EXPLORATORY ANALYSIS OF MULTIVARIATE DATA

Unsupervised Image Segmentation and Data Driven Linear and Nonlinear Decomposition

Klaus Baggesen Hilger

IMM-PHD-2001-89

IMM

Printed by IMM, Technical University of Denmark

Preface

This thesis is prepared at the image analysis group, Informatics and Mathematical Modelling, Technical University of Denmark. It is written in the autumn of 2001 and completes a three year Ph.D. study.

The general framework of this work is exploratory data analysis, multivariate statistics, and digital image analysis. It is assumed that the reader possesses a basic knowledge of these areas.

The treatment of the subjects is by no means exhaustive but is intended to improve the knowledge of exploratory data analysis through selected theory and examples.

This research was sponsored by the GEOSONAR project funded by the Danish National Research Councils under the Earth Observation Program.

Lyngby, November 2001

Klaus Baggesen Hilger

Evaluation Committee:
Professor Henrik Madsen (chairman), Technical University of Denmark
Professor Paul Switzer, Stanford University
Project Manager Kristian Windfeld, Novo Nordisk A/S

Supervisors:
Associate Professor Allan A. Nielsen, Technical University of Denmark
Associate Professor Bjarne K. Ersbøll, Technical University of Denmark

Acknowledgements

I am grateful to a large number of people for support, encouragement, rewarding discussions and assistance.

First of all I would like to thank my supervisors Dr. Allan Aasbjerg Nielsen and Dr. Bjarne Kjær Ersbøll for their involvement, guidance and encouragement during the course of my study, and Professor Knut Conradsen for providing excellent research facilities at the IMM image group. Especially thank you to Allan for his moral support during all phases of my study.

I would like to thank the rest of my co-workers in the image analysis group for providing a pleasant and inspiring atmosphere. In particular thank you to my office-mates Lars Pedersen and Mikkel B. Stegmann and former office-mate Dr. Rune Fisker for their always helpful nature and for many rewarding, and interesting discussions. I also thank Dr. Rasmus Larsen, Dr. Dan Rasmussen, Dr. Per R. Andresen, and Dr. Jens Michael Carstensen for sharing their expertise and for their readiness to provide good advice. Thank you to everyone in the GEOSONAR consortium for a productive collaboration. Especially thank you to Dr. Per Knudsen, Dr. Ole B. Andersen, Dr. Niels K. Højerslev, Dr. Thomas Knudsen, and Professor C. Christian Tscherning for numerous discussions and tutorials. Thank you to Geologist John L. Pedersen for commenting the decompositioning of geophysical data from Greenland.

I am also in debt to Professor Paul Switzer and the rest of his group at the Department of Statistics at Stanford University. I thank you for an academically as well as a personally stimulating and inspiring visit. I thank the Ingeniør Valdemar Selmer Trane og hustru Elisa Tranes Foundation, and the Otto Mønsteds Foundation for the financial support, which made my visit to Stanford possible.

Finally, last but not least, thank you to my family, both new and old members, for all your love, support, and patience when my workload was high.

Abstract

This work describes different methods that are useful in the analysis of multivariate single and multiset data. The thesis covers selected aspects of relevant data analysis techniques in this context. Methods dedicated to handling data of a spatial nature are of primary interest with focus on data driven exploratory methods for i) clustering, and for both ii) linear and iii) nonlinear decompositioning. New extensions are presented in all three fields.

Chapter 1 contains the motivation and objectives of the thesis.

Chapter 2 introduces the notion of fuzzy clustering. To overcome some of the problems of initialization, a simulated annealing controlled stochastic diffusion optimization inspired extension is presented. Spectral clustering is extended to include spatial information into the partitioning of multivariate image data. The spatial awareness is introduced by the construction of two new memberships: the spatial membership and the parental membership. The first of the new memberships is inspired and developed from Markov random field modelling, and the second is obtained from analyses performed on different scales of the image data. A case study on simulated data illustrates the improved robustness of the algorithm and thus the usefulness of the new extensions when handling noise and outlier data and when classes are overlapping in the spectral space. Two additional case studies are presented on unsupervised segmentation of multispectral remotely sensed data and of multichannel x-ray mapping imagery.

Chapter 3 describes the linear decomposition of single and multiple sets. A new extension to the maximum autocorrelation factors (MAF) transform is proposed. The new transform restricts MAF to generate components such that the covariance structure of the noise becomes the identity matrix. This

involves scaling the individual components such that the variance becomes linearly dependent to the signal-to-noise ratio in each component. The new transform is termed the signal MAF (SMAF) transform. The stretching of MAF components, favouring the subspaces rich on autocorrelated signal, gives improved results when used as a preprocessor to the spectral-spatial clustering algorithm presented in Chapter 2 on the simulated data. A case study is presented in which a new application of exploratory methods for single set analysis is presented. The combination of methods provides a decompositioning of multispectral satellite images with suppression of undesired spectra and noise. The presented data driven approach is expected to be useful in obtaining a better spatial and temporal sampling of the ocean colour in the North Sea and adjacent waters. The theory for linear multiset canonical correlations analysis is presented and applied to both a bivariate multitemporal problem, and a two-dimensional multiset shape alignment problem. A relation between the canonical correlations analysis and Procrustes alignment is found.

In Chapter 4 the focus is on nonlinear decompositioning of multiset data. The alternating conditional expectations (ACE) algorithm has previously been extended to perform canonical correlations analysis on two sets. Here ACE is generalized to handle multiple sets. The new algorithm finds estimates of the optimal transformations of the involved variables that maximize the sum of the pair-wise correlations over all sets. The new algorithm is termed multiset ACE (MACE) and can find multiple orthogonal eigensolutions. MACE is a generalization of the linear multiset correlations analysis. It handles multivariate multiset of arbitrary mixtures of both continuous and categorical variables. MACE is applied to maximize the autocorrelation in multispectral remote sensed imagery and to irregularly sampled data of geochemical data collected in South Greenland. MACE is expected to be a useful tool in providing insight in the underlying data distributions of multiset problems. It applies bivariate scatterplot smoothers for which the data analyst may specify appropriate restrictions when performing an exploratory analysis of the data.

Resumé

Dette arbejde beskriver metoder som er nyttige i analysen af multivariate enkelt sæt og multisæt data. Afhandlingen dækker kun udvalgte aspekter med fokus på data drevne eksplorative metoder til i) clusterering, og både ii) lineær og iii) ikke-linær dekomposition. Nye udvidelser præsenteres inden for alle tre felter.

Kapitel 1 omhandler motivationen og målsætningerne for afhandlingen.

Kapitel 2 introducerer fuzzy clustering. For at kompensere for initialiseringsproblemer præsenteres en udvidelse, der er inspireret af simulated annealing kontrolleret stokastisk diffusion optimering. Spektral clustering udvides til at inkludere spatial information i behandlingen af multivariate billeddata. Den spatialle udvidelse introduceres ved to nye medlemsskaber: et spatielt og et forældre-medlemskab. Det første af disse medlemskaber er inspireret af Markov random field modellering, og det andet findes ved analyser af data på forskellige skalaniveauer. Et eksempel med simulerede data præsenteres som demonstrerer den udvidede algoritmes robuste egenskaber til håndteringen af støj, outliers og klasseoverlap i det spektrale rum. To yderligere eksempler demonstrere metoden anvendt på multispektral satellitdata og på multikanal røntgendata.

Kapitel 3 beskriver lineær dekomposition af enkelt- og multisæt data. En udvidelse af maksimum autokorrelationsfaktor (MAF) transformationen præsenteres. Den nye transformation producerer MAF komponenter for hvilken kovariansstrukturen er enhedsmatricen. De enkelte komponenter skaleres så variansen er linært afhængig af signal-støj-forholdet i hver komponent. Transformation kaldes for signal MAF (SMAF) transformationen. SMAF favoriserer underrum med meget signal og er nyttig som en preprocessor. SMAF transformationen forbedrer resultaterne af en cluster analyse af simulerede data. Et eksempel på en ny applikation af eksplorative metoder for enkeltsæt præsenteres. Kombinationen af metoder giver en dekomponering af multispektral satellitdata med undertrykkelse af uønskede spektra og støj. Metoden forventes nyttig til at opnå en bedre sampling af havets spektrelle reflektans i Nordsøen og de omkringliggende farvande. Teorien for linær kanonisk korrelationsanalyse præsenteres og anvendes på et bivariat multitemporalt problem og et multisæt alignment problem for todimensionale punktfordelings modeller. En relation mellem kanonisk korrelationsanalyse og Procrustes analyse gives.

I kapitel 4 fokuseres på ikke-linær dekomposition. Alternating conditional expectations (ACE) algoritmen er tidligere udvidet til brug ved kanonisk korrelationsanalyse for tosæts problemer. ACE generaliseres til håndteringen af multisæt data. Den nye algoritme maksimerer summen af parvise korrelationer over alle sæt og kaldes for multisæt ACE (MACE). Den finder ortogonale løsninger og er en generalisering af linear multisæt kanonisk korrelationsanalyse. MACE kan håndtere arbitrære kombinationer af kontinuerte og kategoriske variable. Et eksempel præsenteres hvor MACE benyttes til at maksimere autokorrelationen i multispektrale satellitbilleder og til analyse af irregulært opsamlede geokemisk data fra Sydgrønland. MACE forventes at være et nyttigt værktøj, når man laver eksplorative analyser og leder efter indsigt i korrelationerne mellem de involverede variable og sæt. Algoritmen anvender bivariate scatterplot smoothers for hvilke specifikke restriktioner kan angives.

Contents

List of Tables

List of Figures

List of Algorithms

Chapter 1

Introduction

This work is expected to be of particular interest to researchers who need methods to apply in the first exploratory probing of multivariate data trying to investigate the joint density function with no preconceived notions or precise questions in mind. This chapter contains a section on the motivation and objectives of this work. It presents a thesis overview along with the applied mathematical notation and a list of the publications made during the Ph.D. study. The last section contains a short introduction to data complexity and dimensionality.

1.1 Motivation and Objectives

Exploratory analysis is an extensive field in modern applied statistics and involves the handling and initial processing of large data sets. The term was first introduced by J. Tukey in 1977 to describe the use of primarily graphical techniques when studying a data set, see [119]. In general, exploratory tools are data driven methods which aim to assist the data analyst in discovering important structures or hidden correlations in the data. In high dimensions it becomes difficult to estimate the joint density function of a multivariate data set, and only the most coarse properties of the density function can be revealed. If the data lumps together in different parts of the vast space, we talk about a clustering effect. Methods for cluster analysis are methods that attempt to discover if clustering is present and how many natural groups exist. In this work we are especially interested in handling data of spatial nature. Motivated by the success of spectral clustering algorithms we wish to extend these methods for handling multivariate image data. Some objectives are therefore to develop and implement methods for unsupervised image segmentation based not only on the spectral characteristics of each pixel but also on the spatial information present in the data. The aim is to improve robustness (and speed) when handling noise and overlapping classes in the spectral space.

Although the data are represented in a high dimensional space, important structures can often be revealed by projecting the observations onto a lower dimensional manifold. Exploratory tools which try to transform the data onto interesting manifolds are known as methods for linear and nonlinear decomposition. Decompositioning is useful when trying to understand the structure of a data set and methods for dimensionality reduction and redundancy analysis are therefore important. Thus, other objectives are the investigation, further development and implementation of data driven methods that can handle both single and multiset scenarios. In particular, methods for maximizing correlations between transformed groups of variables in multisets are of interest. In the literature the methods found for multiset canonical correlations analysis are all linear techniques. One aim is therefore to develop the means of performing nonlinear multiset correlations analysis. A data driven method is proposed for finding the best fitting additive model which maximizes the sum of the pair-wise correlation over all sets. Information on the structure of these models is important and aid in the interpretation and understanding of the relationship between the

underlying variables and the organization of the data.

The amount of literature on exploratory analysis is large. Good approaches to the subject are books on pattern recognition and on applied modern regression and classification techniques, [107, 46, 124, 55]. For classical methods in multivariate statistics see [2]. For a survey paper on statistical pattern recognition see e.g. [71].

1.2 Thesis Overview

The thesis is divided into four chapters:

- Chapter 1 introduces the motivation for and objectives of the work.
- Chapter 2 addresses the problem of performing unsupervised classification. A purely data driven extended fuzzy clustering algorithm is developed for handling image data, and case studies are presented in which the new algorithm is applied.
- Chapter 3 addresses the task of linear decompositioning of single set and multiset data. Relevant theory is presented and new extensions are introduced followed by their application to case studies.
- In Chapter 4 a more general approach is taken. The linear decomposition methods are relaxed to allow for nonlinear transformations of multivariate single and multisets. A new algorithm is developed for decomposing multiset data into new variates that maximize the sum of the pair-wise correlations over all sets. The algorithm can handle arbitrary mixtures of continuous and categorical variables. Relevant theory and additional case studies are presented.

The three main chapters (2-4) have been written as independently as possible allowing parts of the work to be used without requiring an understanding of the whole thesis.

1.3 Mathematical Notation

Vectors are column vectors and typeset in italic lower-case boldface using spaces to separate the elements:

$$\boldsymbol{x} = [\ a \quad b \quad c\ ]^T. \tag{1.1}$$

Matrices are typeset in italic boldface capitals:

$$\boldsymbol{A} = \begin{bmatrix} a & b \\ c & d \end{bmatrix}. \tag{1.2}$$

Sets are typeset using curly braces:

$$\{\boldsymbol{x}_i\}_{i=1}^{N}. \tag{1.3}$$

Note, both capital and lower-case non-boldface italic letters may be applied to represent the number of elements in a set or scalar variables.

The expectation values of a stochastic variable is typeset as $\mathrm{E}\{\boldsymbol{x}\}$ with the dispersion $\mathrm{D}\{\boldsymbol{x}\}$, and in the univariate case the variance $\mathrm{Var}\{x\}$. Covariance is represented by $\mathrm{Cov}\{x, y\}$ and correlation by $\mathrm{Corr}\{x, y\}$.

Inner-products are typeset as

$$< \boldsymbol{x} | \boldsymbol{y} >, \tag{1.4}$$

and the inner-product induced norm is (in squared form) represented by

$$\|\boldsymbol{x}\|^2 =< \boldsymbol{x} | \boldsymbol{x} > . \tag{1.5}$$

1.4 Publications

Some of the work in this thesis is reported in the following papers:

Proceedings

- Klaus Baggesen Hilger, Allan Aasbjerg Nielsen, Ole B. Andersen and Per Knudsen. An ACE-based Nonlinear Extension to Traditional Empirical Orthogonal Function Analysis. MultiTemp2001, Trento, Italy, 13-14 September 2001.
- Allan Aasbjerg Nielsen, Klaus Baggesen Hilger, Ole B. Andersen and Per Knudsen. A Bivariate Extension to Traditional Empirical Orthogonal Function Analysis. MultiTemp2001, Trento, Italy, 13-14 September 2001.
- Allan Aasbjerg Nielsen, Klaus Baggesen Hilger, Ole B. Andersen and Per Knudsen. A Temporal Extension to Traditional Empirical Orthogonal Function Analysis. MultiTemp2001, Trento, Italy, 13-14 September 2001.

- Klaus Baggesen Hilger, Mikkel B. Stegmann, MADCam - The Multispectral Active Decomposition Camera. Proceedings of the 10th Danish Conference on Pattern Recognition and Image Analysis, Copenhagen, Denmark, 2001.
- Klaus Baggesen Hilger, Allan Aasbjerg Nielsen and Rasmus Larsen. A Scheme for Initial Exploratory Data Analysis of Multivariate Image Data. Proceedings of 12th Scandinavian Conference on Image Analysis (SCIA), pp. 717-724. Bergen, Norway, 11-14 June 2001.
- Klaus Baggesen Hilger, Allan Aasbjerg Nielsen, Per Knudsen and Ole B. Andersen. Enhancement of Ocean Related Signal by Suppression of Undesired Spectra in Remotely Sensed Multivariate SeaWiFS Images in the GEOSONAR Project. American Geophysical Union (AGU) Fall Meeting, San Francisco, California, USA, 15-19 December 2000.
- Klaus Baggesen Hilger and Allan Aasbjerg Nielsen. Targeting input data for change detection studies by suppression of undesired spectra. In Proceedings of Seminar on Remote sensing and image analysis techniques for revision of topographic databases, KMS, The National Survey and Cadastre, Copenhagen, Denmark. 29 February 2000.
- Klaus Baggesen Hilger, Allan Aasbjerg Nielsen and Jens Michael Carstensen. Unsupervised classification of x-ray mapping images of polished sections. Proceedings of MVA2000, IAPR Workshop on Machine Vision Applications, pp. 44-47, Tokyo, Japan, 28-30 November 2000.
- Allan Aasbjerg Nielsen and Klaus Baggesen Hilger. Spectral-spatial decomposition of multivariate SeaWiFS images with suppression of undesired spectra and noise. Abstracts from Oceans from Space, Venice 2000, p. 205, Venice, Italy, 9-13 October 2000.
- Klaus Baggesen Hilger and Allan Aasbjerg Nielsen. Unsupervised Fuzzy Clustering of Multivariate Image Data. 11. Havforskermøde (Danish ocean researchers' meeting), Roskilde University Center, Denmark, 26-28 January 2000.
- Klaus Baggesen Hilger, Allan Aasbjerg Nielsen and Per Knudsen. Aspects of remote sensing of the GEOid and Sea level Of the North Atlantic Region (GEOSONAR) project. 11. Havforskermøde (Danish ocean researchers' meeting), Roskilde University Center, Denmark, 26-28 January 2000.
- Per Knudsen, Ole B. Andersen, Thomas Knudsen, Olwijn Leeuwenburgh, Jacob Lorentzen Høyer, Niels Flemming Carlsen, Allan Aasbjerg Nielsen, Klaus Baggesen Hilger, C. Christian Tscherning, Niels Kristian Højerslev, Guilhem Moreaux, Erik Buch og Vibeke Huess. The GEOSONAR Project. 11. Havforskermøde (Danish ocean researchers' meeting), Roskilde University Center, Denmark, 26-28 January 2000.
- Klaus Baggesen Hilger, Allan Aasbjerg Nielsen and Per Knudsen. Aspects of remote sensing of the GEOid and Sea level Of the North Atlantic Region (GEOSONAR) project. In Bjarne Kjær Ersbøll and Peter Johansen (editors) Proceedings of the Scandinavian Image Analysis Conference (SCIA), vol. 2, pp. 881-888, Kangerlussuaq, Greenland, 7-11 June 1999.
- Allan Aasbjerg Nielsen, Klaus Baggesen Hilger Ole B. Andersen, Niels F. Carlsen and Per Knudsen. Remote sensing of the GEOid and Sea level Of the North Atlantic Region (GEOSONAR) project. In Geophysical Research Abstracts, volume 1, number 1, p. 221, European Geophysical Society (EGS), XXIV General Assembly, The Hague, The Netherlands, 19-23 April 1999. Invited contribution.

Technical Reports

- Allan Aasbjerg Nielsen and Klaus Baggesen Hilger (2000). COMB Data Report on Unsupervised Classification of X-Ray Mapping Images of Thin Sections.
- Klaus Baggesen Hilger and Allan Aasbjerg Nielsen (1999). Statistical Transformations of SeaWiFS data: GEOSONAR Data Report 1. Department of Mathematical Modelling, Technical University of Denmark.

1.5 The Curse of Dimensionality

When working with multivariate multiset data problems, it pays to have an intuitive understanding of the character of high-dimensional spaces. This section is an attempt to provide this intuitive understanding by briefly conveying the basics of working in high dimensions in which we find many manifestations of what is known as "the curse of dimensionality."

Bellman, [5], first introduced the term in 1961 and used it to describe the complexity of computations when the number of computations exceeds the

available computing power. Shortly hereafter, in 1966 Elsasser, [29, 111], coined the term "immense" to describe numbers larger than

$$\mathcal{I} = 10^{110}, \tag{1.6}$$

a number which equals the atomic mass of the universe measured in units of the mass of a hydrogen atom (10^{80}) multiplied by the age of the universe measured in picoseconds (10^{30}). Generally, one assumes that any number of items can be put on a list and examined one by one. For an immense number of items, however, this is not possible. First of all, there would not be enough mass in the universe to represent the list, even if only a single atom were used to remember each item on the list. Secondly, if such a list existed, we would not have the time to read through it by any conceivable means.

Another use of the term "curse of dimensionality" relates to the problems associated with the feasibility of density estimation in many dimensions. Samples quickly become lost in the vast space when the dimensionality becomes too large. Local neighbourhoods in a high-dimensional space are likely to be void of observations. When neighbourhoods are extended to include a sufficient number of observations, they become so large that they effectively provide global, rather than local, density estimates. To relieve the problem by filling the space with observations would soon require infeasibly large sample sizes.

The character of high-dimensional spaces can run counter to one's intuition, which tends to be based on low-dimensional Euclidean spaces. Consider a p dimensional hyper-cubic space with edge length r. The proportion of the volume that is contained between the surface of the hypercube and a smaller one with edges of length $r - \epsilon$ is

$$\frac{r^p - (r-\epsilon)^p}{r^p} = 1 - (1 - \frac{\epsilon}{r})^p \to 1 \text{ for } p \to \infty. \tag{1.7}$$

Thus, in a high-dimensional hyper-cubic region most of the available space is spread around the surface of the region. The same is true of regions with other shapes, e.g. spheres.

Consider a p dimensional unit sphere with N uniformly distributed data, $\{\boldsymbol{x}_i\}_{i=1}^N$, centered at the origin. The distance from each observation to the origin is represented by the set $\{d_i\}_{i=1}^N$. The distribution function for the minimum distance $d_{\min}$ of the closest point to the origin in one dimension

is given by

$$\begin{aligned} G(d) &= \mathrm{P}\{d_{\min} \leq d\} \\ &= 1 - \mathrm{P}\{d_{\min} > d\} \\ &= 1 - \mathrm{P}\{d_1 > d, \cdots, d_N > d\} \\ &= 1 - (1 - F(d))^n. \end{aligned} \tag{1.8}$$

where $F(d)$ is the distribution function of d. Then in higher dimensions the median distance from the origin to the closest data point becomes

$$d_{\text{median}} = (1 - (1/2)^{1/N})^{1/p}. \tag{1.9}$$

In Figure 1.1 a contour plot of the median distance from the origin to the nearest point is plotted as a function of observations and dimensions. Notice how rapidly the distance increases with the dimensions and how quickly the sample size must grow in order to preserve the notion of local nearest-point neighbourhoods. In, say, 10 dimensions with 1000 observations we must travel almost half-way to the boundary to find the median distance of the nearest point.

Consider uniformly distributed observations in an unit hypercube. Given a target point from which we wish to capture a fraction f of the neighbouring observations, the edge length of the required hypercube centered at the target point becomes $\tilde{r}_p(f) = f^{1/p}$. In a multivariate case of e.g. ten dimensions we therefore need to cover 63% of the entire range of each variable to capture only 1% of the data. Notice also that the sampling density is proportional to $N^{1/p}$. If we have a density of 1000 observations in a one dimensional space, we need $1000^{10} = 10^{30}$ observations in a ten dimensional space to preserve the sample density. In a 37 dimensional space the required number of observations would already be immense. Therefore, in high dimensions all feasible training samples will sparsely populate the input space.

In summary, the curse of dimensionality must be taken into account when considering the feasibility of performing density estimation in many dimensions. Sometimes multidimensional smoothers can work with a moderate number of inputs, but the curse hinders them in higher dimensions. The manifestations of the curse are that

- local neighbourhoods are empty, or
- nearest-point neighbourhoods are not local,

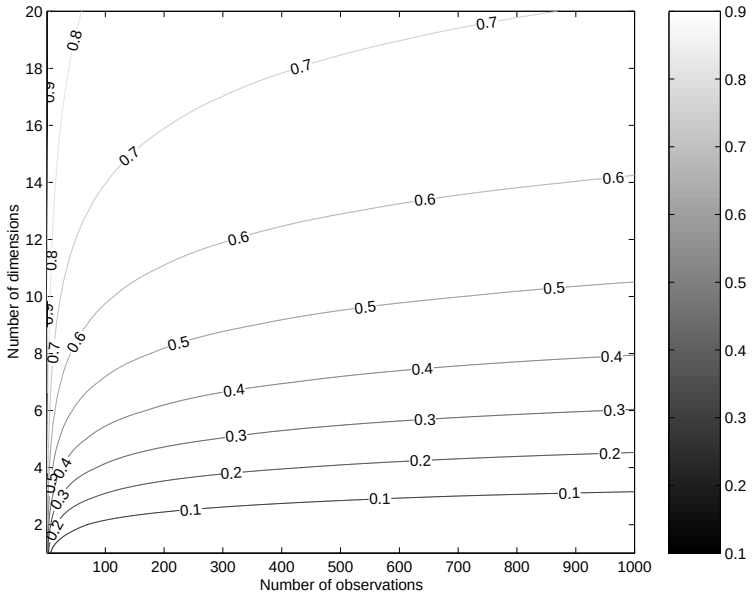


Figure 1.1: A contour plot of the median distance from the origin to the nearest point in a hyper-ball with uniform sample density.

- all points are close to the boundary, and
- the sample size needs to grow exponentially with the dimensions in order to preserve the sampling density.

In addition, high-dimensional functions are difficult to represent and interpret, and high-dimensional procedures should assume some regularizing structure.

In high dimensions the curse of dimensionality makes it impossible to detect all but the very coarsest structure of the joint probability function, and the most one can hope for is to be able to get a general idea of the density function. Cluster analysis is one approach for doing this. The goal is to discover if the density is small everywhere, except for a small number of isolated regions (where it is large). This effect is known as clustering and clustering algorithms attempt to determine when this condition exists and to identify the isolated regions. Dimensionality reducing methods are very useful tools as they can be applied to remove redundancy and to

find subspaces in high-dimensions where interesting correlation structures are present. Linear and nonlinear decomposition methods can thus be applied to locate manifolds onto which the projected data reveal important, otherwise occluded, information.

The performance of an algorithm depends on the interrelationship between both the sample size, the number of features/dimensions, and the complexity of the algorithm. In exploratory analysis the methods for cluster analysis and for decomposition can be useful when applied individually, but they may also be beneficial when used in a pre- or post-processing of each other.

Chapter 2

Unsupervised Classification

This chapter introduces the notion of unsupervised classification. Spectral fuzzy clustering algorithms are presented, and in addition, new fuzzy memberships are introduced such that spatial awareness is included into the partitioning algorithm when handling image data. The new memberships have a regularizing effect and can be advantageous to apply when the traditional spectral clustering approach produces less useful results. Case studies are presented on i) a toy image to evaluate performance, ii) a multispectral satellite image, and iii) a multichannel scanning electron microscope x-ray mapping image.

2.1 Density Estimation and Cluster Analysis

In many applications of pattern recognition it is extremely difficult or even impossible to reliably label a training sample with its true category. Unsupervised classification refers to situations in which the objective is to construct decision boundaries based on unlabeled data. Unsupervised classification is also known as data clustering which is a generic label for a variety of procedures designed to discover natural groupings in multidimensional data based on similarity measures.

Clustering is a difficult problem because data can reveal clusters with different shapes and sizes. Furthermore, the number of clusters in the data often depends on the resolution with which we view the data. A number of functional definitions of a cluster have been proposed which include i) patterns within a cluster are more similar to each other than patterns belonging to different clusters, and ii) a cluster consists of a relatively high density of points separated from other clusters by a relatively low density of points. Even with these functional definitions of a cluster, it is not easy to come up with an operational definition of clusters. One of the challenges is to select an appropriate measure of similarity to define clusters which often is both data and context dependent.

Measures used to quantify clustering algorithms are often related to speed, reliability and consistency. Application of unsupervised classifiers are found in a wide range of fields from data mining, machine learning, and image segmentation, see [71] for selected references. Consequently, several clustering algorithms have been proposed in the literature and new clustering algorithms continue to appear.

Most of the clustering methods can be organized into two types i.e. hierarchical methods and objective function methods.

Hierarchical techniques organize data in tree structures known as dendograms. Often the methods are computationally and conceptually simple, but may produce nice results that are easy to interpret. Graph based methods can be included into the group of hierarchical methods. In these methods each observation is considered as a node in a graph, and the edge weights between pairs of nodes are proportional to the measure of similarity between the observations. Graph theoretic methods are then applied in the clustering strategy breaking edges to form subgraphs.

Objective function methods are often iterative square-error partitioning clustering algorithms which attempt to obtain the partition that minimize the within-cluster scatter or maximize the between-cluster scatter. These methods are often also referred to as partitional clustering techniques. See [9] for an introduction to clustering using objective function algorithms.

When clustering image data a typical approach is to include the spatial information in the feature extraction before proceeding with the feature selection. One can subsequently perform the clustering in the feature space, and in order to further exploit the knowledge of the spatial structure of the data, one may post-process the clustering results with a spatial relaxation algorithm. An example hereof is to do hard clustering (e.g. by the traditional k-means algorithm, [83, 1]) on features found e.g. by using different spatial convolution filters and post-process the result by means of a Potts model, [131, 43, 44].

In this work we will incorporate the spatial element into a spectral clustering algorithm which assumes that the class densities can be described by multidimensional Gaussians. When nothing is known, the Central Limit Theorem tells us that the most reasonable choice for the distribution of a random variable is Gaussian. Furthermore we will use a fuzzy approach assigning every observation memberships to the different classes thus allowing for overlapping classes in the spectral space. The spatial element is introduced by i) applying spatial cliques potentials, known from the Markov random field (MRF) theory, together with ii) information extracted from an analysis of the data scale-space. Including the spatial information into the clustering algorithm has a regularizing effect and enhances the robustness for handling overlapping and non-Gaussian clusters. Naturally, in order for the new algorithm to be successful, the notion of cluster centers must apply to the data across scales.

Users of a clustering algorithm should keep the following issues in mind: i) every clustering algorithm will find clusters in a given data set, whether they exist or not, and ii) there is no "best" clustering algorithm. Therefore, a user is adviced to try several clustering algorithms on a given data set. Issues such as data representation and normalization are as important as the choice of clustering strategy.

References to clustering algorithms and surveys of existing methods can be found in [8, 34, 4], introductory books recommended in the literature are [72, 73]. Finally, the interaction between a data analyst and the applied

algorithms is often an integral part of a full clustering analysis. A user looks at the results of an initial clustering and, as a result, changes the parameters of the clustering algorithm or chooses a new clustering algorithm. Numerous iterations may be required before a suitable partitioning is obtained.

2.2 Fuzzy Partitioning

Given N P-dimensional observations that we wish to classify into C classes, the fuzzy c-means (FCM) clustering algorithm applies. Let $\boldsymbol{U}$ be an $N \times C$ matrix with the elements $u_{ic} = u_c(\boldsymbol{x})$ that describe the membership for observation i at grid point $\boldsymbol{x}$ to class c, and let $\boldsymbol{R}$ be a $P \times C$ matrix that contains the cluster centers, $\{\tilde{\boldsymbol{r}}_c\}_{c=1}^{C}$, i.e. the centroids of the classes. The spectral FCM algorithm minimizes the within class sum of squared errors functional $J(\boldsymbol{U}, \boldsymbol{R})$ under the conditions

$$u_{ic} \in [0,1] \quad \forall\, i = 1, \ldots, N; c = 1, \ldots, C \tag{2.1}$$

$$\sum_{c=1}^{C} u_{ic} = 1 \qquad \forall\, i = 1, \ldots, N \tag{2.2}$$

$$\sum_{i=1}^{N} u_{ic} > 0 \qquad \forall\, c = 1, \ldots, C. \tag{2.3}$$

The condition 2.2 is sometimes referred to as the probabilistic constraint, and FCM as probabilistic clustering.

The functional proposed by Bezdek, [9], is given by

$$J(\boldsymbol{U}, \boldsymbol{R}) = \sum_{i=1}^{N} \sum_{c=1}^{C} u_{ic}^{m} d_{ic}^{2}. \tag{2.4}$$

Here $m > 1$ is a fixed parameter that determines the degree of fuzziness of the final solution, that is the degree of overlap between groups. The degree of fuzziness increases with m. The higher the value of m, the lower the membership degrees of observations which are far from all cluster centers. As m approaches infinity, the solution approaches its highest degree of fuzziness, with $u_{ic} = 1/C$ for every pair of i and c. The squared distance between the ith observation and the cth cluster center is used as difference

(inverse similarity) measure and denoted $d_{ic}^2 = \|\boldsymbol{r}_i - \tilde{\boldsymbol{r}}_c\|^2$, where $\|\cdot\|$ is any inner product induced norm on $\mathbb{R}^P$.

Minimization of the spectral functional

We assume that the minimization can be done by applying a Pichart iteration scheme, [9]. Convergence is thus expected by iteratively keeping i) the memberships fixed while updating the cluster centers, and ii) the centers fixed while updating the memberships. The updating formulas for the memberships and the cluster centers are derived from 2.4 using the Lagrangian multiplier technique.

Assuming fixed cluster centers, $\{\tilde{\boldsymbol{r}}_c\}_{c=1}^C$, we define $j_R(\boldsymbol{U}) = J(\boldsymbol{U}, \boldsymbol{R})$ which we wish to minimize with respect to $\boldsymbol{U}$ under the constraint in 2.2. Thus,

$$\min_{\boldsymbol{U}}[j_R(\boldsymbol{U})] = \min_{\boldsymbol{U}}[\sum_{i=1}^N \sum_{c=1}^C u_{ic}^m d_{ic}^2] \tag{2.5}$$

$$= \sum_{i=1}^N \min_{\boldsymbol{u}_i}[\sum_{c=1}^C u_{ic}^m d_{ic}^2], \tag{2.6}$$

where $\boldsymbol{u}_i$ is the ith row of $\boldsymbol{U}$. In the above we have used that for fixed cluster centers the membership for an observation is independent of the memberships of all other observations. For each inner term in 2.6 let

$$g_i(\boldsymbol{u}_i) = \sum_{c=1}^C u_{ic}^m d_{ic}^2, \tag{2.7}$$

the corresponding Lagrangian can then be constructed as

$$L_i(\boldsymbol{u}_i, \lambda) = \sum_{c=1}^C u_{ic}^m d_{ic}^2 - \lambda(\sum_{c=1}^C u_{ic} - 1), \tag{2.8}$$

for which stationarity is found iif $\nabla_{\boldsymbol{u}_i,\lambda} L_i(\boldsymbol{u}_i, \lambda) = \boldsymbol{0}$. Setting the gradient equal to zero produces

$$\frac{\partial L_i(\boldsymbol{u}_i, \lambda)}{\partial \lambda} = 1 - \sum_{c=1}^C u_{ic} = 0 \tag{2.9}$$

$$\frac{\partial L_i(\boldsymbol{u}_i, \lambda)}{\partial u_{ic}} = m u_{ic}^{m-1} d_{ic}^2 - \lambda = 0. \tag{2.10}$$

From 2.10 we obtain

$$u_{ic} = [\lambda / m d_{ic}^2]^{1/(m-1)} \tag{2.11}$$

using 2.9 we get

$$(\lambda/m)^{1/(m-1)} = 1/\sum_{c=1}^C (1/d_{ic}^2)^{1/(m-1)} \tag{2.12}$$

and returning to 2.11 we obtain the update formula for the memberships of observation i to cluster c for fixed cluster centers:

$$u_{ic} = \frac{1/d_{ic}^{2/(m-1)}}{\sum_{j=1}^C 1/d_{ij}^{2/(m-1)}}. \tag{2.13}$$

For robustness, if one or more observations and a cluster center are identical, the memberships for the observations to the corresponding class will be assigned to one, and zero memberships to the remaining classes. Cases of coinciding observations and class centers hardly ever occur in practice due to machine roundoff.

Assuming fixed memberships, $\{\{u_{ic}\}_{i=1}^N\}_{c=1}^C$, we define $j_U(\boldsymbol{R}) = J(\boldsymbol{U}, \boldsymbol{R})$ which we wish to minimize with respect to $\boldsymbol{R}$

$$\min_{\boldsymbol{R}}[j_U(\boldsymbol{R})] = \min_{\boldsymbol{R}}[\sum_{i=1}^N \sum_{c=1}^C u_{ic}^m d_{ic}^2] \tag{2.14}$$

$$= \sum_{c=1}^C \min_{\tilde{\boldsymbol{r}}_c}[\sum_{i=1}^N u_{ic}^m < \boldsymbol{r}_i - \tilde{\boldsymbol{r}}_c | \boldsymbol{r}_i - \tilde{\boldsymbol{r}}_c >]. \tag{2.15}$$

where $< \cdot | \cdot >$ is the norm-inducing inner product. We have used the fact that, for fixed memberships, the location of each cluster center is independent of the centers of the other clusters. For each c we define

$$g_c(t) = \sum_{i=1}^N u_{ic}^m < \boldsymbol{r}_i - \tilde{\boldsymbol{r}}_c - t\boldsymbol{r'} | \boldsymbol{r}_i - \tilde{\boldsymbol{r}}_c - t\boldsymbol{r'} > \tag{2.16}$$

where $\boldsymbol{r'} \in \mathbb{R}^P$ and $t \in \mathbb{R}$. The minimization is unconstrained, so the directional derivatives $dg_c(t)/dt|_{t=0}$ must vanish for all $\boldsymbol{r'}$. Thus,

$$\frac{d}{dt}[\sum_{i=1}^N u_{ic}^m < \boldsymbol{r}_i - \tilde{\boldsymbol{r}}_c - t\boldsymbol{r'} | \boldsymbol{r}_i - \tilde{\boldsymbol{r}}_c - t\boldsymbol{r'} >]|_{t=0} = 0 \quad \forall \quad \boldsymbol{r'} \tag{2.17}$$

or

$$-2 \quad \sum_{i=1}^{N} u_{ic}^m < \boldsymbol{r}_i - \tilde{\boldsymbol{r}}_c | \boldsymbol{r'} > \quad = 0 \quad \forall \quad \boldsymbol{r'} \tag{2.18}$$

$$\Leftrightarrow \quad \sum_{i=1}^{N} u_{ic}^m (\boldsymbol{r}_i - \tilde{\boldsymbol{r}}_c) = \boldsymbol{0}. \tag{2.19}$$

We are now able to write the update formula for each cluster center c for fixed memberships:

$$\tilde{\boldsymbol{r}}_c = \frac{\sum_{i=1}^{N} u_{ic}^m \boldsymbol{r}_i}{\sum_{i=1}^{N} u_{ic}^m}. \tag{2.20}$$

Spectral Fuzzy Clustering

A general form for a spectral fuzzy algorithm is presented in Algorithm 1. Applying the Equations 2.13 and 2.20 as update formulas for the memberships and the cluster centers respectively results in the traditional c-means algorithm. The spectral distance d_{ic} from the running observation $\boldsymbol{r}_i$ to each cluster center $\tilde{\boldsymbol{r}}_c$ is typically calculated from

$$d_{ic}^2 = (\boldsymbol{r}_i - \tilde{\boldsymbol{r}}_c)^T \boldsymbol{F}_c^{-1} (\boldsymbol{r}_i - \tilde{\boldsymbol{r}}_c). \tag{2.21}$$

With $\boldsymbol{F}_c = \boldsymbol{I}$ the Euclidean distance measure is applied. Using the Mahalanobis distance $\boldsymbol{F}_c$ can be calculated from

$$\boldsymbol{F}_c = \frac{\sum_{i=1}^{N} u_{ic}^m (\boldsymbol{r}_i - \tilde{\boldsymbol{r}}_c)(\boldsymbol{r}_i - \tilde{\boldsymbol{r}}_c)^T}{\sum_{j=1}^{N} u_{ic}^m}. \tag{2.22}$$

The Mahalanobis distance seems the most sensible intuitive choice. However, since all observations in the fuzzy framework contribute in all the class covariance structurs, the application of the Mahalanobis distance measure slows the FCM algorithm down considerably. It also increases the sensitivity to good initialization, and furthermore introduces the risk of encountering numerical problems due to singular matrix inversions.

Cluster Validation

The methods described in this section all share the problem that the algorithms have to be provided with the initial number of classes. Determining

Algorithm 1 Spectral Fuzzy Clustering

1: Initialize cluster centers in the spectral space
2: **repeat**
3: Estimate spectral memberships, $\{\{u_{ic}\}_{i=1}^N\}_{c=1}^C$, from the spectral distances to the cluster centers
4: Estimate cluster centers, $\{\tilde{\boldsymbol{r}}_c\}_{c=1}^C$, using the cluster memberships as spectral weights
5: **until** Convergence

the right number of classes to apply is a difficult problem and often involves a data analyst. However, some success has been achieved applying the spectral FCM for different numbers of classes and evaluating each partitioning by the so called fuzzy partition density.

We assume that good partitioning consists of i) clear separation between the resulting clusters, ii) minimal volume of the clusters, and iii) maximum number of data points concentrated in the vicinity of the cluster center.

Introducing the fuzzy hyper-volume

$$F_{HV} = \sum_{i=1}^{C} [\det(\boldsymbol{F}_i)]^{1/2}, \tag{2.23}$$

where $\boldsymbol{F}_i$ is defined in Equation 2.22, and summing the memberships of the observations within one standard deviation of each cluster center

$$\begin{aligned} S \quad &= \quad \sum_{i=1}^{N} \sum_{c=1}^{C} u_{ic} \\ &\forall \quad \boldsymbol{r}_i \in \{\boldsymbol{r}_i : (\boldsymbol{r}_i - \tilde{\boldsymbol{r}}_c)^T \boldsymbol{F}_c^{-1} (\boldsymbol{r}_i - \tilde{\boldsymbol{r}}_c) < 1\}, \end{aligned} \tag{2.24}$$

the partition density, [41], is defined by

$$P_D = S / F_{HV}. \tag{2.25}$$

Local maxima in the partition density are expected where good partitionings of the data are obtained. Other measures of cluster validation can be found in [106]. Application of the fuzzy partition density can be found in [41, 57].

Hard k-Means

Discarding the fuzzy framework hard clustering (HCM), which does not allow for overlapping clusters, can be obtained when 2.1 is restricted to

$$u_{ic} \in \{0,1\} \quad \forall\, i = 1,\ldots,N; c = 1,\ldots,C. \tag{2.26}$$

In HCM minimization is based on the sum-of-squared-errors function

$$J_{HCM}(\boldsymbol{U},\boldsymbol{R}) = \sum_{i=1}^{N}\sum_{c=1}^{C} u_{ic} d_{ic}^2. \tag{2.27}$$

The update formulas for the memberships and the cluster centers are now

$$u_{ic} = \begin{cases} 1 & d_{ic} < d_{ik} \;\; \forall\, k \neq c \\ 0 & \text{otherwise} \end{cases} \tag{2.28}$$

$$\tilde{\boldsymbol{r}}_c = \frac{\sum_{i=1}^{N} u_{ic}\boldsymbol{r}_i}{\sum_{i=1}^{N} u_{ic}}. \tag{2.29}$$

In 2.28 ties are resolved arbitrarily.

Noise Clustering

The HCM and the FCM clustering methods have a common disadvantage since they are both sensitive to outliers. The membership of a feature vector across classes always sums to one for both clean and noisy data. It would be more sensible if outliers had memberships that where as small as possible, i.e. the sum should be smaller than one. The idea of a noise cluster has been considered in [27, 26, 118] to deal with noisy data or outliers for fuzzy clustering methods. The noise is considered to be a separate class and is represented by a pseudo cluster center with constant distance d_{noise} from all feature vectors. The membership of an observation i in the noise cluster is defined to be

$$u_{i,noise} = 1 - \sum_{c=1}^{C} u_{ic} \quad \forall\, i = 1,\ldots,N. \tag{2.30}$$

Therefore, the probabilistic constraint is relaxed to

$$\sum_{c=1}^{C} u_{ic} < 1 \quad \forall\, i = 1,\ldots,N. \tag{2.31}$$

Outliers and noisy data are now allowed to have arbitrarily small memberships in good clusters. The objective function follows as

$$J_{Noise}(\boldsymbol{U},\boldsymbol{R}) = \sum_{i=1}^{N}\sum_{c=1}^{C} u_{ic}^m d_{ic}^2 + \sum_{i=1}^{N}\sum_{c=1}^{C}(1-u_{ic}^m) d_{noise}^2. \tag{2.32}$$

Minimizing the functional under the constraint in Equation 2.30 with respect to u_{ic} given the update formula

$$u_{ic} = 1/(\sum_{j=1}^{C}(d_{ic}/d_{ij})^{2/(m-1)} + (d_{ic}/d_{noise})^{2/(m-1)}). \tag{2.33}$$

The update formula for the cluster centers for the noise clustering algorithm (NCM) remains unchanged as in 2.20. How to choose d_{noise} relies on the user and must often be based on prior analysis of the data. Other types of "robust" clustering approaches have been proposed; in general they all include the notion of an extra membership class representing noise, see e.g. [18, 78]. A recent approach, [31], tries to avoid the problem of choosing d_{noise} by relaxing the probabilistic constraint to

$$\sum_{c=1}^{C} u_{ic}^p = 1 \quad \forall\, i = 1,\ldots,N \tag{2.34}$$

and uses Equation 2.4 as fuzzy functional. The resulting update formula for the memberships now becomes

$$u_{ic} = [\frac{1/d_{ic}^{2/(m-1)}}{\sum_{j=1}^{C} 1/d_{ij}^{2/(m-1)}}]^p \tag{2.35}$$

and the formula for the cluster centers remains unchanged as in the FCM case. Ad-hoc noise clustering algorithms have also been presented in which the memberships are reduced when more than one class is competing to dominate the same observation.

