 & 0.28 & 0.20 & 0.11 & -0.07 \\
1.0 & 0.42 & 0.51 & 0.49 & 0.52 & 0.31 & 0.36 & 0.21 & 0.09 \\
1.0 & 0.38 & 0.19 & 0.36 & 0.73 & 0.24 & 0.44 & -0.08 \\
1.0 & 0.29 & 0.46 & 0.27 & 0.39 & 0.17 & 0.18 \\
1.0 & 0.34 & 0.17 & 0.23 & 0.13 & 0.39 \\
1.0 & 0.32 & 0.33 & 0.18 & 0.00 \\
1.0 & 0.24 & 0.34 & -0.12 \\
1.0 & 0.24 & 0.17 \\
1.0 & 0.00 \\
1.0 &
\end{bmatrix}
\]

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, $\hat{\delta}_j^2$, 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.

73
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:

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

(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:

\[
\begin{bmatrix}
0.00 & 0.00 \\
0.05 & 0.00 \\
0.06 & 0.12 & 0.00 \\
0.16 & -0.02 & 0.01 & 0.00 \\
0.13 & -0.07 & -0.04 & -0.02 & 0.00 \\
-0.28 & 0.03 & -0.16 & 0.05 & 0.02 & 0.00 \\
0.08 & -0.14 & 0.06 & -0.11 & -0.09 & 0.23 & 0.00 \\
0.11 & -0.06 & -0.039 & -0.05 & -0.02 & 0.02 & 0.06 & 0.00 \\
-0.17 & 0.05 & 0.022 & -0.037 & 0.025 & -0.13 & -0.031 & 0.064 & 0.00 \\
-0.017 & 0.08 & 0.003 & 0.025 & 0.016 & -0.024 & 0.013 & -0.005 & -0.028 & 0.00
\end{bmatrix}
\]

The residual matrix for the maximum likelihood solution is:

\[
\begin{bmatrix}
0.00 & 0.00 \\
0.00 & 0.00 \\
0.00 & 0.00 & 0.00 \\
0.12 & 0.02 & 0.00 & 0.00 \\
0.00 & -0.02 & 0.00 & -0.033 & 0.00 \\
-0.12 & 0.06 & -0.00 & 0.001 & 0.028 & 0.00 \\
0.04 & -0.25 & -0.00 & -0.034 & -0.02 & 0.036 & 0.00 \\
0.00 & -0.09 & -0.00 & 0.006 & 0.08 & -0.012 & 0.043 & 0.00 \\
-0.018 & -0.00 & -0.00 & -0.045 & 0.052 & -0.13 & 0.016 & 0.091 & 0.00 \\
0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.013 & 0.000 & 0.000 & 0.00
\end{bmatrix}
\]

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>Principal Factor</th>
<th>Maximum likelihood</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Estimated Factor loadings</td>
<td>Specific variances</td>
</tr>
<tr>
<td></td>
<td>$F_1$</td>
<td>$F_2$</td>
</tr>
<tr>
<td>1. 100-m run</td>
<td>0.141</td>
<td>0.816</td>
</tr>
<tr>
<td>2. Long jump</td>
<td>0.209</td>
<td>0.473</td>
</tr>
<tr>
<td>3. Shot put</td>
<td>0.851</td>
<td>0.177</td>
</tr>
<tr>
<td>4. High jump</td>
<td>0.201</td>
<td>0.149</td>
</tr>
<tr>
<td>5. 400-m run</td>
<td>0.072</td>
<td>0.751</td>
</tr>
<tr>
<td>6. 100-m hurdles</td>
<td>0.213</td>
<td>0.280</td>
</tr>
<tr>
<td>7. Discuss</td>
<td>0.771</td>
<td>0.138</td>
</tr>
<tr>
<td>8. Pole vault</td>
<td>0.195</td>
<td>0.005</td>
</tr>
<tr>
<td>9. Javelin</td>
<td>0.442</td>
<td>0.029</td>
</tr>
<tr>
<td>10. 1500-m run</td>
<td>0.047</td>
<td>0.021</td>
</tr>
<tr>
<td>Cumulative proportion of total variance explained</td>
<td>0.17</td>
<td>0.33</td>
</tr>
</tbody>
</table>

The 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

![Factor 2 and Factor 3](image)

**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. Yes</td>
</tr>
<tr>
<td>7. Scale “invariance”</td>
<td>No</td>
<td>No</td>
<td></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 χ-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 $x_{ij}$.
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

$$X = P A + E$$

$P: n \times m, A: m \times p$

$$P_{ij} \geq 0, \quad A_{ij} \geq 0$$

(7.2)

and $Q$ defined as

$$Q = \sum_{i=1}^{n} \sum_{j=1}^{p} E_{ij}^2 / S_{ij}^2$$

(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 $A = A_0$: 

80
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 $E = \sqrt{\sum_i \sum_j E_{ij}^2}$)
   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 $(n + p) \times p \times m^2$. 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, $S_y$, 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:

$$S_y = C_1 + C_2 |x_y| + C_3 |x_y|$$

(7.4)

where $C_1$, $C_2$, and $C_3$ are to be selected in advance. The presence of the constant term, $C_1$,
guarantees 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 $S_y$ 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:

$$S_y = C_1 + C_2 \max(|x_y|,|\hat{x}_y|) + C_3 \max(|x_y|,|\hat{x}_y|)$$  \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, $S_y$. 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 $V_y$. It is further assumed that there is measurement error having standard-deviation $T_y$ in each measured value $x_{ij}$. The matrix $\hat{X}$ and $A$ are then determined so that $\hat{X} = PA$ maximizes the likelihood of $X$, given matrices $T$ and $V$. The standard deviation $S_y$ is computed as:

$$S_y = \sqrt{T_y^2 + cV_y^2|\hat{x}_y||x_y|}$$  \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 $\bar{a}_{ij} = a_{ij}/\max_j(a_{ij})$ for $j = 1, 2, \ldots, m$;
2. \[ \sum_{i} \tilde{a}_{ij} = 1, \text{ computed by setting } \tilde{a}_{ij} = a_{ij} / \sum_{i} a_{ij} \text{ for } j = 1, 2, \ldots, m; \]

3. \[ \frac{1}{p} \sum_{i=1}^{p} \tilde{a}_{ij} = 1, \text{ computed by setting } \tilde{a}_{ij} = a_{ij} / \left( \frac{1}{p} \sum_{i=1}^{p} a_{ij} \right) \text{ for } j = 1, 2, \ldots, m; \]