Expectation Maximization

The fuzzy c-means is closely related to the application of the Expectation Maximization (EM) algorithm applied to Gaussian mixture problems. EM

is a general method for carrying out maximum likelihood estimation. Direct maximization is often difficult but made easier by the EM approach through the introduction of so called latent data. In the FCM case the class memberships correspond to the latent data, which is computed in the *Expectation* step of an EM algorithm. The density parameters are then estimated in the *Maximization* step, that is the class means and covariance structures, in effect resulting in the previously presented FCM algorithm.

The continued re-estimation of the class dispersions is a computational burden for high-dimensional large sample size problems and makes the algorithm less attractive since demands of speed of the segmentation come into play. Moreover, it is well known that the EM algorithm needs good initialization in order to converge to the optimal solution and may easily get caught in local minima. Thus, this also becomes the case for the FCM approach, and a fast more rough data-driven segmentation is needed for initialization. Rough methods, e.g. using the same metric for all classes, are typically applied to provide such initializations, but even then initialization of the cluster centers is an important issue.

2.2.1 Initialization

Often a priori knowledge can be applied e.g. including information on the expected relative sizes of the classes, their class means and dispersions. However, data driven initialization schemes are still needed, and these can be organized into random or deterministic methods. Note that one approach is to do repeated segmentation using a carpet bombing brute force approach with different initializations and to evaluate the fuzzy functional from each partitioning in order to find the optimal segmentation.

Typically, a random initialization would consist of the selection of C random different observations as initial class means. A deterministic initialization could be to make equidistant distribution of the class centers in e.g. the subspace with the maximal variation. An interesting initialization is the one proposed in [66] for the SAS fastclus algorithm, in which clusters are added in a heuristic manner such that each cluster center is chosen as an existing observation from the data while trying to obtain maximum separation between the initial clusters. Other approaches include cluster adding methods where one first evolves two clusters and while iterating adds additional cluster centers in a heuristic manner. Methods which apply both

cluster adding and cluster merging can also be found. This it typically done by eliminating relatively small clusters with few observations, and splitting large clusters with a large number of observations.

Simulated Annealing Optimization

Even the clever initialization algorithms can get caught in local minima. Our approach to overcome this problem is to include the clustering algorithm into a scheme inspired by Stochastic Diffusion (SD) optimization controlled by a Simulated Annealing (SA) temperature schedule. The idea is to gradually let the calculated class memberships have more influence on the estimation of the class centers. Such a scheme is presented in Algorithm 2.

A typical Simulated Annealing (SA) scheme, [121, 19], is a successive sampling, using e.g. Gibbs sampling or the Metropolis algorithm, from the density

$$\mathrm{P}_T\{\boldsymbol{x}|\boldsymbol{y}\} \quad \propto \quad [\mathrm{P}\{\boldsymbol{y}|\boldsymbol{x}\}\mathrm{P}\{\boldsymbol{x}\}]^{1/T}, \tag{2.36}$$

which comes from Bayes theorem relating the posterior distribution to the observation model and the prior distribution. The temperature T is initialized by $T_0 > 0$ and drops towards zero in an SA scheme. In the limit Equation 2.36 assigns unit probability to the Maximum A Posteriori (MAP) estimate, if the temperature is reduced sufficiently slowly, [43]. One task is to choose the optimal cooling scheme defined by $T_{t+1} = f(T_t, t, \boldsymbol{c})$ where T_t is the temperature at time t, and $\boldsymbol{c}$ is a parameter vector for the particular SA scheme involved. Popular cooling schedules are $f_{\text{Geometric}} = cT_t$, $f_{\text{Lundy}} = T_t/(1 + cT_t)$, and $f_{\text{Logarithmic}} = c_1/\log(c_2 + c_3 t)$.

Stochastic diffusion, [47, 45], is a method for finding the global minimum energy state of some objective function with many local minima. Its behaviour in the potential field is controlled by a model in which the flow is determined as the direction of steepest decent corrupted by some additive noise term. We want the noise to have decreasing influence on the direction of the flow as the iterations proceed and may therefore use the temperature from the SA schedule as a parameter for controlling the strength of the noise.

For each observation i to each cluster c a noise membership is assigned, $u_{rand,ic}$. By default the noise membership is sampled from a uniform dis-

Algorithm 2 Spectral Fuzzy Clustering with Simulated Annealing

1: Set initial temperature T_0, the SA parameters $\boldsymbol{c}$ and $t = 0$.
2: Initialize cluster centers
3: **repeat**
4: Set $t = t + 1$ and update $T_t = f(T_t, t)$
5: Estimate spectral memberships
6: Estimate random memberships $\{\{u_{rand,ic}\}_{i=1}^{N}\}_{c=1}^{C}$
7: Merge the spectral and the random memberships depending on the temperature
8: Estimate cluster centers from the merged memberships
9: **until** Convergence

tribution from zero to one, $U(0,1)$. Depending on the temperature the noise membership is merged with the spectral membership producing a joint membership for each observation to each cluster. This joint membership is calculated from

$$u_{ic} = \frac{u_{spec,ic} f(u_{rand,ic}, T_t)}{\sum_{j=1}^{C} u_{spec,ij} f(u_{rand,ic}, T_t)} \tag{2.37}$$

where

$$f(u_{rand,ic}, T_t) = 1 - u_{rand,ic}^{1/T_t}. \tag{2.38}$$

The noise memberships are transformed using a temperature dependent transformation $f = f(u_{rand,ic}, T_t)$. With the transform proposed in Equation 2.38, the influence of the noise memberships in the joint membership becomes negligible as the number of iterations increases and the temperature is decreased. The default temperature schedule is the geometric scheme using $T_0 = 1$ and $c = 0.8$. An example of the decay of the transformed noise memberships is shown in Figure 2.1. At each time step 100 simulated transformed noise memberships are shown. Notice that as the number of iterations increases, the variance of the transformed noise memberships decreases and the memberships approach one.

2.2.2 Extended Spectral-Spatial Fuzzy Clustering

Two additional memberships, the spatial and the parental, are introduced that include spatial context information. Before calculating the new cluster

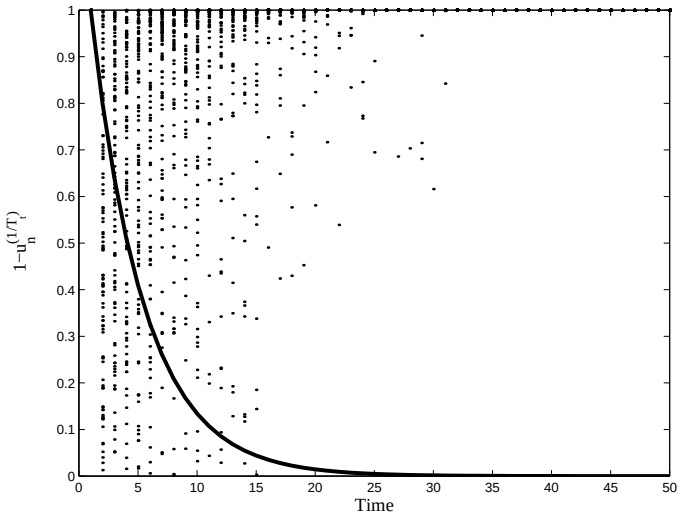


Figure 2.1: The decay of the noise membership as a function of time (no. of iterations). The solid line is the applied temperature and the dots are 100 simulations of transformed noise memberships at each iteration.

centers, we merge the spectral, the spatial, and the parental memberships into joint memberships which are applied as weights for the spectral observations when calculating new cluster centers.

The spatial membership

Applying the Hammersley-Clifford theorem (the Markov random field (MRF) - Gibbs random field (GRF) equivalence theorem, [7]), the posterior probability in Equation 2.36 is related to an MRF energy function, V, defined with respect to a neighbourhood system. A random field $\boldsymbol{x}$ is an MRF with respect to a neighbourhood system $\mathcal{N}$ iif the probability of observing a given outcome of $\boldsymbol{x}$ is a Gibbs distribution with respect to $\mathcal{N}$. Thus the probability of an observation can be determined as a conditional probability given the configuration of the local neighbours. The Gibbs measure is induced by the energy function V

$$\mathrm{P}\{\boldsymbol{x}\} = \frac{1}{Z} \exp[-\frac{1}{T} V(\boldsymbol{x})] \tag{2.39}$$

where T is a control parameter usually referred to as temperature and $Z = \sum \exp[-\frac{1}{T}V(\boldsymbol{x})]$ is the partition function which sums over all possible outcomes of $\boldsymbol{x}$. Geman and Geman, [43], use coupled MRFs for image reconstruction by applying an energy function defined by a sum of clique potentials relating the energy contribution of different pixel configurations. Favouring homogeneous configurations, the MRF can be used as a regularizing factor in, say, noise degraded image reconstruction.

Note, that any random field with a positive probability of all posible configurations is a Markov field, if we can specify a neighbourhood large enough to encompass the conditional dependencies. The structure of the neighbourhood system determines the order of the MRF. For a first order MRF, the four nearest neighbours are included and for a second order MRF the eight nearest neighbours. In Figure 2.2 the order coding of the neighbourhoods up to order five are illustrated. The nth-order neighbourhood of the center pixel contains the pixels with numbers less than or equal to n.

5	4	3	4	5
4	2	1	2	4
3	1	·	1	3
4	2	1	2	4
5	4	3	4	5

Figure 2.2: Order coding of the neighbourhood structure.

Based on the same idea as applied in the MRF theory, we want to include spatial awareness into the clustering algorithm such that the membership of each pixel is estimated not only on its spectral characteristics but also on the spectral characteristics of neighbouring pixels. This will improve the robustness when handling noise and outliers, and it will improve the algorithm's ability to handle overlapping classes in the spectral space. In [126] MRFs energy potentials are also applied to introduce spatial awareness into a clustering algorithm in a similar manner.

We define the spatial membership for a pixel i to a cluster c by

$$u_{spat,ic} = \frac{1}{Z}\exp(-\beta E_{\mathcal{N}}), \tag{2.40}$$

where

$$E_{\mathcal{N}} = \sum_{\mathcal{N}_k \in \mathcal{N}} E(\mathcal{N}_k) \tag{2.41}$$

corresponds to a MRF energy function, $\beta \geq 0$ is a weight parameter, and Z is a normalizing constant. $E_{\mathcal{N}}$ is the energy function which is a sum of potentials depending only on the configurations of the respective cliques $\mathcal{N}_k$ in the neighbourhood $\mathcal{N}$ of an observation. For clique potential we apply

$$E(\mathcal{N}_k) = \beta_k \sum_{j \in \mathcal{N}_k} (1 - u_{jc}). \tag{2.42}$$

One could argue that the clique potentials should be chosen such that a corresponding MRF would be in a supercritical state to an isotropic phase transition at which MRF clustering occurs. However, the spectral part of the extended FCM algorithm preserves multiple clusters in the partitioning and one need not worry about eliminating entire classes when applying high spatial weights.

The choice of clique potentials controls which configurations the resulting spatial membership will favour. The default energy function is a first order neighbourhood system with equal weights for all the cliques, but zero weight for the single node clique. The energy function thus becomes

$$E_{\mathcal{N}} = \frac{1}{|\mathcal{N}|} \sum_{j \in \mathcal{N}} (1 - u_{jc}). \tag{2.43}$$

The sum over $\mathcal{N}$ indicates a sum over the surrounding neighbourhood system of an observation, and $|\mathcal{N}|$ is the number of neighbours in $\mathcal{N}$.

When applying the Equations 2.40 and 2.43, the spatial membership to a class is large, if the observations in the neighbourhood have large memberships to the same class and small, if the neighbours tend to belong to other classes. Homogeneous regions are thus favoured in the segmentation. With $\beta = 0$ no spatial context information is included. In each iteration we estimate the spatial membership degrees from the spectral memberships.

The parental membership

If the image data are very noise corrupted and the clustering algorithm has problems obtaining enough discriminating power even when applying the spatial memberships, we suggest to utilize information drawn from an analysis of the data scale-space.

Using a combination of blurring and subsampling a multiresolution scale-space pyramid is constructed from the input data. Default action is to

apply a nearly Gaussian blurring kernel (a 5-tap filter with values 0.05, 0.25, 0.4, 0.25, and 0.05) and perform subsampling such that a parent pixel at level j has four children at level $j-1$. We use reflection in order to handle border effects. A j level pyramid has $j+1$ levels with level 0 corresponding to the level of highest resolution. The convolution with a Gaussian kernel can be considered as a low-pass filtering of the image removing noise and is separable in the row/column coordinate system of the image. An individual scale-space pyramid is constructed for each variable in the data.

Applying the FCM algorithm, the scale-pyramid is processed top-down. The resulting cluster centers from one level are passed down to the next level as initial cluster centers. This can help avoid local minima and can be applied as a means of initialization by itself.

Passing down through the pyramid the memberships found at a higher level introduces additional spatial awareness into the segmentation algorithm. We introduce the additional membership as an external field that corresponds to an a priori knowledge of the memberships of the given observation to the different classes. The membership of a parent of an observation i to a class c is denoted $u_{parent,ic}$. The default action is to apply linear interpolation among the nearest neighbours of the converged fuzzy field at level $j+1$ when calculating the parental memberships applied at level j.

Including the parental membership into the algorithm should not be considered as a form of feature addition, as the dimensionality of the problem does not increase. The segmentation analysis should rather be viewed as processes on different scales, utilizing information drawn from the coarse levels at the more detailed scale levels. Including the parental memberships aims at further improving the robustness of the algorithm when handling noise and outlier pixels, but may also be applied solely for reasons of increased segmentation speed.

There are several limitations that follow from working in ordinary Gaussian scale-space. Gaussian smoothing first of all not only reduces noise, but also blurs important features such as edges. Ordinary Gaussian scale-space is designed to be completely uncommitted and cannot take any a priori information on which structures are worth preserving. Existing classes can be destroyed and "new" classes created in the image as we travel through scale-space. This brings us to the correspondence problem. When moving from finer to coarser scales, structures which are identified at a coarse scale

do not give the right location in the image and have to be traced back to the original scale. In practice this makes it difficult to carry information over large scales as the passing bifurcations give rise to instabilities.

If a user of the extended FCM algorithm wants to apply the parental membership over a large scale it might be worth considering alternative scale-pyramids i.e. other than the default Gaussian. When applying e.g. inhomogeneous linear diffusion, one can control the diffusion process by the geometry of the image itself. It is then possible only to allow diffusion inside edges of the image, see e.g. [102]. However, the computations are more burdensome and there are several parameters to be chosen before a useful scale-pyramid is available. An important motivation for utilizing the scale-space setup is to increase the speed of the algorithm which is primarily obtained from processing subsampled images at the higher levels. The subsampling and resampling between fine and coarse scales can by itself cause problems similar to the correspondence problem. A detailed treatment of the various aspects of Gaussian scale-space theory can be found in [82, 125]. In [11, 127] multi-resolution aspects are applied to MRF image segmentation.

Merging Memberships

The joint spectral-spatial-parental membership can be calculated as

$$u_{ic} = \frac{u_{spec,ic} u_{spat,ic} u_{parent,ic}}{\sum_{j=1}^{C} u_{spec,ij} u_{spat,ij} u_{parent,ij}}. \tag{2.44}$$

These joint memberships are then applied as weights in the next step of the clustering procedure when calculating new cluster centers.

Implementation of the Extended Clustering Algorithm for Spatially Sampled Data

In Algorithm 3 the extended fuzzy clustering algorithm for spatially sampled data is presented.

Numerous functionals have been proposed for calculating spectral membership apart from those presented in this chapter, however, almost all are related to Bezdek's original formulation. Other functionals than 2.4 can be

Algorithm 3 Extended Fuzzy Clustering for Spatially Sampled Data

```
Blur and subsample the data to an appropriate scale, say level j.
Initialize cluster centers
Set Temperature
repeat
    Estimate spectral memberships
    Estimate spatial memberships
    Generate a random membership field
    Merge the memberships depending on the temperature
    Estimate cluster centers from the joint memberships
until Convergence at level j
repeat
    Resample the resulting memberships from level j to a higher resolution corresponding to the next lower level j − 1.
    Blur and subsample the original data to the level j − 1.
    Initialize cluster centers from the resulting centers found at level j
    Set j = j − 1
    repeat
        Estimate spectral memberships
        Estimate spatial memberships
        Estimate parental memberships
        Estimate the joint memberships
        Estimate cluster centers from the joint memberships
    until Convergence at level j
until Convergence at level j = 0
```

implemented by replacing the spectral membership part of the extended clustering algorithm to meet the necessary criteria.

2.3 Case Studies

Three case studies are presented to evaluate the extended FCM algorithm i) synthetic data are analysed for a performance evaluation of the effect of adding the new memberships, ii) a multispectral satellite image is partitioned, and iii) a multichannel scanning electron microscope x-ray map is analysed.

2.3.1 Performance Evaluation

A toy image is analysed using the extended FCM algorithm. The image consists of two bands (100^2 pixels in each) and contains three classes. There is a different number of pixels in each class as the image consists of 72% a background class ($c = 1$), 24% a cross ($c = 2$), and 4% the center of the cross ($c = 3$). Consider a multivariate data set of P variables with gray levels $r_i(\boldsymbol{x}), i = 1, \cdots, P$, where $\boldsymbol{x}$ is the coordinate vector denoting the grid point of the sample. The two-band ($P = 2$) test image is generated using

$$\begin{aligned} r_1(\boldsymbol{x}) &= s(\boldsymbol{x}) + n_1(\boldsymbol{x})/2 + n(\boldsymbol{x}) \\ r_2(\boldsymbol{x}) &= s(\boldsymbol{x}) + n_2(\boldsymbol{x})/2 + 2n(\boldsymbol{x}) \end{aligned} \tag{2.45}$$

where the signal $s(\boldsymbol{x}) = c(\boldsymbol{x}) - 1$, with the grid point dependent class index $c(\boldsymbol{x}) \in \{1, 2, 3\}$. Choosing n_1, n_2 and n as i.i.d. zero mean Gaussians with unit variance, the covariance matrix for the total noise is given by [1.25 2; 2 4.25] (using semicolon separated rows). In Figure 2.3 one instance of a toy image generated using Equation 2.45 is presented. The image scale-pyramid has been constructed and level two to zero are shown top down. Both bands are shown for the test image.

Results and Discussion

Segmentation of 100 toy images is performed using Euclidean metric with the spectral memberships derived from ordinary c-means clustering, a relatively high $\beta = 4$, and a 3-level scale-pyramid. Results of the analyses on the different levels of the scale-pyramid of the toy image in Figure 2.3 are shown in the Figures 2.4 to 2.6. Hard results are found by selecting the class to which a pixel has the highest membership $u_{ic,max}$. If $u_{ic,max} < u_{reject}$, the observation is assigned to a reject class. We use $u_{reject} = 0.9$ in this study. In order to be able to compare results, deterministic (heuristic based) initialization is applied. Initial cluster centers are chosen from the observations in the data by favouring those that gives maximum separation.

Although only hard results are presented, it should be noted that the memberships could also be inspected and that they sometimes reveal additional information concerning the structure of the clusters. However, for the purpose of performance testing the hard results will suffice.

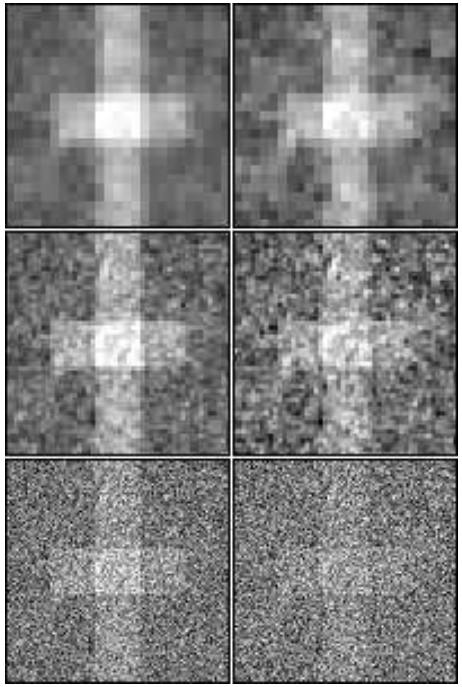


Figure 2.3: Level 2, 1 and 0 (top-down) of the scale pyramid.

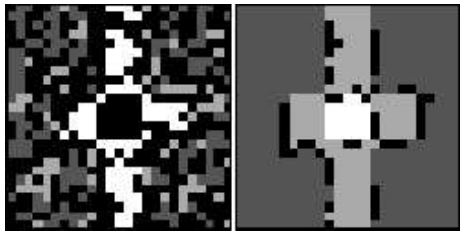


Figure 2.4: Segmentation at level 2. Left: Using spectral membership. Right: Using spectral-spatial memberships. Black=reject class, greyscale $\propto$ class index.

Inspecting Figure 2.4 the segmentation at level two fails for purely spectral driven clustering. Spatial post-relaxation of the spectral clustering

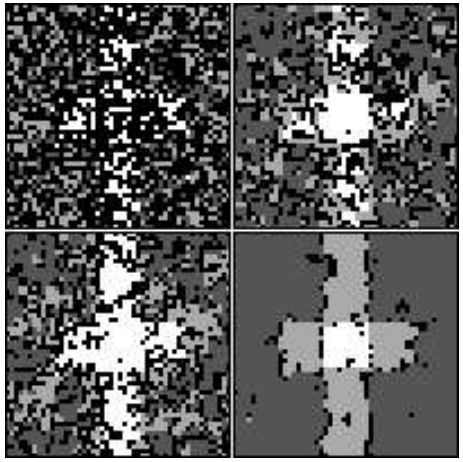


Figure 2.5: Segmentation at level 1. Row-wise) 1) Purely spectral. 2) Spectral-spatial. 3) Spectral-parental. 4) Spectral-spatial-parental. Black=reject class, greyscale $\propto$ class index.

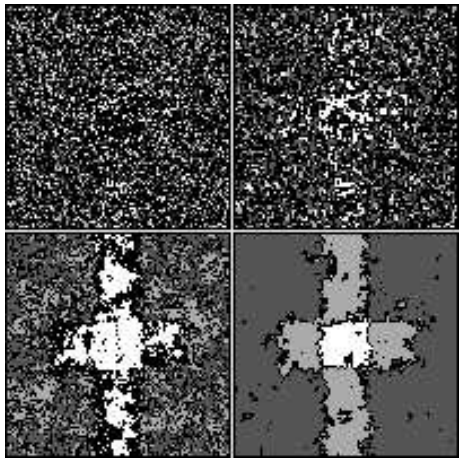


Figure 2.6: Segmentation at level 0. Same ordering as in Figure 2.5.

would be meaningless in cases like this. Including the spatial memberships into the FCM algorithm, however, appears to have the desired effect. It

helps regularize the algorithm and makes the algorithm able to detect the three classes present in the image. Both Figures 2.5 and 2.6 also illustrate that purely spectral clustering does not suffice when handling highly overlapping classes in the spectral space. In both cases the addition of the spatial membership alone does not provide an useful segmentation. Applying spectral and parental memberships does not provide the algorithm with sufficient discriminatory power either. However, when applying both the spectral, the spatial, and the parental memberships the segmentation succeeds. When looking at the bottom right components of the figures, we notice that the reject class is primarily dominated by class border regions which is satisfying. Looking at the results in Table 2.1 we see the average of correctly classified pixels and the standard deviations in the performance of the extended FCM algorithm. Results for applying different combinations of the presented memberships are given for all three levels of the scale pyramid. At level two the spectral-spatial clustering succeeds in producing meaningful results with a performance of $91 \pm 9\%$ (mean $\pm$ one standard deviation) correctly classified pixels. At level one the spatial membership is not sufficient to provide enough spatial awareness in the algorithm, and the performance decreases to $52 \pm 13\%$. When adding the parental membership, the performance increases to $93 \pm 8\%$. At level zero the performance of the spectral-spatial-parental clustering algorithm slightly decreases to $91 \pm 9\%$.

Summary

The segmentation results for the simulated image data are improved when applying the additional memberships. Not only the robustness of the clustering algorithm is enhanced by the new memberships but also the speed of convergence is increased (results not shown). In addition to the traditional spectral membership, the spatial and the parental memberships enhance the textural awareness of the algorithm. The extensions are easily incorporated into other spectral fuzzy clustering algorithms than the FCM applied in this study. The best segmentation results for the simulated image data are obtained by applying a joint spectral-spatial-parental membership.

	u_{spec}		$u_{spec,spat}$	
Level 2	mean %	std %	mean %	std %
Class 1	48.4	16.3	91.4	8.5
Class 2	18.0	19.8	77.1	13.0
Class 3	29.6	21.9	80.7	8.2
The whole image	40.4	17.1	87.5	9.0
Rejected pixels	46.9	8.4	10.1	3.2
Misclassified pixel	12.8	9.2	2.3	5.9
Level 1	mean %	std %	mean %	std %
Class 1	21.8	3.6	52.3	12.9
Class 2	2.8	2.4	37.8	18.2
Class 3	18.5	4.5	91.7	6.7
The whole image	17.1	3.0	50.4	13.3
Rejected pixels	61.5	1.2	32.8	4.3
Misclassified pixels	21.4	3.2	16.9	9.4
Level 0	mean %	std %	mean %	std %
Class 1	15.1	0.4	20.8	1.0
Class 2	5.1	0.5	3.0	0.5
Class 3	17.0	2.0	42.3	4.1
The whole image	12.8	0.3	17.4	0.7
Rejected pixels	65.1	0.5	61.3	0.8
Misclassified pixels	22.1	0.3	21.4	0.5
	$u_{spec,parental}$		$u_{spec,spat,parental}$	
Level 1	mean %	std %	mean %	std %
Class 1	57.7	17.0	93.1	7.9
Class 2	30.0	26.2	82.8	13.6
Class 3	91.0	5.6	86.2	6.3
The whole image	52.4	18.3	90.4	8.8
Rejected pixels	30.0	6.6	7.0	3.0
Misclassified pixels	17.7	12.1	2.6	5.9
Level 0	mean %	std %	mean %	std %
Class 1	49.0	16.0	90.8	8.5
Class 2	22.4	21.6	71.6	12.3
Class 3	80.2	9.8	79.2	6.8
The whole image	43.9	16.2	85.7	8.8
Rejected pixels	39.2	5.8	11.7	3.8
Misclassified pixels	16.9	11.0	2.6	5.0

Table 2.1: Summary of 100 segmentation results using the different extensions to the spectral FCM.

2.3.2 Unsupervised Segmentation of Multispectral Image Data

The Sea-viewing Wide Field-of-view Sensor (SeaWiFS) is an eight channel optical scanner on the SeaStar spacecraft which orbits sun-synchronously at 705 km altitude. On a daily basis SeaWiFS provides 10 bit data in six visual and two near infrared (NIR) bands, see Table 2.2 The pixel size is

SeaWiFS band	Wavelengths [nm]	Spectral ref.
1	402-422	Violet
2	433-453	Blue
3	480-500	Blue-Green
4	500-520	Blue-Green
5	545-565	Green
6	660-680	Red
7	745-785	Near-Infrared
8	845-885	Near-Infrared

Table 2.2: The spectral range of the SeaWiFS sensor bands 1 through 8.

1.1 km $\times$ 1.1 km. See also [112]. This case study provides an example of an initial exploratory clustering of the data applying the default extended FCM algorithm. A 1000 by 1000 SeaWiFS image acquired on the 14th of May 1998 is presented in Figure 2.7. The bands are stretched to mean $\pm$ three standard deviations (std).

Results and Discussion

In Figure 2.8 the results of clustering using two to ten classes are presented for level zero. The analyses are performed using the joint membership of default parental and spatial memberships in combination with the spectral FCM membership. Since this is the initial analysis of the data, the spectral membership applies the Euclidean metric resulting in a fast segmentation of the large multivariate data set. Hard results are generated by selecting the class with the highest membership above a given reject threshold, here 0.75. If the assignment fails, the pixel is assigned to a reject class. The partitioning of the data seems to produce meaningful results. When applying two classes, the data are partitioned into either primarily ocean

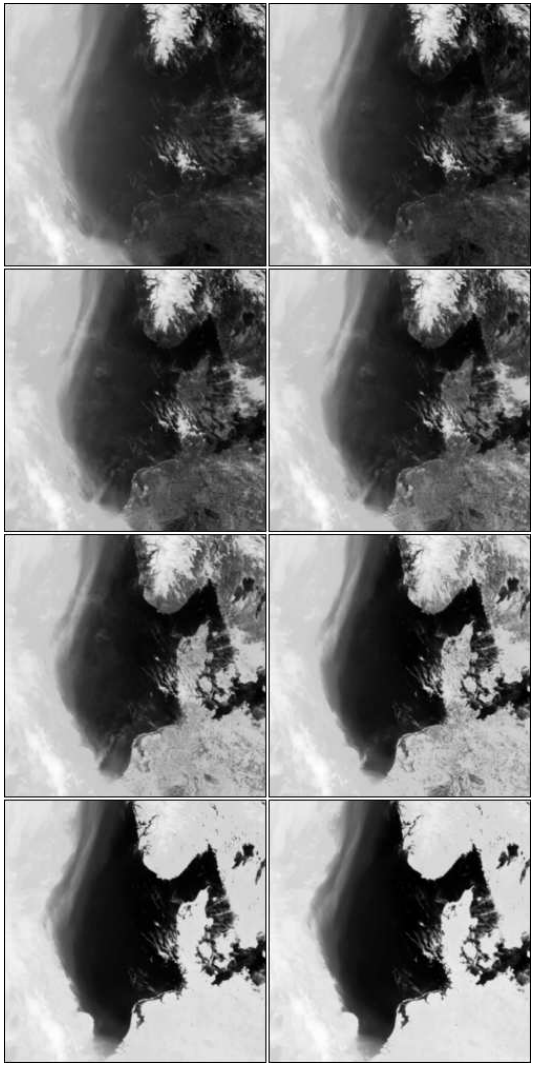


Figure 2.7: The SeaWiFS bands 1-8 (row-wise) stretched to mean $\pm$3 std.

or primarily cloud/vegetated land related signal. Adding additional clusters creates new cloud/ice related classes. When applying six classes, the

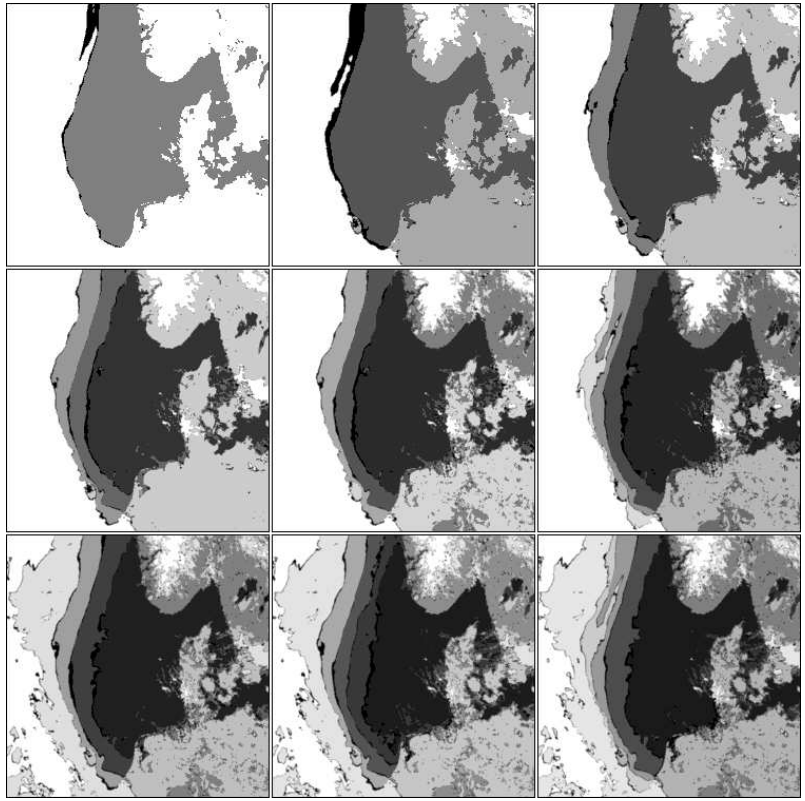


Figure 2.8: Hard clustering results from applying the extended fuzzy algorithm, applying three levels and the default spatial membership. Black=reject class, greyscale $\propto$ class index. Results of applying two to ten clusters are presented (row-wise ordering).

vegetated land related signal is partitioned into two different classes likely related to cultivated farmland and Boreal forest vegetation. The Boreal forest vegetation dominates the Northern part of Scandinavia. The addition of clusters continues to provide meaningful results primarily producing a more detailed partitioning of the cloud and vegetated land related SeaWiFS signal. Notice, that for all the image segmentations the reject class is primarily dominated by the boundaries between classes.

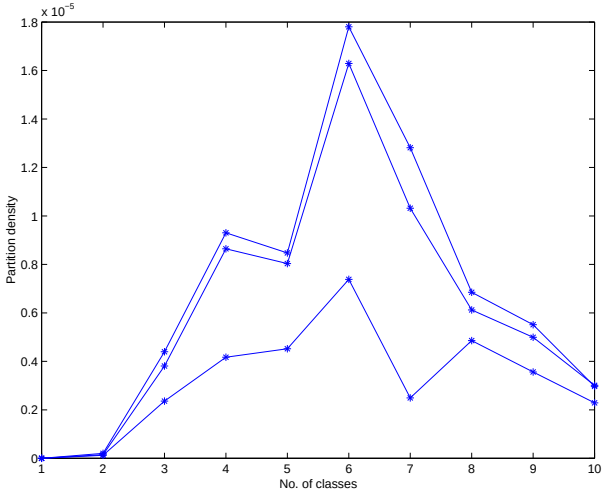


Figure 2.9: The partition density as a function of classes, calculated for level 0 (top), 1 (middle) and 2 (bottom curve).

The partition density (PD) is evaluated for the different segmentations at all levels and shown in Figure 2.9. In general one would expect a decreasing PD as a function of increasing levels as the number of pixels included into the estimation decreases, thus reducing the range of S in Equation 2.24. In the figure the top curve is the PD for level zero. It has a strong maximum at six classes. The same is the case for the PD at level one. At level two the global maximum is still at the partitioning applying six classes, but there is also a local maximum at eight classes. In Figure 2.10 the resulting cluster centers are shown for the analysis applying six classes. The centers can be interpreted as class signatures and provide the data analyst with information on the nature of the individual classes. Empirically cluster 1 is recognized as a water class, 2 and 3 as cloud classes, 4 and 5 as vegetated land classes and 6 as a cloud/ice class. In Figure 2.11 the memberships for the six classes at level zero are shown.

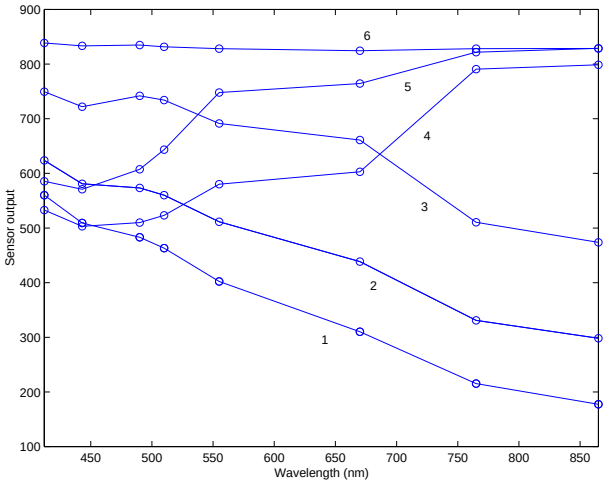


Figure 2.10: The spectral signature of the cluster centers 1 through 6.

Conclusion

The extended FCM algorithm successfully provides a data analyst with an initial analysis of the multivariate image data in terms of data partitioning. The further analysis would include the application of the Mahalanobis distance measure and perhaps relaxation of the regularizing spatial constraints in the fuzzy clustering algorithm. However, the results presented in this study may already be sufficient for generating masks under which more detailed analyses can be performed regarding the nature of the data and the interdependencies of the involved variables.

Acknowledgements

The SeaWiFS data was available through the Danish GEOSONAR project, funding by the joint Danish Research Councils under the Earth Observation Program, [120]. The data are received by the University of Dundee Satellite receiving station and are produced by the SeaWiFS Project at Goddard Space Flight Center, [68]. Use of these data are in accord with the SeaWiFS Research Data Use Terms and Conditions Agreement.

Figure 2.11: The spectral-spatial-parental membership degrees for the six classes at level zero.

2.3.3 Exploration of X-Ray Mapping Images of Polished Sections

This case study deals with unsupervised classification of multichannel scanning electron microscope (SEM) energy dispersive spectroscopy (EDS) image data from polished sections, also known as x-ray mapping imagery. A multivariate image containing 176 rows and 256 columns is segmented. The image consists of ten channels which represent the elements Al, C, Fe, Mg, Na, O, P, S, Si, and Ca, see Figure 2.12. As the data represent counts and

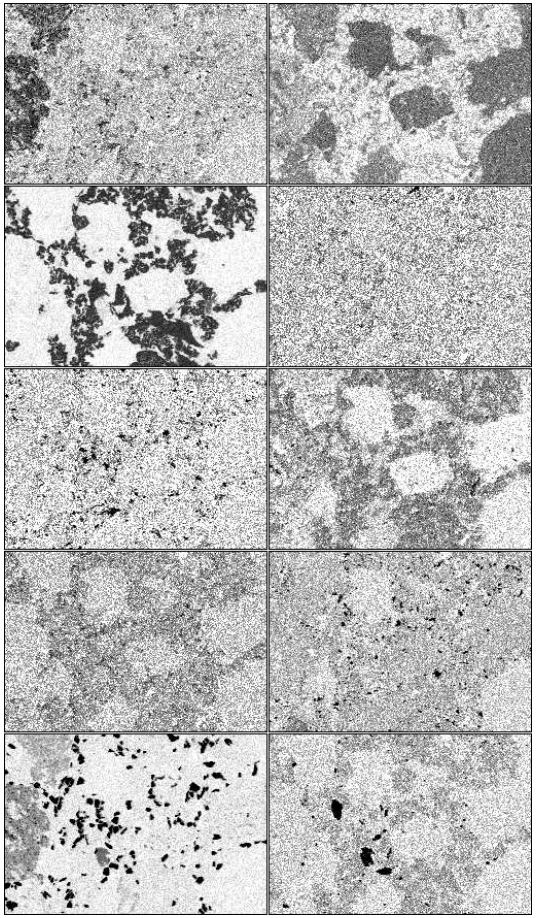


Figure 2.12: The channels 1-10 (row-wise) representing Al, C, Fe, Mg, Na, O, P, S, Si, and Ca. Dark regions represent high counts. Each image is stretched linearly between its mean ±3 standard deviations.

thus ideally follow a Poisson distribution as a variance-stabilising measure, all numbers are square-rooted before the analysis. A priori knowledge is available and we expect the data to be dominated by six major classes.

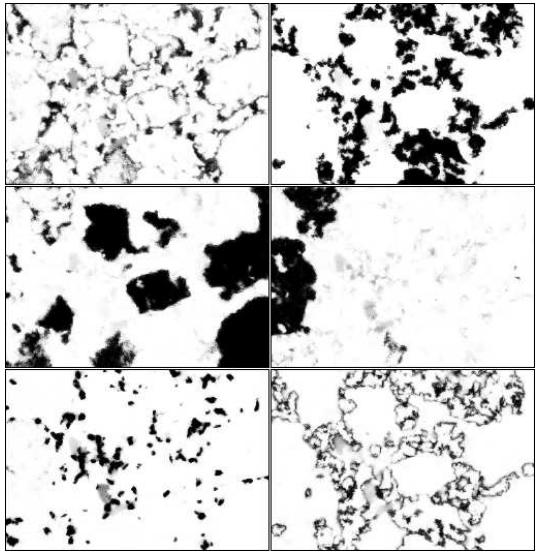


Figure 2.13: Segmentation based on spectral, spatial, and parental memberships. The six cluster membership-degrees (row-wise). Dark regions represent high memberships.

Results

The x-ray mapping images of the polished sections are classified using two unsupervised clustering algorithms. The methods applied are the k-means algorithm and the extended spectral fuzzy c-means (FCM) algorithm.

Applying the c-means algorithm to the data with spectral information only reveals that although the hard classification looks sensible, the membership degrees are all very low (not shown). Hence, the confusion concerning the segmentation is relatively high. Adding spatial and parental context leads to more distinct membership degrees, i.e. membership degrees closer to 0 and 1 indicating a better classification. The algorithm is applied using $m = 2$, $\beta = 1$ in a scale pyramid consisting of four levels. The parental membership is introduced from level two and down, see Figures 2.13 and 2.14.

The k-means algorithm is performed by applying the SAS fastclus proce-

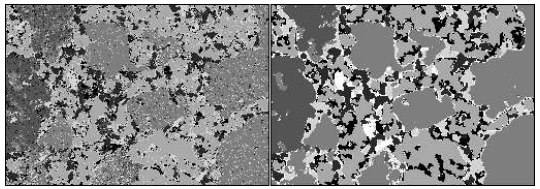


Figure 2.14: Hard results obtained from the membership-degrees. White pixels correspond to the reject class. The image to the left is obtained based on k-means with spectral features. The image to the right is obtained based on c-means using additional spatial and parental features.

dure, [66]. The hard k-means and FCM results are shown in the left and right images of Figure 2.14. Comparing the hard result for the k-means and the extended c-means algorithms, for the latter we notice that the noise-like structures have been avoided. The memberships have been thresholded at 0.5 introducing a reject class (the white regions in the right image of Figure 2.14). The reject class dominates boundary regions between classes and a few smaller regions. Comparing these regions with Figure 2.12 we find that the reject class is dominated by a single element, Ca.

In Figures 2.15 and 2.16 descriptors are calculated for the different classes. The images contain the class centres determined by the k-means and the extended c-means algorithms. The figures are to be compared to Figures 2.12 and 2.14. The spectra in Figure 2.16 are calculated using the memberships from Figure 2.13 without thresholding.

Conclusion

Two types of unsupervised classification have been applied to SEM/EDS or x-ray mapping images, namely k-means and (fuzzy) c-means classification. The FCM algorithm has been extended by incorporating a multi-scale representation of the image data, partially for speed up and partially for carrying spatial information across scale levels. Also, a spatial element at each scale level has been included. Simultaneous inspection of plots of class means (Figure 2.16) and Figures 2.12 and 2.13 provide support for the applications expert in interpreting the contents of the classes found. Comparisons between results from k-means and c-means analyses show that although the class means exhibit some similarities, the visual impression of

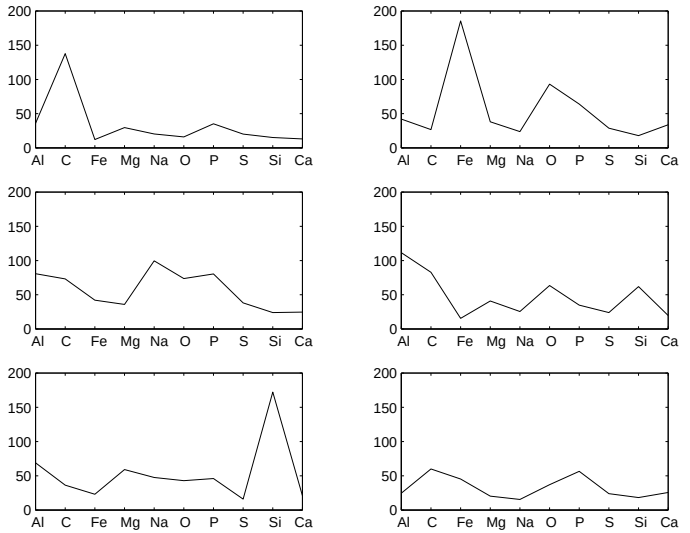


Figure 2.15: Class descriptors obtained applying the k-means algorithm. The class signatures $\tilde{\boldsymbol{r}}_c, c = 1, \ldots, 6$ row-wise.

the c-means result when using the multi-scale approach with spatial information being carried across scale levels, is more pleasing indicating large same-class regions. Also, the degrees of membership are much closer to 0 and 1 when compared to a purely spectral c-means classification indicating a better segmentation result. In spite of the problem with classifying the regions rich in Ca, this type of analysis seems a good exploratory tool for obtaining knowledge of the discriminatory power of the data. It thus constitutes a good preprocessor for a more thoroughly supervised analysis (in which one could explicitly introduce the Ca-rich regions as a separate class).

Acknowledgements

The data used come from the COMB project, the Industrial Centre for Surface Microscopy, Microanalysis and Image Analysis headed by Dr. Leif Højslet Christensen, [70]. The COMB project is funded by Erhvervs-Fremme Styrelsen, the Danish Agency for Trade and Industry.

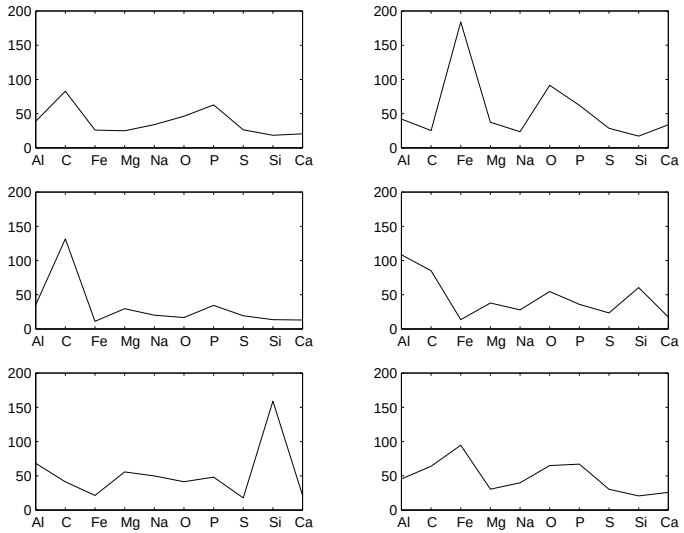


Figure 2.16: Class descriptors obtained applying the c-means algorithm. The class signatures $\tilde{\boldsymbol{r}}_c, c = 1, \ldots, 6$ row-wise.

2.4 Summary

- Spectral clustering algorithms are presented including the fuzzy c-means (FCM) and the hard k-means (HCM) algorithms. Both methods can have problems handling outliers in the data. The HCM can only handle non-overlapping clusters.
- Noise Clustering (NCM) is presented resulting in an algorithm which is more robust to noise and outliers in the data. Noise clustering is obtained by relaxing the probabilistic constraint in FCM.
- The FCM algorithm is related to the Expectation Maximization (EM) algorithm applied to density estimation for a Gaussian mixture problem. Good initialization of the EM algorithm is generally needed. This also applies to the clustering algorithms.
- Different types of initialization approaches are briefly presented including both random and deterministic methods.
- A scheme inspired by stochastic diffusion and simulated annealing (SA) is presented, which can be applied to help avoid local minima when optimizing the clustering functional.

- An extended spectral-spatial fuzzy clustering algorithm is presented. Two new types of memberships are introduced i.e. the spatial and the parental memberships. The spatial memberships are calculated using a Markov Random Field (MRF) energy functional, and the parental memberships from a fuzzy cluster analysis of the data scale-space.
- Including the new memberships into the clustering algorithm has a regularizing effect on the partitioning. It enhances the robustness when handling noisy data with outliers and data with overlapping clusters in the spectral space.
- Three case studies are included performing an analysis of i) simulated data with overlapping Gaussian clusters, ii) a multispectral satellite image, and iii) a multichannel scanning electron microscope x-ray mapping.

Chapter 3

Linear Decomposition

Linear transforms for spectral-spatial decompositioning of single set data are presented in this chapter. The decomposition of multispectral images is motivated by extracting important, otherwise occluded information on e.g. the correlation structures in the data. The transforms are thus designed to maximize different criteria such as the variance or the signal-to-noise ratio (SNR). A scaling of the latter group of transforms is proposed producing a new representation in which the covariance structure of the noise equals the identity matrix. The variance of the resulting components depends on the signal-to-noise ratio contained in each component. Mixtures and hybrids of single set analyses are also proposed. Two- and multiset scenarios are considered in which transforms are applied to find maximum correlation. A relation between the linear multiset canonical correlations analysis and Procrustes shape alignment is found. Case studies are presented including i) a linear decomposition of an eight-band satellite image, ii) two-set canonical correlations analysis of temporal sea surface height and temperature, and iii) a multiset canonical correlations analysis of a collection of landmark registered metacapal II bones.

3.1 Single Sets

Linear statistical transformations are useful tools for the enhancement of remotely sensed multispectral data. Measurements at different wavelengths induce a degree of redundancy in the multivariate data. The redundancy can be explained by the covariation between variables, which can be used for noise reduction and data compression.

The transformations of the image data are constructed to optimize some design criterion producing new variates. The choice of an appropriate transformation is dependent on its application. One criterion is to preserve as much variance as possible. This is done by projecting the data onto the subspace of maximum data variation, thus obtaining the traditional Principal Components (PCs), [64]. The PC transformation performs pixel-wise operations that do not take the spatial nature of image data into account, and it does not necessarily separate important information from noise.

As opposed to the PC transform the Minimum Noise Fraction (MNF) transform can take the spatial nature of the image into account. The MNF transform was proposed as a transformation for ordering multispectral data in terms of image quality with implications for noise removal. It was introduced by Green et al. in 1988, [51], inspired by earlier work on the Maximum Autocorrelation Factors (MAF) transform by Switzer and Green in 1984, [114]. The MNF/MAF transform is further described in [20, 115, 30, 89]. Whereas the PC transform only requires knowledge or an estimate of the dispersion matrix, the MNF transform requires the dispersion matrix of the noise structure as additional information. The MNF/MAF transformation can be considered as a spatial extension of PC analysis. For MAF the design criterion is to produce new variates with maximum autocorrelation between neighbouring pixels. Hence the transformation takes the spatial nature of the image data into account and new variates are constructed by projecting the data onto the linear subspace containing maximum autocorrelation.

Most of the methods considered in this section are linear transformations for which the solutions can be written in closed form as a generalized eigenproblem. If more complex measures are specified Exploratory Projection Pursuit (EPP), see [38, 128, 21], can be applied to obtain the linear combinations that maximize the desired criteria. EPP is based on the idea of determining the optimal projection by gradient descent, followed by a

transformation that deflates the previous maximum, such that successive solutions may be found.

Consider the following generalized eigenproblem:

$$\boldsymbol{A}\boldsymbol{v} = \lambda \boldsymbol{B}\boldsymbol{v} \quad \text{or} \quad \boldsymbol{B}^{-1}\boldsymbol{A}\boldsymbol{v} = \lambda \boldsymbol{v}. \tag{3.1}$$

Here $\boldsymbol{A}$ and $\boldsymbol{B}$ are matrices consisting of components which are expectation values from stochastic processes. Furthermore, both matrices are symmetric and $\boldsymbol{B}$ is positive definite. λ is the eigenvalue and $\boldsymbol{v}$ the corresponding eigenvector.

Solving the eigenproblem is related to locating extremum points of a ratio of quadratic forms:

$$\lambda = \frac{\boldsymbol{v}^T \boldsymbol{A}\boldsymbol{v}}{\boldsymbol{v}^T \boldsymbol{B}\boldsymbol{v}}. \tag{3.2}$$

This ratio is also known as the Rayleigh quotient. The critical points correspond to the solutions of the generalized eigenproblem. Thus, the eigenvalues of $\boldsymbol{B}^{-1}\boldsymbol{A}$ are the extremum values of the quotient, and the eigenvectors are the corresponding vectors of the quotient.

The Power Method

There are many methods for solving an eigenproblem. Later in this thesis, when dealing with nonlinear transformations, we shall make use of one method in particular known as the Power Method (PM).

The PM is a classical method for computing the eigenvector associated with the eigenvalue of the largest magnitude. Assume $\boldsymbol{A}$ is an $N \times N$ matrix with real eigenvalues $\lambda_1 > \lambda_2 \geq \lambda_3 \geq \cdots \geq \lambda_N$ and associated eigenvectors $\{\boldsymbol{v}_i\}_{i=1}^N$. The power method is given in Algorithm 4 for finding the right-eigenvector with the largest eigenvalue and consists of repeated application of the matrix $\boldsymbol{A}$ as an operator on a state vector $\boldsymbol{z}$ followed by normalization.

The state vector becomes an increasingly better estimate of the eigenvector corresponding to the largest eigenvalue as the iterations proceed. One condition is that the initial choice of $\boldsymbol{z}$ must not be orthogonal to the eigenvector we wish to find. In practice this hardly never happens due

Algorithm 4 The Power Method

1: Initialize $\boldsymbol{z}^0; ||\boldsymbol{z}^0|| = 1$
2: **repeat**
3: Calculate $\boldsymbol{z}^{k+1} = \boldsymbol{A}\boldsymbol{z}^k$
4: Normalize $\boldsymbol{z}^{k+1} = \boldsymbol{z}^{k+1}/||\boldsymbol{z}^{k+1}||$
5: **until** Convergence

to machine precession and round-off error. Let the initial state ($\boldsymbol{z}^0$) be represented by a linear combination of the eigenvectors of $\boldsymbol{A}$

$$\boldsymbol{z}^0 = \sum_{i=1}^{N} c_i \boldsymbol{v}_i. \tag{3.3}$$

Then

$$\boldsymbol{z}^k = \boldsymbol{A}^k \boldsymbol{z}^0 / ||\boldsymbol{A}^k \boldsymbol{z}^0||, \tag{3.4}$$

and

$$\boldsymbol{A}^k \boldsymbol{z}^0 = \sum_{i=1}^{N} c_i \lambda_i^k \boldsymbol{v}_i \tag{3.5}$$

$$= \lambda_1^k (c_1 \boldsymbol{v}_1 + \sum_{i=2}^{N} c_i (\frac{\lambda_i}{\lambda_1})^k \boldsymbol{v}_i). \tag{3.6}$$

Thus, as $k \to \infty$, then $(\lambda_i/\lambda_1)^k \to 0$ and $\boldsymbol{z}^k$ approximate the true eigenvector $\boldsymbol{v}_1$ with an error of $\mathcal{O}(|\lambda_2/\lambda_1|^k)$. The estimate of the eigenvalue can e.g. be found from $(\boldsymbol{z}^k)^{\mathrm{T}} \boldsymbol{A}\boldsymbol{z}^k / (\boldsymbol{z}^k)^{\mathrm{T}} \boldsymbol{z}^k$.

If the power method is applied to $\tilde{\boldsymbol{A}} = (\boldsymbol{A} - \alpha \boldsymbol{I})$, where α is a real number, the eigenvectors will remain unchanged. However the eigenvalues will be shifted by α, and the method will find the eigensolution of $\boldsymbol{A}$ that maximizes $\delta\lambda = |\lambda_j - \alpha|$. Moreover, applying the power method to $\tilde{\boldsymbol{A}} = (\boldsymbol{A} - \alpha \boldsymbol{I})^{-1}$ will find the eigensolution that minimizes $\delta\lambda$. If you wish to avoid the matrix inversion, the Inverse Power Method can be applied to approximate the eigenvalue closest to α, [105]. This is done by repeatedly solving for $\boldsymbol{z}^k$ from $(\boldsymbol{A} - \alpha \boldsymbol{I})\boldsymbol{z}^k = \boldsymbol{z}^{k-1}$ followed by normalization. The eigenvalues found by the inverse power method are shifted by α and should also be corrected. The Subspace Iteration Method, [122] can be

applied to find the M largest eigenvalues and eigenvectors using the power method and is good to apply, if the gaps between the target eigenvalues are relatively small. Finally, if $\boldsymbol{A}$ is symmetric, the task simplifies and successive eigenvalues and eigenvectors can be found by e.g. applying the power method to $\tilde{\boldsymbol{A}} = \boldsymbol{A} - \sum_{i=1}^{M} \lambda_i \boldsymbol{v}_i \boldsymbol{v}_i^T$ to find the $M+1$ largest eigensolution.

The power method can also be extented to handle eigenproblems involving complex eigenvalues. This involves registrating the rotation and rate of divergence of the state vector, as the operator $\boldsymbol{A}$ is applied.

3.1.1 Principal Components Transformation

Consider a multivariate data set of P variables with grey levels $r_i(\boldsymbol{x}), i = 1, \cdots, P$, where $\boldsymbol{x}$ is the coordinate vector denoting the grid point of the sample.

Let

$$\boldsymbol{r}(\boldsymbol{x}) = \begin{bmatrix} r_1(\boldsymbol{x}) \\ \vdots \\ r_P(\boldsymbol{x}) \end{bmatrix} \tag{3.7}$$

and assume first and second order stationarity such that

$$\mathrm{E}\{\boldsymbol{r}(\boldsymbol{x})\} = \boldsymbol{0} \tag{3.8}$$
$$\mathrm{D}\{\boldsymbol{r}(\boldsymbol{x})\} = \boldsymbol{\Sigma}. \tag{3.9}$$

Determining the direction of maximum variation means finding the direction $\boldsymbol{a}$, subject to $\boldsymbol{a}^T\boldsymbol{a} = 1$, such that the linear combination $z(\boldsymbol{x}) = \boldsymbol{a}^T\boldsymbol{r}(\boldsymbol{x})$ possesses maximum variance.

Applying the Lagrangian technique we define $\mathrm{L}(\boldsymbol{a}, \lambda) = \boldsymbol{a}^T\boldsymbol{\Sigma}\boldsymbol{a} - \lambda(\boldsymbol{a}^T\boldsymbol{a} - 1)$. When looking for stationarity of $\mathrm{L}(\boldsymbol{a}, \lambda)$ we find $\boldsymbol{\Sigma}\boldsymbol{a} - \lambda\boldsymbol{a} = \boldsymbol{0}$. The PC transformation thus chooses P linear transformations

$$z_i(\boldsymbol{x}) = \boldsymbol{a}_i^T\boldsymbol{r}(\boldsymbol{x}), \quad i = 1, \cdots, P \tag{3.10}$$

such that the variance for $z_i(\boldsymbol{x})$ is maximum among all linear transforms orthogonal to $z_j(\boldsymbol{x}), j = 1, \cdots, i-1$. The variance is given by

$$\lambda_i = \boldsymbol{a}_i^T\boldsymbol{\Sigma}\boldsymbol{a}_i. \tag{3.11}$$

We see that the basis for the PCs is identified as the conjugate eigenvectors of the dispersion matrix. Let $\lambda_1 \geq \cdots \geq \lambda_P \geq 0$ be the eigenvalues with the corresponding conjugate eigenvectors $\boldsymbol{a}_1, \cdots, \boldsymbol{a}_P$. Then $z_i(\boldsymbol{x})$ is the ith PC (PCi) with variance λ_i.

The amount of variance explained by the ith PC is determined by

$$\frac{\boldsymbol{a}_i^T\boldsymbol{\Sigma}\boldsymbol{a}_i}{\sum_{j=1}^{P} \boldsymbol{a}_j^T\boldsymbol{\Sigma}\boldsymbol{a}_j} = \frac{\lambda_i}{\sum_{j=1}^{P} \lambda_j}. \tag{3.12}$$

By construction the PC transform depends on the unit of measurement of the original variables. This problem can be circumvented by considering the correlation matrix instead of the covariance matrix when solving the eigenvalue problem and in effect applying standardized variables.

The most numerically stable method for determining the PCs of a data set is not to solve the eigenvalue problem in Equation 3.11. If enough memory is available one can calculate the principal components by applying Singular Value Decomposition (SVD). We define $\boldsymbol{R} = [\boldsymbol{r}_1\ \boldsymbol{r}_2\ \cdots\ \boldsymbol{r}_N]$, where $\boldsymbol{r}_i, i = 1, \cdots, N$ is the spectrum for the ith observation. When solving the SVD of $\boldsymbol{R} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^T$, $\boldsymbol{U}$ will be an orthogonal matrix which includes the PC eigenvectors. The corresponding eigenvalues can be found by squaring the diagonal elements of the matrix $\boldsymbol{D}$ and dividing it by $N-1$. SVD is more robust as the the condition number of the covariance matrix $\boldsymbol{\Sigma} = \boldsymbol{R}\boldsymbol{R}^T/(N-1)$ is the squared condition number of $\boldsymbol{R}$.

3.1.2 Minimum Noise Fractions Transformation

Let us again consider the random signal variable $\boldsymbol{r}(\boldsymbol{x})$ and assume first and second order stationarity by imposing Equation 3.8 and 3.9. When we assume that an additive noise structure applies

$$\boldsymbol{r}(\boldsymbol{x}) = \boldsymbol{s}(\boldsymbol{x}) + \boldsymbol{n}(\boldsymbol{x}), \tag{3.13}$$

the dispersion structure can be separated into

$$\mathrm{D}\{\boldsymbol{r}(\boldsymbol{x})\} = \boldsymbol{\Sigma} = \boldsymbol{\Sigma}_s + \boldsymbol{\Sigma}_n. \tag{3.14}$$

The Minimum Noise Fractions (MNF) transformation chooses P linear transformations

$$z_i(\boldsymbol{x}) = \boldsymbol{a}_i^T\boldsymbol{r}(\boldsymbol{x}), \quad i = 1, \cdots, P \tag{3.15}$$

which maximize the signal-to-noise ratio (SNR) for the ith component defined by

$$\mathrm{SNR}_i = \frac{\mathrm{V}\{\boldsymbol{a}_i^T \boldsymbol{s}(\boldsymbol{x})\}}{\mathrm{V}\{\boldsymbol{a}_i^T \boldsymbol{n}(\boldsymbol{x})\}}. \tag{3.16}$$

Combining Equation 3.14 and 3.16 we find

$$\mathrm{SNR}_i = \frac{\boldsymbol{a}_i^T \boldsymbol{\Sigma} \boldsymbol{a}_i}{\boldsymbol{a}_i^T \boldsymbol{\Sigma}_n \boldsymbol{a}_i} - 1, \tag{3.17}$$

and the problem is reduced to solving a generalized eigenproblem, say

$$\boldsymbol{\Sigma}_n \boldsymbol{a}_i = \lambda_i \boldsymbol{\Sigma} \boldsymbol{a}_i. \tag{3.18}$$

Let $\lambda_1 \leq \cdots \leq \lambda_P$ be the eigenvalues of $\boldsymbol{\Sigma}_n$ with respect to $\boldsymbol{\Sigma}$ with the corresponding conjugate eigenvectors $\boldsymbol{a}_1, \cdots, \boldsymbol{a}_P$. Then $z_i(\boldsymbol{x})$ is the ith MNF (MNFi). A high order component has a high noise fraction and thus little signal. A low order component has a high SNR, hence the name Minimum Noise Fraction transform.

The central issue in obtaining good MNF components is the estimation of the dispersion matrix for the noise. In [80, 101, 90] several models are presented for estimating noise in images based on spatial characteristics. Using the difference between the current pixel and its neighbours, the MNF reduces to the MAF transform which is presented in the following section. When the covariance structure for the noise is proportional to the identity matrix, the MNF transform reduces to the PC transform. The traditional MNF utilizes the spatial information in the image for estimating the noise structure. In [58] we handle real time decomposition of streaming three band colour images and propose methods for utilizing the temporal dimension when estimating the correlation structure of the noise.

When the structure of the noise is estimated in each pixel, the most numerically stable method for calculating the MNF transform is to apply the Quotient Singular Value Decomposition. See [87].

By applying the logarithm transform to the signal $\boldsymbol{r}(\boldsymbol{x})$, the MNF transform is able to handle a multiplicative noise structure.

3.1.3 Maximum Autocorrelation Factors Transformation

Let us again consider the random signal variable $\boldsymbol{r}(\boldsymbol{x})$ and assume first and second order stationarity, imposing 3.8 and 3.9. The MAF transformation chooses P linear transformations

$$z_i(\boldsymbol{x}) = \boldsymbol{a}_i^T \boldsymbol{r}(\boldsymbol{x}), \quad i = 1, \cdots, P \tag{3.19}$$

such that the spatial autocorrelation for $z_i(\boldsymbol{x})$ is maximum among all linear transforms orthogonal to $z_j(\boldsymbol{x}), j = 1, \cdots, i-1$.

Let $\boldsymbol{\Delta}^T = [\Delta_1 \quad \Delta_2]$ represent a spatial shift. Then the spatial covariance function is defined by

$$\mathrm{Cov}\{\boldsymbol{r}(\boldsymbol{x}), \boldsymbol{r}(\boldsymbol{x}+\boldsymbol{\Delta})\} = \boldsymbol{\Gamma}(\boldsymbol{\Delta}). \tag{3.20}$$

$\boldsymbol{\Gamma}$ has the following properties $\boldsymbol{\Gamma}(\boldsymbol{0}) = \boldsymbol{\Sigma}$ and $\boldsymbol{\Gamma}(\boldsymbol{\Delta})^T = \boldsymbol{\Gamma}(-\boldsymbol{\Delta})$.

Looking at the correlations between projections of the variables and the shifted variables we find

$$\begin{aligned} &\mathrm{Cov}\{\boldsymbol{a}_i^T \boldsymbol{r}(\boldsymbol{x}), \boldsymbol{a}_i^T \boldsymbol{r}(\boldsymbol{x}+\boldsymbol{\Delta})\} \\ &\quad = \frac{1}{2}\, \boldsymbol{a}_i^T (\boldsymbol{\Gamma}(\boldsymbol{\Delta}) + \boldsymbol{\Gamma}(-\boldsymbol{\Delta})) \boldsymbol{a}_i. \end{aligned} \tag{3.21}$$

Introducing $\boldsymbol{\Sigma}_\Delta$ gives

$$\begin{aligned} \boldsymbol{\Sigma}_\Delta &= \mathrm{D}\{\boldsymbol{r}(\boldsymbol{x}) - \boldsymbol{r}(\boldsymbol{x}+\boldsymbol{\Delta})\} \\ &= \quad 2\, \boldsymbol{\Sigma} - (\boldsymbol{\Gamma}(\boldsymbol{\Delta}) + \boldsymbol{\Gamma}(-\boldsymbol{\Delta})) \end{aligned} \tag{3.22}$$

which, when considered as a function of $\boldsymbol{\Delta}$, is a multivariate variogram. Combining 3.21 and 3.22 we obtain

$$\mathrm{Cov}\{\boldsymbol{a}_i^T \boldsymbol{r}(\boldsymbol{x}), \boldsymbol{a}_i^T \boldsymbol{r}(\boldsymbol{x}+\boldsymbol{\Delta})\} = \boldsymbol{a}_i^T (\boldsymbol{\Sigma} - \frac{1}{2}\, \boldsymbol{\Sigma}_\Delta) \boldsymbol{a}_i. \tag{3.23}$$

Dividing this by the variance of the data projected on the subspace corresponding to $\boldsymbol{a}_i$, the expression for the correlation is given by

$$\mathrm{Corr}\{\boldsymbol{a}_i^T \boldsymbol{r}(\boldsymbol{x}), \boldsymbol{a}_i^T \boldsymbol{r}(\boldsymbol{x}+\boldsymbol{\Delta})\} = 1 - \frac{1}{2}\, \frac{\boldsymbol{a}_i^T \boldsymbol{\Sigma}_\Delta \boldsymbol{a}_i}{\boldsymbol{a}_i^T \boldsymbol{\Sigma} \boldsymbol{a}_i}. \tag{3.24}$$

If we want to maximize the correlation, we must minimize the Rayleigh coefficient

$$\lambda = \frac{\boldsymbol{a}^T \boldsymbol{\Sigma}_\Delta \boldsymbol{a}}{\boldsymbol{a}^T \boldsymbol{\Sigma} \boldsymbol{a}}. \tag{3.25}$$

Consider solving the equivalent generalized eigenproblem

$$\boldsymbol{\Sigma}_\Delta \boldsymbol{a}_i = \lambda_i \boldsymbol{\Sigma} \boldsymbol{a}_i. \tag{3.26}$$

Let $\lambda_1 \leq \cdots \leq \lambda_P$ be the eigenvalues of $\boldsymbol{\Sigma}_\Delta$ with respect to $\boldsymbol{\Sigma}$ with the $\boldsymbol{a}_1, \cdots, \boldsymbol{a}_P$ corresponding conjugate eigenvectors. Then $z_i(\boldsymbol{x})$ is the ith MAF.

It is possible to calculate MAFs using two-set canonical correlations analysis (CCA), see Section 3.2.1, comparing the original data to a spatially shifted set. In CCA the resulting components are conditioned to have unit variance, [2], and the MAFs have traditionally been imposed the same restriction.

In order to obtain an estimate of $\boldsymbol{\Sigma}_\Delta$, Switzer and Green, [114], recommend the formation of two sets of difference images. The two sets are $\boldsymbol{r}(\boldsymbol{x}+\boldsymbol{\Delta}_{\boldsymbol{h}})$ and $\boldsymbol{r}(\boldsymbol{x}+\boldsymbol{\Delta}_{\boldsymbol{v}})$, where $\boldsymbol{\Delta}_{\boldsymbol{h}}$ is an unit horizontal shift and $\boldsymbol{\Delta}_{\boldsymbol{v}}$ is an unit vertical shift. Calculate the dispersion matrices for both sets of images and pool them as an estimate of $\boldsymbol{\Sigma}_\Delta$. One important condition in choosing a reasonable shift is that we must not exceed the range-of-influence of the signal. Often the noise can be split into several parts, and one may try to apply additional shifts before pooling them in a combined estimate of the noise structure. Strategies may even be applied when looking for the optimal combination of shifts that maximize e.g. i) the gap between the first two eigenvalues or ii) the number of components with very large or small eigenvalues.

A comparative study of the PC and the MNF/MAF transforms applied to high dimensional data can be found in [89]. The MAF transform does a much better job of separating signal from noise than the PC transform. It produces a nice ordering of the new components, which can often be perceived as a decomposition of spatial frequency.

A recent result is that the MAF transform qualifies as an Independent Components Analysis (ICA) transform. In fact MAF is identical to the Molgedy-Schuster ICA decomposition proposed later in 1994 by [86]. A proof of correspondence is given in [81].

3.1.4 SMAF/SMNF Transformations

Unlike the PC, the MAF transform is invariant to linear rank preserving transformations of the original data. Consider a linear transformation of the original data, $\boldsymbol{R} = [\boldsymbol{r}_1\ \boldsymbol{r}_2\ \cdots\ \boldsymbol{r}_N]$, by a non-singular transformation matrix $\boldsymbol{T} \in \mathbb{R}^{P\times P}$ producing $\tilde{\boldsymbol{R}} = \boldsymbol{T}\boldsymbol{R} = [\tilde{\boldsymbol{r}}_1\ \tilde{\boldsymbol{r}}_2\ \cdots\ \tilde{\boldsymbol{r}}_N]$. The dispersion structures are now transformed to

$$\mathrm{D}\{\tilde{\boldsymbol{r}}(\boldsymbol{x})\} = \boldsymbol{T}\boldsymbol{\Sigma}\boldsymbol{T}^T, \tag{3.27}$$

and

$$\mathrm{D}\{\tilde{\boldsymbol{r}}(\boldsymbol{x}) - \tilde{\boldsymbol{r}}(\boldsymbol{x}+\boldsymbol{\Delta})\} = \boldsymbol{T}\boldsymbol{\Sigma}_\Delta\boldsymbol{T}^T \tag{3.28}$$

for the transformed data and the transformed noise respectively. The generalized eigenproblem to solve is now

$$\boldsymbol{T}\boldsymbol{\Sigma}_\Delta\boldsymbol{T}^T\tilde{\boldsymbol{a}} = \lambda\boldsymbol{T}\boldsymbol{\Sigma}\boldsymbol{T}^T\tilde{\boldsymbol{a}}, \tag{3.29}$$

reducing to

$$\boldsymbol{\Sigma}_\Delta\boldsymbol{T}^T\tilde{\boldsymbol{a}} = \lambda\boldsymbol{\Sigma}\boldsymbol{T}^T\tilde{\boldsymbol{a}}. \tag{3.30}$$

Comparing the generalized eigenvalue problems in the Equations 3.26 and 3.30, we see that the eigenvalues remain unchanged and that the relation between the eigenvectors for the original and the transformed data fulfill $\boldsymbol{a_i} = \boldsymbol{T}^T\tilde{\boldsymbol{a}}_i$. The MAFs of the transformed problem becomes $\tilde{\boldsymbol{a}}_i^T\tilde{\boldsymbol{R}} = \boldsymbol{a}_i^T\boldsymbol{R}, i = 1, \cdots, P$. Hence, the MAF transformation is invariant to linear transformations. The same is true for the MNF transforms iif the estimate of the noise structure is transformed by $\boldsymbol{T}\boldsymbol{\Sigma}_n\boldsymbol{T}^T$.

In principal components analysis the new variates are scaled according to the amount of variance they explain. Why not apply a similar property to the MNF/MAF transforms? In the above we showed that the eigenvalues from the MAF problem are invariant to linear transformations. Furthermore, we find a relation of the SNR as a function of the eigenvalues, namely

$$\mathrm{SNR}_i = \frac{2}{\lambda_i} - 1, \quad i = 1, \cdots, P \tag{3.31}$$

where we have used that

$$\boldsymbol{\Sigma}_n = \boldsymbol{\Sigma}_\Delta/2. \tag{3.32}$$

The variance in each MAF component can now be scaled relative to the amount of signal in each component. A natural choice is to scale the new components such that the covariance structure of the noise becomes the indentity matrix. Thus, we impose the condition $\mathrm{V}\{\boldsymbol{a}_i^T\boldsymbol{n}(\boldsymbol{x})\} = 1$ and obtain new components with variance $\mathrm{SNR}_i + 1$. We call the new components signal-MAFs (SMAFs). In order to discard components that have little signal and to reduce the dimensionality of the data, we suggest to eliminate SMAFs that have negative SNRs that can occur due to model breakdown.

The SMAF transformation works as a feature selector. It preserves the MAF components rich on signal and discards those corrupted by noise. It simultaneously stretches the subspaces rich in autocorrelated signal and low in noise.

A similar scaling of the MNF transformation can be done. However for the new components to qualify as SMNFs, the underlying MNF transformation must be invariant to linear transformation.

Both the MAF and the MNF transformations may encounter problems when handling matrices that are not of full rank. Singular problems can be handled by first performing a principal components analysis and discarding the null space before proceeding with the MNF/MAF analysis. This approach is possible due to the invariance to linear transformations.

Case study 1.1 revisited

The PC and the SMAF transformations are applied to a toy image generated as described in Section 2.3.1. The image and the resulting PC and MAF components are shown in Figure 3.1. In Figure 3.2 (top-left) is presented a scatter plot of the test image. With the test image being two-dimensional the scatter plot can also serve to illustrate the spectral space of the calculated PCs.

Comparing the MAFs to the PCs, we notice that they do a much better job of separating the signal from noise. The reason why MAF performs so much better than the PC is related to the fact, that the noise is constructed to be highly correlated in the two bands of the test image. The signal contained in MAF1 has an autocorrelation of 0.26 and in MAF2 of -0.01. In Figure 3.2 (top-right) is included a scatter plot of the traditional MAFs

shown in Figure 3.1. Using Equation 3.32 the estimate of the covariance structure for the noise of the test image is found as [1.24 1.98; 1.98 4.23] which can be compared to the one presented in Section 2.3.1.

One hundred realizations of the toy image are analysed together with the corresponding SMAFs by means of the extended FCM algorithm. A parameter setting similar to the one in Section 2.3.1 is applied using spectral-spatial-parental memberships. The results are shown in Table 3.1. Using

Region of interest	Original data/PCs mean %	std %	SMAFs mean %	std %
Class 1	91.7	5.4	96.3	0.5
Class 2	71.8	7.9	84.7	1.5
Class 3	78.9	6.6	85.1	3.5
The whole image	86.4	5.5	93.0	0.4
Rejected pixels	11.4	2.4	5.4	0.5

Table 3.1: Summary of 100 segmentation results using the original data and using SMAFs as input. The percentage (mean and std) of correctly classified pixels in the different class regions and of rejected pixels are included. Spectral-spatial-parental memberships are applied and hard results are obtained using a reject threshold of 0.9. Using the SMAF transform as a preprocessor improves the performance of the clustering algorithm.

the SMAF transformation as a preprocessor improves the classifications. The segmentations improve from an average of 86% to 93% correctly classified pixels, and the percentage of rejected pixels is reduced from 11% to 5%. Evidence can be found of improved reliability of the algorithm's performance by inspecting the standard deviations in the table. Applying the PC transformation as a preprocessor does not alter the similarity measures used in the clustering algorithm, and the segmentations do thus not benefit from the preprocessing apart from the fact that redundancy and the null space can be removed. Merely applying MAF as a preprocessor will produce degraded results, since the components with noise will have as much influence as the components containing the signal. In Figure 3.3 segmentations of a toy image using the SMAF transform are shown. See [57] for application of the SMAF transform on a multivariate remote sensed image prior to an initial cluster analysis. Note, segmentation of the individual components of MAF may also be useful when trying to label an multivariate image.

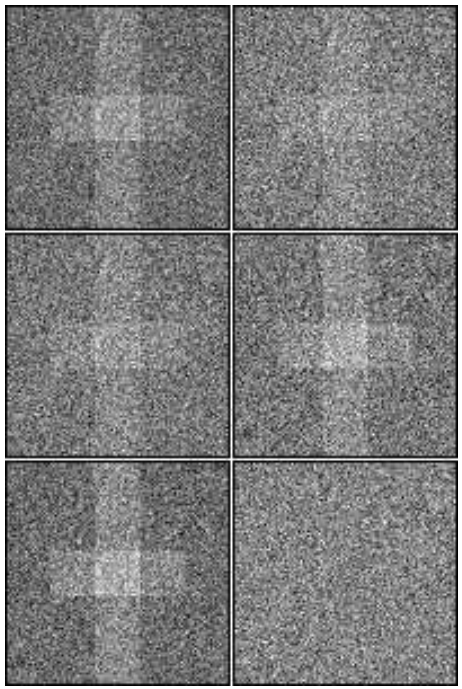


Figure 3.1: The top row contains the test image. The middle row contains the calculated PCs and the bottom row contains the MAFs. All images are scaled to mean ±3 std. Notice that the MAF transform performs a much better job in separating signal from noise when compared to the PC transform.

3.1.5 Combined Spatial and Temporal Autocorrelation Analysis

When working with multitemporal image data it can be desirable not to maximize the spatial autocorrelation but rather the temporal autocorrelation. Let one observation in a multitemporal image data set be represented by $r = r(\boldsymbol{x}, t)$ where $\boldsymbol{x}$ denotes a grid point and t is time, and let $\boldsymbol{r}(t)$ represent one image at time t. Consider estimating $\boldsymbol{\Sigma}$ and $\boldsymbol{\Sigma}_{\delta t}$ as the dispersion matrices of the original variables $\boldsymbol{r}(t)$ and, the difference between the original and the temporally shifted variables $\boldsymbol{r}(t) - \boldsymbol{r}(t + \delta t)$ respectively.

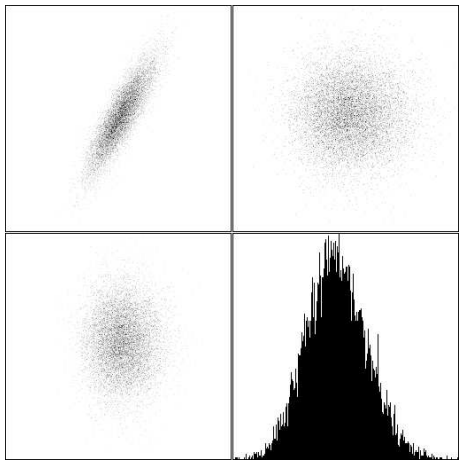


Figure 3.2: Scatterplots 1-3 and the histogram of SMAF1, row-wise ordering. The scatterplots are calculated from 1) the original data, and MAFs constrained to 2) $\boldsymbol{a}^T\boldsymbol{\Sigma}\boldsymbol{a} = 1$, and 3) $\boldsymbol{a}^T\boldsymbol{a} = 1$. Notice that classes are not necessarily well separated in the SMAF space.

The temporal maximum autocorrelation factors (TMAFs) of $\boldsymbol{r}$ are then given by $\boldsymbol{a}_i^T\boldsymbol{r}$ where the $\boldsymbol{a}_i$'s are determined from the eigenvalue problem $\boldsymbol{\Sigma}_{\delta t}\boldsymbol{a}_i = \lambda_i\boldsymbol{\Sigma}\boldsymbol{a}_i$. See [94] for a TMAF analysis of remote-sensed multitemporal image data. As for the ordinary MAF analysis, several shifts may be pooled together. One may even merge the spatially and the temporally estimated covariance structures thus maximizing the autocorrelation with respect to both space and time. The covariance structure of the noise is thus estimated from

$$\boldsymbol{\Sigma}_n = \sum_{i=1}^{N_s} \alpha_i \boldsymbol{\Sigma}_{\Delta_i} + \sum_{j=1}^{N_t} \beta_j \boldsymbol{\Sigma}_{\delta t_j}, \tag{3.33}$$

where N_s and N_t represent the number of spatial and temporal shifts respectively and the α_is and the β_js are weights for the noise matrices. Exactly how to combine the temporal and the spatial dimension can involve prior analysis of the data, before reasonable arguments on how to pool the covariance structures are found. Optimization schemes can also be applied maximizing e.g. the largest eigenvalue. See the previous discussion in Section 3.1.3.

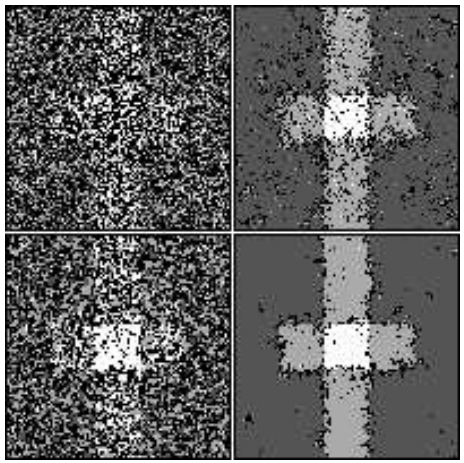


Figure 3.3: Image segmentation 1-4 (row-wise, using SMAF as a preprocessor. The results are based on the memberships: 1) spectral, 2) spectral-parental, 3) spectral-spatial, 4) spectral-spatial-parental. Black=reject class, greyscale $\propto$ class index.

3.1.6 Orthogonal Subspace Projection

Orthogonal subspace projection (OSP) provides the means of projecting high-dimensional data onto a basis which is orthogonal to specific undesired spectra which are often specified by the data analyst, see [85, 53, 95, 96, 98]. OSP is more related to multiple regression, than to the methods of linear decompositioning described in the previous sections, and can be applied in linear mixing problems for e.g. partial unmixing.

If we want to predict the running $P \times 1$ vector observation $\boldsymbol{r}$ by means of a set of variables written as columns in a matrix $\boldsymbol{M}$, the linear mixing model applies

$$\boldsymbol{r} = \boldsymbol{M}\boldsymbol{\gamma} + \boldsymbol{\epsilon}. \tag{3.34}$$

The signal measured at each pixel is thus assumed to consist of a linear combination (determined by the weight vector $\boldsymbol{\gamma}$) of so-called end-members (the columns of $\boldsymbol{M}$) plus some noise term $\boldsymbol{\epsilon}$, i.e. the variation in $\boldsymbol{r}$ that is not explained by the model.

We split the end-members into two groups of desired and undesired end-members. Thus

$$\boldsymbol{M} = [\boldsymbol{D} \;\; \boldsymbol{U}] \tag{3.35}$$

and write

$$\boldsymbol{r} = \boldsymbol{D}\boldsymbol{\alpha}_{\boldsymbol{d}} + \boldsymbol{U}\boldsymbol{\alpha} + \boldsymbol{\epsilon} \tag{3.36}$$

where the vectors $\boldsymbol{\alpha}_{\boldsymbol{d}}$ and $\boldsymbol{\alpha}$ contains the desired and undesired spectral weights respectively. In partial unmixing we assume that only one group of end-members are known. Say that the undesired end-members are given, then the term representing the desired end-members can be merged with the noise term thus introducing $\boldsymbol{n} = \boldsymbol{D}\boldsymbol{\alpha}_{\boldsymbol{d}} + \boldsymbol{\epsilon}$.

When describing the data by means of the undesired spectra only we arrive at the model

$$\boldsymbol{r} = \boldsymbol{U}\boldsymbol{\alpha} + \boldsymbol{n}. \tag{3.37}$$

We want to minimize $n^2 = \boldsymbol{n}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{n} = (\boldsymbol{r} - \boldsymbol{U}\boldsymbol{\alpha})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{r} - \boldsymbol{U}\boldsymbol{\alpha})$. $\boldsymbol{\Sigma}$ is the dispersion of the residuals $\boldsymbol{n}$.

Setting the partial derivative $\partial n^2/\partial\boldsymbol{\alpha} = \boldsymbol{0}$ we get

$$\boldsymbol{\alpha} = (\boldsymbol{U}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{U})^{-1}\boldsymbol{U}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{r}. \tag{3.38}$$

For the residual we now obtain

$$\boldsymbol{n} = (\boldsymbol{I} - \boldsymbol{U}(\boldsymbol{U}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{U})^{-1}\boldsymbol{U}^T\boldsymbol{\Sigma}^{-1})\boldsymbol{r}. \tag{3.39}$$

Applying the transformation matrix

$$\boldsymbol{T} = \boldsymbol{I} - \boldsymbol{U}(\boldsymbol{U}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{U})^{-1}\boldsymbol{U}^T\boldsymbol{\Sigma}^{-1} \tag{3.40}$$

on $\boldsymbol{r}$ results in the orthogonal subspace projection of the data. By nature OSP is a rank reducing transformation. Performing OSP on K spectra reduces the dimensionality of the data from P to $P - K$, assuming full rank of the original data. As serious problem is how to estimate $\boldsymbol{\Sigma}$ which is often assumed proportional to the identity matrix.

3.2 Multiple Sets

Two-set canonical correlations analysis is presented followed by multiset canonical correlations analysis. Both methods can be regarded as exploratory data driven tools which can be applied when looking for linear combinations of the input variables with maximum correlation.

3.2.1 Two-Set Canonical Correlations Analysis

Two-set canonical correlations analysis maximizes the correlation between linear combination of two multivariate groups of variables, see [65, 22]. Consider two sets of variables $\boldsymbol{x}$ and $\boldsymbol{y}$. Let the dimensions of the variables be respectively p and q, with $p \leq q$. Let the variables be described by the the $p+q$ dimensional variable $\boldsymbol{z} = [\boldsymbol{x}^T \quad \boldsymbol{y}^T]^T$ and assume

$$\begin{aligned} \mathrm{E}\{\boldsymbol{z}\} &= \boldsymbol{0} && (3.41) \\ \mathrm{D}\{\boldsymbol{z}\} &= \boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{bmatrix} && (3.42) \end{aligned}$$

with $\boldsymbol{\Sigma}_{11}$ and $\boldsymbol{\Sigma}_{22}$ non-singular and $\boldsymbol{\Sigma}_{12} = \boldsymbol{\Sigma}_{21}^T$.