4. \[ \sum_{i} \tilde{a}_{ij} = 1 - \phi_j, \text{ computed by setting } \tilde{a}_{ij} = (1 - \phi_j) a_{ij} / \sum_{i} a_{ij} \text{ for } j = 1, 2, \ldots, m, \]

where \( \phi_j \) is the residual variation ratio of column \( j \) defined by \( \phi_j = \sum_{i} E_{ij}^2 / S_{ij}^2 / \sum_{i} x_{ij}^2 / S_{ij}^2 \). 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\(_3\); (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:

\[
X_{ijk} = \sum_{k=1}^{m} A_{ik} B_{jk} C_{kk} + E_{ijk} \tag{7.8}
\]

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:

\[
V = \Lambda R
\]  

(8.1)

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

\[ Q_i^* = \frac{1}{m} \sum_{j=1}^{m} (a_{ij}^2)^2 - \frac{1}{m^2} \left( \sum_{j=1}^{m} a_{ij}^2 \right)^2 \]  \hspace{1cm} (8.2)

where the \( a_{ij} \) represent the new loadings, and \( i = 1, 2, \ldots, p \). \( Q_i^* \) represents the "variance" of the squared loadings, which in turn represents the contribution to the variance of the variables. The quartimax maximizes

\[ Q = \sum_{i=1}^{p} Q_i^* \]  \hspace{1cm} (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

\[ V_j^* = \frac{1}{p} \sum_{i=1}^{p} (a_{ij}^2)^2 - \frac{1}{p^2} \left( \sum_{i=1}^{p} a_{ij}^2 \right)^2 \]  \hspace{1cm} (8.4)

for \( j = 1, 2, \ldots, m \). 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

\[ V^* = \sum_{j=1}^{m} V_j^* \]  \hspace{1cm} (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

\[
V_j = \frac{1}{p} \sum_{i=1}^{p} \left( \frac{a_{ij}^2}{h_i^2} \right)^2 - \frac{1}{p^2} \left( \sum_{i=1}^{p} a_{ij}^2 \right)^2
\]

(8.6)

where \( h_i^2 \) is the proportion of variance of the \( i \)th 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 communalities, all variables are given equal weight in the rotation. Note, however, that several authors have argued against such a scaling, especially when some communalities 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:

\[ K = \sum_{i=1}^{p} \sum_{j=1}^{m} a_{ij}^4 \left/ \left( \sum_{i=1}^{p} \sum_{j=1}^{m} a_{ij}^2 \right)^4 \right. \]  \hspace{1cm} (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:

\[ Q = \sum_{\ell=q+1}^{m} \sum_{j=1}^{p} \frac{a_{\ell j}^2}{a_{qq}^2} \]  \hspace{1cm} (8.10)

where the \( a_{\ell j} \) 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

\[ C = \sum_{\ell=q+1}^{m} \left\{ p \sum_{j=1}^{p} \left( \frac{a_{\ell j}^2}{h_{\ell}^2} \right) \left( \frac{a_{qq}^2}{h_q^2} \right) - \left( \sum_{j=1}^{p} \frac{a_{\ell j}^2}{h_{\ell}^2} \right) \left( \sum_{j=1}^{p} \frac{a_{qq}^2}{h_q^2} \right) \right\} \]  \hspace{1cm} (8.11)

where the \( a_{\ell j} \) 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:

\[ B = Q + \frac{C}{p} \]  \hspace{1cm} (8.12)
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:

\[
B' = \beta_1 Q + \frac{\beta_2 C}{P} \tag{8.13}
\]

where \( \beta_1 \) and \( \beta_2 \) are two constants. Writing \( \gamma = \beta_1/(\beta_1 + \beta_2) \), \( B' \) can be written as:

\[
B' = \sum_{\ell, q=1}^{m} \left\{ p \sum_{j=1}^{p} \left( \frac{a_{\ell j}^2}{h_j^2} \right) \left( \frac{a_{q j}^2}{h_j^2} \right) - \gamma \left( \sum_{j=1}^{p} \frac{a_{\ell j}^2}{h_j^2} \right) \left( \sum_{j=1}^{p} \frac{a_{q j}^2}{h_j^2} \right) \right\} \tag{8.14}
\]

where the \( a_{ij} \) are the new loadings, and \( \ell \) and \( q \) are the \( \ell \)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:

\[
B'' = \sum_{\ell, q=1}^{m} \left\{ \frac{p \sum_{j=1}^{p} \left( \frac{a_{\ell j}^2}{h_j^2} \right) \left( \frac{a_{q j}^2}{h_j^2} \right)}{\left( \sum_{j=1}^{p} \frac{a_{\ell j}^2}{h_j^2} \right) \left( \sum_{j=1}^{p} \frac{a_{q j}^2}{h_j^2} \right)} \right\} \tag{8.15}
\]

where the \( a_{ij} \) are the new loadings, and \( \ell \) and \( q \) are the \( \ell \)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:

$$CF = K_1 \sum_{r,s} \sum_{i=1}^{n} \alpha_{i}^{2} \alpha_{s}^{2} + K_2 \sum_{i \neq j, r \neq s}^{m} \alpha_{i}^{2} \alpha_{j}^{2}$$

(8.16)

where the $\alpha_{ij}$ 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, \( \hat{\theta} \), 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
\[ \hat{\theta}_i^* = n\hat{\theta}_n - (n-1)\hat{\theta}_i \]  

(9.1)

The jackknife estimate of \( \theta \) is

\[ \hat{\theta}^* = \frac{1}{n} \sum_{i=1}^{n} \hat{\theta}_i^* \]  

(9.2)

or the mean of the \( n \) pseudovalues. The variance of the pseudovalues is

\[ s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (\hat{\theta}_i^* - \hat{\theta}^*)^2 \]  

(9.3)

and the variance of \( \hat{\theta}^* \) 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 \( \mathbf{X} \) be the samples matrix. One first generates \( n \) submatrices, \( \mathbf{X}_n \), by systematically eliminating each row in turn. A PCA, retaining \( m \) principal components, is made of each matrix \( \mathbf{X}_n \), giving \( n \) sets of eigenvalues, loadings matrices, and scores matrices. Using the \( n \) loadings matrices, \( \mathbf{A}_n \), the \( n \) pseudo-loadings matrices, \( \mathbf{A}_n^* \), are calculated using the relation:

\[ \mathbf{A}_n^* = n\mathbf{A}_n - (n-1)\mathbf{A}_n \]  

(9.4)

where \( \mathbf{A}_n \) is the loadings matrix obtained using all the data.