We wish to determine the transforms

$$\begin{aligned} u &= \boldsymbol{a}^T\boldsymbol{x} && (3.43) \\ v &= \boldsymbol{b}^T\boldsymbol{y} && (3.44) \end{aligned}$$

under the constraints

$$\begin{aligned} \mathrm{Var}\{u\} &= \boldsymbol{a}^T\boldsymbol{\Sigma}_{11}\boldsymbol{a} = 1 && (3.45) \\ \mathrm{Var}\{v\} &= \boldsymbol{b}^T\boldsymbol{\Sigma}_{22}\boldsymbol{b} = 1 && (3.46) \end{aligned}$$

such that the correlation

$$\rho = \mathrm{Corr}\{u,v\} = \frac{\mathrm{Cov}\{u,v\}}{\sqrt{\mathrm{Var}\{u\}\mathrm{Var}\{v\}}} = \mathrm{Cov}\{u,v\} = \boldsymbol{a}^T\boldsymbol{\Sigma}_{12}\boldsymbol{b} \qquad (3.47)$$

is maximized.

We introduce the Lagrange multipliers $\lambda/2$ and $\mu/2$ and maximize

$$\psi = \boldsymbol{a}^T\boldsymbol{\Sigma}_{12}\boldsymbol{b} - \frac{\lambda}{2}(\boldsymbol{a}^T\boldsymbol{\Sigma}_{11}\boldsymbol{a} - 1) - \frac{\mu}{2}(\boldsymbol{b}^T\boldsymbol{\Sigma}_{22}\boldsymbol{b} - 1) \qquad (3.48)$$

with respect to $\boldsymbol{a}$ and $\boldsymbol{b}$.

We find the partial derivatives of ψ as

$$\begin{aligned} \frac{\partial\psi}{\partial\boldsymbol{a}} &= \boldsymbol{\Sigma}_{12}\boldsymbol{b} - \lambda\boldsymbol{\Sigma}_{11}\boldsymbol{a} && (3.49) \\ \frac{\partial\psi}{\partial\boldsymbol{b}} &= \boldsymbol{\Sigma}_{21}\boldsymbol{a} - \mu\boldsymbol{\Sigma}_{22}\boldsymbol{b}. && (3.50) \end{aligned}$$

Setting the partial derivatives equal zero and multiplying by respectively $\boldsymbol{a}^T$ and $\boldsymbol{b}^T$, we obtain

$$\begin{aligned} \boldsymbol{a}^T\boldsymbol{\Sigma}_{12}\boldsymbol{b} - \lambda\boldsymbol{a}^T\boldsymbol{\Sigma}_{11}\boldsymbol{a} &= 0 && (3.51) \\ \boldsymbol{b}^T\boldsymbol{\Sigma}_{21}\boldsymbol{a} - \mu\boldsymbol{b}^T\boldsymbol{\Sigma}_{22}\boldsymbol{b} &= 0. && (3.52) \end{aligned}$$

Using the constraints in Equation 3.45 and 3.46, and the fact that $\boldsymbol{a}^T\boldsymbol{\Sigma}_{12}\boldsymbol{b} = (\boldsymbol{a}^T\boldsymbol{\Sigma}_{12}\boldsymbol{b})^T = \boldsymbol{b}^T\boldsymbol{\Sigma}_{12}^T\boldsymbol{a} = \boldsymbol{b}^T\boldsymbol{\Sigma}_{21}\boldsymbol{a}$, we find that

$$\lambda = \mu = \boldsymbol{a}^T\boldsymbol{\Sigma}_{12}\boldsymbol{b} = \rho \qquad (3.53)$$

which is the correlation between the transformed variables.

We can now solve for the extrema of the function ψ. Writing the partial derivatives as

$$\begin{aligned} \boldsymbol{\Sigma}_{12}\boldsymbol{b} - \rho\boldsymbol{\Sigma}_{11}\boldsymbol{a} &= \boldsymbol{0} && (3.54) \\ \boldsymbol{\Sigma}_{21}\boldsymbol{a} - \rho\boldsymbol{\Sigma}_{22}\boldsymbol{b} &= \boldsymbol{0} && (3.55) \end{aligned}$$

then multiplying Equation 3.54 by ρ and Equation 3.55 by $\boldsymbol{\Sigma}_{22}^{-1}$ gives

$$\begin{aligned} \rho\boldsymbol{\Sigma}_{12}\boldsymbol{b} - \rho^2\boldsymbol{\Sigma}_{11}\boldsymbol{a} &= \boldsymbol{0} && (3.56) \\ \boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}\boldsymbol{a} - \rho\boldsymbol{b} &= \boldsymbol{0}. && (3.57) \end{aligned}$$

Substituting Equation 3.57 in 3.56 we obtain for $\boldsymbol{a}$

$$\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}\boldsymbol{a} = \rho^2\boldsymbol{\Sigma}_{11}\boldsymbol{a} \qquad (3.58)$$

which is recognized as an eigenvalue problem. Similarly, we obtain for $\boldsymbol{b}$

$$\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}\boldsymbol{b} = \rho^2\boldsymbol{\Sigma}_{22}\boldsymbol{b}. \qquad (3.59)$$

Determining the u and the v with maximum correlation is done by projecting $\boldsymbol{x}$ respectively $\boldsymbol{y}$ onto the subspaces spanned by the eigenvectors

$\boldsymbol{a}$ respectively $\boldsymbol{b}$ with the corresponding largest eigenvalue equal to the squared correlation.

If $p = q$ we obtain $u_i,\ i = 1, \cdots, p$, by projecting $\boldsymbol{x}$ onto the subspaces spanned by the eigenvectors $\boldsymbol{a}_1, \cdots, \boldsymbol{a}_p$ corresponding to the eigenvalues $\rho_1^2 \geq \cdots \geq \rho_p^2$ of $\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}$. Projecting $\boldsymbol{y}$ onto the subspaces spanned by the eigenvectors $\boldsymbol{b}_1, \cdots, \boldsymbol{b}_p$ of $\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}$ corresponding to the same eigenvalues we obtain $v_i,\ i = 1, \cdots, p$. If $p < q$ the eigenvalue problem in Equation 3.59 degenerates since the last eigenvalue will equal zero with $(q-p)$th multiplicity.

The CCA transformation chooses $p+q$ linear transforms

$$u_i = \boldsymbol{a}_i^T\boldsymbol{x},\ i = 1, \cdots, p \tag{3.60}$$

$$v_i = \boldsymbol{b}_i^T\boldsymbol{y},\ i = 1, \cdots, q, \tag{3.61}$$

such that the variates are uncorrelated within groups expressed by

$$\mathrm{Corr}\{\boldsymbol{a}_i^T\boldsymbol{x}, \boldsymbol{a}_j^T\boldsymbol{x}\} = \delta_{ij} \tag{3.62}$$

$$\mathrm{Corr}\{\boldsymbol{b}_i^T\boldsymbol{y}, \boldsymbol{b}_j^T\boldsymbol{y}\} = \delta_{ij} \tag{3.63}$$

and between groups expressed by

$$\mathrm{Corr}\{\boldsymbol{a}_i^T\boldsymbol{x}, \boldsymbol{b}_j^T\boldsymbol{y}\} = \rho_j\delta_{ij} \tag{3.64}$$

where δ is the Kronecker delta.

CCA thus jointly analyses two sets of variables. It finds two sets of linear combinations (called canonical variates, CVs) of the zero mean original variables that maximize correlation between the two. The two CV1s are linear combinations of the original data (one from each set) that are maximally correlated. Higher order CVs are maximally correlated subject to orthogonality or uncorrelatedness with lower order CVs. The correlations obtained between corresponding CVs are termed canonical correlations.

Consider linear transformations of the x and the y data sets by the matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ respectively. The eigenvalue problem in e.g. 3.58 thus becomes

$$\boldsymbol{A}\boldsymbol{\Sigma}_{12}\boldsymbol{B}^T(\boldsymbol{B}\boldsymbol{\Sigma}_{22}\boldsymbol{B}^T)^{-1}\boldsymbol{B}\boldsymbol{\Sigma}_{21}\boldsymbol{A}^T\tilde{\boldsymbol{a}} = \rho^2\boldsymbol{A}\boldsymbol{\Sigma}_{11}\boldsymbol{A}^T\tilde{\boldsymbol{a}} \tag{3.65}$$

$$\Rightarrow \boldsymbol{\Sigma}_{12}{\boldsymbol{\Sigma}_{22}}^{-1}\boldsymbol{\Sigma}_{21}\boldsymbol{A}^T\tilde{\boldsymbol{a}} = \rho^2\boldsymbol{\Sigma}_{11}\boldsymbol{A}^T\tilde{\boldsymbol{a}}. \tag{3.66}$$

We see that the canonical correlations remain unchanged and that $\boldsymbol{A}^T\tilde{\boldsymbol{a}} = \boldsymbol{a}$. Thus the CVs remain unchanged, and CCA is therefore invariant to linear transformations of the individual data sets.

Partial Least Squares

CCA is closely related to the method of partial least squares (PLS) in which $R = \mathrm{Cov}\{\boldsymbol{a}^T\boldsymbol{x}, \boldsymbol{b}^T\boldsymbol{y}\} = \boldsymbol{a}^T\boldsymbol{\Sigma}_{12}\boldsymbol{b}$ (often with $\boldsymbol{y}$ as a scalar response variable) is maximized with another choice of constraints, namely $\boldsymbol{a}^T\boldsymbol{a} = \boldsymbol{b}^T\boldsymbol{b} = 1$ leading to

$$R^2 = \frac{\boldsymbol{a}^T\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{21}\boldsymbol{a}}{\boldsymbol{a}^T\boldsymbol{a}} = \frac{\boldsymbol{b}^T\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{12}\boldsymbol{b}}{\boldsymbol{b}^T\boldsymbol{b}}, \tag{3.67}$$

[10]. We see that in this case matrix inversion is not needed which is good if we have many variables and few observations. Only the first pair of canonical variates (or latent variables) are calculated and the response CV is regressed on the predictor CV. If more information is present in the residuals these are subtracted from the original response variables, the predictor variables are projected into a subspace orthogonal to the solution found, and more iterations are performed, see also [129, 42, 63, 35].

3.2.2 Multiset Canonical Correlations Analysis

The linear multiset canonical correlations analysis (MCCA) is a technique which maximizes the sum of the pair-wise correlations over all linear combinations of the multivariate sets. Work done applying the MCCA can be found in [74, 89, 97].

Let $\boldsymbol{x}$ be an $m = m_1 + m_2 + \cdots + m_n$ dimensional variable, and without loss of generality assume $\mathrm{E}\{\boldsymbol{x}\} = \boldsymbol{0}$. We furthermore assume that the sets are ordered according to rank and that the minimum rank is m_{min}.

The signal variable is represented by

$$\boldsymbol{x} = \begin{bmatrix} \boldsymbol{x}_1 \\ \vdots \\ \boldsymbol{x}_n \end{bmatrix}, \tag{3.68}$$

containing the n sets of variables

$$\boldsymbol{x}_1 = \begin{bmatrix} x_{11} \\ \vdots \\ x_{1m_1} \end{bmatrix},\ \boldsymbol{x}_2 = \begin{bmatrix} x_{21} \\ \vdots \\ x_{2m_2} \end{bmatrix}, \ldots, \boldsymbol{x}_n = \begin{bmatrix} x_{n1} \\ \vdots \\ x_{nm_n} \end{bmatrix}, \tag{3.69}$$

where each $\boldsymbol{x}_i$ is the m_i dimensional signal variable from set i. We assume that the covariance functions can be represented by the covariance matrices $\text{Cov}\{\boldsymbol{x}_i, \boldsymbol{x}_j\} = \boldsymbol{\Sigma}_{ij} \in \mathbb{R}^{m_i \times m_j}$ and may thus write

$$\text{D}\{\boldsymbol{x}\} = \boldsymbol{\Sigma}_{\boldsymbol{x}} = \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} & \cdots & \boldsymbol{\Sigma}_{1n} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} & \cdots & \boldsymbol{\Sigma}_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{\Sigma}_{n1} & \boldsymbol{\Sigma}_{n2} & \cdots & \boldsymbol{\Sigma}_{nn} \end{bmatrix}. \tag{3.70}$$

We are looking for nm_{min} linear combinations

$$z_{ij} = \boldsymbol{a}_{ij}^T \boldsymbol{x}_i, \quad i = 1, \cdots, n; j = 1, \cdots, m_{min} \tag{3.71}$$

such that the sum of the correlation

$$\begin{aligned} R &= \sum_{k=1}^{n} \sum_{l=1}^{n} \rho(z_{kj}, z_{lj}) && (3.72) \\ &= \sum_{k=1}^{n} \sum_{l=1}^{n} \frac{\boldsymbol{a}_{kj}^T \boldsymbol{\Sigma}_{kl} \boldsymbol{a}_{lj}}{\sqrt{(\boldsymbol{a}_{kj}^T \boldsymbol{\Sigma}_{kk} \boldsymbol{a}_{kj} \ \boldsymbol{a}_{lj}^T \boldsymbol{\Sigma}_{ll} \boldsymbol{a}_{lj})}} && (3.73) \end{aligned}$$

is maximum among all linear transforms orthogonal to z_{pq}, $p = 1, \cdots, n; q = 1, \cdots, j-1$.

The dispersion structure of the transformed variables with maximal correlations, say $\boldsymbol{z} = [z_1 \ \ z_2 \ \ \cdots \ \ z_n]^T$ determined by the weight vector $\boldsymbol{a} = [\boldsymbol{a}_1^T \ \ \boldsymbol{a}_2^T \ \ \cdots \ \ \boldsymbol{a}_n^T]^T$ can be written

$$\text{D}\{\boldsymbol{z}\} = \boldsymbol{\Sigma}_{\boldsymbol{z}} = \begin{bmatrix} \boldsymbol{a}_1^T \boldsymbol{\Sigma}_{11} \boldsymbol{a}_1 & \boldsymbol{a}_1^T \boldsymbol{\Sigma}_{12} \boldsymbol{a}_2 & \cdots & \boldsymbol{a}_1^T \boldsymbol{\Sigma}_{1n} \boldsymbol{a}_n \\ \boldsymbol{a}_2^T \boldsymbol{\Sigma}_{21} \boldsymbol{a}_1 & \boldsymbol{a}_2^T \boldsymbol{\Sigma}_{22} \boldsymbol{a}_2 & \cdots & \boldsymbol{a}_2^T \boldsymbol{\Sigma}_{2n} \boldsymbol{a}_n \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{a}_n^T \boldsymbol{\Sigma}_{n1} \boldsymbol{a}_1 & \boldsymbol{a}_n^T \boldsymbol{\Sigma}_{n2} \boldsymbol{a}_2 & \cdots & \boldsymbol{a}_n^T \boldsymbol{\Sigma}_{nn} \boldsymbol{a}_n \end{bmatrix}. \tag{3.74}$$

Maximizing the sum of the elements in $\boldsymbol{\Sigma}_{\boldsymbol{z}}$ under different constraints will maximize the sum of the pair-wise correlations. [74] and [89] list measures to maximize, including the one just presented, when looking for maximal correlations are given i) maximize the sum of elements, ii) maximize the sum of squared elements, iii) maximize the largest eigenvalue, iv) minimize the smallest eigenvalue, and v) minimize the determinant of $\boldsymbol{\Sigma}_{\boldsymbol{z}}$.

Natural constraints are also given and include

1) $\boldsymbol{a}^T \boldsymbol{a} = 1$
2) $\boldsymbol{a}_i^T \boldsymbol{a}_i = 1 \ \ \forall \ \ i$
3) $\boldsymbol{a}_i^T \boldsymbol{\Sigma}_{ii} \boldsymbol{a}_i = 1 \ \ \forall \ \ i$
4) $\text{trace}\{\boldsymbol{\Sigma}_{\boldsymbol{z}}\} = \sum_{i=1}^{n} \boldsymbol{a}_i^T \boldsymbol{\Sigma}_{ii} \boldsymbol{a}_i = 1 \ \ \forall \ \ i.$

In [62] maximizing the sum of elements is examined, and [74] examines all the above measures using the constraint of unit-variance of the new components. In [89] the measures are examined under the remaining constraints and other natural constraints are proposed.

We shall limit this presentation to maximizing the sum of the elements of $\boldsymbol{\Sigma}_{\boldsymbol{z}}$ under constraints 2, 3 and 4. Under constraint 1 the problem reduces to principal components analysis of $\boldsymbol{x}$. We use the Lagrange multiplier technique and for the three constraints define the following functions

$$\begin{aligned} L_2 &= \sum_{i=1}^{n} \sum_{j=1}^{n} \boldsymbol{a}_i^T \boldsymbol{\Sigma}_{ij} \boldsymbol{a}_j - \sum_{i=1}^{n} \lambda_i (\boldsymbol{a}_i^T \boldsymbol{a}_i - 1) && (3.75) \\ L_3 &= \sum_{i=1}^{n} \sum_{j=1}^{n} \boldsymbol{a}_i^T \boldsymbol{\Sigma}_{ij} \boldsymbol{a}_j - \sum_{i=1}^{n} \lambda_i (\boldsymbol{a}_i^T \boldsymbol{\Sigma}_{ii} \boldsymbol{a}_i - 1) && (3.76) \\ L_4 &= \sum_{i=1}^{n} \sum_{j=1}^{n} \boldsymbol{a}_i^T \boldsymbol{\Sigma}_{ij} \boldsymbol{a}_j - \lambda (\sum_{i=1}^{n} \boldsymbol{a}_i^T \boldsymbol{\Sigma}_{ii} \boldsymbol{a}_i - 1). && (3.77) \end{aligned}$$

Solving $\partial L_c / \partial \lambda = 0, c = 2, 3, 4$ produce the expressions for the corresponding constraints. Furthermore we have that

$$\frac{\partial}{\partial \boldsymbol{a}_i} \sum_{i=1}^{n} \sum_{j=1}^{n} \boldsymbol{a}_i^T \boldsymbol{\Sigma}_{ij} \boldsymbol{a}_j = 2 \sum_{j=1}^{n} \boldsymbol{\Sigma}_{ij} \boldsymbol{a}_j \tag{3.78}$$

and by solving $\partial L_c / \partial \boldsymbol{a}_i = \boldsymbol{0}, c = 2, 3, 4$ respectively produces the following systems

$$\begin{aligned} \sum_{j=1}^{n} \boldsymbol{\Sigma}_{ij} \boldsymbol{a}_j &= \lambda_i \boldsymbol{a}_i, \quad c = 2, i = 1, \cdots, n && (3.79) \\ \sum_{j=1}^{n} \boldsymbol{\Sigma}_{ij} \boldsymbol{a}_j &= \lambda_i \boldsymbol{\Sigma}_{ii} \boldsymbol{a}_i, \quad c = 3, i = 1, \cdots, n && (3.80) \\ \sum_{j=1}^{n} \boldsymbol{\Sigma}_{ij} \boldsymbol{a}_j &= \lambda \boldsymbol{\Sigma}_{ii} \boldsymbol{a}_i, \quad c = 4, i = 1, \cdots, n && (3.81) \end{aligned}$$

of which only the latter is a generalized eigensystem. The first two systems of equations can be solved using an iterative approach which we shall account for in the following section. Note, solutions to generalized eigensystems are also iterative. When written in matrix form Equations 3.79 and 3.80 become

$$\begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} & \cdots & \boldsymbol{\Sigma}_{1n} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} & \cdots & \boldsymbol{\Sigma}_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{\Sigma}_{n1} & \boldsymbol{\Sigma}_{n2} & \cdots & \boldsymbol{\Sigma}_{nn} \end{bmatrix} \begin{bmatrix} \boldsymbol{a}_1 \\ \boldsymbol{a}_2 \\ \vdots \\ \boldsymbol{a}_n \end{bmatrix} = \begin{bmatrix} \lambda_1 \boldsymbol{a}_1 \\ \lambda_2 \boldsymbol{a}_2 \\ \vdots \\ \lambda_n \boldsymbol{a}_n \end{bmatrix}, \tag{3.82}$$

and

$$\begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} & \cdots & \boldsymbol{\Sigma}_{1n} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} & \cdots & \boldsymbol{\Sigma}_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{\Sigma}_{n1} & \boldsymbol{\Sigma}_{n2} & \cdots & \boldsymbol{\Sigma}_{nn} \end{bmatrix} \begin{bmatrix} \boldsymbol{a}_1 \\ \boldsymbol{a}_2 \\ \vdots \\ \boldsymbol{a}_n \end{bmatrix} = \begin{bmatrix} \lambda_1 \boldsymbol{\Sigma}_{11} & 0 & \cdots & 0 \\ 0 & \lambda_2 \boldsymbol{\Sigma}_{22} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_n \boldsymbol{\Sigma}_{nn} \end{bmatrix} \begin{bmatrix} \boldsymbol{a}_1 \\ \boldsymbol{a}_2 \\ \vdots \\ \boldsymbol{a}_n \end{bmatrix}. \tag{3.83}$$

Once m_{min} orthogonal solutions have been found, the corresponding sets can be removed from the analysis, and additional canonical variates can be found between the remaining sets by applying a restricted search until only the null space is left.

Multiset canonical correlations analysis can be applied to simultaneously maximize several criteria. Maximum autocorrelation factors may be generated applying MCCA to sets generated using different spatial shifts. Shifting the sets in the spatial domain provides analyses similar to applying the traditional MAF transform, but also provides the means to simultaneously maximize the similarity between all groups of variables.

Multiset Partial Least Squares

Multiset or multiblock partial least squares (PLS) can be based on Equation 3.82 with $\boldsymbol{\Sigma}_{ii}, i = 1, \cdots, n$ replaced by the null matrix. In multiblock

PLS we want to maximize covariance and thus do not want to include the diagonal terms of $\boldsymbol{\Sigma}_{\boldsymbol{z}}$ in 3.74 in the maximization. See also [97].

3.2.3 Procrustes Alignment

In shape analysis, where estimates of the similarity between shapes, their average shape and variability is of interest, Procrustes analysis is an useful tool, [50, 6, 49, 28].

Alignment is done to remove all Euclidean effects i.e. translation, rotation, and isotropic scale. When using the L_2-norm, the term Generalized Procrustes Analysis (GPA) is applied when handling multiset scenarios of landmark registered shapes with known correspondences. The alignment of only two such shapes is called Ordinary Procrustes Analysis (OPA).

Ordinary Procrustes Analysis

Consider the case of aligning two landmark registered shapes in a two-dimensional space. Thus each landmark in set i consists of a coordinate vector $\boldsymbol{x}_i = [x_{i1} \;\; x_{i2}]^T$. Assuming $\mathrm{E}\{\boldsymbol{x}_i\} = 0, i = 1, 2$ all translation has been removed and only rotation and scale must be determined. Ordinary Procrustes analysis minimizes

$$e^2_{OPA} = \mathrm{E}\{(\|\boldsymbol{x}_1 - \boldsymbol{z}_2\|)^2\} = \mathrm{E}\{(\|\boldsymbol{x}_1 - \boldsymbol{A}\boldsymbol{x}_2\|)^2\} \tag{3.84}$$

where $\|\cdot\|$ is the L_2-norm. Set 1 is considered as a reference set and OPA typically estimates the transformation, $\boldsymbol{z}_2 = \boldsymbol{A}\boldsymbol{x}_2$, of set 2 in an orthogonal multiple regression setting minimizing the mean of the squared residuals. Notice that $\boldsymbol{A} = [a \; -b \; ; \; b \;\; a]$ can be split into two parts $\boldsymbol{A} = \boldsymbol{A}_1\boldsymbol{A}_2$ both Euclidean rotation and scaling matrices. Minimizing

$$\mathrm{E}\{(\|\boldsymbol{x}_1 - \boldsymbol{A}_1\boldsymbol{A}_2\boldsymbol{x}_2\|)^2\} \tag{3.85}$$

corresponds to minimizing

$$\mathrm{E}\{(\|\boldsymbol{A}_1^{-1}\boldsymbol{x}_1 - \boldsymbol{A}_2\boldsymbol{x}_2\|)^2\}, \tag{3.86}$$

where $\boldsymbol{A}_1^{-1} = \boldsymbol{A}_1^T$ is a rotation and scaling matrix as well. We are looking for the transformations of both sets that minimize the mean of the squared

residuals thus maximizing correlation. The alignment can thus be obtained by solving a canonical correlations problem.

Consider constructing the sets

$$\tilde{x}_{i1} = \{\{x_{k1}\}_{k=1}^{N}, \{-x_{l2}\}_{l=1}^{N}\} \quad (3.87)$$
$$\tilde{x}_{i2} = \{\{x_{k2}\}_{k=1}^{N}, \{x_{l1}\}_{l=1}^{N}\} \quad (3.88)$$

and define

$$\tilde{\boldsymbol{x}}_i = [\tilde{x}_{i1} \;\; \tilde{x}_{i2}]^T, \; i = 1, 2. \quad (3.89)$$

When solving the canonical correlations problem of minimizing the variance of the resisuals

$$\mathrm{E}\{(\boldsymbol{p}_1^T \tilde{\boldsymbol{x}}_1 - \boldsymbol{p}_2^T \tilde{\boldsymbol{x}}_2)^2\}, \quad (3.90)$$

then the basis (the eigenvectors) for the first pair of canonical variates will contain estimates of a_1, b_1, a_2, and b_2 that defines $\boldsymbol{A}_1$ and $\boldsymbol{A}_2$. In addition to the first solution which maximize the correlation of the two shapes a second canonical pair is also obtained providing an alignment solution which OPA does not find.

Performing the CCA analysis on the sets $\boldsymbol{x}_i, i = 1, 2$ will minimize the squared residual of the two shapes when projected onto one-dimensional subspaces. In shape analysis in higher than two dimensions complications arise due to nonlinearities in the rotation and scaling matrices. This applies to both the OPA and the CCA approach. To solve the problem one may apply iterative strategy maximizing correlation in alternating subspaces, but no clear solution appears yet to exist.

Generalized Procrustes Analysis

Often we wish to align a multiset of shapes for which the Generalized Procrustes Analysis (GPA) is developed. Assuming n sets with N landmarks and $\mathrm{E}\{\boldsymbol{x}_i\} = 0, i = 1, \cdots, n$, GPA minimizes

$$e_{GPA}^2 = \sum_{ij}^{n} \mathrm{E}\{(\|\boldsymbol{z}_i - \boldsymbol{z}_j\|)^2\} = \sum_{ij}^{n} \mathrm{E}\{(\|\boldsymbol{A}_i \boldsymbol{x}_j - \boldsymbol{A}_i \boldsymbol{x}_j\|)^2\} \quad (3.91)$$

by applying an iterative scheme. Defining the mean shape by

$$\bar{\boldsymbol{z}} = \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{z}_i \quad (3.92)$$

the scheme iteratively i) applies OPA on each shape using the mean shape as a reference, and ii) re-estimate the mean shape while preserving its variance. The procedure is presented in Algorithm 5. Normalization of mean shape is important to prevent the algorithm from drifting towards null transformations.

Algorithm 5 Generalized Procrustes Alignment

1: Initialize $\bar{\boldsymbol{z}}^{k=0} = 1/n \sum_{i=1}^{n} \boldsymbol{x}_i$ and $w_j^0 = 1/N, j = 1, \cdots, N$
2: Normalize $\bar{\boldsymbol{z}}^0 = \bar{\boldsymbol{z}}^0 / \|\bar{\boldsymbol{z}}^0\|$
3: **repeat**
4: Using weighted OPA align each element in the set $\{\boldsymbol{x}_i\}_{i=1}^{n}$ to $\bar{\boldsymbol{z}}^{k-1}$ producing $\{\boldsymbol{z}_i\}_{i=1}^{n}$
5: Calculate new weights $w_j = 1/\mathrm{Var}\{\boldsymbol{z}_i(j)\}$, $j = 1, \cdots, N$, where $\mathrm{Var}\{\boldsymbol{z}_i(j)\}$ is the variance of the jth landmark over all aligned shapes.
6: Estimate mean shape from the aligned set $\bar{\boldsymbol{z}}^k = 1/n \sum_{i=1}^{n} \boldsymbol{z}_i$
7: Normalize $\bar{\boldsymbol{z}}^k = \bar{\boldsymbol{z}}^k / \|\bar{\boldsymbol{z}}^k\|$
8: **until** Convergence

Generally GPA is very robust and converges in few (less than five, often only two to three) iterations. Extensions have been proposed including landmark weights that are re-estimated in each iteration. A weighting scheme is included in Algorithm 5. Typically, weights are chosen proportional to the inverse variance of position of the landmark over all sets. The weight selection scheme is a variant of the Boosting schemes, [55, 36], applied in classification problems.

MCCA revisited

The measure minimized by generalized Procrustes fitting corresponds to maximizing the sum of the pair-wise correlations, that is the elements in $\boldsymbol{\Sigma}_{\boldsymbol{z}}$ from 3.74, under variance preserving constraints for the mean shape, typically normalized such that $\mathrm{Var}\{1/n \sum_{i=1}^{n} \boldsymbol{a}_i^T \boldsymbol{x}_i\} = 1$. Alignments can thus also be obtained by applying the MCCA system of equations given in

3.79 and 3.80. The only difference to ordinary GPA is that the variance preserving restrictions on the mean shape is replaced by the appropriate constraint. A Generalized Multiset Canonical Correlations Analysis scheme is presented in Algorithm 6 under MCCA constraints 2 and 3 for finding the canonical variate set with maximal correlation. Under these constraints the scale of each transformed set is determined within sets by the variance perserving conditions.

Algorithm 6 Generalized Multiset Canonical Correlations Algorithm

1: Initialize $\boldsymbol{a}_i = \mathbf{1}, i = 1, \cdots, n$
2: Normalize $\boldsymbol{a}_i = \begin{cases} \boldsymbol{a}_i/\sqrt{(}\boldsymbol{a}_i^T\boldsymbol{a}_i), & \text{Preserving } \boldsymbol{a}_i^T\boldsymbol{a}_i = 1 \ \forall \ i \\ \boldsymbol{a}_i/\sqrt{(}\boldsymbol{a}_i^T\boldsymbol{\Sigma}_{ii}\boldsymbol{a}_i), & \text{Preserving } \boldsymbol{a}_i^T\boldsymbol{\Sigma}_{ii}\boldsymbol{a}_i = 1 \ \forall \ i \end{cases}$
3: Calculate $z_i^{k=0} = \boldsymbol{a}_i^T\boldsymbol{x}_i$, $\bar{z}^0 = 1/n\sum_{i=1}^n \boldsymbol{z}_i^0$ and set $w_j = 1/N, j = 1, \cdots, N$
4: **repeat**
5: Apply weighted multiple regression on $\{\boldsymbol{z}_i^{k-1}\}_{i=1}^n$ to $\bar{z}^{k-1}$ updating the set $\{\boldsymbol{a}_i\}_{i=1}^n$
6: Normalize $\boldsymbol{a}_i = \begin{cases} \boldsymbol{a}_i/\sqrt{(}\boldsymbol{a}_i^T\boldsymbol{a}_i), & \text{Preserving } \boldsymbol{a}_i^T\boldsymbol{a}_i = 1 \ \forall \ i \\ \boldsymbol{a}_i/\sqrt{(}\boldsymbol{a}_i^T\boldsymbol{\Sigma}_{ii}\boldsymbol{a}_i), & \text{Preserving } \boldsymbol{a}_i^T\boldsymbol{\Sigma}_{ii}\boldsymbol{a}_i = 1 \ \forall \ i \end{cases}$
7: Calculate $z_i^{k=0} = \boldsymbol{a}_i^T\boldsymbol{x}_i$
8: Calculate new weights $w_j = 1/\mathrm{Var}\{\boldsymbol{z}_i(j)\}$, $j = 1, \cdots, N$, where $\mathrm{Var}\{\boldsymbol{z}_i(j)\}$ is the variance of the jth landmark over all aligned shapes.
9: Estimate mean shape from the aligned set $\bar{z}^k = 1/n\sum_{i=1}^n \boldsymbol{z}_i^k$
10: **until** Convergence

Successive solutions may be found by constraining the "eigenvectors" to the subspace orthogonal to the previously determined solutions. There are several possible extensions to MCCA that could be introduced, especially on how to choose the weights for the observations. Moreover, one could construct alternative algorithms applying different reference shapes other than the mean shape e.g. the median shape across sets.

3.2.4 Change Detection

In change detection often only two sets are involved. For handling multivariate sets there are several approaches. If we are working with well calibrated and noise free data, the difference between each variable in the

sets is often sufficient for change detection, and the redundancy in the data may be removed by principal components analysis. However, based on the previous discussion it should be clear, that when handling image data e.g. the MAF transform is a better approach. It all depends on whether the change one is interested in detecting is characterized by spatial homogeneous regions in the data or not. Applying MAF on the difference data makes good sense when the sets are well calibrated and removes uninteresting change characterized by speckled noise. In [115] case studies are presented applying both the PC and the MAF transform on bitemporal multivariate data. In cases where the data are not well calibrated or even measured from different sources, two-set CCA may be a better approach. It is invariant to linear transformations and thus qualifies as a good tool for data fusion in change detection studies. Differences of the CCA pairs can be applied resulting in the so-called Maximum Alteration Detection (MAD) transformation, [99]. The MAD transform is a CCA analysis of the two sets involved followed by taking the differences of the CV pairs and reversing the ordering of the resulting components. The first MAD component thus corresponds to the difference between the last CV pair that has minimal correlation, hence containing maximum difference between the CV pairs. The MAF transform can of course be applied as a post-processing step of the MAD transform, [92]. For change detection in a multiset case, MCCA is applied in [89].

3.3 Case Studies

Three case studies are presented in this section. First the SeaWiFS example from Section 2.3.2 is revisited performing linear decomposition of the data under the FCM obtained water mask. The second case study contains a linear two-set CCA analysis of multitemporal global sea surface temperature and height data. Finally, a short example is included performing linear MCCA on a multiset of landmark registered shapes.

3.3.1 Linear Decomposition of Multispectral Image Data

The acquisition and analysis of the ocean colour is important in many environmental studies. The concentration of substances and particles in

the lighted zone of the upper ocean influences the apparent colour of the ocean. The colour can range from deep blue to varying shades of green and ruddy brown. Living phytoplankton (which contains chlorophyll and associated photosynthetic pigments), inorganic sediments, and dissolved organic matter all contribute to the colour of the ocean. Remote sensing in ocean studies can provide information on the large-scale ocean configuration. However, both the temporal and the spatial sampling is limited by e.g. cloud coverage. In the study of the North Sea, the data coverage is particularly poor and the sampling of useful optical remote sensed images sparse.

In this case study linear decomposition is performed on the SeaWiFS scene, previously presented in Section 2.3.2, in an attempt to enhance the ocean related signal in the data. The SeaWiFS images are shown in Figure 3.4 and are stretched to mean ± three standard deviations (std) under the water mask obtained from the cluster analysis when using six classes (see Figure 2.8). Notice that all the images seem to include very little dynamics.

Results and Dicussion

The previous initial cluster analysis partitioned the SeaWiFS image into six classes of which two are recognized as related to signals dominated by clouds. This allows us to perform partial unmixing of signals by means of othogonal subspace projection. The cloud spectra for the cluster classes (no. 2 and 3) are applied as undesired spectra, see Figure 2.10. The results of the OSP analysis is shown in Figure 3.5 stretching the images to mean ± three std. Notice that the dynamics in the images increase. Thus, there is ocean related signal contained in the SeaWiFS signal, but it has very little variance and is hard to identify primarily due to the corrupting and dominating cloud signal in the data.

Post-processing of the OSP cloud reduced signal is done by a PC and a MAF analysis. The results are shown in the Figures 3.6 and 3.7 and result in six components since OSP is a rank reducing transformation. The transformations are calculated on the basis of the observations under the water mask but are applied to the whole image. Both the PC and the MAF transforms seem successful in separating part of the signal from salt-and-pepper-noise. Comparing the PCs and the MAFs we notice that MAF seems to produce a nicer ordering of the components in terms of image qual-

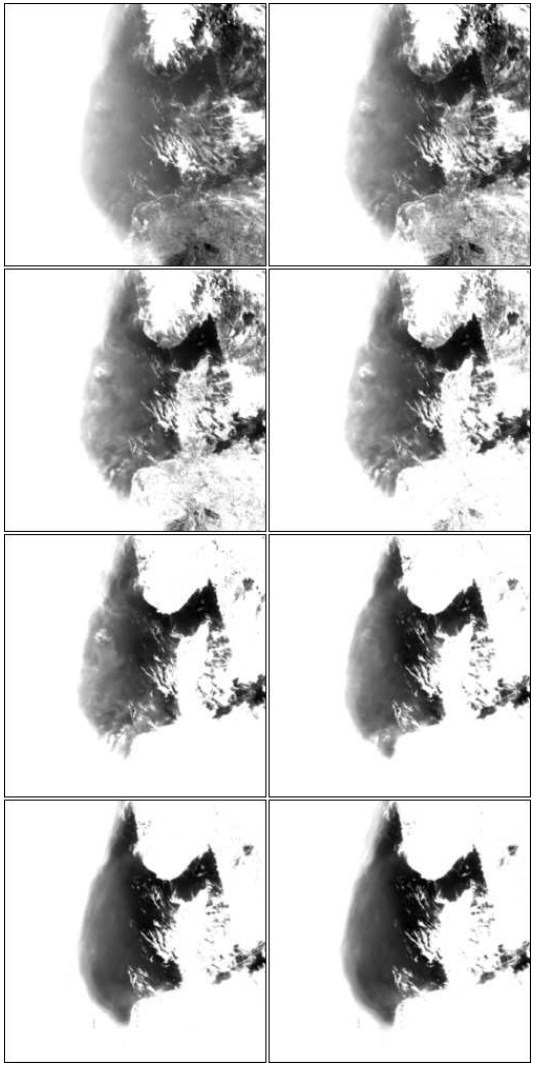


Figure 3.4: The original SeaWiFS bands 1-8 row-wise. The data are stretched to mean ±3 std under the water mask. Notice that there is little dynamics in the individual components.

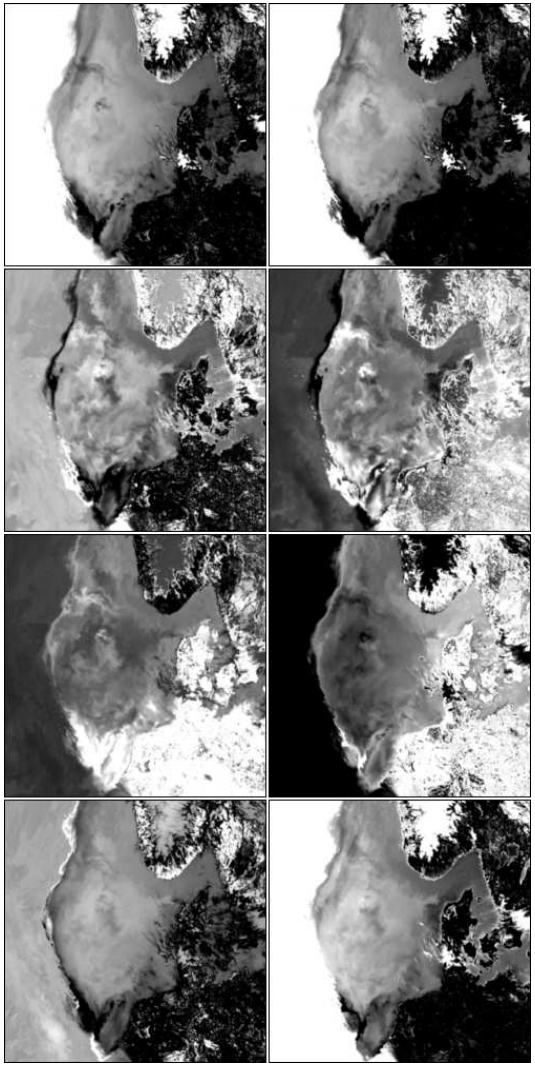


Figure 3.5: The SeaWiFS bands 1-8 row-wise after OSP cloud signal reduction. The data are stretched to mean ±3 std under the water mask. Notice the increase in the dynamics under the ocean water mask.

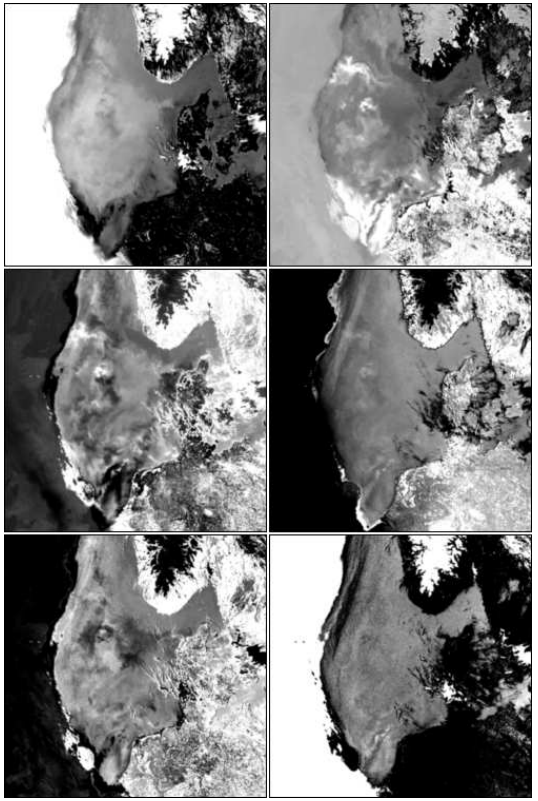


Figure 3.6: Principal components of the SeaWiFS bands 1-8 row-wise after OSP cloud signal reduction. The data are stretched to mean ±3 std under the water mask.

ity. Correlations, weigths and loading for the calculated PCs and MAFs can form the basis for a further, more supervised analysis and interpretation of the individual components. Concentrating on the MAFs the first four components seem to collect most of the signal remaining in the data after the OSP analysis. The last two MAFs seem related primarily to noise and contains little autocorrelation. In Figure 3.8 the correlations between the MAF components and the original SeaWiFS bands are shown. Concentrating on the first three OSPMAF components, we see that they all have a maximum in the *absolute* value of the correlation to wavelengths

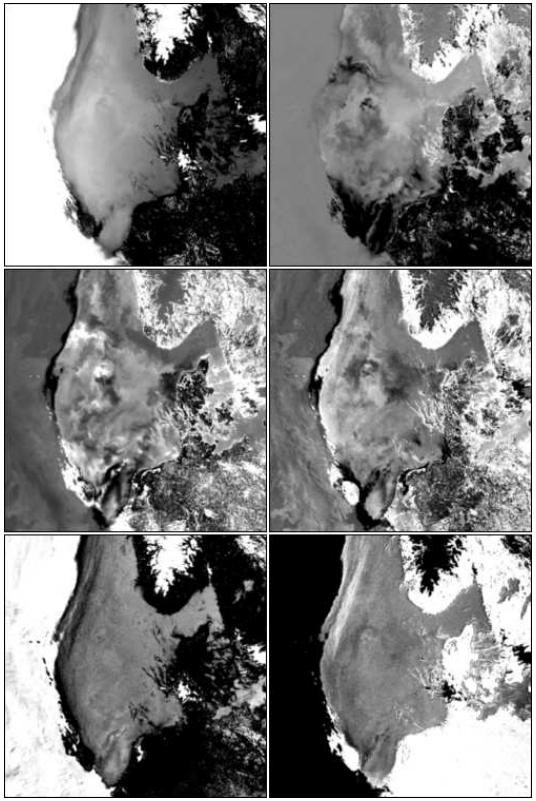


Figure 3.7: Maximum autocorrelations factors of the SeaWiFS bands 1-8 row-wise after OSP cloud signal reduction. The data are stretched under the water mask.

related to visible light (SeaWiFS bands 1-6). The first component has a maximum for violet (412 nm) light with monotone decreasing correlations for higher wavelengths. The second OSPMAF has a maximum at green (555 nm) combined with low absolute correlations to the wavelengths in the NIR range. The third componet has a maximum at blue-green (490 nm) combined with increased correlations in the NIR range. The spectra for the components 1-3 are similar to the reflectance spectra one can obtain for the general water classes known as i) open-ocean water (Atlantic Ocean Water), ii) coastal-water (Baltic Sea/German Bight Water) with

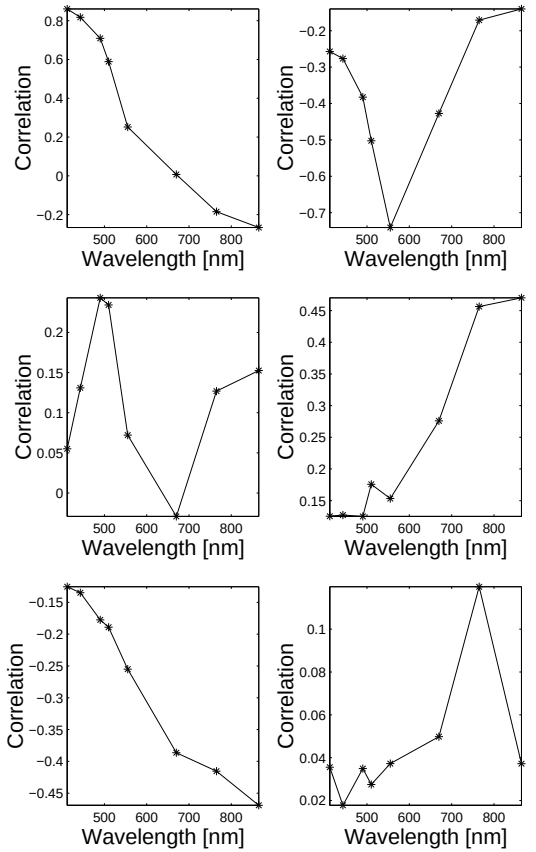


Figure 3.8: Correlations of the OSPMAFs 1-6 (row-wise) and the original SeaWiFS data. The stars indicate the center wavelengths of the individual SeaWiFS bands 1-8.

plankton and chlorophyll, and iii) costal-water with higher concentration of suspended matter producing higher reflectance of high wavelengths.

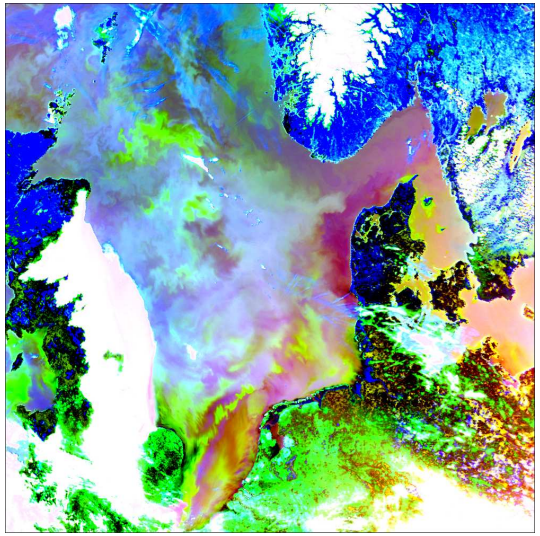


Figure 3.9: OSPMAFs 1-3 as RGB for 13th of May 1998.

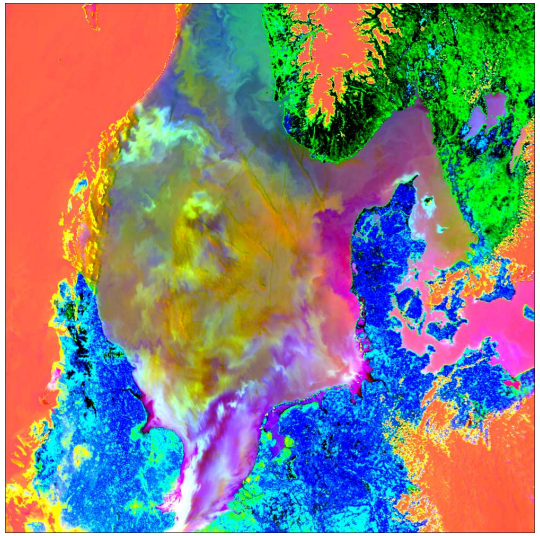


Figure 3.10: OSPMAFs 1-3 as RGB for 15th of May 1998.

Conclusion

In general the combined cluster analysis and OSPMAF transformation succeed in decomposing the signal into few components that are rich on ocean related signal with reduced undesired cloud and noise signal. The structures and dynamics contained in the first components of both the PCs and the MAFs are identifiable in similar analysis of neighbouring days [59, 60, 100]. The first three OSPMAFs from these analyses are shown in the Figures 3.9 and 3.10 as RGB images. Concentrating on Figure 3.10 it depicts the large scale configuration of the water masses in the North Sea. Atlantic Ocean Water arrives from the South through the English Channel. It mixes primarily with the continental coastal water, consisting of ocean water and river run-off, and is represented by its own characteristic purple colour. A different type of Atlantic Ocean Water, the primarily yellow/green coloured water masses, arrives from the North-West and dominates the central and Northern part of the North Sea. Baltic brackish water, red coloured water masses, is transported towards the North Sea primarily in the surface layers. It becomes gradually mixed with different ocean waters encountered in the Skagerak. From here it continues up the west coast of Norway as the rather permanent Norwegian Coastal Current.

The exploratory analysis presented produces components that are in good agreement with the expected ocean configuration during summer in the North Sea. When comparing the OSPMAFs and the RGB images to the class-map containing the general summer water class configuration, Figure 3.11, and the generalized near-surface patterns of water movements, Figure 3.12, we see noteworthy good agreement. Oceanographer N. K. Højerslev (pers. comm.) confirms the promising potential of the data driven approach. The sampling of applicable optical remotely sensed images is known to be very sparse for the North Sea and the adjacent waters. On average only 15 to 17 days a year give full spatial data coverage. Moreover, the inter-annual variation is very high and some years may not contain any such large-scale samples. By utilizing the temporal dimension composite images can be constructed to compensate for the poor spatial resolution. However, the composite technique cannot be applied when a combined temporal and large-scale spatial analysis is needed. Thus, the development of new tools for extracting information on the ocean configuration is important. The presented spectral-spatial decomposition of the remotely sensed

optical data, with suppression of undesired spectra and noise, can help obtain a better temporal and spatial sampling of the ocean. The application and new combination of the exploratory methods is thus expected to be useful in the future analysis and understanding of the ocean dynamics and mixing, (pers. comm. N. K. Højerslev).

Acknowledgements

The SeaWiFS data where available through the Danish GEOSONAR project funded by the joint Danish Research Councils under the Earth Observation Program, [120]. The data are received by the University of Dundee Satellite receiving station and are produced by the SeaWiFS Project at Goddard Space Flight Center, [68]. Use of this data are in accord with the SeaWiFS Research Data Use Terms and Conditions Agreement. Dr. Per Knudsen, the National Survey and Cadastre, Denmark heads GEOSONAR. Oceanographer Dr. N. K. Højerslev is acknowledged for comments on the results of data analysis and for providing the class map of the water masses of the North Sea in the summer, and the figure of generalized of near-surface pattern water movement.

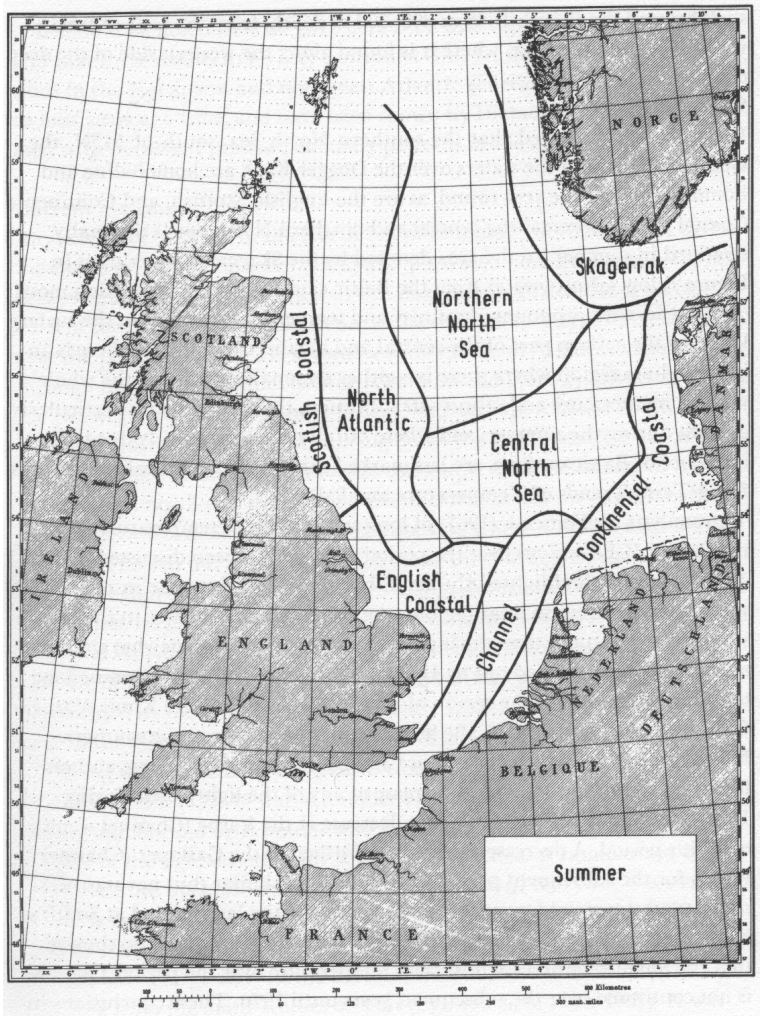


Figure 3.11: Water masses of the North Sea in the summer, [48].

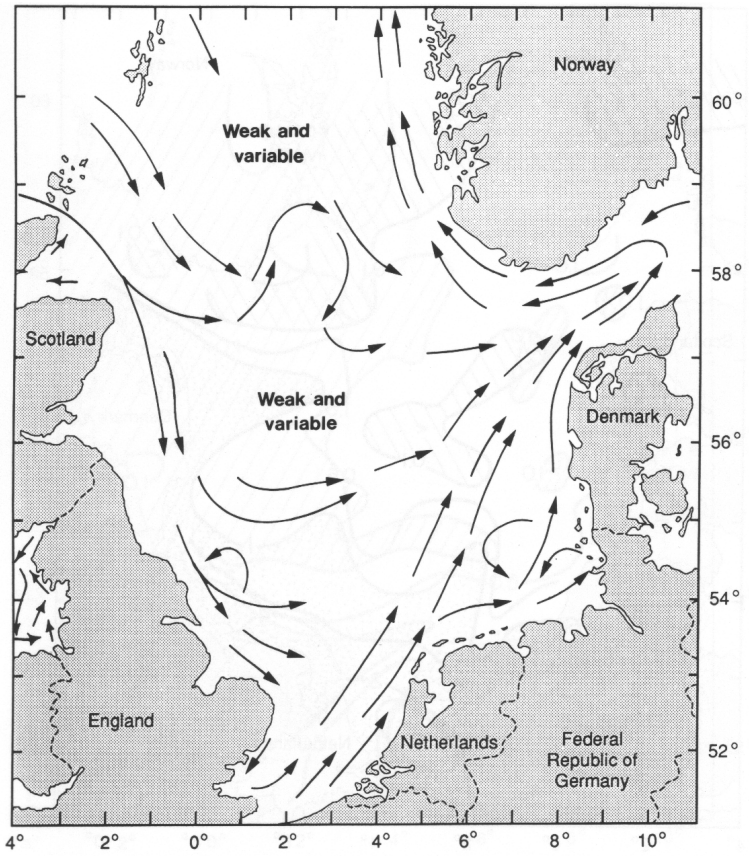


Figure 3.12: Generalized near-surface pattern of water movement, [61].

3.3.2 Canonical Correlations Analysis of Multitemporal Global Sea Surface Height and Temperature

This case study describes the application of canonical correlations analysis to the joint analysis of global monthly mean values of 1996-1997 sea surface temperature (SST) and height anomaly (SSH) data. The SST data are considered as one set and the SSH anomaly data as another set of multivariate observations, both with 24 variables. This type of analysis can be considered as an extension of traditional empirical orthogonal function (EOF) analysis that provides a marginal analysis of one variable over time. The motivation for using a bivariate extention stems from the fact that the two fields are interrelated as i.e. an increase in the SST will lead to an increase in the SSH. An extension of the method of empirical orthogonal functions (EOF) is presented, [104, 75]. EOF analysis is often used in oceanography and other geophysical sciences to analyse temporal sequences of scalar (image) data. In this study we extend this type of analysis by applying canonical correlations analysis to two temporal sequences of scalar image data, namely global sea surface temperature (SST) and global sea surface height (SSH). This type of analysis can be extended further to more than two sets of data. The data used to illustrate the analysis carried out comes from the Ocean Pathfinder programmes set up by NASA/NOAA, [67, 69]. The data chosen represent relevant oceangraphic properties related to one of the largest El Niño events ever recorded, [32]. El Niño is a large-scale warm ocean event in the Pacific off the coast of Peru and Ecuador caused by eastward drifting toward the west coast of South America of the pool of warm waters normally residing in the western part of the Pacific. This event is not local but may influence weather conditions worldwide.

EOF analysis is a name often used in geophysical data processing for principal component (PC) analysis, [64, 2]. Often the usual PCA assumption on variables with mean zero is replaced by an assumption of temporal means of zero.

Results and Discussion

The data used are global monthly mean values of 1996-1997 SST data from the NOAA/NASA Oceans Pathfinder AVHRR SST database, [69], and global monthly mean values of 1996-1997 SSH data from the NASA/GSFC Ocean Altimeter Pathfinder database, [67]. The SSH data are interpo-

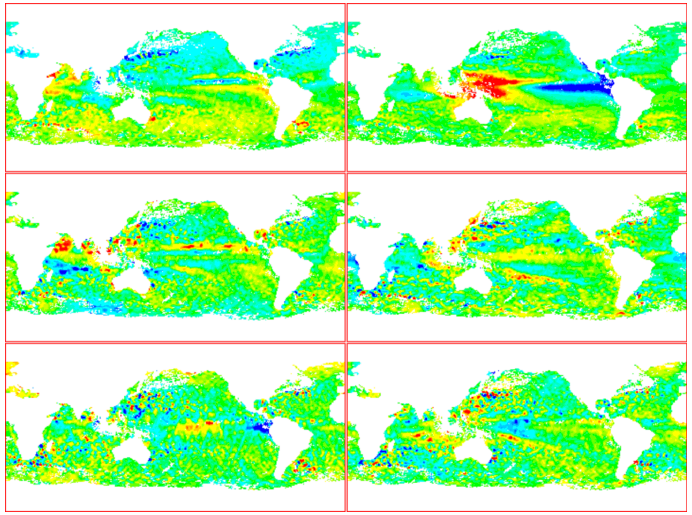


Figure 3.13: SST CVs 1-6 row-wise.

lated point observations from the TOPEX/Poseidon radar altimeter mission. The SST come as 360 rows by 720 columns half degree data starting at 180^o longitude, the SSH come as 179 rows by 360 columns one degree data starting at 0^o longitude. Consequently, the SST data have been resampled to the SSH grid. The AVHRR instrument is influenced by cloud coverage whereas the radar altimeter provides uninterrupted data.

Figure 3.13 shows the first six SST CVs resulting from an CCA analysis where the 24 months of SST data are considered as one set and the 24 months of SSH data are considered as the other set. The CVs are stretched linearly from mean three standard deviations, the pseudo-colour scale goes from blue (minimum) over cyan, green, yellow to red (maximum). Figure 3.14 shows the first six SSH CVs (same analysis, stretch and colouring). Statistics for the CCA are calculated only where both variables have non-missing values for all 24 months.

The first pair of CVs has a correlation coefficient of 0.7587 and accordingly the two exhibit very similar spatial patterns. Generally speaking, we see highs in the Southern Hemisphere and lows in the Northern Hemisphere. This is in accordance with the observed maximally negative correlations

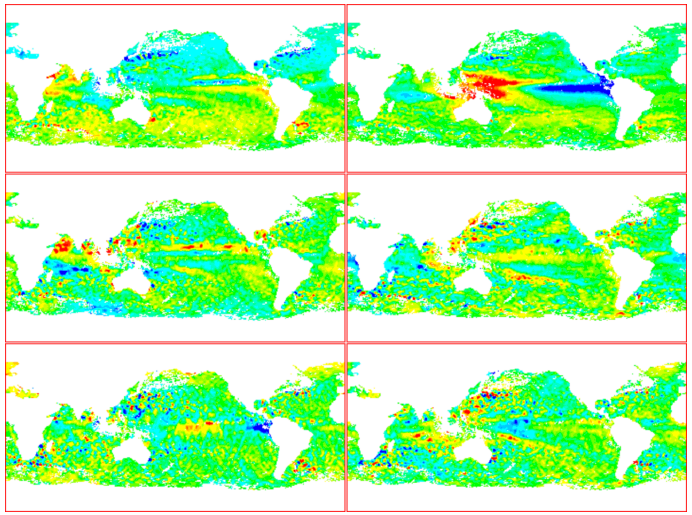


Figure 3.14: SSH CVs 1-6 row-wise.

between both SST and SSH and both CV1s in the winter months. A high in SSH CV1 and a strong high in SST CV1 off the Equatorial west coast of South America is seen. This is in accordance with the El Niño phenomenon. The apparent oscillation is clearly at an annual period, but for the correlation between the CV1 and the SST remains negative and does not alternate around zero, which should be expected. The highest correlations occur in month 9 and 21 for both the SST and the SSH. Normally the highest correlation for the SST would occur one month earlier, but because of the joint analysis of the SST and SSH, this get dominated by the SSH. This consequently answers the previous negative correlations between the CV1 and the SST. In Figure 3.15 the canonical correlations (black-curve), and correlations between original SST (left of the dashed vertical line, index 1-24) and SSH variables (right of the dashed vertical line, index 25-48), and SST CVs 1-3 (red, green and blue curves, left figure) and SSH CVs 1-3 (RGB curves, right figure). The second pair of CVs has a correlation coefficient of 0.7179 and again the two exhibit similar spatial patterns. For this CV pair we see strong lows in both SSH CV2 and SST CV2 off the Equatorial west coast of South America which again is in accordance with the El Niño event. Also, a high in SST CV2 and a very strong high in

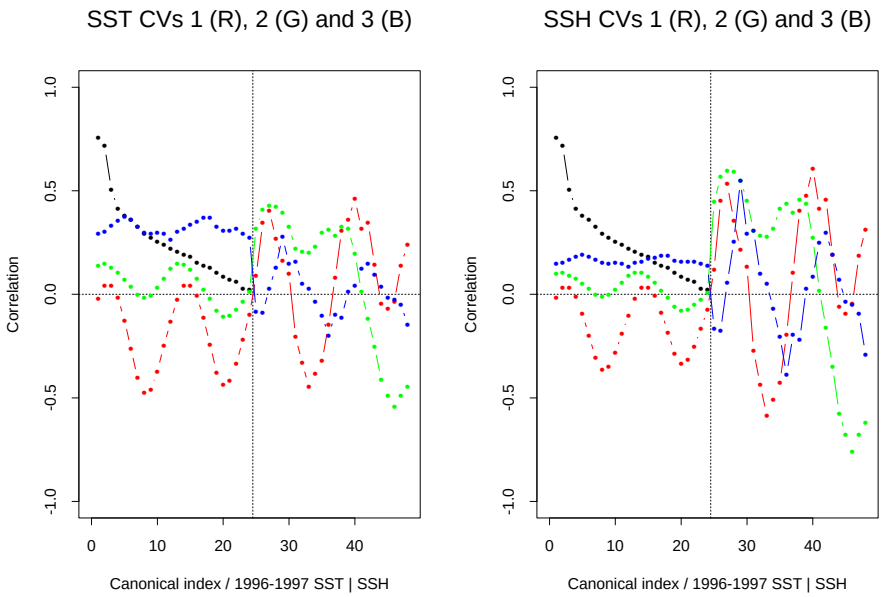


Figure 3.15: Correlations between the canonical CVs and the original monthly data.

SSH CV2 east of Indonesia and the Philippines is seen. The correlations between the original data and the CVs indicate that both CVs 1 and CVs 2 reflect strong annual oscillations [75]. Especially for SSH and CVs 2 but also for SSH and CVs1 and to a lesser degree for SST and CVs2 these oscillations seem to be disturbed in the second half of 1997. This is in agreement with the spatial patterns observed in especially CVs 2 and the El Niño build-up during the last eight months of 1997. The third canonical correlation is 0.5061, and the fourth is 0.4148. This is still relatively high, but the fact that the first two canonical correlations are much higher than the remaining ones shows that the joint variation of the two geophysical fields over the two-year period is well explained by only two CVs. On the other hand the two highest canonical correlations are only 0.7587 and 0.7179, and therefore there are also variations in the two fields that are not well explained by a joint (linear) analysis. For a temporal analysis of the same data see [94]. For a nonlinear analysis of the same data, see [56].

Conclusion

In spite of the very short time span of the data and the associated risk of over-interpreting the results, simultaneous inspection of spatial patterns of the CVs and the correlations between the original and transformed variables from the analysis gives good indications of an anomaly off the South American west coast taking place in the second half of 1997. This is in good agreement with established oceanographic knowledge on the build-up of one of the largest El Niño events on record.

Future analysis should include longer time series in order to establish whether 1997 (and 1998) really represent anomalous events in terms of global SST and SSH. Investigations should also be made in which the SST field is shifted temporally to the SSH field in order to investigate, if the correlations between the CV and the individual SST and SSH fields could be increased. The apparent phase lag of one month between the highs of the SST and the SSH is explained by the fact that the SST represents the instant temperature of the sea surface, whereas the steric expansion causing the sea level to rise is more of an integrated effect.

Acknowledgements

The Pathfinder SST data provision at JPL is due to J. Vazques, R. Sumagaysay and co-workers. The Pathfinder SSH data provision at GSFC is due to V. Zlotnicki and co-workers. This work is done as a part of the GEOSONAR project funded by the Danish National Research Councils under the Earth Observation Program. Dr. Per Knudsen, the National Survey and Cadastre, Denmark heads GEOSONAR.

3.3.3 Two-Dimensional Multiset Shape Alignment

This is a short case study which is included to illustrate the effect of applying MCCA to problems related to multiset shape alignment. A general accepted definition of shape is “all that remains after Euclidean transformations have been removed”. Shape alignment and representation are important tasks that come before performing e.g. linear decomposition of the variability in the shape space. The data are 24 sets of 50 two-dimensional landmark registered metacarpal II bones, see the Figure 3.16. The un-

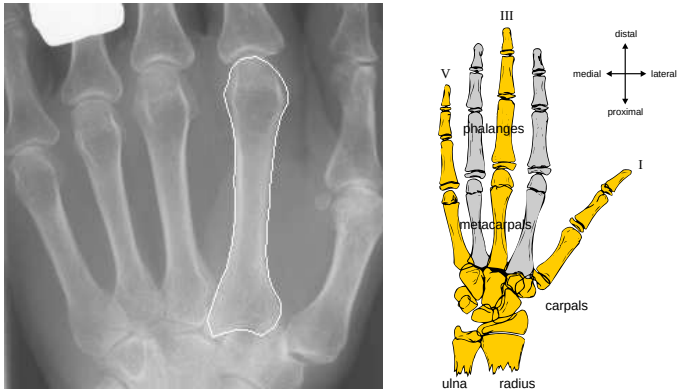


Figure 3.16: A radiograph of the human hand. The metacarpal II has been marked.

aligned data are presented in Figure 3.17.

Results and Discussion

The results of alignment using different methods are shown in the Figure 3.18. The MCCA analysis provides two different alignments, whereas the GPA only provides one. Two different weight schemes have been applied in the GPA analyses i.e. using equal weights and using weights that are inversely proportional to the variance of the corresponding landmark over all sets. Notice that the MCCA CV1s and both the GPA results are very similar. MCCA is performed under the constraint of unit variance within each set and thus differes from GPA in which the variance of the mean shape is restricted. This explains why the first MCCA solution differs slightly from the GPA result using equal weights. Thus, to include scale into each transformation matrix of the individual shapes MCCA should be applied under e.g. the constraint of constant sum of variance over all sets. The pair-wise correlation for the second MCCA alignment is very high and the solution is orthogonal to the fist MCCA alignment. The average pair-wise correlations are 0.999 and 0.991 for the first and the second MCCA solutions respectively. For the second alignment it appears as if the algorithm manages to produce a high correlation mainly based on the observation related to landmarks on the mid-section of the bone at the

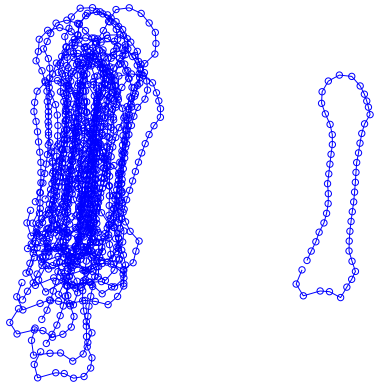
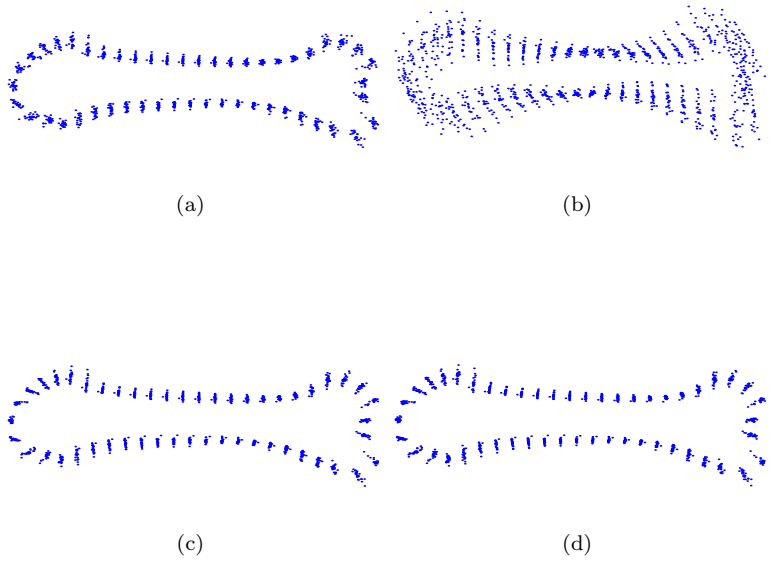


Figure 3.17: A set of 24 unaligned bones. Notice the position outlier to the right.

expense of increased variance at the ends of the bone. It is clear that the alignment of the second MCCA solution does not remove rotation and thus does not qualify as a traditional shape alignment.

The aligned shapes are analysed using the PC transform. This is done in order to obtain a model that spans most of the variation in the high dimensional shape space with only few components. In Figure 3.19 the dynamics included in the first three PCs is illustrated for the first and the second MCCA solutions. Notice, that the PCs for the second MCCA solution differ from solutions found for the first MCCA alignment. The first component in the second MCCA set appears to have separated the rotational residue effect in its alignment. The higher order PCs for the second MCCA solution may therefore be interesting to study for important features concerning the dynamics in the traditional shape space. In Figure 3.20 the fraction of variance explained by each eigenmode of the PC components is given. For the first MCCA solution the first 10 eigenmodes explain 93% of the shape variation and for the second MCCA the first 10 eigenmodes explain 99%. Decomposition could also be obtained using e.g. the MAF transform which would be more robust to annotation noise, [81].



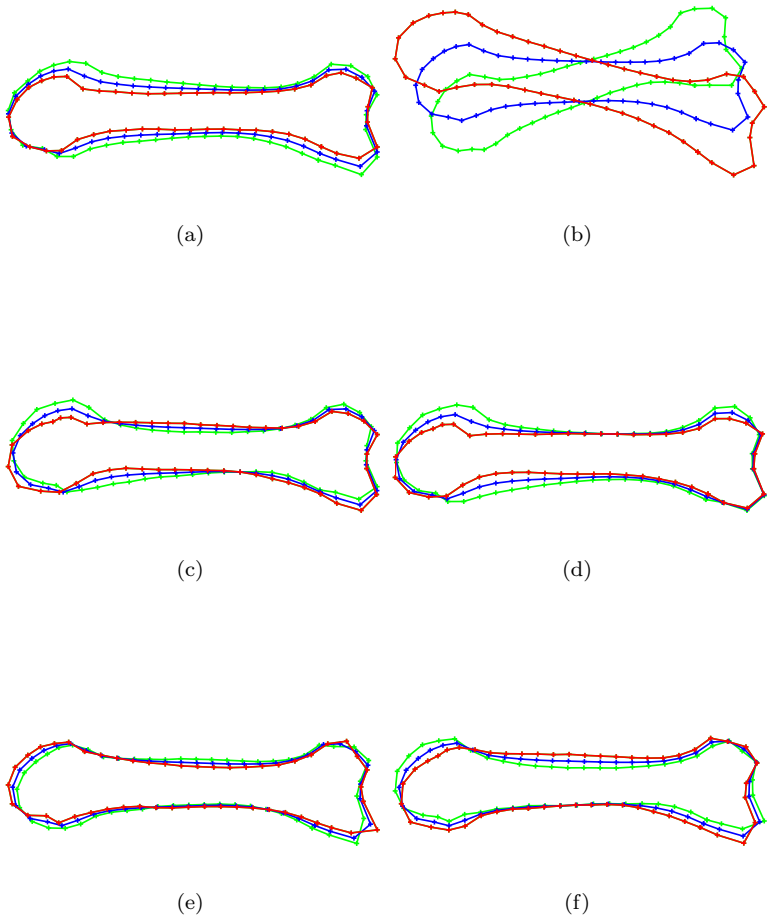
(a) (b)

(c) (d)

Figure 3.18: Alignment using (a) The first MCCA alignment solution, (b) the second MCCA alignment solution, (c) Procrustes fit with equal weights, (d) Procrustes fit with weights with variable weights.

Conclusion

The multiset of 24 two-dimensional shapes is aligned using GPA and MCCA alignment. The MCCA alignment produces two solutions, both invariant to linear transformations of the individual unaligned data sets. The first MCCA solution is similar to the traditional alignments obtained through GPA. The second MCCA alignment clearly includes some rotational effects, and one can therefore argue against applying it in a further analysis. However, if more complex shapes are processed, it is expected that the suboptimal alignments of MCCA may reveal additional information concerning the structures and correlations in the data. The presented decompositioning corresponds to the model building phase in Active Shape Models (ASM),

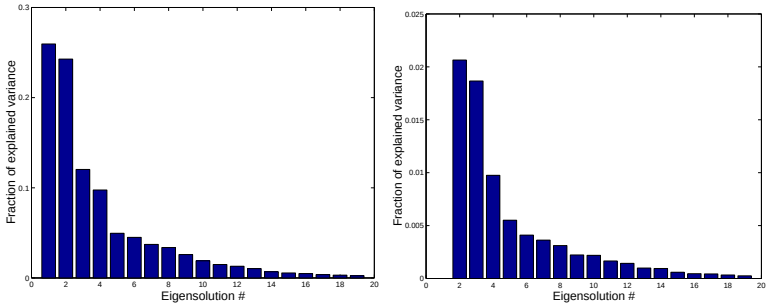


(a) (b)

(c) (d)

(e) (f)

Figure 3.19: Shape dynamics from a PC analysis using (1st column) the first MCCA alignment, and (2nd column) the second MCCA alignment solution. The ith row contains the ith eigensolution. Blue is the mean shape, red/green is the mean shape deformation using $\pm$ 3 std.



(a) (b)

Figure 3.20: Fraction of variance explained by each eigenmode for (a) the first MCCA alignment solution, (b) the second MCCA alignment solution, here the first eigenmode explains 92.3% of the total variation (not shown in the bar plot).

[25, 113]. The models in ASM can be generalized in a straightforward manner to include textural information as proposed in Active Appearance Models (AAM), [23, 24, 113].

Acknowledgements

The data set was supplied by M.D. Lars Hyldstrup, Dept. of Endocrinology, H:S Hvidovre University Hospital, Copenhagen, Denmark, and annotated by Hans Henrik Thodberg, Ph.D., Pronosco A/S, Denmark.

3.4 Summary

- Methods for linear decomposition of single and multisets are presented.
- For single sets the methods include the principal components (PC) transform, the minimum noise fractions (MNF) transform, the maximum autocorrelation factors (MAF) analysis, a new signal minimum/maximum noise/autocorrelation factors transform (SMNF/SMAF), and the orthogonal subspace projection (OSP) transform for partial unmixing.
- The new SMNF/SMAF transform scales the components depending on the signal-to-noise ratio contained in each subspace. The motivation for the stretching/compression is to increase the distance between natural clusters in the data. Improved unsupervised segmentation results for simulated image data are obtained using the SMAF as a preprocessor.
- An approach for maximizing both the spatial and the temporal autocorrelation in multitemporal image data is described.
- Two-set canonical correlations analysis (CCA) and multiset linear canonical correlations analysis (MCCA) are presented.
- The ordinary (OPA) and generalized Procrustes alignment (GPA) methods are presented. A relation to the canonical correlations transformation is found.
- A case study of multispectral image data is presented. Both the PC and the MAF transform are applied on OSP could signal reduced data. The data driven analysis succeeds in extracting components rich in ocean related signal, which can form the basis for a further supervised investigation on the ocean dynamics and configuration.
- Two-set CCA is performed on multitemporal global sea surface height and temperature data. The analysis finds components highly correlated to the El Niño build-up in the last part of 1997.
- Two-dimensional multiset shape alignment is performed on 24 metacarpal II bones each registered by 50 landmarks. MCCA analysis and GPA analyses are performed for aligning the shapes. Decomposition of the MCCA alignments is performed by PC analysis and results in a compact low dimensional model, which includes most of the data variation of the high dimensional shape space.

Chapter 4

Nonlinear Decomposition

This chapter addresses the problem of nonlinear decomposition of multivariate data. Traditional approaches containing explicit and implicit methods are briefly described and followed by a section on nonlinear additive models. In the latter section several methods are presented including the general additive models and the alternating conditional expectations algorithm. These models provide the foundation for a new algorithm which finds the optimal transformations that maximize correlation for multivariate multiset problems. The transformations can be nonlinear even non-monotone mappings. The new algorithm can handle multisets of mixed types of categorical and continuous variables. Case studies are presented including i) a decomposition of bivariate multitemporal sea surface height and temperature data, ii) a multiset analysis of the autocorrelation in Landsat TM multivariate image data, iii) a multiset nonlinear principal components analysis of MSS data, and iii) an analysis of irregularly sampled multivariate stream sediment geochemistry data treated as a multiset problem.

4.1 Traditional Approaches

Projecting the data into a higher-dimensional nonlinear space may reveal important information less evident in the original representation. In the following explicit and implicit methods are presented. Both approaches apply the traditional linear techniques for decomposition. Other noteworthy nonlinear methods of representing the data in a lower dimensional target space are e.g. Multi-Dimensional Scaling (MDS), [117, 79, 108], and the Self-Organizing Maps by Kohonen, [76, 77]. These shall not be dealt with here but should be mentioned as useful tools in exploratory analysis. MDS aims at minimizing a stress coefficient which measures how much the distances between the observation in the target space differs from the original data representation. SOM is related to artificial neural networks and clustering. It assigns the observations to the nodes of the applied net and attempts to preserve the topography of the input data letting similar observations dominate neighbourning parts of the net. In contrast to the SOM algorithm the MDS method provides a measure of the quality of the representation of the data in the target space.

4.1.1 Explicit Methods

For performing nonlinear decomposition of multivariate data the most simple approach is first explicitly to represent the data on a new basis generated by nonlinear transformations of the original variables. Secondly, ordinary linear transformations can be applied to perform the decomposition. The explicit approach may thus be interpreted as a feature generation, combination and selection procedure. Polynomial regression is one method that may apply the explicit approach. The predictors are expanded onto a polynomial basis, and the prediction problem is solved by means of ordinary linear multiple regression. Similar approaches can be applied when performing the principal component analysis or the maximum autocorrelation factor analysis on multivariate data. An obvious problem faced by the explicit approach is the curse of dimensionality which limits the number of terms we can include in the new basis. The general aim is to find a new basis onto which the projected data reveal interesting, otherwise occluded structures. Construction of the new space is therefore very critical, since we run the risk of destroying, say, true clustering in the data and may risk imposing artificial correlation structures between variables. When restricted

transformations such as nonlinear rank mapping transforms are applied, more robust analyses are often obtained.

4.1.2 Implicit Methods

Implicit methods also attempt to analyze the data in a higher nonlinear dimension. The methods apply the "kernel idea" which is also used in e.g. Support Vector Machines [123] and other kernel based algorithms. Most linear decomposition problems (and some clustering algorithms) can be formulated solely by the inner-product of the involved features. Exploiting this fact nonlinear decomposition can be performed by implicitly performing the analysis in the so-called "kernel"-space calculating all the inner products by the corresponding kernel function $< \boldsymbol{x}|\boldsymbol{y} >= k(\boldsymbol{x}, \boldsymbol{y})$. Examples of kernels are Gaussian Radial Basis Functions $k(\boldsymbol{x}, \boldsymbol{y}) = \exp(-\|\boldsymbol{x} - \boldsymbol{y}\|^2/c)$ or polynomial kernels $k(\boldsymbol{x}, \boldsymbol{y}) = (\boldsymbol{x}^T\boldsymbol{y})^d$, for some positive constants c and d respectively. In [84] ordinary Fisher discriminant analysis with kernels is successfully performed and found to be competitive to e.g. support vector machines on a wide range of two-class problems. In [110] the conditions which the class of functions must satisfy are presented for them to qualify as appropriate kernels. Performing an implicit PC analysis is also referred to as kernel PC analysis, [109]. To the best of the author's knowledge, the implicit approach has yet to be applied to MNF/MAF and CCA problems in image analysis.

4.2 Nonlinear Additive Models

Methods belonging to the class of nonlinear additive models are described in this section. The models are commonly applied in regression problems, and we will briefly introduce the most common ones before arriving at a new method for decomposition of multiset multivariate problems in which the optimal transformations that maximize the sum of the pairwise correlations are estimated. For a given response and predictor data set $\{y_i, \boldsymbol{x}_i\}_{i=1}^N, \boldsymbol{x} \in \mathbb{R}^m$ a common formulation of the regression problem involves finding the optimal target function $f(\boldsymbol{x})$ which minimizes $\mathrm{E}\{(y - f(\boldsymbol{x}))^2\}$. With the model $y = f(\boldsymbol{x}) + \epsilon$ the optimal transformation of the predictors is $f(\boldsymbol{x}) = \mathrm{E}\{y|\boldsymbol{x}\}$, and estimating this target function solves the regression problem.

4.2.1 Generalized Additive Models

Due to their flexibility, the generalized additive models (GAM) by Hastie and Tibshirani are often applied to identify and characterize nonlinear regression effects. See [54, 52]. In the regression setting, a generalized additive model has the form

$$f(\boldsymbol{x}) = \alpha + f_1(x_1) + f_2(x_2) + \cdots f_m(x_m). \tag{4.1}$$

The set $\{x_i\}_{i=1}^m$ represents the predictors and y the response variable, the f_i's are unspecified "non-parametric" functions.

In GAMs the individual transformation of the predictor variables are found by applying scatterplot smoothers. The choice of smoother depends on the data type, and if a priori knowledge is available appropriate restrictions on the individual transformations may be imposed. GAM can be implemented to handle mixtures of continuous variables and categorical variables. For a well posed problem, regularizing smoothness restrictions are often imposed on the predictor functions.

The optimal transformation for each predictor is the conditional expectation

$$f_i(x_i) = \mathrm{E}\{y - \alpha - \sum_{j \neq i} f_j(x_j)|x_i\}. \tag{4.2}$$

Notice that the target in the conditional expectation expression is a residual, $\delta_i = y - (\alpha + \sum_{j \neq i} f_j(x_j))$. Thus, for each predictor x_i we try to estimate a transformation to explain the part of variance in the response variable not explained by the other predictor functions $f_j(x_j), j \neq i$.

Mixed models can be constructed by including linear and other parametric forms in combination with nonlinear terms. When both non-parametric and parametric terms are included the model is often referred to as a semi-parametric GAM. For continuous predictors the transforms may include e.g. splines, kernel methods, polynomial regression, and nonlinear non-parametric estimators. For categorical variables the conditional expectation estimates are straightforward

$$\hat{\mathrm{E}}\{\delta_i|x_i = z\} = \sum_{x_i = z} \delta_i / \sum_{x_i = z} 1. \tag{4.3}$$

The general GAM has the form (assuming that the responce has been transformed by its canonical link function)

$$y = \alpha + \sum_{j=1}^{m} f_j(x_j) + \epsilon \tag{4.4}$$

where ϵ is a residual error with mean zero. The constant α is not identifiable, since we can add or subtract an arbitrary constant to each of the functions $f_i(x_i)$, and adjust α accordingly. Therefore, the standard convention is to assume that $\mathrm{E}\{f_i(x_i)\} = 0, i = 1, \cdots, m$ and to choose $\alpha = \mathrm{E}\{y\}$. An iterative procedure exists for finding the solution. The appropriate conditional expectation is calculated for each target, δ_i, generating a new estimate for the ith transformation $f_i(x_i)$. The process is continued until the estimates stabilize. The procedure is known as "backfitting." The procedure is given in Algorithm 7.

Algorithm 7 The Backfitting Algorithm

1: Set $\alpha = \mathrm{E}\{y\}$ and $f_i(x_i) = 0 \ \ \forall\ i$
2: **repeat** $\{\forall\ i\}$
3: $\quad f_i(x_i) = \mathrm{E}\{y - \alpha - \sum_{j \neq i} f_j(x_j) | x_i\}$
4: $\quad f_i(x_i) = f_i(x_i) - \mathrm{E}\{f_i(x_i)\}$
5: **until** Convergence

The second step (step 4) in the loop of the backfitting algorithm is included to avoid problems due to machine round off which may cause slippage.

The backfitting algorithm is known under several other names including iterative residual fitting and the restricted alternating conditional expectations algorithm. The latter was introduced by Breiman and Friedman some time before GAMs and is the inner loop of their alternating conditional expectations algorithm which will be described later. Analytic proofs of convergence of GAMs is conditioned on which scatterplot smoothers are applied. Known proofs include e.g. cubic splines, and nearest neighbours smoothers, [13, 54].

4.2.2 Projection Pursuit Regression

Projection Pursuit Regression (PPR), [39], has the form

$$y = \alpha + \sum_{j=1}^{m} f_j(\boldsymbol{w}_j^T \boldsymbol{x}) + \epsilon. \tag{4.5}$$

By applying linear combinations of the input variables, PPR may compensate for the problems GAMs can have in higher-dimensions, and when interaction effects are present between the input variables. The weight vectors $\boldsymbol{w}_j$, α and the dimension m are chosen by the user or may be the tuning parameters in a model building setting where we are looking for "good" projections of the data. The f-functions can be chosen as conditional expectations using e.g. nonlinear scatterplot smoothers and handled in a similar manner as in GAMs.

4.2.3 Artificial Neural Networks

Artificial Neural Networks (ANN), [130, 107], can be considered a special kind of repeated and coupled layers of PPR schemes in which the f-functions are imposed certain restrictions. If $\mathrm{E}\{y\} = 0$, an ANN with one hidden layer can be expressed in the form

$$y = f_2(\boldsymbol{b}^T \boldsymbol{h}) + \epsilon \tag{4.6}$$

where

$$\boldsymbol{h} = \begin{bmatrix} f_{11}(\boldsymbol{a}_1^T \boldsymbol{x}) \\ \vdots \\ f_{1m}(\boldsymbol{a}_m^T \boldsymbol{x}) \end{bmatrix} \tag{4.7}$$

with $\boldsymbol{x} \in \mathbb{R}^n$, $\boldsymbol{a}_i \in \mathbb{R}^n, i = 1, \cdots, m$, and $\boldsymbol{b} \in \mathbb{R}^m$. In ANN the f-functions are often called activation functions and are used as "squashing" functions to keep the response bounded. A common choice is the logistic function. With identical activation functions, the expression for the output in Equation 4.6 can be written

$$y = f(\sum_{j=1}^{m} b_j f(\sum_{k=1}^{n} a_{jk} x_k)) + \epsilon. \tag{4.8}$$

Naturally, ANN can have more than one layer, and the "connections" in the net may even skip layers. Training of ANN consists of determining the network weights and possibly the parameters for the activation functions.