The jackknife estimate of the loadings matrix is:

\[ \mathbf{A}^* = \frac{1}{n} \sum_{i=1}^{n} \mathbf{A}_n^* \]  

(9.5)

with a variance given by:

\[ s^2 = \frac{1}{n(n-1)} \sum_{i=1}^{n} (\mathbf{A}_n^* - \mathbf{A}^*)^2. \]  

(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 \( \mathbf{A}_i \) and \( \mathbf{A}_j \). 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 ways to make sure that the loadings matrices \( \mathbf{A}_i \) are aligned so that each column corresponds to those in matrix \( \mathbf{A}_n \). 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 \( \mathbf{A}_i \) with \( \mathbf{A}_n \) (see Section 9.2), one first creates the \((m \times m)\) matrix \( \mathbf{M} \) with elements equal to:

\[
m_{ij} = \min\left(\sum_{i=1}^{n} (a_{ik}^* - a_{jk})^2, \sum_{i=1}^{n} (a_{ik}^* + a_{jk})^2\right)
\]

(9.7)

where \( a_{ij}^* \) and \( a_{ij} \) are the elements of \( \mathbf{A}_i \) and \( \mathbf{A}_n \), respectively.

If \( \mathbf{M}^* \) is a matrix obtained by permuting the two or more columns of \( \mathbf{M} \), its trace is equal to:

\[
\text{trace}(\mathbf{M}^*) = \sum_{i=1}^{m} m_{ii}^*
\]

(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>0.915±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>NH(_4)(^+)</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>NO(_3)(^-)</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>SO(_2)(^{3-})</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>