4.2.4 The Alternating Conditional Expectations Algorithm

If the task is regression and we want to predict y, it may be better to transform y as well. This is exactly what is proposed in the Alternating Conditional Expectations (ACE) algorithm using the model

$$f_0(y) = \sum_{j=1}^{m} f_j(x_j) + \epsilon. \tag{4.9}$$

The ACE algorithm aims at minimizing the fraction of variance not explained by a regression of $f_0(y)$ on $\sum_{j=1}^{m} f_j(x_j)$. ACE thus minimizes

$$e^2(f_0, f_1, \cdots, f_m) = \frac{\mathrm{E}\{[f_0(y) - \sum_{j=1}^{m} f_j(x_j)]^2\}}{\mathrm{E}\{f_0(y)^2\}}. \tag{4.10}$$

Without loss of generality, let $\mathrm{E}\{f_0(y)^2\} = \|\mathrm{E}\{f_0(y)\}\|^2 = 1$ and assume that all functions have expectation zero. ACE then amounts to minimizing

$$e^2(f_0, f_1, \cdots, f_m) = \mathrm{E}\{[f_0(y) - \sum_{j=1}^{m} f_j(x_j)]^2\}. \tag{4.11}$$

It can be shown that optimal transformations exist, and that they satisfy a complex system of integral equations, [13]. The name ACE arises from the fact that an iterative algorithm can be applied to find estimates of the optimal transformations using only bivariate conditional expectations.

Univariate response setting

To illustrate we first look at the bivariate case. To minimize

$$e^2(f_0(y), f_1(x)) = \mathrm{E}\{[f_0(y) - f_1(x)]^2\} \tag{4.12}$$

the *basic* ACE algorithm iterates until convergence:

$$f_1(x) = \mathrm{E}\{f_0(y)|x\} \text{ and} \tag{4.13}$$
$$f_0(y) = \mathrm{E}\{f_1(x)|y\}/\|\mathrm{E}\{f_1(x)|y\}\|. \tag{4.14}$$

The first step minimizes e^2 with respect to $f_1(x)$ for a given transformation $f_0(y)$. The second step minimizes e^2 with respect to $f_0(y)$ for a given transformation $f_1(x)$ and preserves $\mathrm{E}\{f_0(y)^2\} = 1$. This basic ACE algorithm for the bivariate case thus decreases e^2 in Equation 4.12 at each step by alternatingly minimizing with respect to one function and holding the other fixed at its previous evaluation. There are various ways of initializing the procedure. One possible initialization is with $f_0(y) = y/\|\mathrm{E}\{y\}\|$ and $f_1(x) = 0$. The necessary convergence critia are examined in [13].

In direct analogue to the basic ACE algorithm the method can be expanded to include several predictors such that Equation 4.10 is minimized. First say we have a given set of transforms for the predictors, then the optimal transformation for the response will be

$$f_0(y) = \mathrm{E}\{\sum_{j=1}^{m} f_j(x_j)|y\}/\|\mathrm{E}\{\sum_{j=1}^{m} f_j(x_j)|y\}\|. \tag{4.15}$$

Similarly we find that for a given set of transforms $f_0, f_1, f_2, f_3 \cdots, f_{i-1}, f_{i+1}, \cdots, f_m$, the optimal function for transforming the x_i variable is

$$f_i(x_i) = \mathrm{E}\{f_0(y) - \sum_{j \neq i} f_j(x_j)|x_i\}. \tag{4.16}$$

Now that there are several predictors we need to repeat the estimation of the individual transforms of the x_i's. This step is called the *restricted* ACE algorithm and corresponds to the backfitting algorithm. The *full* ACE algorithm thus includes a repeated loop over all the predictors applying Equation 4.16 until convergence.

Multivariate response setting

The ACE method can be further generalized to handle multiple response variables as indicated in [12, 128]. Extending the algorithm to handle such scenarios makes it useful for two-set nonlinear canonical correlations

analysis. Say we have m_x variables and m_y variables in the two sets. The *generalized* ACE algorithm will maximize the correlation between the sums of the transformations within each set by estimating the transformations $\{f_{xj}\}_{j=1}^{m_x}, f_{xj} = f_{xj}(x_j)$ and $\{f_{yi}\}_{i=1}^{m_y}, f_{yi} = f_{yi}(y_i)$ that minimize

$$\mathrm{E}\{[\sum_{i=1}^{m_y} f_{yi}(y_i) - \sum_{j=1}^{m_x} f_{xj}(x_j)]^2\} \tag{4.17}$$

under variance preserving constraints for the sum of the transformations of the variables in either the x or the y set. In fact, constraints on the variance could be introduced on both sets, but this is not really necessary in order to maximize correlation. The constraint prevents the algorithm from converging to null transformations.

In analogue to the extension of the basic ACE algorithm the restricted ACE is modified to handle multivariate response scenarios by introducing an additional inner backfitting loop over the y-set. It thus applies

$$f_{yi}(y_i) = \mathrm{E}\{\sum_{l=1}^{m_x} f_{xl}(x_l) - \sum_{j\neq i} f_{yj}(y_j)|y_i\} \tag{4.18}$$

while iterating over all the variables in the y-set until convergence, similar to backfitting in GAMs. Equation 4.16 is generalized to

$$f_{xi}(x_i) = \mathrm{E}\{\sum_{l=1}^{m_y} f_{yl}(y_l) - \sum_{j\neq i} f_{xj}(x_j)|x_i\}. \tag{4.19}$$

It should be stressed that ACE in this form, including several response variables, is more related to canonical correlations analysis than to regression problems.

In Algorithm 8 the generalized ACE algorithm is presented. It includes the extensions of finding suboptimal solutions to the maximum correlation problem. That is solutions that posses maximum correlations subject to the constraint that they must be orthogonal to the previously determined solutions. First the transformations are initialized. After initialization the algorithm finds N solutions by iterating the outer loop and the inner loops of the ACE algorithm until convergence. By applying multiple regression and subtracting previously determined solutions, the algorithm is able to find suboptimal solutions that maximize correlation. This approach was

Algorithm 8 The Generalized Alternating Conditional Expectations Algorithm

initialize $f_{yi}(y_i) = y_i/\|\mathrm{E}\{\sum_{j=1}^{m_y} y_j\}\|,\ i = 1,\cdots,m_y$ and $f_{xj}(x_j) = 0,\ j = 1,\cdots,m_x$.
for $n = 1$ to N **do**
  repeat {Outer loop}
    repeat {1st inner loop}
      Set $f_{xi}(x_i) = \mathrm{E}\{\sum_{l=1}^{m_y} f_{yl}(y_l) - \sum_{j\neq i} f_{xj}(x_j)|x_i\},\ i = 1,\cdots,m_x$
    until Convergence
    repeat {2nd inner loop}
      Set $f_{yi}(y_i) = \mathrm{E}\{\sum_{l=1}^{m_x} f_{xl}(x_l) - \sum_{j\neq i} f_{yj}(y_j)|y_i\},\ i = 1,\cdots,m_y$
    until Convergence
    for $k = 1$ to $n-1$ **do**
      Solve for weights, $\{\beta_j\}_{j=1}^{m_y}$, in multiple regression of $\sum_{i=1}^{m_y} f_{yi}(y_i)$ on $\{f_{kij}\}_{j=1}^{m_y}, = \sum_{j=1}^{m_y} \beta_j g_{kj}(y_j)$
      Subtract previous solution by $f_{yi}(y_i) = f_{yi}(y_i) - \beta_i g_{ki}(y_i),\ i = 1,\cdots,m_y$
    end for
    normalization $f_{yi}(y_i) = f_{yi}(y_i)/\|\mathrm{E}\{\sum_{j=1}^{m_y} f_{yj}(y_j)\}\|,\ i = 1,\cdots,m_y$
  until Convergence
  Store $g_{ni}(y_i) = f_{yi}(y_i),\ i = 1,\cdots,m_y$ and $h_{ni}(x_i) = f_{xi}(x_i),\ i = 1,\cdots,m_x$
end for

first suggested by [17] for handling univariate response settings and by [128] for handling multivariate response settings. The latter is adopted here. The algorithm stores the resulting transformations in the functions $g_{ni}(y_i)$ and $h_{ni}(x_i)$, and on exit contains the components $G_n = \sum_{i=1}^{m_y} g_{ni}$ and $H_n = \sum_{i=1}^{m_x} h_{ni}$ for $n = 1,\cdots,N$ the canonical variates for the nth solution that maximizes correlation and satisfies the following equations

$$\mathrm{Corr}\{G_i, G_j\} = \delta_{ij}, \quad \forall \quad i,j = 1,\cdots,N \tag{4.20}$$
$$\mathrm{Corr}\{H_i, H_j\} = \delta_{ij}, \quad \forall \quad i,j = 1,\cdots,N \tag{4.21}$$
$$\mathrm{Corr}\{G_i, H_j\} = \rho_j\delta_{ij}, \quad \forall \quad i,j = 1,\cdots,N \tag{4.22}$$
$$\tag{4.23}$$

where δ_{ij} is the Kronecker delta.

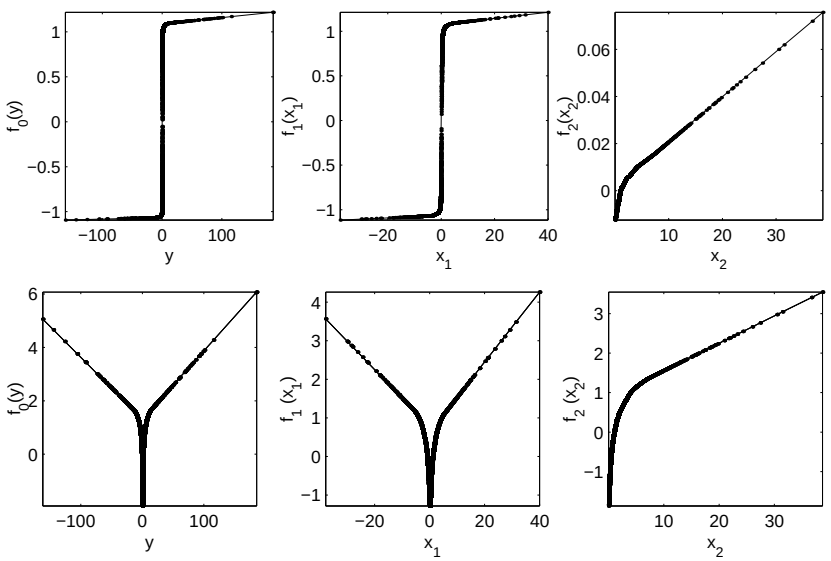


Figure 4.1: Eigensolutions 1 and 2 (top/bottom row) found by ACE.

A case study

To illustrate the ACE algorithm we here reproduce a case study first given in [14]. Consider the problem

$$y = x_1 x_2 \exp(a\epsilon), \tag{4.24}$$

where $a > 0$ is a constant and $\epsilon \in N(0,1)$ and $\|x_1\|, x_2 \in LN(0,1)$. The full ACE algorithm is applied to a data set consisting of 1000 samples generated using $a = 0.5$. The results of the first two eigensolutions are shown in Figure 4.1.

Inspecting the first eigensolution we notice that ACE first focuses on the fact that the sign of variable y is equal to the sign of variable x_1. The correlations for this solution, i.e. $\text{sign}(y) = \text{sign}(x_1) + 0 \cdot x_2$, is 1 and for the solution numerically determined by ACE we obtain a correlation of 0.99. The second solution found by ACE indicates the relation $\log(\|y\|) = \log(\|x_1\|) + \log(x_2)$ with the transforms $f_0(y), f_1(x_1)$, and $f_2(x_2)$ having normal distributions (not shown). Since noise is present, the expected

correlation for the second solutions is expected to be less than one, and ACE finds a correlation of 0.94. The two first solutions found by ACE pin down the structure of the data quite well. They are both informative since they give the sign and the absolute value of y respectively. Note that ACE is able to find additional suboptimal solutions (not shown) to the problem, but they appear to be trivial solutions that do not contribute with any additional information on the structure of the data.

Characteristics of ACE

ACE can handle arbitrary mixtures of continuous and categorical variables. The method is data-driven and non-parametric with minimal assumptions concerning the data distribution and the form of the optimal transformations. Finding the best fitting additive model can often aid in the interpretation and understanding of the relationship between the variables in the data sets. An alternative to ACE is AVAS (additivity and variance stabilizing transformation), [116]. In AVAS the transformations of the response set are restricted to monotone transformations, and the method aims to achieve constant variance of the residuals. The additional restrictions can make AVAS appear more reliable than ACE, [107]. For further comments on the ACE see the discussion of [13] in [103, 17, 33, 14], see also [16, 3].

4.2.5 The Multiset Alternating Conditional Expectations Algorithm

Let $\boldsymbol{x}$ be an $m = m_1 + m_2 + \cdots + m_n$ dimensional variable,

$$\boldsymbol{x} = \begin{bmatrix} \boldsymbol{x}_1 \\ \vdots \\ \boldsymbol{x}_n \end{bmatrix} \tag{4.25}$$

containing the n sets of variables

$$\boldsymbol{x}_1 = \begin{bmatrix} x_{11} \\ \vdots \\ x_{1m_1} \end{bmatrix}, \quad \boldsymbol{x}_2 = \begin{bmatrix} x_{21} \\ \vdots \\ x_{2m_2} \end{bmatrix}, \ldots, \boldsymbol{x}_n = \begin{bmatrix} x_{n1} \\ \vdots \\ x_{nm_n} \end{bmatrix}. \tag{4.26}$$

Define the set of mappings, $\varphi = \{\{\varphi_{ij}(x_{ij})\}_{j=1}^{m_i}\}_{i=1}^{n}$, that determines the transformations

$$\Psi_i = \sum_{j=1}^{m_i} \varphi_{ij}(x_{ij}), \quad i = 1, ..., n \tag{4.27}$$

constrained such that

$$\begin{aligned} \mathrm{E}\{\varphi_{ij}(x_{ij})\} &= 0 \quad \forall \ i, j & (4.28)\\ \mathrm{E}\{\Psi_i^2\} &= 1 \quad \forall \ i. & (4.29) \end{aligned}$$

The mappings that minimize

$$e^2 = \sum_{i=1}^{n}\sum_{j=1}^{n} \mathrm{E}\{[\Psi_i - \Psi_j]^2\} = \sum_{i=1}^{n}\sum_{j=1}^{n} 2(1 - \rho_{ij}) \tag{4.30}$$

maximize the sum of the pair-wise correlations over all combinations.

Let the transformations have the following properties

$$\varphi_{ij}(x_{ij}) : \varphi_{ij} \text{ is measurable, } \mathrm{E}\{\varphi_{ij}(x_{ij})\} = 0, \mathrm{E}\{\varphi_{ij}^2(x_{ij})\} < \infty \tag{4.31}$$

and define the following Hilbert space

$$\mathcal{H} = \mathrm{span}\{\varphi_{11}, .., \varphi_{1m_1}, \varphi_{21}, .., \varphi_{2m_2}, .., \varphi_{nm_n}\}, \tag{4.32}$$

consisting of the following subspaces

$$\begin{aligned} \mathcal{H}_1 &= \mathrm{span}\{\varphi_{11}, .., \varphi_{1m_1}\} & (4.33)\\ \mathcal{H}_2 &= \mathrm{span}\{\varphi_{21}, .., \varphi_{2m_2}\} & (4.34)\\ &\vdots & (4.35)\\ \mathcal{H}_n &= \mathrm{span}\{\varphi_{n1}, .., \varphi_{nm_n}\} & (4.36) \end{aligned}$$

with inner product and norm

$$< g|f >= \mathrm{E}\{gf\}, \ \|f\|^2 = \mathrm{E}\{f^2\}, \quad f, g \in \mathcal{H}. \tag{4.37}$$

Looking for maximum correlation in the simple bivariate case, that is two sets with one variable in each, we wish to maximize

$$\rho_{12} = \frac{< \varphi_1|\varphi_2 >}{\sqrt{(}\|\varphi_1\|^2\|\varphi_2\|^2)} \tag{4.38}$$

under the constraints that $\|\varphi_1\|^2 = \|\varphi_2\|^2 = 1$. Constructing the Lagrangian we get

$$L(\varphi_1, \varphi_2) =< \varphi_1|\varphi_2 > -\frac{\lambda_1}{2}(< \varphi_1|\varphi_1 > -1) - \frac{\lambda_2}{2}(< \varphi_2|\varphi_2 > -1). \tag{4.39}$$

For a given φ_2 we want to find the optimal φ_1. Thus, we define

$$\begin{aligned} H_1(t) &= L(\varphi_1 + t\varphi_1^*, \varphi_2) & (4.40)\\ &= < \varphi_1 + t\varphi_1^*|\varphi_2 > -\frac{\lambda_1}{2}(< \varphi_1 + t\varphi_1^*|\varphi_1 + t\varphi_1^* > -1)\\ &\quad -\frac{\lambda_2}{2}(< \varphi_2|\varphi_2 > -1) & (4.41) \end{aligned}$$

where $\varphi_1^* \in \mathcal{H}_1$ and $t \in \mathbb{R}$. For stationarity of $H_1(t)$ the directional derivatives with respect to t must vanish for all φ_1^*. Thus,

$$[dH_1(t)/dt]|_{t=0} = 0 \quad \forall \quad \varphi_1^* \tag{4.42}$$

such that

$$< \varphi_1^*|\varphi_2 > -\lambda_1 < \varphi_1^*|\varphi_1 >= 0 \quad \forall \quad \varphi_1^* \tag{4.43}$$

which can also be written as

$$< \varphi_2 - \lambda_1\varphi_1|\varphi_1^* >= 0 \quad \forall \quad \varphi_1^*. \tag{4.44}$$

Similarly for a fixed φ_1 we want to find the optimal φ_2 and define $H_2(t) = L(\varphi_1, \varphi_2 + t\varphi_2^*)$ with $\varphi_2^* \in \mathcal{H}_2$ and $t \in \mathbb{R}$. Evaluating

$$[dH_2(t)/dt]|_{t=0} = 0 \quad \forall \quad \varphi_2^* \tag{4.45}$$

we find

$$< \varphi_1|\varphi_2^* > -\lambda_2 < \varphi_2|\varphi_2^* >= 0 \quad \forall \quad \varphi_2^* \tag{4.46}$$

or

$$< \varphi_1 - \lambda_2\varphi_2|\varphi_2^* >= 0 \quad \forall \quad \varphi_2^*. \tag{4.47}$$

Using the constraints of unit variance and choosing $\varphi_1^* = \varphi_1$ and $\varphi_2^* = \varphi_2$ we find from Equations 4.43 and 4.46 that the Lagrange multipliers are

$$\lambda_1 = \lambda_2 = \rho. \tag{4.48}$$

Equations 4.44 and 4.47 can now be written as

$$< \varphi_2 - \rho\varphi_1 | \varphi_1^* >= 0 \quad \forall \quad \varphi_1^* \tag{4.49}$$
$$< \varphi_1 - \rho\varphi_2 | \varphi_2^* >= 0 \quad \forall \quad \varphi_2^*. \tag{4.50}$$

The solution with maximal correlation is when $\varphi_2 = \rho\varphi_1$ in $\mathcal{H}_1$, that is when $\rho\varphi_1$ is the projection of φ_2 onto $\mathcal{H}_1$. Similarly, we get that $\rho\varphi_2$ must be the projection of φ_1 onto $\mathcal{H}_2$. Introducing the projection operators P_1, P_2 the Equations 4.49 and 4.50 become

$$P_1\varphi_2 = \rho\varphi_1 \tag{4.51}$$
$$P_2\varphi_1 = \rho\varphi_2 \tag{4.52}$$

which may be decoupled into the eigenvalue problems

$$P_1P_2\varphi_1 = \rho^2\varphi_1 \tag{4.53}$$
$$P_2P_1\varphi_2 = \rho^2\varphi_2. \tag{4.54}$$

When the combined operators $U = P_1P_2$ and $V = P_2P_1$ are compact, self-adjoint and non-negative definite, they have the same eigenvalues and eigenspaces. For an elaborate analysis of finding optimal transformations in the function space see [13, 3], in which the convergence conditions are analysed as well. See [16] for remarks on functional canonical variates and on the relation between the alternating least squares methods and ACE. The analysis presented here was inspired by [128] who presents an analysis of the maximization of the squared correlation in the Hilbert space.

In ACE the projection operators are the conditional expectations

$$P_1 = \mathrm{E}\{\cdot|x_1\} \tag{4.55}$$
$$P_2 = \mathrm{E}\{\cdot|x_2\} \tag{4.56}$$

and the alternating algorithm is essentially the power method applied to either U or V. Thus the successive application of the projection operators followed by normalization of the transforms on the data will minimize the variance of the residuals i.e. $\mathrm{E}\{(\varphi_1(x_1) - \varphi_2(x_2))^2\}$.

For completeness we present the projection operators in the linear case for multivariate two-set cases:

$$P_1 = \mathbf{\Sigma}_{11}^{-1}\mathbf{\Sigma}_{12} \tag{4.57}$$
$$P_2 = \mathbf{\Sigma}_{22}^{-1}\mathbf{\Sigma}_{21}. \tag{4.58}$$

In two-set ACE problems with more than one variable in each set, we are aiming to maximize

$$\rho_{12} = \mathrm{Corr}\{\sum_{j=1}^{m_1}\varphi_{1j,}\sum_{j=1}^{m_2}\varphi_{2j}\}, \tag{4.59}$$

and the backfitting algorithm can be applied to find each individual transformation φ_{ij} (see the previous section). With n sets we are aiming to maximize

$$R = \sum_{kl}\rho_{kl} = \sum_{kl}\mathrm{Corr}\{\sum_{i=1}^{m_k}\varphi_{ki,}\sum_{j=1}^{m_l}\varphi_{lj}\}, \tag{4.60}$$

and applying the backfitting algorithm the projection operator for e.g. the ijth variable we propose to operate on the residual

$$\delta\varphi_{ij} = \sum_{k\neq i}\sum_{l=1}^{m_k}\varphi_{kl} - \sum_{m\neq j}^{m_i}\varphi_{im}. \tag{4.61}$$

Notice that the residual is a sum of functions in the closed subspaces of the Hilbert space and thus $\delta\varphi_{ij} \in \mathcal{H}$.

The residual can be written as

$$\delta\varphi_{ij} = \sum_{k\neq i}\Psi_k - \sum_{m\neq j}^{m_i}\varphi_{im}. \tag{4.62}$$

Thus, in the multiset case, when applying the backfitting algorithm to find the optimal transformations for ith set, we can define a reference set $\Psi_{0\backslash i}$ and consider it as the response in a regression setting with the variables in the ith set being the predictors. We define

$$\Psi_{0\backslash i} = \sum_{k\neq i}\Psi_k/\|\sum_{k\neq i}\Psi_k\| \tag{4.63}$$
$$= \sum_{k\neq i}\sum_{j=1}^{m_k}\varphi_{kj}/\|\sum_{k\neq i}\sum_{j=1}^{m_k}\varphi_{kj}\|. \tag{4.64}$$

In practice when looking for optimal transformations that maximize the sum of the pair-wise correlations, the normalization of the reference set

is not really necessary since after application of the backfitting algorithm we will constrain the transformations to preserve condition 4.29 such that $\mathrm{Var}\{\Psi_i\} = 1$. The normalization is thus only motivated for speeding up the convergence of the backfitting algorithm by preserving and restricting the variance.

Minimizing e^2 from Equation 4.30 under $\mathrm{Var}\{\Psi_i\} = 1$ and $\mathrm{E}\{\varphi_{ij}\} = 0$ corresponds to minimizing

$$\tilde{e}^2 = \sum_{i=1}^{n} \mathrm{E}\{(\Psi_{0\setminus i} - \Psi_i)^2\} \tag{4.65}$$

$$= \sum_{i=1}^{n} [\mathrm{Var}\{\Psi_{0\setminus i}\} + 1 - 2\mathrm{Cov}\{\Psi_{0\setminus i}, \Psi_i\}]. \tag{4.66}$$

When writing

$$\mathrm{Cov}\{\Psi_{0\setminus i}, \Psi_i\} = \sum_{j\neq i} \mathrm{Cov}\{\Psi_j, \Psi_i\} \tag{4.67}$$

$$= \sum_{j\neq i} \sqrt{(}\mathrm{Var}\{\Psi_j\})\mathrm{Corr}\{\Psi_j, \Psi_i\} \tag{4.68}$$

and constructing $\Psi_{0\setminus i}$ using $\Psi_j, j \neq i$ with equal variance σ^2, we obtain

$$\tilde{e}^2 = \sum_{i=1}^{n} [\mathrm{Var}\{\Psi_{0\setminus i}\} + 1 - 2\sigma \sum_{j\neq i} \mathrm{Corr}\{\Psi_j, \Psi_i\}], \tag{4.69}$$

and thus finding the optimal transformation for the ith set that minimize $\tilde{e}^2$ will maximize the correlation to $\Psi_{0\setminus i}$ and the sum of the pair-wise correlations over all the transformed sets.

The new algorithm is called the *multiset* ACE algorithm (MACE). The eigensolutions found by MACE are called canonical variates, and they have the following properties

$$< \Psi_{ki} | \Psi_{lj} >= \mathrm{E}\{\Psi_{ki}\Psi_{lj}\} = \rho_{ijk}\delta_{kl} \quad \forall \quad i, j, k, l \tag{4.70}$$

where $i, j = 1, .., n$ are set indexes, and k, l are indexes to the eigensolutions to 4.53. Thus Ψ_{ki} is the canonical variate for the ith set in the kth eigensolution, ρ_{ijk} is the canonical correlation between the ith and the jth set in the kth eigensolution, and δ is the Kronecker delta.

Minimizing correlation

Consider again the two-set case of finding optimal transformations that maximize the correlation, ρ. Since ρ^2 is the eigenvalue of the eigenproblem solved by ACE, the range is between zero and one. Knowing the range of the eigenvalues to the eigenproblem allows us to look for solutions with minimal correlation. Returning to the Power Algorithm presented in Algorithm 4, we apply the trick of shifting the eigenvalues. Instead of applying e.g. the operator $U = P_1P_2$, we apply $\tilde{U} = U - \alpha$, where $\alpha = 1$ is chosen as the largest possible eigenvalue. Iterating

$$\Psi^k = \tilde{U}\Psi^{k-1} = P_1P_2\Psi^{k-1} - \Psi^{k-1} \tag{4.71}$$

and normalizing the transforms at each step will produce the eigensolution with the eigenvalue closest to zero thus maximizing $\delta\lambda = |\lambda_j - 1|$.

In [128] the same operator for estimating the optimal transformations with minimal correlation is proposed. In this case the operator is found by trying to transform set 1 to maximize the correlation to a residual consisting of the difference of the transformed set 1 and the tranformed set 2 projected on $\mathcal{H}_1$.

When minimizing correlation one is actually looking for the noise in the data. This raises the question of how meaningful is minimum correlation? It can be interesting to look for minimal similarity, but for it to be meaningful, one must probably introduce a relatively high degree of regularizing restrictions on the applied scatterplot smoothers.

Looking for eigensolutions close to a specific expected eigenvalue

In theory a variant of the Inverse Power Method could be applied when looking for an eigensolution close to a specific expected eigenvalue. However, it would require that we find the solution of Ψ^k at the kth iteration to

$$\tilde{U}\Psi^k = P_1P_2\Psi^k - \alpha\Psi^k = \Psi^{k-1}, \tag{4.72}$$

which we currently do not know how to obtain.

Algorithm 9 The Multiset Alternating Conditional Expectations Algorithm

1: **initialize** $\varphi_{ij}(x_{ij}) = x_{ij}/\|\mathrm{E}\{\sum_{j=1}^{m_i} x_{ij}\}\|,\ i = 1, \cdots, n;\ j = 1, \cdots, m_i$
2: Set $\Psi_i = \sum_{j=1}^{m_i} \varphi_{ij}(x_{ij}),\ i = 1, ..., n$
3: **for** $m = 1$ to M **do**
4: **repeat** {Outer loop}
5: **for** $i = 1$ to n **do**
6: Set $\Psi_{0\backslash i} = \sum_{k \neq i} \Psi_k / \| \sum_{k \neq i} \Psi_k \|$
7: **repeat** {Inner loop}
8: Set $\varphi_{ij}(x_{ij}) = \mathrm{E}\{\Psi_{0\backslash i} - \sum_{k \neq j} \varphi_{ik}(x_{ik}) | x_{ik}\},\ j = 1, \cdots, m_x$
9: **until** Convergence
10: **for** $k = 1$ to $m - 1$ **do**
11: Solve for weights, $\{\beta_j\}_{j=1}^{m_i}$, in multiple regression of Ψ_i on $\{f_{kij}\}_{j=1}^{m_i}$, $= \sum_{j=1}^{m_i} \beta_j f_{kij}(x_{ij})$
12: Subtract previous solution by $\varphi_{ij}(x_{ij}) = \varphi_{ij}(x_{ij}) - \beta_j f_{kij}(x_{ij}),\ j = 1, \cdots, m_i$
13: **end for**
14: **normalization** $\varphi_{ij}(x_{ij}) = \varphi_{ij}(x_{ij}) / \|\mathrm{E}\{\sum_{j=1}^{m_i} \varphi_{ij}(x_{ij})\}\|,\ j = 1, \cdots, m_i$
15: Update $\Psi_i = \sum_{j=1}^{m_i} \varphi_{ij}(x_{ij})$
16: **end for**
17: **until** Convergence
18: Store $f_{mij}(x_{ij}) = \varphi_{ij}(x_{ij}),\ i = 1, ..., n;\ j = 1, \cdots, m_i$
19: **end for**

Implementation of MACE

In algorithm 9 the Multiset Alternating Conditional Expectations algorithm is implemented. In step 1 the transformations are initialized. This is a default initialization and could possibly be performed otherwise. Notice that the initialization transforms constrain all the sets to unit variance. In step 2 we sum over the transforms in each set and store the result in the Ψ_i's. When solving for M eigensolutions, the algorithm iterates an outer loop (steps 4-17) inside which each set is processed by an inner loop. Before applying the inner loop, $\Psi_{0\backslash i}$ is constructed when handling the ith set. The inner loop (steps 7-9) consists of performing backfitting to find the optimal transformation to the variable in the ith set. In steps 10-13 orthogonality to previously determined eigensolutions is preserved. This

is done by performing multiple regression and substituting the individual transformations with the residual from the regression. The remaining part of the outer loop consists of normalization and updating of the ith set (steps 14-15). Upon convergence of the outer loop the mth eigensolution is stored in the components $f_{mij}(x_{ij})$.

On exit we may construct $\Psi_{ki} = \sum_{j=1}^{m_i} f_{kij},\ k = 1, \cdots, M;\ i = 1, \cdots, n$ which will be estimates of the canonical variates that maximize the sum of pair-wise correlations and satisfy Equation 4.70.

Extensions to MACE

One natural extension to MACE would be to include weights on the observations. The weights could be controlled in a similar fashion to the proposed generalized linear MCCA algorithm letting it depend on the variance of the transformed observations over all sets.

MACE can be applied to generate a type of nonlinear principal components. Consider the case of a n-dimensional multivariate single set and think of the data as an univariate multiset scenario. Principal components are generated by looking for the optimal transformation of each set or variable $\varphi_i(x_i), i = 1, \cdots, n$ such that the sum of the transformed sets have maximum variance. We are thus trying to find the set of transforms φ that produces

$$\max_{\varphi}[\mathrm{E}\{(\varphi_1(x_1) + \varphi_2(x_2) + \cdots + \varphi_n(x_n))^2\}]. \tag{4.73}$$

If the functions are restricted to the group of linear regression transformations i.e. $\varphi(x) = constant \cdot x$, the problem reduces to ordinary PC analysis. When looking for nonlinear principal components we will require that there is some dynamics in each transformation. We thus impose the MACE restriction $\mathrm{E}\{\varphi_i(x_i)^2\} = 1\ \forall\ i$ and maximizing the variance thus reduces to a problem of maximizing the sum of the pair-wise correlations over all sets

$$\max_{\varphi}[\mathrm{E}\{(\sum_{i}^{n} \varphi_i(x_i))^2\}] = \max_{\varphi}[\sum_{i}^{n}\sum_{j}^{n} \mathrm{Corr}\{\varphi_i(x_i), \varphi_j(x_j)\}]. \tag{4.74}$$

The nonlinear principal component with maximum variance is then found as the sum of the MACE CVs over all sets. Nonlinear principal component with minimum variance can also be generated. This is done by applying a

modified version of MACE in which the sign of the reference set is changed before entering the inner loop. Similarly to the linear PC analysis MACE will attempt to group the variables or sets together to cancle each other in order to obtain minimum variance. The analysis of nonlinear principal components with minimum variance may help a data analyst in determing which sets or variables naturally group together when appling nonlinear transformations.

We stress that the choice of how to calculate the conditional expectations is crucial in MACE. Friedman's supersmoother, see [37], provides the means of repeately estimating the required bivariate conditional expectations in an effective data driven manner. Without any a priori knowledge it is chosen as the default smoother since it is fast and can adapt to nonlinearities in the data. There are many other kinds of relevant smoothers which could be reasonable to include into the MACE algorithm. One important effect of applying unrestricted smoothers is that they often result in non-monotone transformations. The MACE transform then becomes non-invertible and hard to return to the original space of data representation making interpretation difficult. For monotone restrictions on scatterplot smoothers see e.g. [40]. See [15] for a comparative study of popular automatic smoother techniques i.e. smoothing splines, kernel smooths, running linear smoothers with adaptive window size, and regression splines. The smoothers presented here can all automatically adapt their internal parameters to the data. Thus they can all be applied in MACE preserving the data driven aspect of the algorithm.

4.3 Case Studies

Four case studies are presented. The generalized ACE algorithm is applied to maximize the correlation in a bivariate multitemporal data set. The data are the global ocean temperature and height of 1996 and 1997 previously analysed using linear CCA in Section 3.3.1. The second case study consists of applying MACE in a nonlinear maximum autocorrelation factors analysis of six band multispectral satellite image data. In the third case study nonlinear principal components that maximize and minimize variance are generated of four band satellite image data. The last case study is an analysis of irregularly sampled stream sediments geochemistry data.

4.3.1 An ACE-Based Nonlinear Extension to Traditional Empirical Orthogonal Function Analysis

The aim of this case study is to present an extension to the traditional empirical orthogonal functions (EOF) analysis [75, 104]. EOF is often applied in geophysical sciences to analyze temporal sequences. The EOF analysis is closely related to the principal components analysis (PCA) [64, 2]. Often the usual PCA assumption on variables with mean zero is replaced by an assumption of temporal means of zero. In the two-set case the traditional EOF analysis can be extended by the means of linear canonical correlations analysis (CCA) [65, 22, 93]. This study presents a nonlinear correlations analysis of two-set data using the alternating conditional expectations (ACE) transform. ACE was originally proposed for nonlinear multiple regression [13]. Traditionally ACE is used to handle data which naturally separates into two sets often with only one variable in one of the sets. The algorithm can, however, be generalized to handle several variables in both sets and even multiset scenarios. By allowing for nonlinear transformations, it is possible to determine CVs with higher correlations than in the linear CCA analysis. However, it becomes more challenging to interpret the CV sets, since the degree of freedom for the transformations is very high. The ACE transform is applied using Friedman's local nonlinear supersmoother [37] and applying the EOF analysis as a preprocessor. Two temporal sequences of scalar image data representing global sea surface temperature (SST) and sea surface height (SSH) are analysed, since these quantities are physically related. The data period covers 1996 and 1997 on a monthly basis. This particular period was chosen, because the build-up of one of the largest El Niño events [32] ever recorded occurred during the last half of 1997.

Data and Results

The data used are global monthly mean values of 1996-1997 SST data from the NOAA/NASA Oceans Pathfinder AVHRR SST database [69] and global monthly mean values of 1996-1997 SSH data from the NASA/GSFC Ocean Altimeter Pathfinder database [67, 69]. The SSH data are interpolated point observations from the TOPEX/Poseidon radar altimeter mission. The SST data comes as 360 rows by 720 columns half degree data starting at 180^o longitude, the SSH data comes as 179 rows by 360 columns

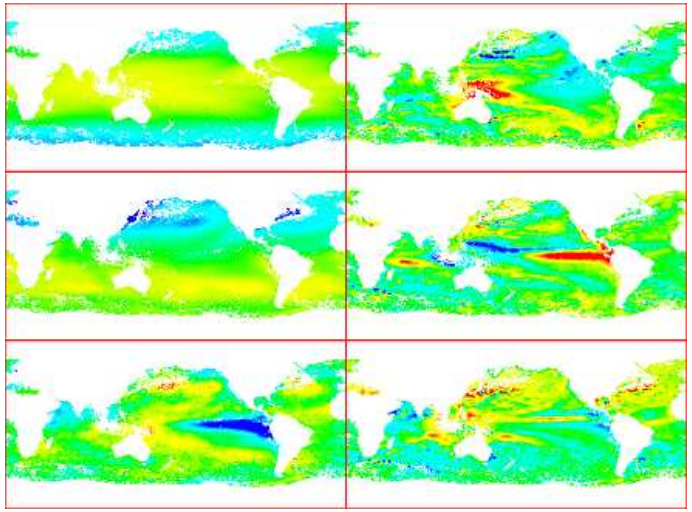


Figure 4.2: First column (top-down): The first three principal components of the SST data, the percentage of variance explained by each componet is 95%, 4% and 0.2%. Second column: The first three principal components of the SSH data. The percentage of variance explained by each component is 29%, 23%, and 7%. The images are stretched linearly from mean ± three standard deviations and shown in pseudocolour, blue is minimum and red is maximum.

one degree data starting at 0^o longitude. The SST data have been resampled to the SSH grid. The AVHRR instrument is influenced by cloud coverage, whereas the radar altimeter provides uninterrupted data. For a temporal analysis of the same data see [94]. Statistics for the SST analysis are calculated where SST and the SSH have non-missing values for all 24 months. In Figure 4.2 the first three principal components (PCs) are shown for both the SST and the SSH data. The first three SST-PCs explain approximately 99% of the total variation of the data. The first three SSH-PCs explain 59% of the variation in the data. By studying the principal components, one can obtain valuable information about the dynamics of the ocean. In Figure 4.3 is included the correlations between the original SST and the SSH variables and the PCs. The PC transformed SST and SSH data are used as input to ACE. The orthogonal transformation is helpful when we wish to evaluate the results of the nonlinear correla-

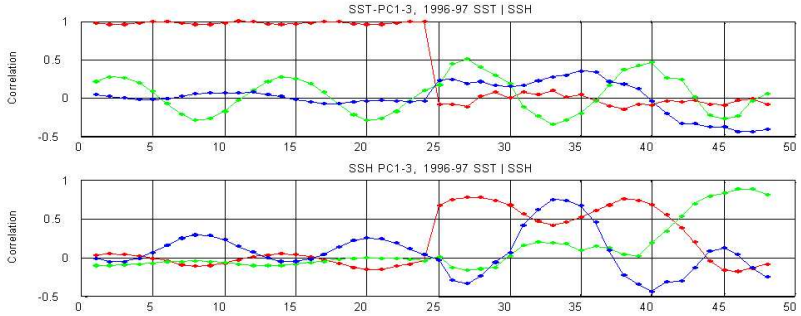


Figure 4.3: Top: The correlations between original SST (labeled 1-24) and SSH variables (labeled 25-48) and SST-PCs 1 through 3 (shown as red, green, and blue). Bottom: The correlations between original SST (labeled 1-24) and SSH variables (labeled 25-48) and SSH-PCs 1 through 3 (shown as red, green, and blue).

tions analysis. Inference problems are avoided, which can occur when the original data are cross-correlated. The ACE algorithm applied generates nonlinear mappings of each input variable. The sum of all the transformed variables in each set determines the first ACE CV pair. In Figure 4.4 the first two ACE CV pairs are shown. The correlation for the first pair is 0.88 and for the second pair 0.63. The two highest correlations found using a linear approach are 0.76 and 0.72 [93]. In the figure is also shown the squared differences for each CV pair. The ACE analysis is performed on a complete set using all PCs for both SST and SSH. In Figure 4.5, for the first ACE pair, is shown the first three ACE transformations of the PCs together with a bar plot of the variances of all the ACE transforms. The bar plot may be useful when determining which transforms dominate each ACE CV.

Conclusion

Inspecting the correlations between the SST-PCs and the sea surface temperature in Figure 4.3 reveals that the annual cycle is very dominant. SST-PC1 represents a temporal mean signal of the SST and has high positive semi-annually oscillating correlation with all of the SST months. The com-

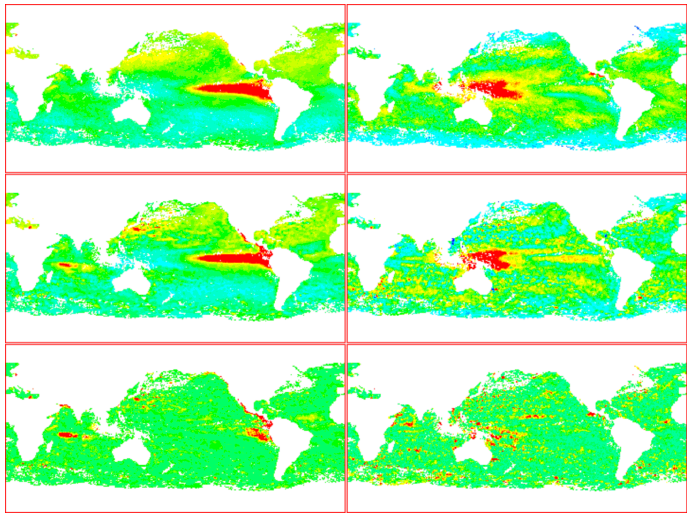


Figure 4.4: The columns contain the first respectively the second ACE CV pairs and their squared differences. Each column (top-down): The ACE CV of the SST-PCs, the ACE CV of the SSH-PCs, and the squared differences. The images are stretched linearly from mean ± three standard deviations and shown in pseudocolour, blue is minimum and red is maximum.

ponent represents the fact that it is warm at the Equator and colder near the poles. There is no apparent correlation with the SSH field because the mean of this field has been removed prior to the analysis. In SST-PC2 we see highs in the Southern Hemisphere and lows in the Northern Hemisphere. In the correlations the annual cycle is clearly present with positive correlations in the months from December to May and negative correlations from June to November. In SST-PC3 a strong signal off the Equatorial west coast of South America is present. The component does contain some annual signal but has correlations very close to zero to the SST data. Near the end of 1997, where the El Niño is starting to build up, we see a slight divergence from the zero into negative correlations. The negative correlations indicate that during the El Niño the temperature of the west coast of South America is high. Focusing on the El Niño event we see that especially the signal in SSH-PC2 is related to the phenomenon with high correlations to SSH in the last half of 1997. SSH-PC1 appears

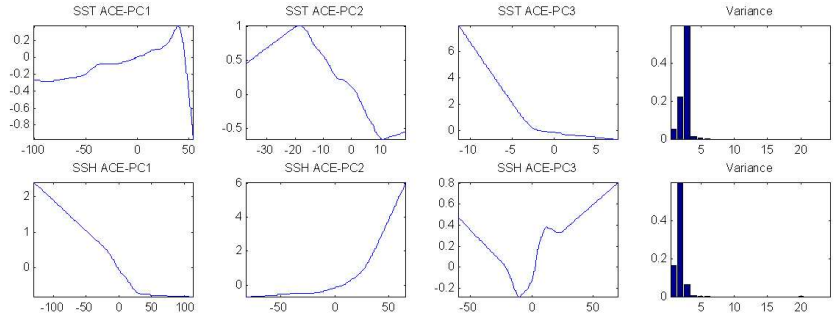


Figure 4.5: Top row: The first three ACE transformations of the SST-PCs - transformed value vs. input value. The last column contains a bar plot of the variances of all the ACE transformations for the SST PC1-24. Bottom row: The first three ACE transformations of the SSH-PCs - transformed value vs. input value, and a bar plot of the variances of the ACE transformations for the SSH PC1-24.

to be the mean height signal and is highly correlated with the SSH except during the El Niño where it has negative correlations closer to zero. SSH-PC3 seems related to the same annual cycle as found for SST-PC2.

In Figure 4.4, for the first ACE CV pair, we see that the variance of the first two to three ACE-PCs dominate the SST and the SSH data. The remaining ACE transforms are all very close to zero. Inspecting SST-PC3 and SSH-PC2 in Figure 4.2, we would expect ACE to change the sign of one of the components when looking for maximum correlation. This is exactly what is happening with SST ACE-PC3 being transformed using a monotonely decreasing function and transforming SSH ACE-PC2 with a function that is monotonely increasing. Thus the ACE CVs focus on the El Niño event as part of the signal containing the highest correlation between SST and SSH. The nonlinear transformations of the SST-PC1 and 2 and the SSH-PC1 and 3 appear to be included in the ACE CVs to explain additional spatial patterns present primarily within the Pacific Ocean. In the CVs in Figure 4.4 the pattern can be seen as a low signal in the western and southern parts of the Pacific Ocean. Inspecting the ACE CVs for the first ACE pair, the El Niño appears very strong. Highs in the squared difference image show where the global model is less successful in correlating sea surface temperature to the sea surface height. This is particularly conspicuous near the west coast

of Central and South America. There are also contributing regions in the Atlantic and Indian Oceans. Looking at the second ACE CV pair we notice an interesting signal in the western part of the Pacific Ocean. The signal is recognized as being related to the usual ocean temperature and height configuration. A full interpretation of the lower order ACE pairs is beyond the scope of this study. The ACE analysis seems to be able to separate relevant ocean configurations from the temporal data. It looks for high correlations between the involved variables through nonlinear mappings and finds components with higher correlations in comparison to those found by a linear analysis. Furthermore, the ACE analysis is purely data-driven and thus constitutes an useful exploratory tool for a data analyst when looking for insight into the structure of data. The analysis presented in this study is in good agreement with the established oceanographic knowledge on the build-up of one of the largest El Niño events on record.