<table>
<thead>
<tr>
<th></th>
<th>Bootstrap 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>0.438±0.149</td>
<td>0.673±0.197</td>
<td>0.264±0.176</td>
<td>0.102±0.262</td>
</tr>
<tr>
<td>Fe</td>
<td>0.956±0.022</td>
<td>0.023±0.097</td>
<td>-0.153±0.066</td>
<td>0.079±0.035</td>
</tr>
<tr>
<td>Ca</td>
<td>0.871±0.060</td>
<td>0.298±0.183</td>
<td>-0.018±0.073</td>
<td>0.170±0.073</td>
</tr>
<tr>
<td>K</td>
<td>0.226±0.387</td>
<td>-0.043±0.073</td>
<td>-0.089±0.037</td>
<td>0.965±0.451</td>
</tr>
<tr>
<td>Si</td>
<td>0.911±0.080</td>
<td>0.156±0.130</td>
<td>-0.108±0.076</td>
<td>0.092±0.077</td>
</tr>
<tr>
<td>Al</td>
<td>0.962±0.014</td>
<td>0.062±0.107</td>
<td>-0.129±0.062</td>
<td>0.057±0.035</td>
</tr>
<tr>
<td>Na(^+)</td>
<td>0.113±0.135</td>
<td>0.965±0.021</td>
<td>0.030±0.091</td>
<td>-0.054±0.060</td>
</tr>
<tr>
<td>NH(_4)(^+)</td>
<td>0.114±0.138</td>
<td>0.977±0.015</td>
<td>0.085±0.091</td>
<td>-0.030±0.072</td>
</tr>
<tr>
<td>Cl(^-)</td>
<td>-0.131±0.137</td>
<td>0.446±0.245</td>
<td>0.730±0.282</td>
<td>-0.081±0.103</td>
</tr>
<tr>
<td>NO(_3)(^-)</td>
<td>0.124±0.168</td>
<td>0.684±0.196</td>
<td>0.643±0.184</td>
<td>-0.009±0.124</td>
</tr>
<tr>
<td>SO(_2)(^{3-})</td>
<td>-0.255±0.089</td>
<td>-0.062±0.122</td>
<td>0.912±0.047</td>
<td>-0.056±0.035</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\(_b\), is already aligned. The aligned matrix, A\(_b\)'\(_r\), is the matrix obtained from A\(_b\) 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 0.7 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$_4^{2-}$, 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.](image-url)
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 these 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>0.287</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>Number of Component Retained</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>SO₄²⁻</td>
<td>0.879</td>
<td>0.914</td>
<td>0.922</td>
<td>0.927</td>
<td>0.932</td>
<td>0.958</td>
<td>0.963</td>
<td></td>
</tr>
<tr>
<td>H⁺</td>
<td>0.767</td>
<td>0.845</td>
<td>0.845</td>
<td>0.857</td>
<td>0.875</td>
<td>0.931</td>
<td>0.966</td>
<td></td>
</tr>
<tr>
<td>Br</td>
<td>0.653</td>
<td>0.754</td>
<td>0.801</td>
<td>0.801</td>
<td>0.978</td>
<td>0.992</td>
<td>0.995</td>
<td></td>
</tr>
<tr>
<td>NH₄⁺</td>
<td>0.876</td>
<td>0.887</td>
<td>0.890</td>
<td>0.890</td>
<td>0.898</td>
<td>0.899</td>
<td>0.942</td>
<td></td>
</tr>
<tr>
<td>NO₃⁻</td>
<td>0.531</td>
<td>0.787</td>
<td>0.797</td>
<td>0.798</td>
<td>0.909</td>
<td>0.989</td>
<td>0.995</td>
<td></td>
</tr>
<tr>
<td>Na</td>
<td>0.889</td>
<td>0.898</td>
<td>0.900</td>
<td>0.914</td>
<td>0.932</td>
<td>0.935</td>
<td>0.944</td>
<td></td>
</tr>
<tr>
<td>Cl</td>
<td>0.878</td>
<td>0.890</td>
<td>0.890</td>
<td>0.899</td>
<td>0.936</td>
<td>0.940</td>
<td>0.970</td>
<td></td>
</tr>
<tr>
<td>K⁺</td>
<td>0.846</td>
<td>0.851</td>
<td>0.852</td>
<td>0.852</td>
<td>0.852</td>
<td>0.863</td>
<td>0.939</td>
<td></td>
</tr>
<tr>
<td>Pb</td>
<td>0.859</td>
<td>0.868</td>
<td>0.868</td>
<td>0.869</td>
<td>0.877</td>
<td>0.906</td>
<td>0.911</td>
<td></td>
</tr>
<tr>
<td>V</td>
<td>0.838</td>
<td>0.847</td>
<td>0.848</td>
<td>0.867</td>
<td>0.867</td>
<td>0.906</td>
<td>0.936</td>
<td></td>
</tr>
<tr>
<td>Mg</td>
<td>0.880</td>
<td>0.921</td>
<td>0.931</td>
<td>0.935</td>
<td>0.936</td>
<td>0.936</td>
<td>0.944</td>
<td></td>
</tr>
<tr>
<td>Zn</td>
<td>0.891</td>
<td>0.892</td>
<td>0.895</td>
<td>0.899</td>
<td>0.899</td>
<td>0.911</td>
<td>0.921</td>
<td></td>
</tr>
<tr>
<td>Cu</td>
<td>0.910</td>
<td>0.923</td>
<td>0.931</td>
<td>0.943</td>
<td>0.943</td>
<td>0.953</td>
<td>0.954</td>
<td></td>
</tr>
<tr>
<td>Ca</td>
<td>0.873</td>
<td>0.889</td>
<td>0.902</td>
<td>0.923</td>
<td>0.927</td>
<td>0.944</td>
<td>0.946</td>
<td></td>
</tr>
<tr>
<td>Mn</td>
<td>0.447</td>
<td>0.475</td>
<td>0.730</td>
<td>0.995</td>
<td>0.997</td>
<td>0.998</td>
<td>1.000</td>
<td></td>
</tr>
<tr>
<td>I</td>
<td>0.597</td>
<td>0.691</td>
<td>0.768</td>
<td>0.907</td>
<td>0.993</td>
<td>0.993</td>
<td>0.994</td>
<td></td>
</tr>
<tr>
<td>Al</td>
<td>0.884</td>
<td>0.890</td>
<td>0.897</td>
<td>0.910</td>
<td>0.914</td>
<td>0.918</td>
<td>0.925</td>
<td></td>
</tr>
<tr>
<td>MSA</td>
<td>0.552</td>
<td>0.608</td>
<td>0.871</td>
<td>0.962</td>
<td>0.967</td>
<td>0.998</td>
<td>0.998</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>0.933</td>
<td>0.812</td>
<td>0.753</td>
</tr>
<tr>
<td>±0.240</td>
<td>±0.141</td>
<td>±0.508</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>
</tr>
<tr>
<td>Loadings</td>
<td>SO_4(^{2-})</td>
<td>H(^+)</td>
<td>Br</td>
<td>NH_4(^+)</td>
<td>NO_3(^-)</td>
<td>Na</td>
<td>Cl</td>
<td>K(^+)</td>
<td>Pb</td>
<td>V</td>
</tr>
<tr>
<td>±0.787</td>
<td>±0.083</td>
<td>0.574</td>
<td>±0.127</td>
<td>0.183</td>
<td>±0.048</td>
<td>±0.028</td>
<td>±0.279</td>
<td>±0.079</td>
<td>±0.052</td>
<td></td>
</tr>
<tr>
<td>±0.040</td>
<td>±0.024</td>
<td>-0.231</td>
<td>±0.022</td>
<td>0.024</td>
<td>±0.026</td>
<td>±0.028</td>
<td>±0.049</td>
<td>±0.049</td>
<td>±0.043</td>
<td></td>
</tr>
<tr>
<td>-0.023</td>
<td>±0.032</td>
<td>±0.71</td>
<td>±0.026</td>
<td>-0.27</td>
<td>±0.039</td>
<td>±0.051</td>
<td>±0.075</td>
<td>±0.075</td>
<td>±0.034</td>
<td></td>
</tr>
<tr>
<td>0.120</td>
<td>±0.031</td>
<td>0.198</td>
<td>±0.027</td>
<td>0.943</td>
<td>±0.015</td>
<td>±0.027</td>
<td>±0.001</td>
<td>-0.003</td>
<td>±0.047</td>
<td></td>
</tr>
<tr>
<td>0.253</td>
<td>±0.027</td>
<td>0.082</td>
<td>±0.026</td>
<td>0.943</td>
<td>±0.021</td>
<td>±0.027</td>
<td>±0.001</td>
<td>-0.003</td>
<td>±0.047</td>
<td></td>
</tr>
<tr>
<td>0.134</td>
<td>±0.069</td>
<td>0.043</td>
<td>±0.040</td>
<td>1.04</td>
<td>±0.024</td>
<td>±0.027</td>
<td>±0.001</td>
<td>-0.003</td>
<td>±0.047</td>
<td></td>
</tr>
<tr>
<td>0.115</td>
<td>0.002</td>
<td>-0.041</td>
<td>±0.034</td>
<td>±0.083</td>
<td>±0.019</td>
<td>±0.029</td>
<td>±0.001</td>
<td>-0.003</td>
<td>±0.047</td>
<td></td>
</tr>
<tr>
<td>0.022</td>
<td>±0.007</td>
<td>-0.167</td>
<td>±0.031</td>
<td>0.13</td>
<td>±0.002</td>
<td>±0.019</td>
<td>±0.001</td>
<td>-0.003</td>
<td>±0.047</td>
<td></td>
</tr>
<tr>
<td>±0.022</td>
<td>±0.012</td>
<td>-0.017</td>
<td>±0.012</td>
<td>±0.012</td>
<td>±0.012</td>
<td>±0.012</td>
<td>±0.012</td>
<td>±0.012</td>
<td>±0.012</td>
<td></td>
</tr>
<tr>
<td>±0.008</td>
<td>±0.028</td>
<td>-0.012</td>
<td>±0.001</td>
<td>±0.001</td>
<td>±0.001</td>
<td>±0.001</td>
<td>±0.001</td>
<td>±0.001</td>
<td>±0.001</td>
<td></td>
</tr>
<tr>
<td>±0.011</td>
<td>±0.001</td>
<td>±0.001</td>
<td>±0.001</td>
<td>±0.001</td>
<td>±0.001</td>
<td>±0.001</td>
<td>±0.001</td>
<td>±0.001</td>
<td>±0.001</td>
<td></td>
</tr>
</tbody>
</table>

1.000 0.098 1.000
-0.118 -0.064 1.000
0.023 0.040 0.081 1.000
0.104 0.067 0.330 0.173 1.000
0.052 0.281 0.050 0.099 0.268 1.000
0.056 0.017 -0.033 0.026 0.102 0.051 1.000
0.031 -0.162 0.071 -0.054 -0.085 -0.064 0.018 1.000
-0.006 0.057 0.214 0.062 0.080 0.060 0.011 0.810 1.000
-0.009 -0.011 0.143 0.099 0.115 0.089 0.033 0.094 0.196 1.000

101
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></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><strong>Ions</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>SO(_4^2-)</strong></td>
<td>0.067</td>
<td>±0.056</td>
<td>±0.749</td>
<td>0.104</td>
<td>0.163</td>
<td>0.092</td>
<td>-0.011</td>
<td>0.088</td>
<td>0.191</td>
<td>0.098</td>
</tr>
<tr>
<td><strong>H(^+)</strong></td>
<td>±0.028</td>
<td>±0.053</td>
<td>±0.059</td>
<td>±0.025</td>
<td>±0.073</td>
<td>±0.029</td>
<td>±0.046</td>
<td>±0.026</td>
<td>±0.029</td>
<td>±0.030</td>
</tr>
<tr>
<td><strong>Br</strong></td>
<td>±0.046</td>
<td>±0.039</td>
<td>±0.049</td>
<td>±0.021</td>
<td>±0.060</td>
<td>±0.054</td>
<td>±0.043</td>
<td>±0.029</td>
<td>±0.037</td>
<td>±0.024</td>
</tr>
<tr>
<td><strong>NH(_4^+)</strong></td>
<td>±0.012</td>
<td>±0.017</td>
<td>±0.028</td>
<td>±0.013</td>
<td>±0.027</td>
<td>±0.015</td>
<td>±0.014</td>
<td>±0.011</td>
<td>±0.012</td>
<td>±0.014</td>
</tr>
<tr>
<td><strong>NO(_3^-)</strong></td>
<td>±0.073</td>
<td>±0.016</td>
<td>±0.489</td>
<td>±0.078</td>
<td>0.425</td>
<td>0.213</td>
<td>±0.005</td>
<td>0.062</td>
<td>0.020</td>
<td>0.214</td>
</tr>
<tr>
<td><strong>Na</strong></td>
<td>±0.030</td>
<td>±0.094</td>
<td>±0.114</td>
<td>±0.026</td>
<td>±0.141</td>
<td>±0.052</td>
<td>±0.115</td>
<td>±0.031</td>
<td>±0.036</td>
<td>±0.044</td>
</tr>
<tr>
<td><strong>Cl</strong></td>
<td>±0.019</td>
<td>±0.033</td>
<td>±0.019</td>
<td>±0.012</td>
<td>±0.033</td>
<td>±0.025</td>
<td>±0.018</td>
<td>±0.014</td>
<td>±0.013</td>
<td>±0.014</td>
</tr>
<tr>
<td><strong>K</strong></td>
<td>±0.024</td>
<td>±0.033</td>
<td>±0.026</td>
<td>±0.027</td>
<td>±0.051</td>
<td>±0.035</td>
<td>±0.036</td>
<td>±0.021</td>
<td>±0.021</td>
<td>±0.022</td>
</tr>
<tr>
<td><strong>Pb</strong></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><strong>V</strong></td>
<td>±0.112</td>
<td>0.609</td>
<td>0.075</td>
<td>0.023</td>
<td>0.315</td>
<td>0.217</td>
<td>-0.029</td>
<td>-0.333</td>
<td>0.162</td>
<td>0.035</td>
</tr>
<tr>
<td><strong>Mg</strong></td>
<td>±0.065</td>
<td>±0.154</td>
<td>±0.126</td>
<td>±0.024</td>
<td>±0.225</td>
<td>±0.074</td>
<td>±0.061</td>
<td>±0.033</td>
<td>±0.043</td>
<td>±0.040</td>
</tr>
<tr>
<td><strong>Zn</strong></td>
<td>±0.256</td>
<td>-0.034</td>
<td>0.191</td>
<td>0.133</td>
<td>±0.696</td>
<td>0.209</td>
<td>0.034</td>
<td>-0.099</td>
<td>0.083</td>
<td>0.009</td>
</tr>
<tr>
<td><strong>Cu</strong></td>
<td>±0.049</td>
<td>±0.047</td>
<td>±0.060</td>
<td>±0.045</td>
<td>±0.120</td>
<td>±0.055</td>
<td>±0.099</td>
<td>±0.036</td>
<td>±0.035</td>
<td>±0.033</td>
</tr>
<tr>
<td><strong>Ca</strong></td>
<td>±0.093</td>
<td>0.708</td>
<td>±0.036</td>
<td>±0.048</td>
<td>±0.028</td>
<td>±0.150</td>
<td>±0.061</td>
<td>±0.018</td>
<td>±0.029</td>
<td>±0.027</td>
</tr>
<tr>
<td><strong>Mn</strong></td>
<td>±0.054</td>
<td>±0.078</td>
<td>±0.063</td>
<td>±0.016</td>
<td>±0.068</td>
<td>±0.033</td>
<td>±0.058</td>
<td>±0.025</td>
<td>±0.032</td>
<td>±0.029</td>
</tr>
<tr>
<td><strong>I</strong></td>
<td>±0.016</td>
<td>0.026</td>
<td>0.031</td>
<td>0.740</td>
<td>±0.385</td>
<td>0.054</td>
<td>0.051</td>
<td>-0.044</td>
<td>0.020</td>
<td>0.042</td>
</tr>
<tr>
<td><strong>Al</strong></td>
<td>±0.037</td>
<td>±0.044</td>
<td>±0.041</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><strong>MSA</strong></td>
<td>±0.025</td>
<td>±0.022</td>
<td>±0.024</td>
<td>0.024</td>
<td>±0.011</td>
<td>0.060</td>
<td>0.145</td>
<td>0.003</td>
<td>-0.011</td>
<td>-0.036</td>
</tr>
<tr>
<td><strong>O<del>1</del></strong></td>
<td>±0.018</td>
<td>0.094</td>
<td>±0.059</td>
<td>±0.037</td>
<td>±0.011</td>
<td>±0.030</td>
<td>±0.039</td>
<td>±0.085</td>
<td>±0.022</td>
<td>±0.021</td>
</tr>
<tr>
<td><strong>Mo</strong></td>
<td>±0.029</td>
<td>±0.023</td>
<td>±0.015</td>
<td>±0.020</td>
<td>±0.146</td>
<td>±0.021</td>
<td>±0.482</td>
<td>±0.039</td>