Acknowledgements

The Pathfinder SST data provision at JPL is due to J. Vazques, R. Sumagaysay and co-workers. The Pathfinder SSH data provision at GSFC is due to V. Zlotnicki and co-workers. This work is done as a part of the GEOSONAR project funded by the Danish National Research Councils under the Earth Observation Program. GEOSONAR is headed by Dr. Per Knudsen at the National Survey and Cadastre, Denmark.

4.3.2 Nonlinear Maximum Autocorrelation Factors Analysis of TM Data, Ymer Ø.

A Landsat-5 Thematic Mapper (TM) scene has an instantaneous field of view of 30 meters by 30 meters in bands 1 through 5 and band 7. See Table 4.1 for the spectral range the individual TM bands. Figure 4.6 presents a raw TM (220x220) image (bands 1-5, and 7 row-wise ordering). The image has not yet been processed with the forward transformation that removes geometric distortions. For the raw TM data we expect a difference in the row-wise vs. column-wise noise structure due to the row-wise method of data collection. Each variable in the set is standardized and saturated between ± three standard deviations before any further analysis is performed. The scene reveals a part of Ymer Ø in Greenland. The

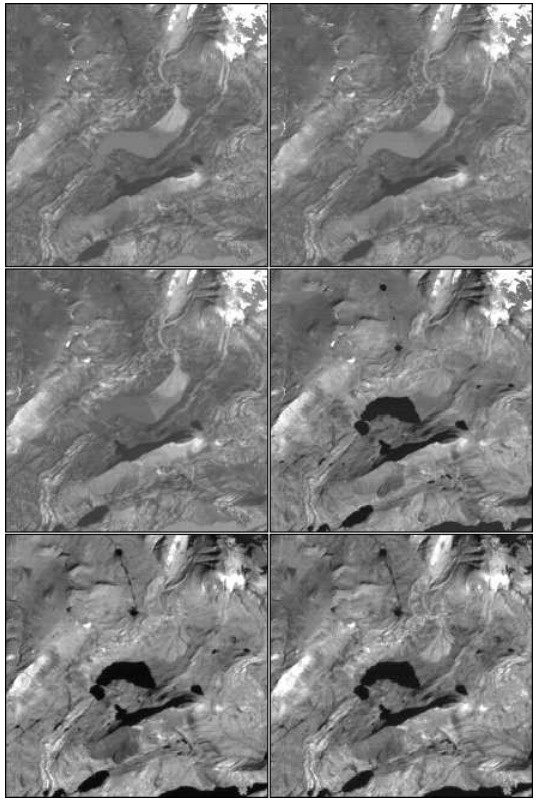


Figure 4.6: The raw TM data bands 1-5, and 7.

Danish Chief Geologist John L. Pedersen, Mountain Resources ApS, gives the following description of the TM scene: "The test ground at Ymer Ø comprises an area where the upper half of the sedimentary, Neoproterozoic Eleonore Bay Supergroup makes up the bedrock. The area is intersected by E-W trending faults, along which the northern block has moved down. The area is partly covered by a thin veneer of basal till, locally talus cones and fluviatile deposits, and also a few lakes. A meagre but persistent cover of vegetation comprising grass, flowers and low level scrubs is developed. The spectral signature for each pixel is thus an aggregate of contributions from various sources."

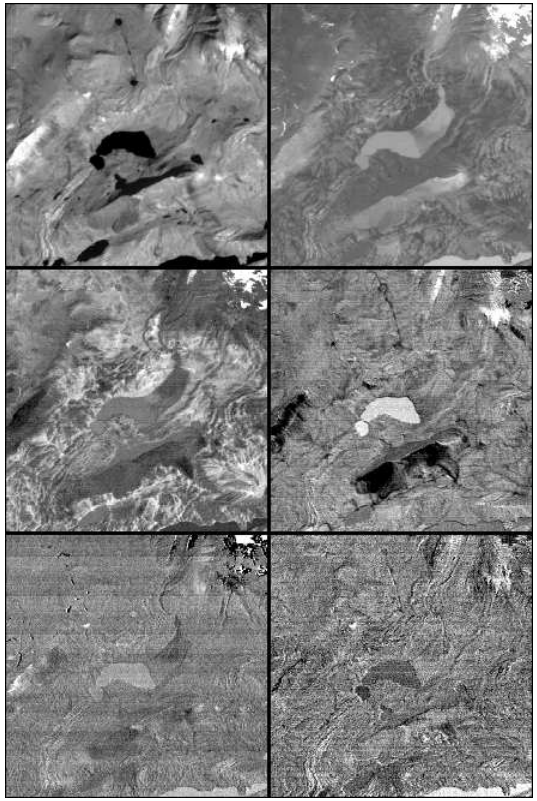


Figure 4.7: Linear MAF decomposition of the TM data.

In Figure 4.8 a geological map of Ymer Ø is presented. The colours indicate the following geological structures: blue-dotted regions are undifferentiated quaternary deposits, bright-purple regions are orange shale and sandstone and white dolomite, mid-purple regions are dark limestone and red and yellow dolomitic shale, dark-purple regions are red and yellow dolomitic shale, bright-red/brown regions are sandstone and siltstone, dark-red/brown regions are dark shale and siltstone, bright-blue regions are ice, mid-blue regions are white dolomite and dark limestone, dark-blue regions are dark limestone, bright-brown regions are light sandstone, and dark-brown are light sandstone and dark shale.

TM Band	Wavelengths [μm]	Spectral ref.
1	0.45-0.52	Blue
2	0.52-0.60	Green
3	0.63-0.69	Red
4	0.76-0.90	Near-Infrared
5	1.55-1.75	Near-Infrared
7	2.08-2.35	Mid-Infrared

Table 4.1: The spectral range of the Thematic Mapper sensor bands 1 through 5, and 7.

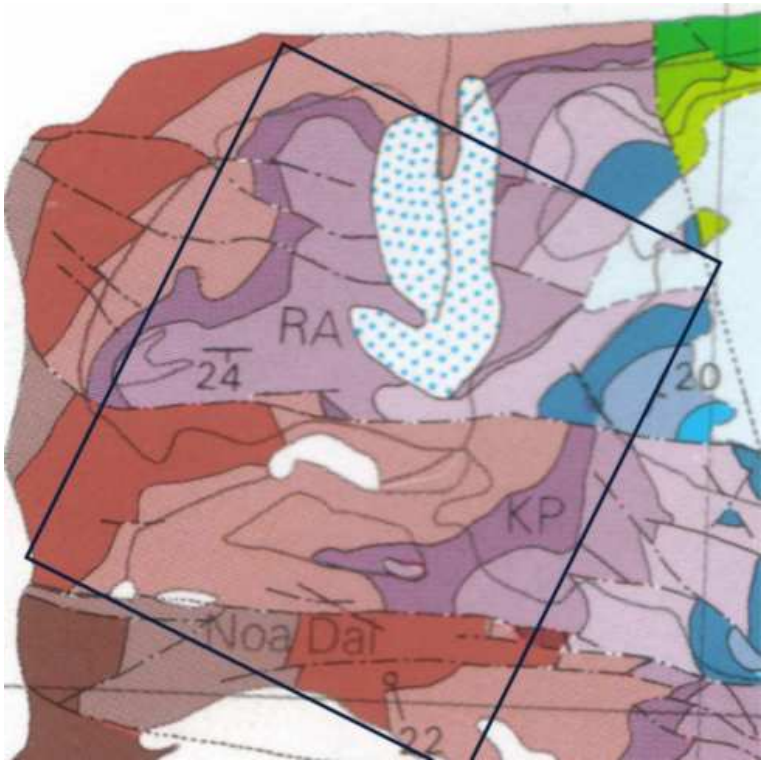


Figure 4.8: A map of the geological features of Ymer Ø. A black box indicates the location of TM scene under study.

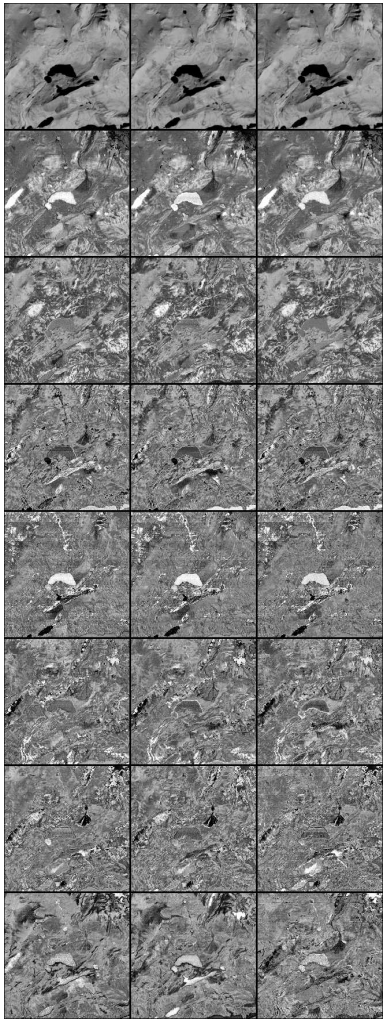


Figure 4.9: MACE decomposition using two shifts i.e. a three set nonlinear canonical correlations analysis (column-wise). The first column represents a center pixel set, the second a horizontally shifted set, and the third column a vertically shifted set. The first eight MACE CV triplets are shown. The images are scaled between mean ±3 std.

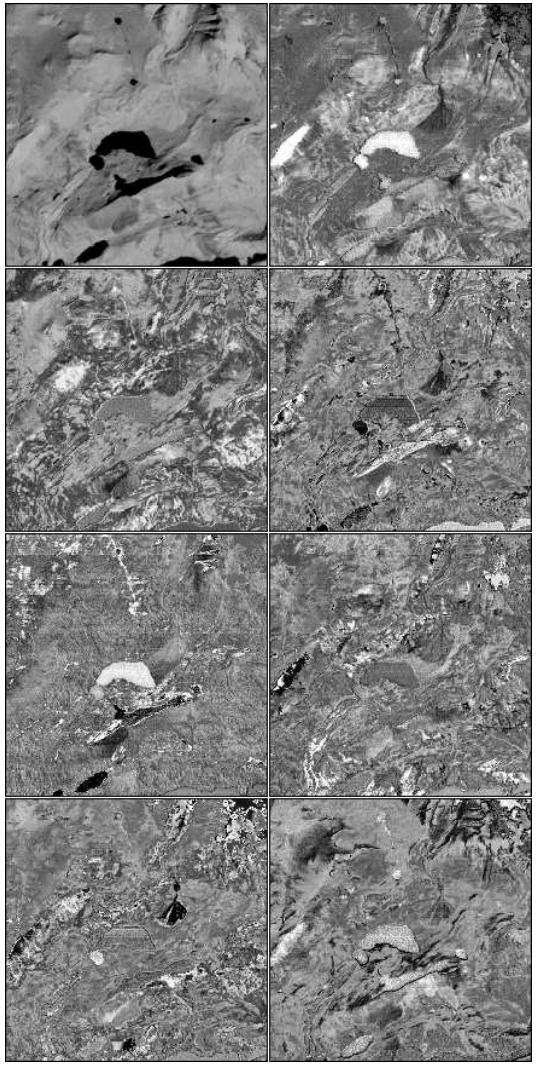


Figure 4.10: Nonlinear MACE decomposition of the TM center pixel set. The first eight MACE CVs are shown in a row-wise ordering. The images are scaled between mean ±3 std.

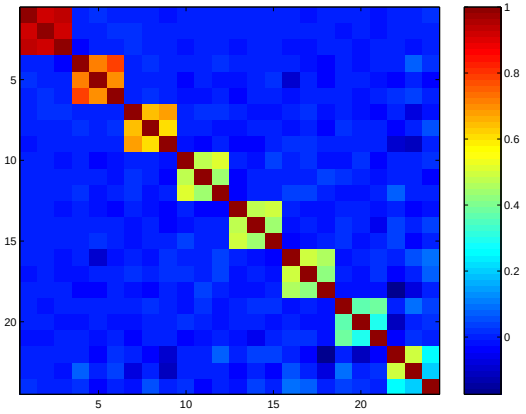


Figure 4.11: Correlations between the individual MACE CVs.

Results and Discussion

In Figure 4.7 the linear MAFs have been calculated using the standard spatial shifts. Again the MAFs appear successful in decomposing signal from noise, and we see salt-and-pepper-noise and some sensor row-wise striping in the high order MAFs. Nonlinear MAFs can be generated applying MACE. This is done using a north and east shift around each pixel generating three sets on which a nonlinear multiset correlations analysis is performed. The first eight CVs are presented in Figure 4.9 of all three sets. The first column contains the center pixel set, the second a horizontal neighbour set, and the third a vertical neighbour set. Notice that the CVs are very similar between sets. In Figure 4.10 the CV solutions for the first set are shown in a higher resolution. The canonical correlations of the MACE CV triplets are presented in Figure 4.11. Notice that they are uncorrelated between eigensolutions. The sum of the pair-wise correlations for the three sets decreases along the 3x3 block diagonal in the correlations matrix. The autocorrelation in the MACE CVs thus decreases for higher order components. Comparing MACE to MAF it manages to decompose the data into a large number of components before the image quality becomes degraded by speckled noise and striping.

Focusing on the high order MACE components and looking at the pair-wise

correlations, we notice that the method weights the horizontal neighbouring sets more when trying to obtain maximal pair-wise correlation. In the Figures 4.12 through 4.17 RGB colour combinations of the CVs for the first set are presented. Looking at the high order components some striping is present, but it seems as if MACE favours the sets which are horizontal neighbours. This reflects the means of collecting of the data, namely by a Thematic Mapper which scans the images row-wise.

From studying the MACE CVs with high correlations Geologist John L. Pedersen makes the following observations: “The RGB plot of the MACE CVs 1-3, return a distinct segmentation (see Figure 4.12). Talus cones and fluviatile fans are identified as discrete features, each with it own colour code. The remaining area can be divided into blue-red segments and yellow-white-red segments. The blue-red segment corresponds reasonably well with a distinctive stratigraphic sequence within the Eleonore Bay Supergroup consisting of alternating shales and massive, quartzite beds - named bedgroups 4 through 7 in the old stratigraphic terminology.

Due to the aggregate signature of the spectral signal, it is surprising that the method can create a diffuse segmentation with similarity to the bed-rock lithology. However, borders between the segments are not definitive, and the applicability of the methods with respect to classification of lithological units requires extensive additional testing.”

Conclusion

MACE generates a different decomposition of the data in comparison to the traditional MAFs. In comparison to the linear analysis the high order MACE-CVs have less striping due to sensor noise. It appears as if the MACE analysis manages to compensate for the noise structures not modelled by pooling the default spatial shifts in the MAF analysis.

The MACE CVs are uncorrelated between the eigensolutions but correlated within. MACE maximizes the sum of the pair-wise correlations and attempts to make the relation between the resulting CVs as linear as possible. In Figures 4.18 and 4.19 scatterplots are presented for the first two MACE CV triplets. Notice that they indicate a linear relation between the individual CVs for each eigensolution. In Figures 4.20 and 4.21 the transformations for the original six TM bands are shown for the first and second MACE solutions for the first set. Such plots can help a data analyst decide

how to proceed with the analysis when applying MACE. The data analyst controls e.g. the amount of smoothness constraints for the non-parametric scatterplot smoothers and may even choose to parameterize some of the transformations. This example works to illustrate that MACE can find multiple orthogonal eigensolutions to the correlation problem. Here MACE is applied to maximize autocorrelation, and it appears to have a potential ability to automatically compensate for anisotropic noise in the data.

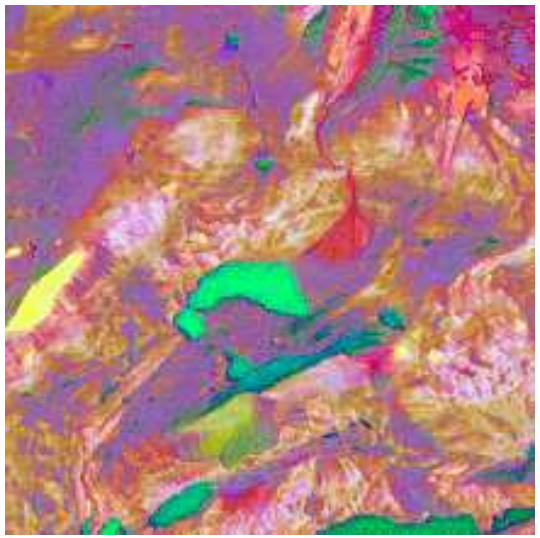


Figure 4.12: Colour combination of MACE CVs 1-3 of set 1.

Acknowledgements

Geologist John L. Pedersen, Mountain Resources ApS, is acknowledged for comments on the TM data set and on the results of the MACE analysis, and for providing the map of geological features of Ymer Ø.

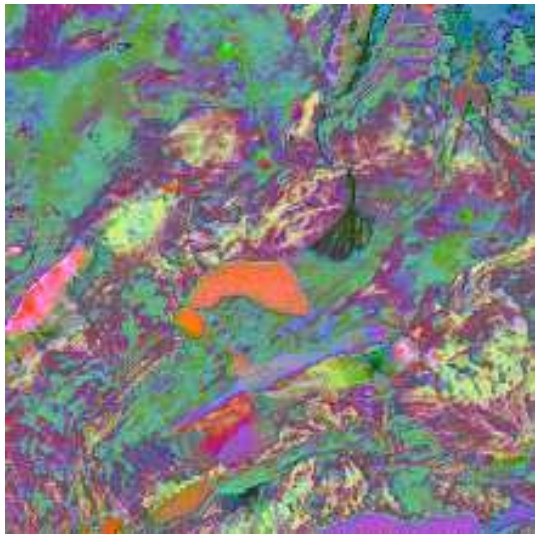


Figure 4.13: Colour combination of MACE CVs 2-4 of set 1.

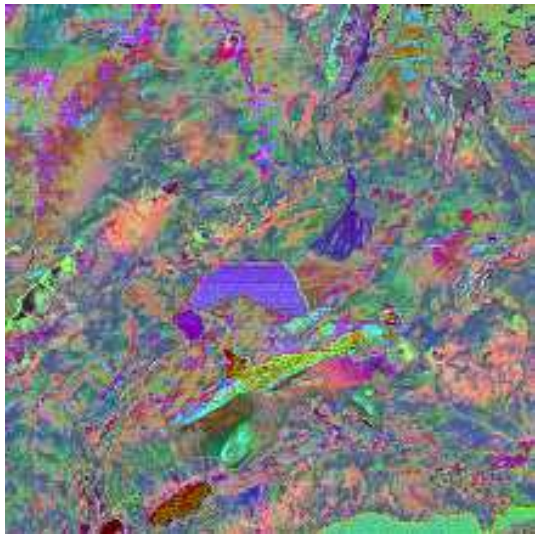


Figure 4.14: Colour combination of MACE CVs 3-5 of set 1.

Figure 4.15: Colour combination of MACE CVs 4-6 of set 1.

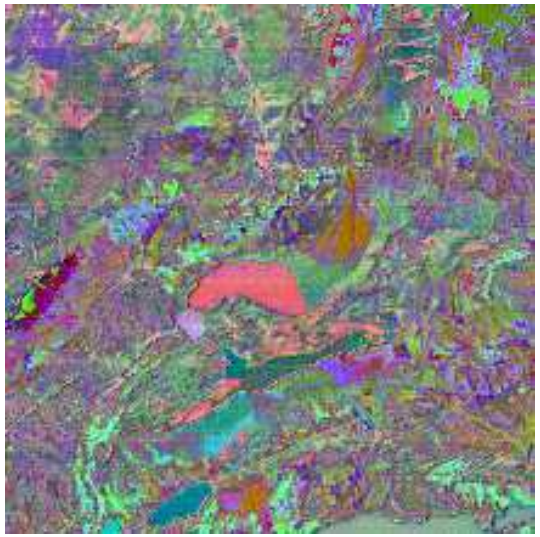


Figure 4.16: Colour combination of MACE CVs 5-7 of set 1.

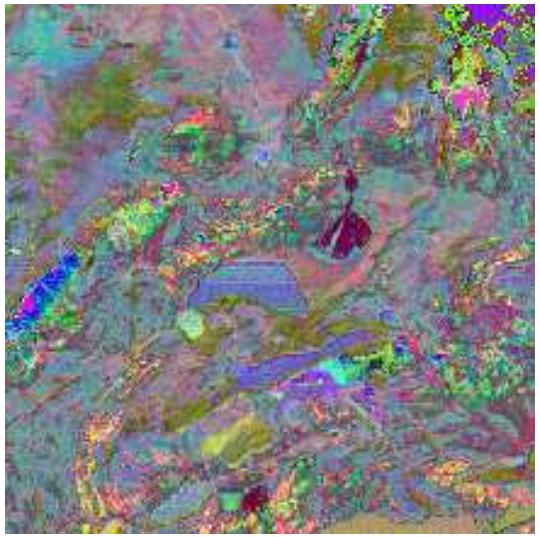


Figure 4.17: Colour combination of MACE CVs 6-8 of set 1.

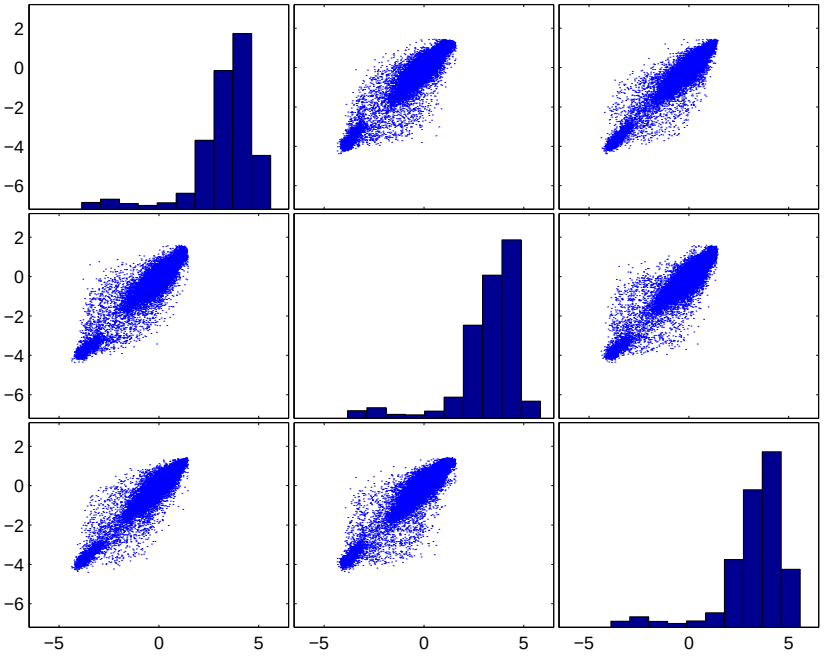


Figure 4.18: Scatterplots of the first MACE solution.

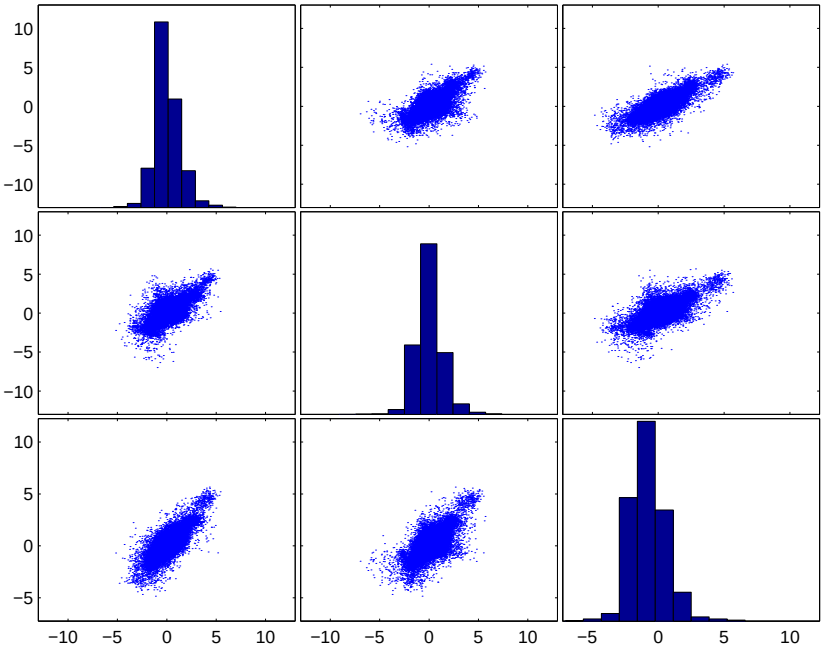


Figure 4.19: Scatterplots of the second MACE solution.

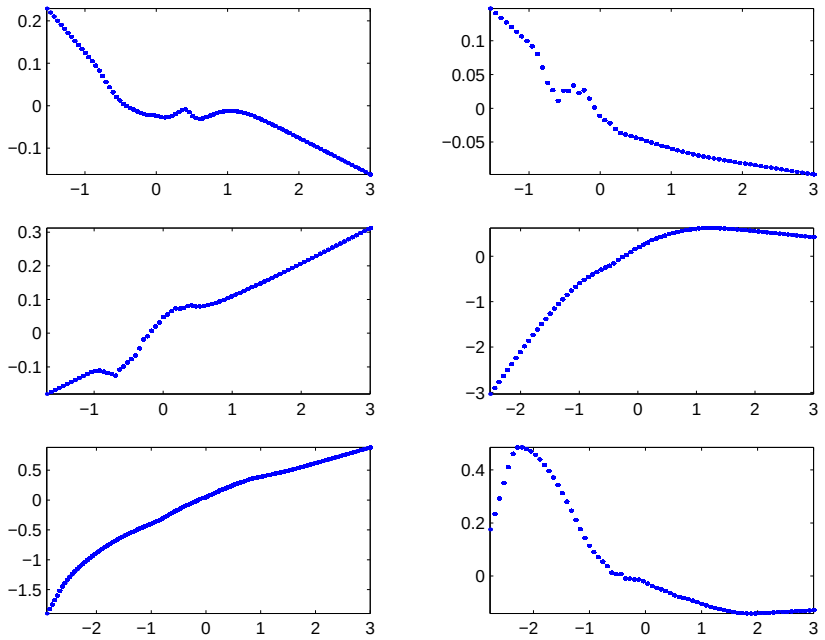


Figure 4.20: Transformations of the variables of the first set for the first MACE solutions.

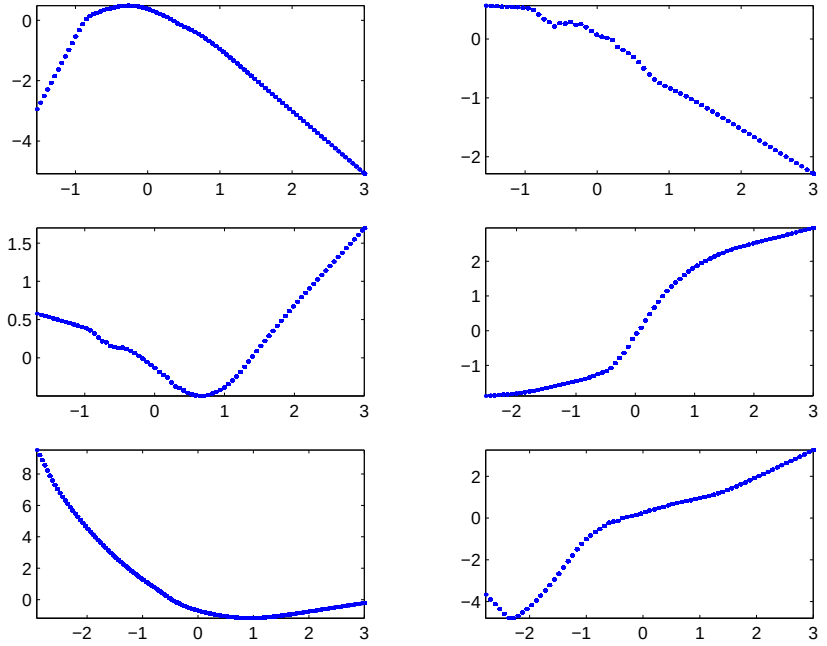


Figure 4.21: Transformations of the variables of the first set for the second MACE solutions.

4.3.3 Nonlinear Principal Components of MSS Data, Ymer Ø.

An analysis of raw image data from the four spectral bands of the MultiSpectral Scanner (MSS) is performed. The raw data are show in Figure 4.22 stretched to mean ± three standard deviations and depict parts of the same geographical region of Ymer Ø as in the previous case study. A histogram of the original data is shown in Figure 4.23. In Figures 4.24 and 4.25 are the linear PC and MAF decomposition presented. The spectral characteristics for the four MSS bands are shown in Table 4.2. The spatial resolution for the MSS scanner is 80 meters by 80 meters and the images shown in Figure 4.22 contain 100 by 100 pixels. A nonlinear correlations analysis is performed considering each band to be a single set containing one variable. Thus a four set canonical correlations analysis is performed. When applying only one variable in each set the resulting sum of CVs, each constrained to zero mean and unit variance, can be interpreted as a nonlinear principal component. MACE is applied to either maximize or minimize the variance of the sum of the CVs.

MSS Band	Wavelengths [μm]	Spectral ref.
1	0.5-0.6	Green
2	0.6-0.7	Red
3	0.7-0.8	Near-Infrared
4	0.8-1.1	Near-Infrared

Table 4.2: The spectral range of the Multispectral Scanner sensor bands 1 through 4.

The results of the linear principal components analysis and the maximum autocorrelations factors analysis are summarized in Table 4.3.

PC of MAF component index	1	2	3	4
Variance (in %) explained by the ith PC	89.67	9.24	0.58	0.51
Signal-to-noise ratio in the ith MAF	8.07	7.02	1.29	0.40

Table 4.3: Summary of the PC and the MAF analysis of the four MSS bands.

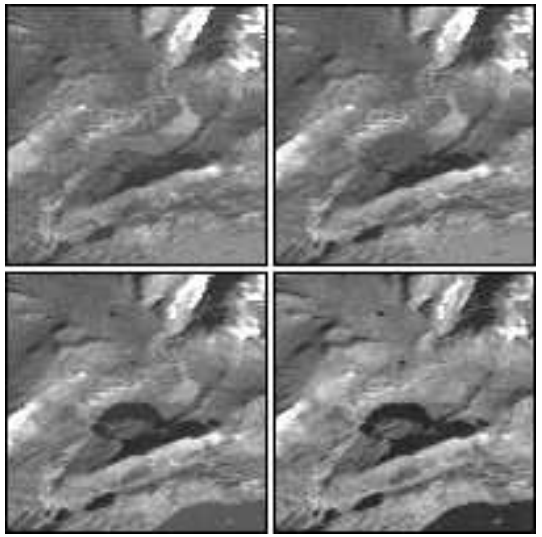


Figure 4.22: The raw MSS data bands 1-4.

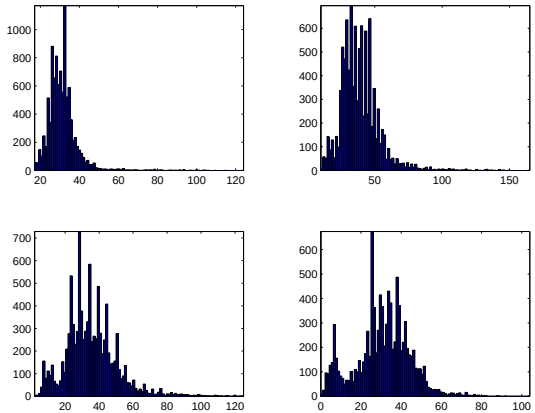


Figure 4.23: The histograms of the raw MSS data bands 1-4.

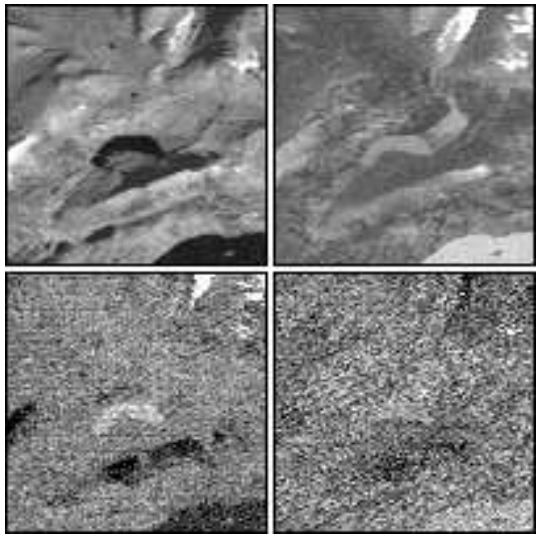


Figure 4.24: Linear PC decomposition of the MSS data.

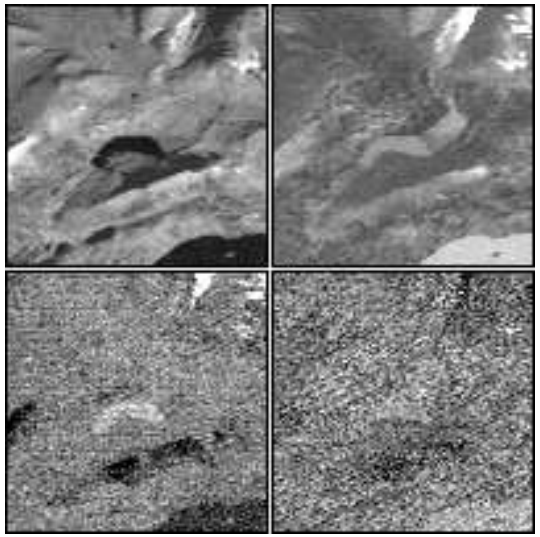


Figure 4.25: Linear MAF decomposition of the MSS data.

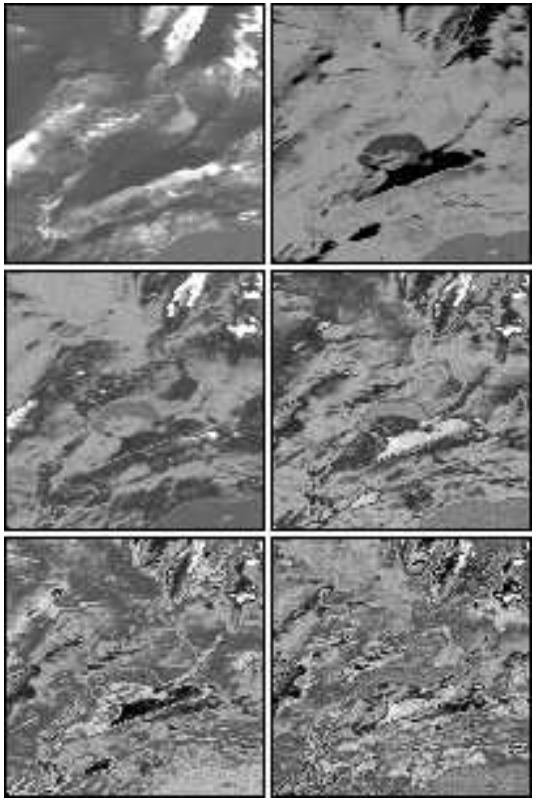


Figure 4.26: The nonlinear principal components with maximum variance.

Results

Nonlinear Principal Components with Maximum Variance

MACE is applied on the four spectral bands of the MSS data using the operator which maximizes correlation. The results are shown as nonlinear principal components with maximum variance in Figure 4.26. The first six nonlinear principal components are shown. Each component is the sum of the MACE CVs within each eigensolution. In Figures 4.27, 4.29, and 4.31 we show the MACE CVs of all four sets. In Figures 4.28, 4.30, and 4.32 we show transformations that produce the individual MACE CVs.

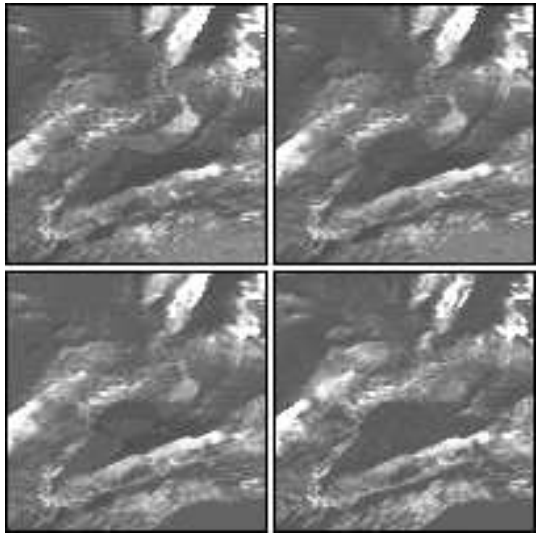


Figure 4.27: Images of the individually transformed variables in the 1st maximum nonlinear principal component.

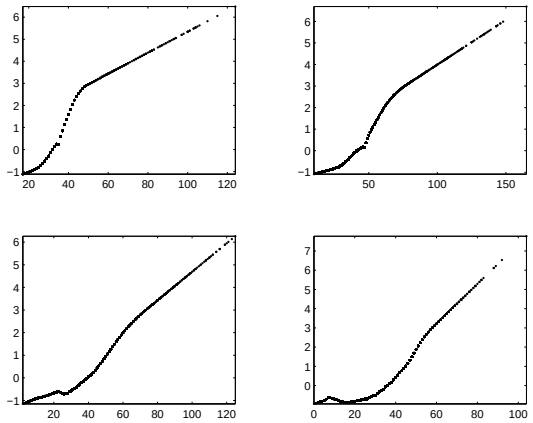


Figure 4.28: The transformations of the 1st maximum nonlinear principal component.

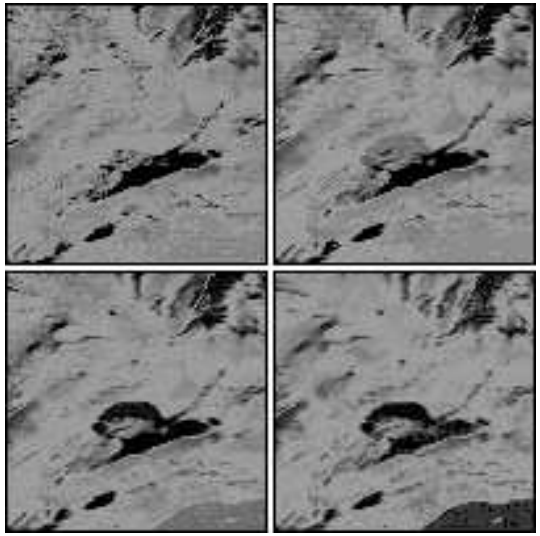


Figure 4.29: Images of the individually transformed variables in the 2nd maximum nonlinear principal component.

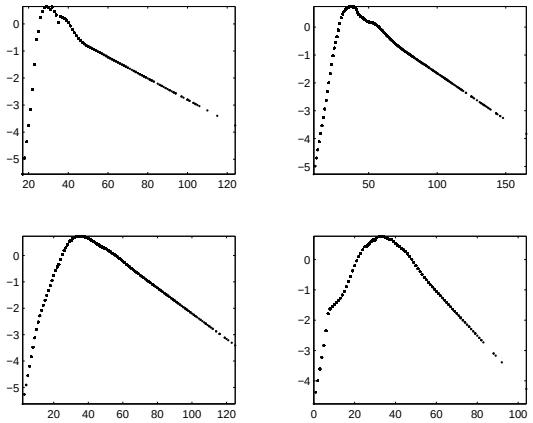


Figure 4.30: The transformations of the 2nd maximum nonlinear principal component.

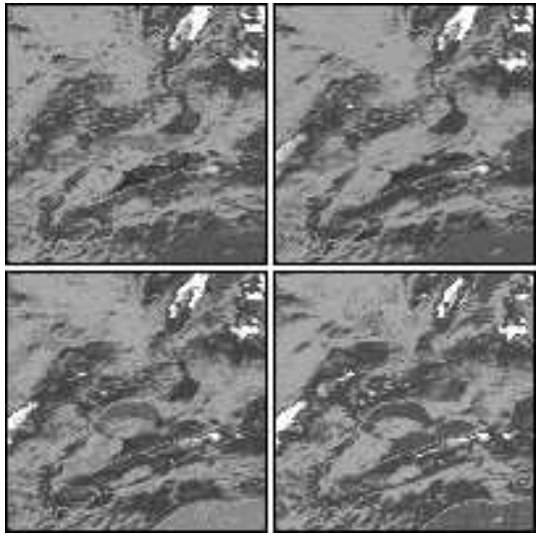


Figure 4.31: Images of the individually transformed variables in the 3rd maximum nonlinear principal component.

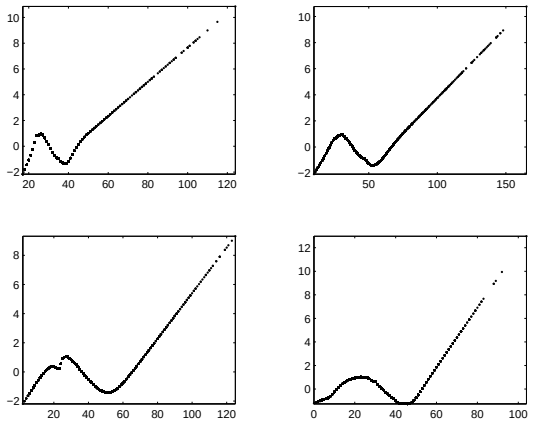


Figure 4.32: The transformations of the 3rd maximum nonlinear principal component.

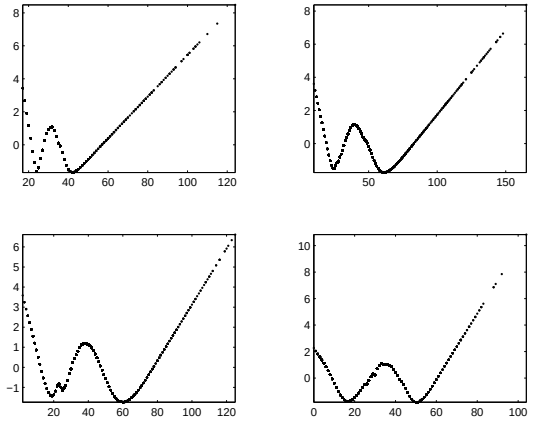


Figure 4.33: The transformations of the 4th maximum nonlinear principal component.

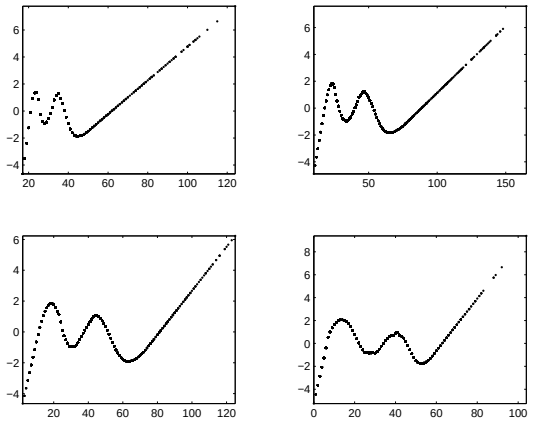


Figure 4.34: The transformations of the 5th maximum nonlinear principal component.

In Figures 4.33 through 4.35 the remaining transformations for the higher order MACE CVs are shown. In Table 4.4 the variance contained in the first six orthogonal nonlinear principal components which maximize variance are presented.