<td>±0.032</td>
<td>±0.035</td>
</tr>
<tr>
<td><strong>Si</strong></td>
<td>±0.001</td>
<td>±0.005</td>
<td>±0.037</td>
<td>±0.012</td>
<td>±0.950</td>
<td>±0.003</td>
<td>0.009</td>
<td>0.041</td>
<td>0.086</td>
<td>0.972</td>
</tr>
<tr>
<td><strong>H<del>2</del></strong></td>
<td>±0.015</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>±0.012</td>
<td>±0.017</td>
<td>±0.010</td>
</tr>
<tr>
<td><strong>H<del>3</del></strong></td>
<td>±0.039</td>
<td>±0.075</td>
<td>±0.037</td>
<td>±0.017</td>
<td>±0.099</td>
<td>±0.059</td>
<td>±0.070</td>
<td>±0.034</td>
<td>±0.022</td>
<td>±0.037</td>
</tr>
<tr>
<td><strong>Mg</strong></td>
<td>±0.006</td>
<td>0.023</td>
<td>0.008</td>
<td>0.007</td>
<td>0.017</td>
<td>0.016</td>
<td>0.010</td>
<td>0.992</td>
<td>0.047</td>
<td>0.039</td>
</tr>
<tr>
<td><strong>Fe</strong></td>
<td>±0.019</td>
<td>±0.015</td>
<td>±0.017</td>
<td>±0.006</td>
<td>±0.020</td>
<td>±0.018</td>
<td>±0.010</td>
<td>±0.006</td>
<td>±0.010</td>
<td>±0.011</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~3+* factor, but here also the loadings are not statistically significant from
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>Ions</th>
<th>Loadings</th>
<th>Factor 1</th>
<th>Factor 2</th>
<th>Factor 3</th>
<th>Factor 4</th>
<th>Factor 5</th>
<th>Factor 6</th>
<th>Factor 7</th>
<th>Factor 8</th>
<th>Factor 9</th>
<th>Factor 10</th>
<th>Communality</th>
</tr>
</thead>
<tbody>
<tr>
<td>SO$_4^{2-}$</td>
<td>0.890</td>
<td>0.053</td>
<td>-0.202</td>
<td>0.119</td>
<td>0.228</td>
<td>0.216</td>
<td>0.228</td>
<td>0.188</td>
<td>-0.022</td>
<td>0.067</td>
<td>1.000</td>
<td></td>
</tr>
<tr>
<td>H$^+$</td>
<td>0.666</td>
<td>-0.256</td>
<td>-0.271</td>
<td>-0.032</td>
<td>0.191</td>
<td>0.127</td>
<td>0.562</td>
<td>-0.010</td>
<td>-0.038</td>
<td>-0.013</td>
<td>0.953</td>
<td></td>
</tr>
<tr>
<td>Br</td>
<td>0.196</td>
<td>0.016</td>
<td>-0.017</td>
<td>0.042</td>
<td>0.920</td>
<td>0.329</td>
<td>0.051</td>
<td>0.031</td>
<td>0.014</td>
<td>0.027</td>
<td>1.000</td>
<td></td>
</tr>
<tr>
<td>NH$_4^+$</td>
<td>-0.101</td>
<td>0.031</td>
<td>0.028</td>
<td>0.041</td>
<td>0.845</td>
<td>0.141</td>
<td>0.035</td>
<td>0.253</td>
<td>-0.061</td>
<td>0.020</td>
<td>0.118</td>
<td>-0.083</td>
</tr>
<tr>
<td>NO$_3^-$</td>
<td>0.104</td>
<td>-0.036</td>
<td>-0.034</td>
<td>-0.037</td>
<td>0.203</td>
<td>0.094</td>
<td>0.138</td>
<td>-0.281</td>
<td>0.042</td>
<td>0.067</td>
<td>0.046</td>
<td></td>
</tr>
<tr>
<td>Na</td>
<td>0.184</td>
<td>0.059</td>
<td>0.037</td>
<td>0.035</td>
<td>0.069</td>
<td>0.059</td>
<td>0.137</td>
<td>0.226</td>
<td>0.044</td>
<td>0.039</td>
<td>0.047</td>
<td></td>
</tr>
<tr>
<td>Cl</td>
<td>0.033</td>
<td>0.015</td>
<td>0.031</td>
<td>0.026</td>
<td>0.027</td>
<td>0.047</td>
<td>0.065</td>
<td>0.073</td>
<td>0.044</td>
<td>0.030</td>
<td>0.027</td>
<td></td>
</tr>
<tr>
<td>K$^+$</td>
<td>0.084</td>
<td>0.043</td>
<td>0.054</td>
<td>0.029</td>
<td>0.104</td>
<td>0.049</td>
<td>0.066</td>
<td>0.066</td>
<td>0.017</td>
<td>0.024</td>
<td>0.090</td>
<td></td>
</tr>
<tr>
<td>Pb</td>
<td>0.115</td>
<td>0.014</td>
<td>0.010</td>
<td>0.032</td>
<td>0.026</td>
<td>0.073</td>
<td>0.059</td>
<td>0.030</td>
<td>0.044</td>
<td>0.039</td>
<td>0.038</td>
<td></td>
</tr>
<tr>
<td>V</td>
<td>0.068</td>
<td>0.097</td>
<td>0.043</td>
<td>0.097</td>
<td>0.048</td>
<td>0.048</td>
<td>0.073</td>
<td>0.137</td>
<td>0.335</td>
<td>0.404</td>
<td>0.038</td>
<td></td>
</tr>
<tr>
<td>Mg</td>
<td>0.094</td>
<td>0.045</td>
<td>0.787</td>
<td>0.055</td>
<td>0.067</td>
<td>0.088</td>
<td>0.037</td>
<td>0.089</td>
<td>0.039</td>
<td>0.032</td>
<td>1.000</td>
<td></td>
</tr>
<tr>
<td>Zn</td>
<td>-0.055</td>
<td>-0.060</td>
<td>-0.051</td>
<td>-0.016</td>
<td>0.072</td>
<td>0.042</td>
<td>0.043</td>
<td>0.057</td>
<td>0.040</td>
<td>0.046</td>
<td>0.038</td>
<td></td>
</tr>
<tr>
<td>Cu</td>
<td>0.097</td>
<td>0.049</td>
<td>0.012</td>
<td>0.001</td>
<td>0.002</td>
<td>0.028</td>
<td>0.065</td>
<td>0.021</td>
<td>0.051</td>
<td>1.000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ca</td>
<td>-0.124</td>
<td>-0.031</td>
<td>-0.028</td>
<td>-0.006</td>
<td>-0.050</td>
<td>0.055</td>
<td>0.091</td>
<td>0.177</td>
<td>0.028</td>
<td>0.096</td>
<td>0.093</td>
<td></td>
</tr>
<tr>
<td>Mn</td>
<td>0.079</td>
<td>0.003</td>
<td>0.511</td>
<td>0.018</td>
<td>0.016</td>
<td>0.050</td>
<td>0.049</td>
<td>0.071</td>
<td>-0.019</td>
<td>-0.062</td>
<td>0.279</td>
<td></td>
</tr>
<tr>
<td>I</td>
<td>0.101</td>
<td>0.042</td>
<td>0.275</td>
<td>0.040</td>
<td>0.053</td>
<td>0.035</td>
<td>0.053</td>
<td>0.053</td>
<td>0.017</td>
<td>0.037</td>
<td>0.401</td>
<td></td>
</tr>
<tr>
<td>Al</td>
<td>0.028</td>
<td>0.021</td>
<td>0.141</td>
<td>0.233</td>
<td>0.547</td>
<td>0.043</td>
<td>0.043</td>
<td>0.068</td>
<td>0.436</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MSA</td>
<td>0.039</td>
<td>-0.065</td>
<td>0.105</td>
<td>-0.098</td>
<td>0.102</td>
<td>0.047</td>
<td>0.071</td>
<td>-0.114</td>
<td>-0.058</td>
<td>0.059</td>
<td>0.374</td>
<td></td>
</tr>
</tbody>
</table>
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>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>-0.9123</td>
<td>0.1066</td>
<td>0.0199</td>
<td>0.7832</td>
<td>0.7193</td>
<td>0.5797</td>
<td>0.5289</td>
<td>0.7710</td>
<td>0.1421</td>
<td>0.7379</td>
</tr>
<tr>
<td></td>
<td>±0.1728</td>
<td>±0.8866</td>
<td>±0.8299</td>
<td>±0.4473</td>
<td>±0.1551</td>
<td>±0.8415</td>
<td>±1.1763</td>
<td>±0.3523</td>
<td>±0.0962</td>
<td>±0.7623</td>
</tr>
<tr>
<td>H$^+$</td>
<td>-0.0180</td>
<td>0.0001</td>
<td>0.0006</td>
<td>0.0003</td>
<td>0.0000</td>
<td>0.0008</td>
<td>0.0024</td>
<td>0.0001</td>
<td>0.0000</td>
<td>0.0004</td>
</tr>
<tr>
<td>Br</td>
<td>±0.0011</td>
<td>±0.0001</td>
<td>±0.0014</td>
<td>±0.0009</td>
<td>±0.0003</td>
<td>±0.0017</td>
<td>±0.0053</td>
<td>±0.0007</td>
<td>±0.0001</td>
<td>±0.0013</td>
</tr>
<tr>
<td>NH$_4^+$</td>
<td>±0.0006</td>
<td>±0.0005</td>
<td>±0.0045</td>
<td>±0.0024</td>
<td>±0.0006</td>
<td>±0.0049</td>
<td>±0.0026</td>
<td>±0.0012</td>
<td>±0.0003</td>
<td>±0.0032</td>
</tr>
<tr>
<td>NO$_3^-$</td>
<td>±0.0220</td>
<td>±0.0002</td>
<td>±0.0400</td>
<td>±0.1600</td>
<td>0.0579</td>
<td>0.2655</td>
<td>0.0334</td>
<td>0.1759</td>
<td>0.0059</td>
<td>0.0512</td>
</tr>
<tr>
<td>Na</td>
<td>±0.0154</td>
<td>±0.0068</td>
<td>±0.0932</td>
<td>±0.0306</td>
<td>±0.0149</td>
<td>±0.1049</td>
<td>±0.1704</td>
<td>±0.0380</td>
<td>±0.0069</td>
<td>±0.0701</td>
</tr>
<tr>
<td>Cl</td>
<td>±0.0001</td>
<td>±0.0226</td>
<td>±0.6863</td>
<td>0.0396</td>
<td>0.0207</td>
<td>0.0508</td>
<td>0.1268</td>
<td>0.0005</td>
<td>0.0006</td>
<td>0.0117</td>
</tr>
<tr>
<td>K$^+$</td>
<td>±0.0024</td>
<td>±0.0082</td>
<td>±0.0907</td>
<td>±0.0277</td>
<td>±0.0089</td>
<td>±0.0711</td>
<td>±0.1156</td>
<td>±0.0173</td>
<td>±0.0078</td>
<td>±0.0402</td>
</tr>
<tr>
<td>Pb</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>±0.0921</td>
<td>±0.0328</td>
<td>±0.0115</td>
<td>±0.0961</td>
</tr>
<tr>
<td>V</td>
<td>±0.0002</td>
<td>±0.0455</td>
<td>±0.0229</td>
<td>±0.0200</td>
<td>±0.0180</td>
<td>±0.0140</td>
<td>±0.0086</td>
<td>±0.0286</td>
<td>±0.0025</td>
<td>±0.0052</td>
</tr>
<tr>
<td>Mg</td>
<td>±0.0022</td>
<td>±0.0032</td>
<td>±0.0224</td>
<td>±0.0097</td>
<td>±0.0032</td>
<td>±0.0313</td>
<td>±0.0650</td>
<td>±0.0140</td>
<td>±0.0065</td>
<td>±0.0039</td>
</tr>
<tr>
<td>Zn</td>
<td>±0.0008</td>
<td>±0.0002</td>
<td>±0.0001</td>
<td>±0.0000</td>
<td>±0.0000</td>
<td>±0.0000</td>
<td>±0.0000</td>
<td>±0.0000</td>
<td>±0.0000</td>
<td>±0.0000</td>
</tr>
<tr>
<td>Cu</td>
<td>±0.0009</td>
<td>±0.0002</td>
<td>±0.0002</td>
<td>±0.0002</td>
<td>±0.0002</td>
<td>±0.0002</td>
<td>±0.0002</td>
<td>±0.0002</td>
<td>±0.0002</td>
<td>±0.0002</td>
</tr>
<tr>
<td>Ca</td>
<td>±0.0003</td>
<td>±0.0001</td>
<td>±0.0010</td>
<td>±0.0004</td>
<td>±0.0001</td>
<td>±0.0011</td>
<td>±0.0018</td>
<td>±0.0004</td>
<td>±0.0001</td>
<td>±0.0031</td>
</tr>
<tr>
<td>Mn</td>
<td>±0.0003</td>
<td>±0.0001</td>
<td>±0.0001</td>
<td>±0.0004</td>
<td>±0.0001</td>
<td>±0.0009</td>
<td>±0.0010</td>
<td>±0.0003</td>
<td>±0.0004</td>
<td>±0.0005</td>
</tr>
<tr>
<td>I</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>±0.0004</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>±0.0004</td>
</tr>
<tr>
<td>MSA</td>
<td>±0.0002</td>
<td>±0.0001</td>
<td>±0.0018</td>
<td>±0.0042</td>
<td>±0.0002</td>
<td>±0.0023</td>
<td>±0.0054</td>
<td>±0.0005</td>
<td>±0.0003</td>
<td>±0.0009</td>
</tr>
</tbody>
</table>

met in this case. In fact, the Alert data are not independent, as Siros 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

104
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 $\text{H}^+$ 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:

$$
\begin{bmatrix}
1.000 \\
0.199 & 1.000 \\
0.039 & 0.209 & 1.000 \\
0.991 & 0.190 & 0.075 & 1.000 \\
0.986 & 0.274 & 0.107 & 0.980 & 1.000 \\
0.934 & 0.196 & 0.118 & 0.957 & 0.932 & 1.000 \\
0.849 & 0.186 & 0.218 & 0.850 & 0.842 & 0.804 & 1.000 \\
0.983 & 0.189 & 0.030 & 0.992 & 0.975 & 0.970 & 0.831 & 1.000 \\
0.258 & 0.102 & 0.060 & 0.258 & 0.283 & 0.375 & 0.228 & 0.256 & 1.000 \\
0.994 & 0.241 & 0.072 & 0.987 & 0.995 & 0.936 & 0.847 & 0.981 & 0.282 & 1.000
\end{bmatrix}
$$

The large correlations, higher than 0.8, are all associated with very large values of the loadings for $\text{SO}_2^-$. 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:

$$
\begin{bmatrix}
1.000 \\
0.000 & 1.000 \\
0.000 & 0.000 & 1.000 \\
0.000 & 0.000 & 0.000 & 1.000 \\
0.981 & 0.088 & 0.000 & 0.000 & 1.000 \\
0.000 & 0.000 & 0.000 & 0.000 & 0.000 & 1.000 \\
0.000 & 0.000 & 0.000 & 0.000 & 0.079 & 0.000 & 0.000 & 1.000 \\
0.000 & 0.026 & 0.000 & 0.010 & 0.000 & 0.000 & 0.000 & 0.000 & 1.000 \\
0.000 & 0.000 & 0.000 & 0.000 & 0.000 & 0.000 & 0.000 & 0.029 & 0.000 & 1.000
\end{bmatrix}
$$

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></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>1.306</td>
<td>0.042</td>
<td>0.566</td>
</tr>
<tr>
<td>5</td>
<td>2.083</td>
<td>0.306</td>
<td>0.037</td>
<td>0.602</td>
</tr>
<tr>
<td>6</td>
<td>1.629</td>
<td>0.454</td>
<td>0.029</td>
<td>0.631</td>
</tr>
<tr>
<td>7</td>
<td>1.503</td>
<td>0.126</td>
<td>0.026</td>
<td>0.567</td>
</tr>
<tr>
<td>8</td>
<td>1.392</td>
<td>0.110</td>
<td>0.024</td>
<td>0.681</td>
</tr>
<tr>
<td>9</td>
<td>1.383</td>
<td>0.009</td>
<td>0.025</td>
<td>0.706</td>
</tr>
<tr>
<td>10</td>
<td>1.185</td>
<td>0.198</td>
<td>0.021</td>
<td>0.726</td>
</tr>
<tr>
<td>11</td>
<td>1.139</td>
<td>0.046</td>
<td>0.020</td>
<td>0.746</td>
</tr>
<tr>
<td>12</td>
<td>1.081</td>
<td>0.057</td>
<td>0.019</td>
<td>0.765</td>
</tr>
<tr>
<td>13</td>
<td>0.918</td>
<td>0.163</td>
<td>0.016</td>
<td>0.782</td>
</tr>
<tr>
<td>14</td>
<td>0.838</td>
<td>0.080</td>
<td>0.015</td>
<td>0.796</td>
</tr>
<tr>
<td>15</td>
<td>0.783</td>
<td>0.055</td>
<td>0.014</td>
<td>0.810</td>
</tr>
<tr>
<td>16</td>
<td>0.748</td>
<td>0.034</td>
<td>0.013</td>
<td>0.823</td>
</tr>
<tr>
<td>17</td>
<td>0.718</td>
<td>0.031</td>
<td>0.013</td>
<td>0.836</td>
</tr>
<tr>
<td>18</td>
<td>0.644</td>
<td>0.074</td>
<td>0.011</td>
<td>0.847</td>
</tr>
<tr>
<td>19</td>
<td>0.620</td>
<td>0.024</td>
<td>0.011</td>
<td>0.858</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 x 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 10.4](image1.png)  
**Figure 10.4** Scree graph for the sulphate concentration data

![Figure 10.5](image2.png)  
**Figure 10.5** Variation of the minimum communality with the number of components retained.
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
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>

110
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 10.7 The 1985 total annual emission density distribution of anthropogenic SO$_2$ in North America on the gridding scale of 1° of longitude by 2/3° of latitude. The emission densities are in units of kg SO$_2$ ha$^{-1}$ yr$^{-1}$. The shading interval is logarithmic base 3 (From Luechen et al., 1992).
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.
Figure 10.9 Spatial variation of the loadings for the 10 PMF factors for the sulphate concentration in the precipitation example.

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.
Figure 10.10 Temporal variation of scores for PMF analysis for the sulphate concentration in the precipitation example.

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 SO$_2$ 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 Rummel (1970). Rummel, 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

Figure 11.1 Flow chart of guideline.
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 interact 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; Dillon 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.