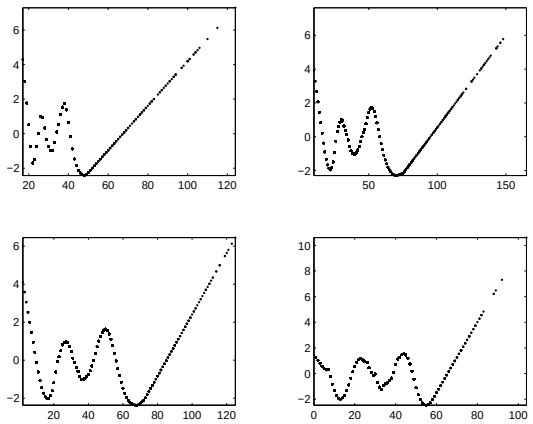


Figure 4.35: The transformations of the 6th maximum nonlinear principal component.

Nonlinear Principal Components with Minimum Variance

MACE is applied to look for minimum variation in the sum of the CVs when applied to the four MSS bands. The results of determining nonlinear principal components with minimum variance are shown in Figure 4.36. The first six nonlinear principal components are presented. Each component is the sum of the MACE CVs within each eigensolution. Notice that they are all highly noise corrupted. This is expected since noise is often characterized by low variance. In Figures 4.37, 4.39, and 4.41 we show the MACE CVs of all four sets. In Figures 4.38, 4.40, and 4.42 we show the transformations that produce the individual MACE CVs. In Figures 4.43 through 4.45 the remaining transformations for the higher order MACE CVs are shown. The variance contained in each nonlinear principal component which minimizes variance is presented in Table 4.4.

Component	1	2	3	4	5	6
Max-NPCA	14.718	13.636	13.113	11.686	10.854	9.522
Min-NPCA	0.188	0.195	0.514	0.785	1.085	1.608

Table 4.4: The variance explained by first six nonlinear principal components (NPCA) that maximize and minimize variance respectively of the four spectral MSS bands.

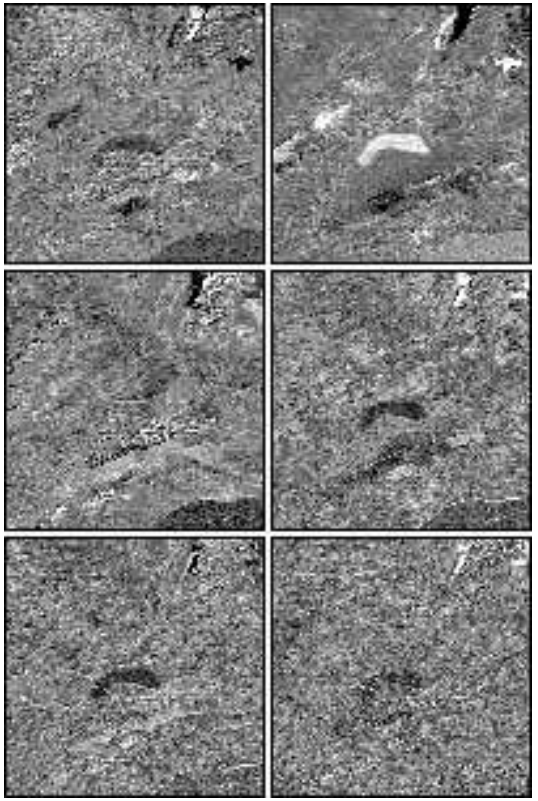


Figure 4.36: The nonlinear principal components with minimum variance.

Conclusion

This case study illustrates that MACE can be applied when trying to either maximize or minimize the sum over all the pair-wise correlations in a multiset scenario. When the sets consist of only one variable in each set, the sum of the MACE CVs can be interpreted as nonlinear principal components. We find multiple orthogonal eigensolutions in both ends of the variance spectrum. When maximizing variance, MACE generates components of the individual sets that are as similar as possible. A data analyst can study the resulting transformation curves and the resulting CVs when

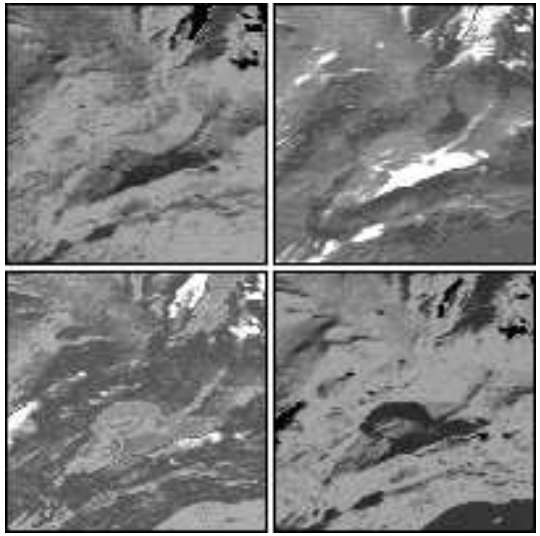


Figure 4.37: Images of the individually transformed variables in the 1st minimum nonlinear principal component.

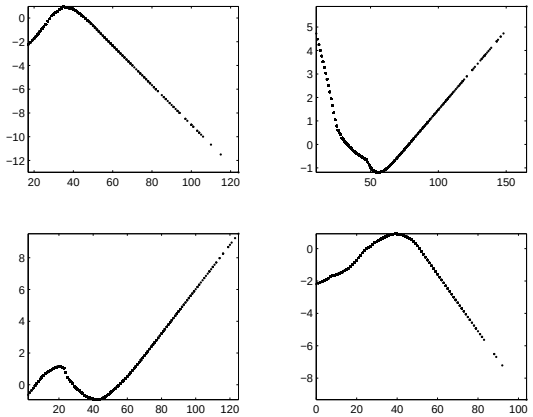


Figure 4.38: The transformations of the 1st minimum nonlinear principal component.

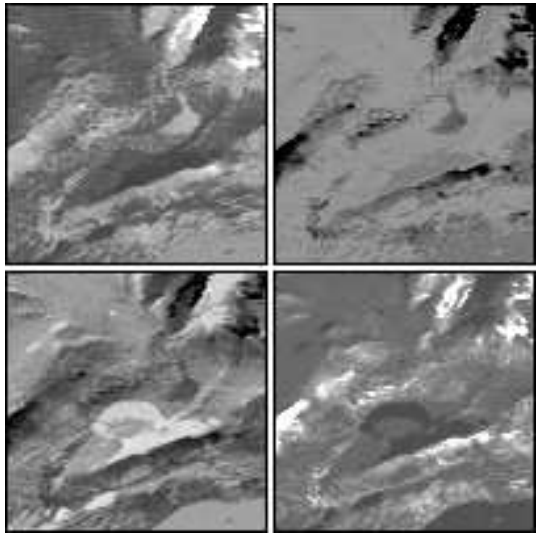


Figure 4.39: Images of the individually transformed variables in the 2nd minimum nonlinear principal component.

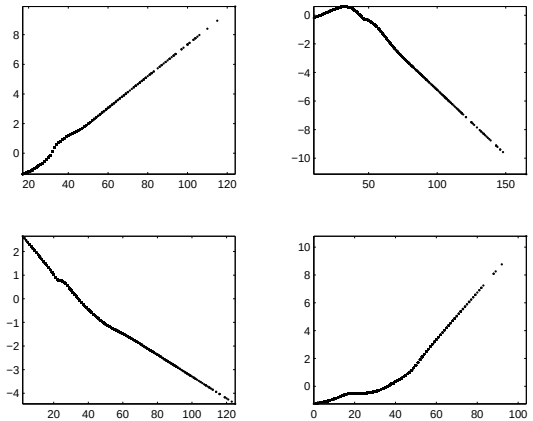


Figure 4.40: The transformations of the 2nd minimum nonlinear principal component.

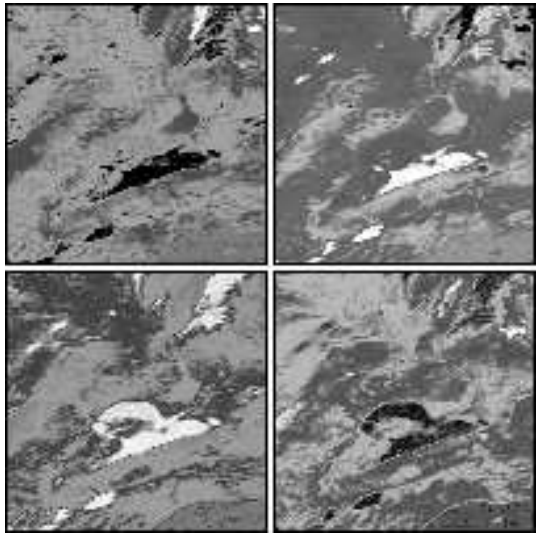


Figure 4.41: Images of the individually transformed variables in the 3rd minimum nonlinear principal component.

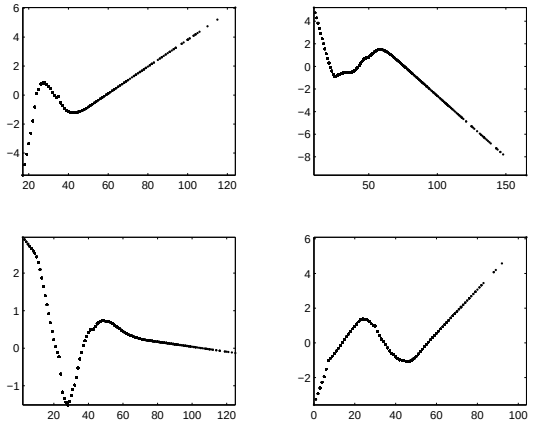


Figure 4.42: The transformations of the 3rd minimum nonlinear principal component.

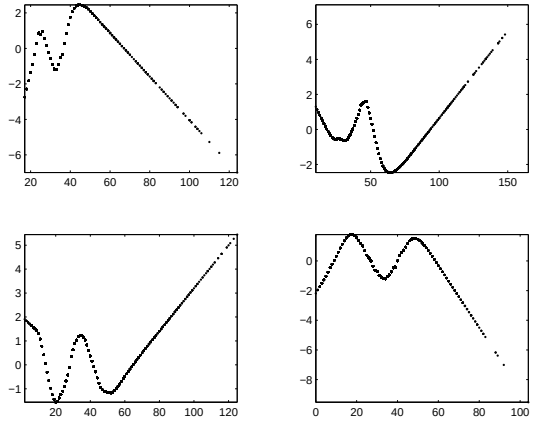


Figure 4.43: The transformations of the 4th minimum nonlinear principal component.

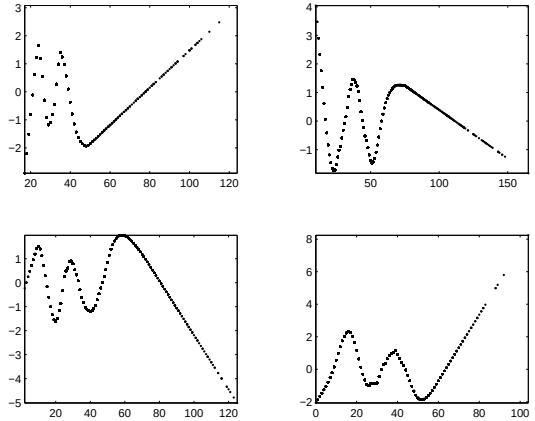


Figure 4.44: The transformations of the 5th minimum nonlinear principal component.

interpreting the analysis. In the presented case study the curves indicate how the pixel values in each band of the MSS data must be transformed to maximize correlation. Looking at the CVs which determine the first nonlinear principal component, we see the individual mappings for the MSS

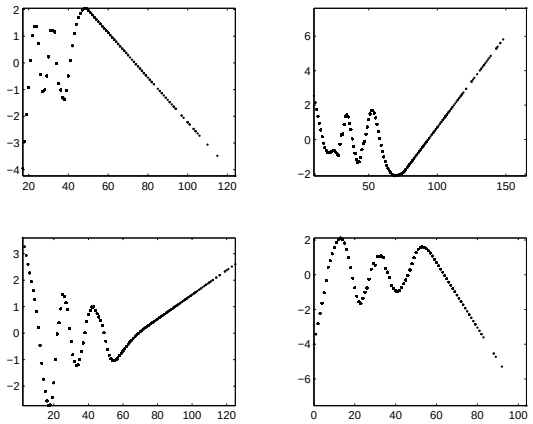


Figure 4.45: The transformations of the 6th minimum nonlinear principal component.

bands. The transformation curves are all different, but there seems to be a general agreement on how to transform the data between groups. Looking at the transformations within each eigensolution, there is a good agreement on how many strong local extrema exist for each individual mapping. However, when interpreting solutions found by MACE, one must be careful not to interpret trivial suboptimal solutions. In the trivial solutions the form of the resulting mapping is more dependent on the previously determined solutions through the orthogonal preserving criteria in MACE than on the remaining structures in the data.

Minimizing the variance of the nonlinear principal components produces CVs that minimize the sum of the pair-wise correlations over all sets. Similarly to linear principal components, the algorithm tries to determine the transformations of the individual components such that they cancel each other in the resulting principal variate i.e. the sum of the transformed individual bands. The nonlinear principal component decompositioning of the data can be applied to study, if some sets naturally group together when attempting to compensate for each other. It can thus assist a data analyst in obtaining information on the correlation structure of the involved variables through nonlinear transformations. Similarly to the case of maximum nonlinear principal components trivial, suboptimal solutions can occur when looking for minimum variance, which must be taken into

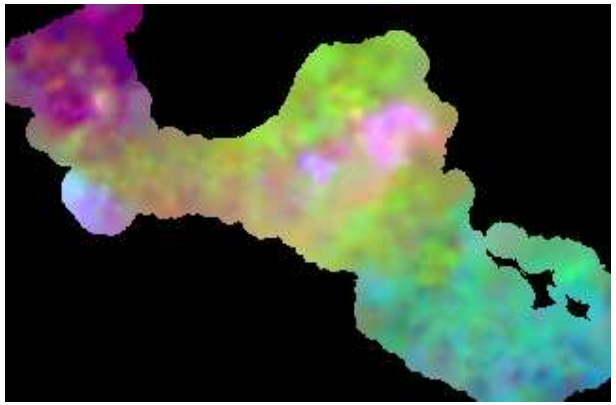


Figure 4.46: The first three CCA CVs as RGB. Neighbouring observation have been applied in a two-set linear CCA analysis.

consideration when interpreting the results. One way to avoid the trivial eigensolutions is to discard the orthogonalizing steps in MACE and replace them with some transformation that deflates the previous optimal correlations solution in the Hilbert space. This approach is inspired by the means of locating multiple solutions in projection pursuit described in [38, 128, 21]. Developing such a transformation for structure removal is not trivial, and it is not obvious whether this is at all possible.

4.3.4 Nonlinear Canonical Correlations Analysis of Stream Sediments Geochemistry Data, South Greenland

Under the Syduran project stream sediments geochemistry samples were collected and analysed for the content of 41 elements. The elements are: Au, Ag, As, Ba, Br, Ca, Co, Cr, Cs, Cu, Fe, Ga, Hf, K, Mn, Mo, Na, Ni, Pb, Rb, Sb, Sc, Se, Sr, Nb, Ta, Th, Ti, U, W, Y, Zn, Zr, La, Ce, Nd, Sm, Eu, Tb, Yb, and Lu. The data set consists of 2097 samples. The results shown in this section are kriged images using the same first order spherical semivariogram with nugget effect of 0.1, a sill of 1, and the search radius equal to the range of influence of 40 km.

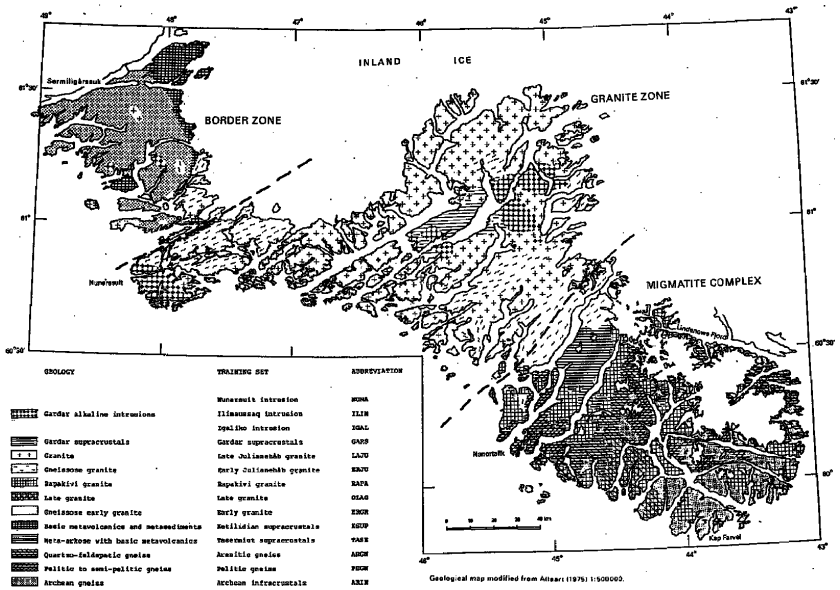


Figure 4.47: A geological map of South Greenland.

Results and Discussion

Linear decomposition is performed by analyzing the data in a two-set CCA. The two sets involved are constructed from neighbouring observations. The results of the first three CVs are shown as RGB in Figure 4.46.

The linear CCA CVs produce a decompositioning that distinguishes the major lithotectonic units of South Greenland: the Border Zone, the Granite Zone and the Migmatite Complex, see Figure 4.47 in which a geological map of South Greenland is presented. A more thorough analysis is done by applying the MAF transform on the irregular data set, see [89, 91].

MACE is applied, performing nonlinear decomposition of the data, to generate maximum autocorrelations factors. We analyze both a two, three and four set scenario of neighbouring observations. The results are shown in the Figures 4.48 through 4.50.

When visually comparing the results of the different MACE analyses to each

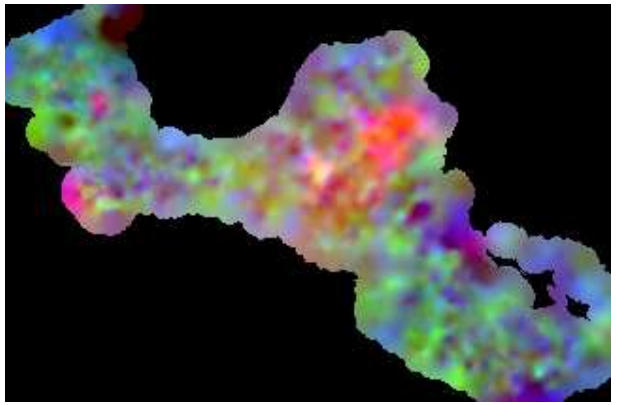


Figure 4.48: The first three MACE maximum autocorrelation factors as RGB. Neighbouring observations are applied in a two-set MACE analysis.

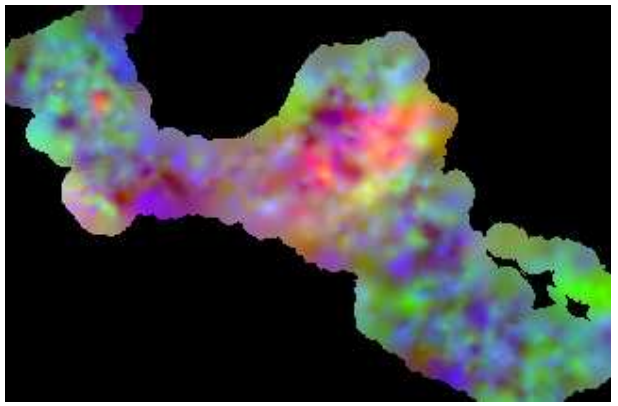


Figure 4.49: The first three MACE maximum autocorrelation factors as RGB. Neighbouring observations are applied in a three-set MACE analysis.

other, they appear to be in relatively good correspondence. In paticular the Garder intrusion appears very conspicuous as bright red regions.

Geologist John L. Pedersen proposes to divide the data into three groups:

- Rock-forming elements: Ba, Ca, Co, Fe, Ga, K , Mn, Na, Rb, Sc, Sr,

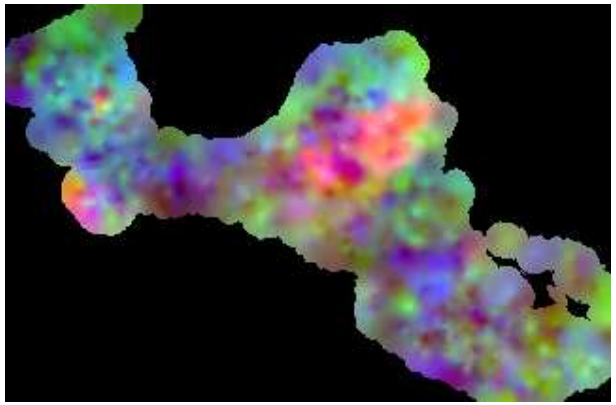


Figure 4.50: The first three MACE maximum autocorrelation factors as RGB. Neighbouring observations are applied in a four-set MACE analysis.

and Ti.

- Mineralization related elements: Au, Ag, As, Cr, Cs, Cu, Mo, Ni, Sb, Se, U, W, and Zn.
- Elements with strong correlations to factor one in a factor analysis of all elements ([88]): Ga, Hf, Nb, Ta, Th, Y, Zn, Zr, La, Ce, Nd, Sm, Eu, Tb, Yb and Lu.

When applying MACE to the three groups and inspecting the results, the decompositioning is hard to interpret from a geological perspective. The first three MACE CVs of each set are shown in Figure 4.51 through 4.53 as RGB images. If the data are highly interdependent, MACE is expected to encounter problems when looking for maximum correlation. For data compression we therefore perform principal component analysis of each group and discard the high order components which explain a little fraction (less than 5%) of the total variation of each set. The three groups are thus compressed into 8, 10 and 8 principal components respectively.

The truncated PC basis of the three sets is analysed using MACE. The first three canonical correlations eigensolutions are shown for the three groups in the Figures 4.54 through 4.56. The CVs are shown as RGB colour images.

Focusing on the CVs for the group of mineralization elements we see some interesting effects. The strong red regions are relatively young geological

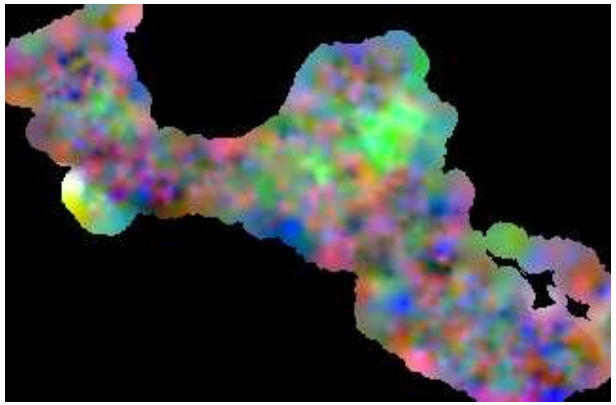


Figure 4.51: The first three MACE CVs of group 1 as RGB.

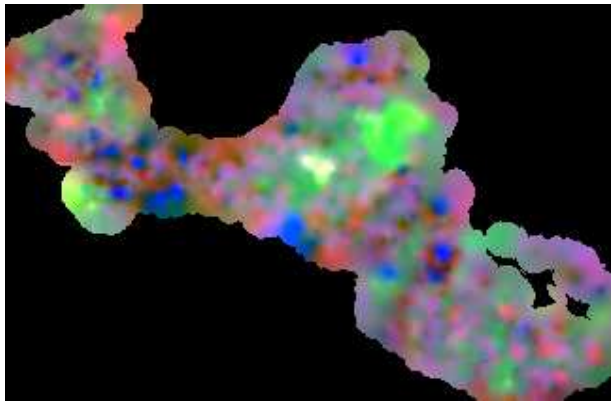


Figure 4.52: The first three MACE CVs of group 2 as RGB.

structures. The areas indicated in the RGB component are in good correspondence with the a priori knowledge on the individual regions. Moreover, the blue region, located in the boundary between the Granite Zone and the Migmatite Complex, is interesting from a geological perspective. In particular the angle at which the blue region intersects the boundary of the two zones is relevant and interesting when trying to obtain insight on the geological configuration (John L. Pedersen, pers. comm.).

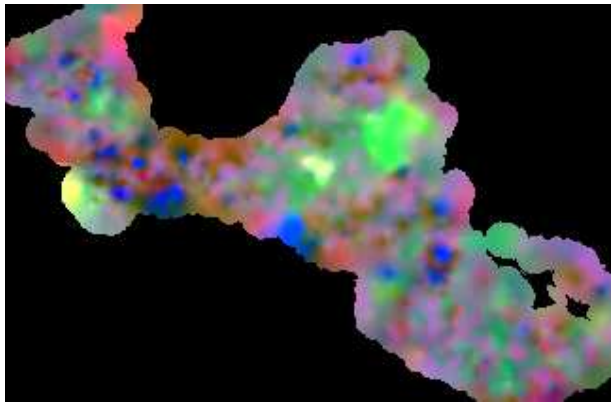


Figure 4.53: The first three MACE CVs of group 3 as RGB.

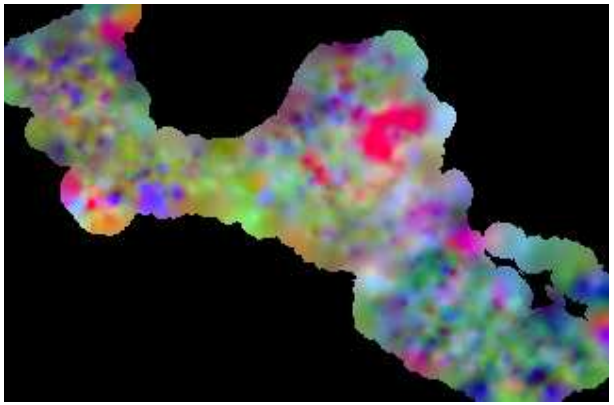


Figure 4.54: The first three MACE CVs of the truncated PCs of group 1 as RGB.

Conclusion

Analyses of irregularly sampled stream sediments geochemistry data from South Greenland are presented. The concluding comments of Geologist John L. Pedersen are the following: "Linear correlations analyses of geochemical data from regional stream sediments samples give geologically meaningful results. RGB plots of the variables with the largest autocorrelation structure reveal a segmentation of the area which reflects the various Archean and Proterozoic components. The ability of the method to create such segmentation is seen as the key advantage from a geological perspective.

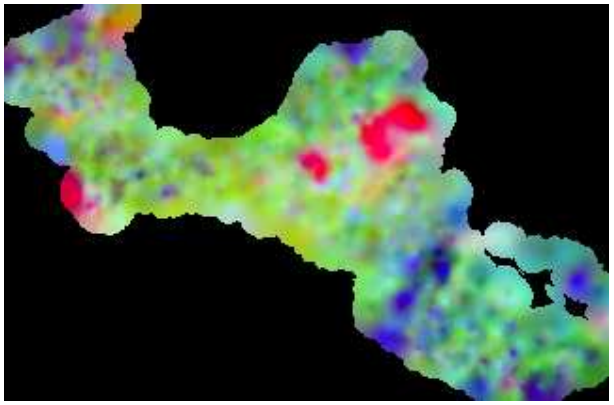


Figure 4.55: The first three MACE CVs of the truncated PCs of group 2 as RGB.

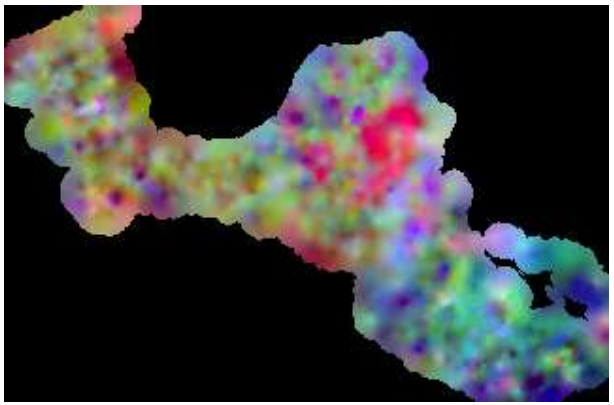


Figure 4.56: The first three MACE CVs of the truncated PCs of Group 3 as RGB.

Various methods for nonlinear analyses have been tested. RGB plot of

variables obtained from such analyses are, on a visual basis, compared with results from linear analyses. Even the best results are less efficient in yielding a geological meaningful segmentation. Geological units with distinct characteristics e.g. the Meso-Proterozoic alkaline intrusions are well identified, but the discrimination between various Archean and Proterozoic terrain is less distinct."

From a geological perspective the linear CCA analysis thus appears to decompose the data into meaningful components. Furthermore, the CCA decompositioning differs from a linear non-spatial PC analysis (results not shown). The presented nonlinear components seem to produce less useful results. However, it would be interesting to slowly relax the linear transformations to more nonlinear transforms using MACE. A data analyst would then have the possibility of better controlling and interpreting the nonlinear decomposition when looking for revealing projection of the data.

Acknowledgements

Senior Geologist Agnete Steenfelt (Geological Survey of Denmark and Greenland - GEUS) is acknowledged for providing the data, and Geologist John L. Pedersen, Mountain Resources ApS, is acknowledged for comments on the results of the data decompositioning.

4.4 Summary

- Nonlinear methods for decomposition of multivariate single and multiset problems are presented.
- The explicit and the implicit approaches are briefly described.
- Nonlinear additive models are presented. This includes the generalized additive model (GAM), the projection pursuit regression (PPR) model, artificial neural networks (ANN) models, and finally the alternating conditional expectations (ACE) model.
- Non-parametric GAM (on canonical link transformed responce data) is in effect identical to the restricted ACE which is the inner loop of the ACE algorithm.
- PPR is an extension of GAM which simply applies linear combinations of the input variables to the non-parametric transformations.
- ANN can be perceived as coupled layers of PPR schemes in which the nonlinear transformations are restricted to activation functions which are primarily constrained to provide bounded responses.
- ACE is similar to GAM but is more related to canonical correlations analysis than multiple regression. The ACE algorithm estimates the transformations of two sets of variables, which maximize the correlation. It thus attempts to produce new canonical variates between which the relation is as linear as possible.
- ACE is generalized to handle multiple sets. The new algorithm is termed multiset ACE (MACE).
- MACE can handle arbitrary mixtures of multivariate sets, including both continuous and categorical variables.
- MACE can be considered as a generalization of the linear MCCA. When the transformations are restricted to linear combinations, MACE reduces to MCCA.
- MACE is non-parametric and applies minimal assumptions concerning the underlying data structures.
- The only requirements on the transformations applied in MACE are i) they have zero mean, and ii) the sum of the transformations within each set has a variance of one. These restrictions originate from ACE and are meant to prevent the algorithm from drifting and to ensure that there is dynamics in the resulting canonical variates i.e. to avoid null transformations.
- The user may introduce additional restrictions on the transformations applied in MACE. This could be monotone restrictions or the application of parametric or even linear transformations thus resulting in semi-parametric MACE.
- Case studies are presented applying the non-parametric MACE on a bivariate multitemporal data set, multispectral remotely sensed data, and on irregularly sampled data of geochemical variables collected in South Greenland.
- In the analysis of bivariate ocean temperature and height, ACE obtained an additive model in which the El Niño shows up as a strong component and ACE is able to obtain a solution with a higher correlation in comparison with a linear decomposition.
- When applied in a nonlinear maximum autocorrelation factor analysis, MACE produced a different decompositioning in comparison to the linear MAF analysis. It produced a large number of components before degrading due to the influence of the noise in the data.
- MACE is applied to generate nonlinear principal components of a

multispectral four band image. Components are found which maximize and minimize variance.

- A case study of irregularly sampled geochemical data is given applying MACE. MACE appears to encounter problems due to the high dimensionality of the problem. Examples are given maximizing autocorrelation in a combined analysis of all elements in the data and of a canonical correlations analysis between the data partitioned into three groups.
- MACE provides the means of performing nonlinear decomposition of multiset problems and is expected to be useful when applied by a data analyst, who controls how to regularize the algorithm while looking for interesting correlation structures in the data.

Chapter 5

Conclusion

This thesis presents exploratory methods which all originate in traditional data analysis. Some of the methods belong to the group of well-established multivariate statistical methods and others to the more modern applied statistical group of methods. An attempt is made to describe the relation between the individual approaches and to present the ideas and objectives behind the algorithms.

This work can be organized into three major parts concerning methods for data driven exploratory analysis

- Cluster analysis of regular spatially sampled multivariate image data.
- Linear decomposition of single and multiple multivariate data sets.
- Nonlinear decomposition and generalization of the classical multiset canonical correlations analysis.

The methods are meant to be applied in the initial probing of the data when trying to reveal interdependencies between the involved variables. In high dimensions the curse of dimensionality makes it impossible to detect all but the very coarsest structure of the joint probability function, and the most one can hope for is to be able to get a general idea of the density function. In exploratory analysis the methods for cluster analysis and for decomposition can be useful when applied individually, but they may also be beneficial when used in a pre- or post-processing of each other.

Based on the success of the established spectral fuzzy c-means (FCM) algo-

rithm, the method is chosen for unsupervised classification purposes. The algorithm is extended to provide better segmentation of multivariate image data. The FCM algorithm applies memberships and differs from hard clustering in allowing for overlapping classes. Hard k-means (HCM) and spectral noise clustering (NCM) are also presented. The FCM algorithm is extended by spatially related memberships which enhance the robustness of the algorithm to outliers and noise and the situations where ordinary FCM fails due to the degree of overlap of the clusters in the spectral space. The new memberships include a spatial membership and a parental membership. The spatial membership is constructed from Markov random field energy potentials. It favours segmentation into homogeneous regions. The parental membership is obtained by analyzing the data at more coarse scales. The ordinary, uncommitted, Gaussian scale-space is applied as default. The parental membership passes down information through the data scale-space, and one must be careful not to travel over "to large scales" due to the correspondence problem. Results on simulated data show that the use of a joint spectral-spatial-parental membership gives the best segmentation result. Applying the spatial membership introduces additional calculations within each iteration of the clustering algorithm, but it improves the robustness and often leads to fewer overall iterations and thus faster convergence. Applying the scale-space framework is also an effective way to speed up the performance of the clustering algorithm. Results on speed and convergence of the new algorithm are not included in this presentation. It should be stressed that the new fuzzy clustering algorithm assumes that the data naturally partition into homogeneous same class regions. The data analyst controls the degree of spatial regularization applied. Case studies are presented on i) simulated image data for evaluation, ii) a multispectral eight band Sea-viewing Wide Field-of-view Sensor (SeaWiFS) image of Northern Europe, and iii) a multichannel scanning electron microscope x-ray mapping image of ten elements. In the presented case studies, the use of Euclidean distance seems to be compensated for by the application of the new memberships, and the algorithm provides useful segmentations for further analysis of the data.

For linear decomposition of multivariate single sets, the principal components (PC), the minimum noise fractions (MNF), the maximum autocorrelation factors (MAF) transformations are presented. The latter two transforms are more dedicated in decomposing multivariate image data, since they can take the spatial nature af image data into account. An extension of the MNF/MAF transforms is proposed restricting the covariance

structure of the noise in the new representation to the identity matrix. The variance of the resulting components depends on the signal-to-noise ratio contained in each component. The new transformation is termed the signal-MNF/MAF transformation (SMNF/SMAF) and is restricted to the group of transforms that are invariant to linear transformations of the data. When the SMAF is applied as a preprocessor to the simulated image data, the extended FCM algorithm produces enhanced results. The new transformation works as a feature selector. It stretches the data in the subspaces rich on autocorrelated signal and compresses (or even discards) those corrupted by noise. For linear decomposition of multiple sets, the canonical correlations analysis (CCA) of two sets is presented. It is generalized to handle multiple sets (MCCA), and a relation to Procrustes shape alignment of point distribution models is found. CCA and MCCA produce so-called canonical variates that are uncorrelated between but correlated within eigensolutions. Case studies are presented including an application and new combination of the methods for decomposing the multispectral SeaWiFS scene previously evaluated by the clustering algorithm. The analysis presented produces a decomposition by suppression of undesired cloud spectra and speckled noise The decomposition is obtained by means of the rank reducing orthogonal subspace projection (OSP) for performing the partial unmixing, followed by the MAF transformation. The suggested approach appears to be able to collect interesting ocean related signal from the data into few components and is expected to be useful in analyzing the ocean colour. Two-set canonical correlations analysis is performed on a bivariate multitemporal data set. The data describe the global sea surface height and temperature of 1996 and 1997. The two fields are interrelated, i.e. an increase in the temperature will lead to an increase in the sea height. The analysis gives good indications of an anomaly off the South American west coast taking place in the second half of 1997. This is in good agreement with established knowledge on the build-up of one of the largest El Niño events on record. Multiset canonical correlations analysis is performed on 24 metacarpal II two-dimensional registered bones. The analysis consists of alignment of the shapes. A decomposition of the shape dynamics in the shape space of the two eigensolutions found by the canonical analysis is also presented. In the presented case study the second eigensolution does not remove all rotation, which is a conventional requirement when estimating pose.

Methods for nonlinear decomposition are presented. The explicit and the implicit approaches are briefly described. The first of these explicitly ex-

pands the variables onto a new basis and then applies the ordinary linear methods for decomposition. The second approach accomplishes the task by implicitly expanding the data onto a new basis. This is possible for methods that can be formulated such that only inner products occur in the calculations. Kernels are applied to calculate the inner-products implicitly in the large high-dimensional space. Both methods have several shortcomings. The primary problem is how to choose the new basis representation. This is entirely up to the user and there seems to be no clear data driven approach for model selection in this regard. Nonlinear additive models are presented. This includes the generalized additive models (GAM), the projection pursuit regression (PPR) model, artificial neural networks (ANN), and finally the alternating conditional expectations (ACE) model. Non-parametric GAM is closely related to ACE and in fact corresponds to the inner loop of the ACE algorithm. GAM applies the backfitting algorithm when estimating the optimal transformations of the input variables. PPR is an extension of GAM, and includes the introduction of linear transformation of the input variables before applying the nonlinear mappings. PPR thus attempts to compensate for the problems that are encountered in higher-dimensions. How to choose the number and weights for the linear combinations can be incorporated into a model building process. ANN can be perceived as coupled layers of PPR schemes. The linear combination determines the weights in the neural net, and the nonlinear transformations are the activation functions. ANN typically applies the logistic function as activation function. The net can consist of several PPR layers, and connections may even skip layers, allowing for a wide range of nonlinear models. The restricted ACE algorithm is similar to GAM, but the full ACE algorithm also allows transformation on the "response" set and is related to two-set canonical correlations analysis. It can be generalized to handle two sets of multivariate data and can find several eigensolutions. We generalize ACE to handle multiset problems. The new algorithm is termed multiset-ACE (MACE). MACE is non-parametric and applies minimal assumptions concerning the underlying data structures. It is a data driven algorithm, which estimates the optimal transformations of the input variables such that the sum of the pair-wise correlations over all sets is maximized. It handles arbitrary numbers and mixtures of continuous and categorical variables within each set. The user may specify restrictions on the involved transformations that may even take parametric forms, resulting in semi-parametric MACE when mixtures are applied. MACE can find suboptimal eigensolutions that are orthogonal. The non-parametric

MACE algorithm using Friedman's supersmoother is applied in case studies on multispectral remotely sensed data and on irregularly sampled geochemical data. Two-set nonlinear canonical correlations analysis is performed on ocean temperature and height multitemporal data. Inspecting the ACE CVs for the first ACE pair the El Niño appears very strongly, and the algorithm seems to be able to separate relevant ocean configurations from the temporal data. When MACE is applied on spatially shifted data sets, it maximizes the spatial autocorrelation and thus provides the means for obtaining nonlinear MAFs. A case study, comparing the linear MAF and MACE, shows that MACE gives a different decomposition in which a higher number of components can be generated. For the high order components, MACE favours the sets which are horizontal neighbours reflecting the construction of the instrument collecting the data, namely the Landsat Thematic Mapper (TM) that scans the images row-wise. Nonlinear principal components are generated using MACE on a four dimensional spectral (MultiSpectral Scanner, MSS) remotely sensed image. The data are partitioned into four sets with one variable in each set. When maximizing correlation over all sets the variance of the sum of the MACE canonical variates is maximized. MACE can also be applied to minimize the variance. The sum of the canonical variates can be interpreted as nonlinear principal components. Multiple orthogonal nonlinear principal components are found in both ends of the variance spectrum. The components which minimize variance are useful when studying whether sets naturally group together when trying to cancel each other. MACE can thus assist a data analyst in obtaining information on the structure and correlations of the involved variables by including nonlinear effects. Various nonlinear analyses of irregularly sampled stream sediment data are applied. From a geological perspective it seems as if a linear CCA that maximizes autocorrelation provides a better decomposition of the data. MACE may have problems due to the interdependencies of the involved variables or because the data density is to sparse. However, MACE is expected to be useful in (future) analyses in which the linear decompositioning is slowly relaxed to nonlinear transformations by means of decreasing the degree of regularization on the scatterplot smoothers applied by the algorithm.

Bibliography

[1] Alsabti, Ranka, and Singh. An efficient parallel algorithm for high dimensional similarity join. In *Proc. IPPS: 11th International Parallel Processing Symposium*. IEEE Computer Society Press, 1998.

[2] T. W. Anderson. *An Introduction to Multivariate Statistical Analysis*. John Wiley, New York, second edition, 1984. 675 pp.

[3] A. Baja. What criterion for a power algorithm? In *Festschrift on the Occasion of Peter J. Huber's 60'th Birthday*. Springer, 1995.

[4] A. Baraldi and P. Blonda. A survey of fuzzy clustering algorithms for pattern recognition. II. *IEEE Transactions on Systems, Man and Cybernetics, Part B (Cybernetics)*, 29:786–801, 1999.

[5] R. E. Bellman. *Adaptive Control Processes*. Princeton University Press, Princeton, NJ, 1961.

[6] J. M. F. Ten Berge. Orthogonal Procrustes rotation for two or more matrices. *Psychometrika*, 42:267–276, 1977.

[7] J. Besag. Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society, Series B*, 36:192–236, 1974.

[8] J. C. Bezdek, E. C. K. Tsao, and N. R. Pal. Fuzzy Kohonen clustering networks. In *Proc. IEEE International Conference on Fuzzy Systems*, pages 1035–1043, Piscataway, NJ, 1992. IEEE Service Center.

[9] James C. Bezdek. *Pattern Recognition with Fuzzy Objective Function Algorithms*. Plenum Press, New York, 1981.

[10] Magnus Borga. *Learning Multidimensional Signal Processing*. PhD thesis, Department of Electrical Engineering, Linköping University, Linköping, Sweden, 1998.

[11] C. A. Bouman and M. Shapiro. A multiscale random field model for Bayesian image segmentation. *IEEE Transactions on Image Processing*, 3(2):162–177, 1994.

[12] L. Breiman and J. H. Friedman. Estimating optimal transformations for multiple regression. In *Computer Science and Statistics. Proceedings of the Sixteenth Symposium on the Interface*, pages 121–34, 1985.

[13] L. Breiman and J. H. Friedman. Estimating optimal transformations for multiple regression and correlation. *Journal of the American Statistical Association*, 80(391):580–598, 1985.

[14] L. Breiman and J. H. Friedman. Rejoinder (to Pregibon and Vardia, Buja and Kass, and Fowlkes and Kettenring on Estimating optimal transformations for multiple regression and correlation). *Journal of the American Statistical Association*, 80(391):614–619, 1985.

[15] L. Breiman and S. Peters. Comparing automatic smoothers (a public service enterprise). *International Statistical Review*, 60(3):271–290, 1992.

[16] A. Buja. Remarks on functional canonical variates, alternating least squares methods and ACE. *The Annals of Statistics*, 18:1032–1069, 1990.

[17] A. Buja and R. E. Kass. Some observations on ACE methodology (discussion of Breiman and Friedman, Estimating optimal transformations for multiple regression and correlation). *Journal of the American Statistical Association*, 80:602–607, 1985.

[18] K. K. Chintalapudi and M. Kam. A noise-resistant fuzzy c-means algorithm for clustering. *IEEE Fuzzy Systems Proceedings*, 2:1458–1463, 1998.

[19] H. Cohn and M. Fielding. Simulated annealing: searching for an optimal temperature schedule. *SIAM Journal of Optimization*, 9:779–802, 1999.

[20] Knut Conradsen, Bjarne Kjær Nielsen, and Tage Thyrsted. A comparison of min/max autocorrelation factor analysis and ordinary factor analysis. In *Proceedings from Symposium in Applied Statistics*, pages 47–56, Lyngby, Denmark, January 1985.

[21] D. Cook, A. Buja, and J. Cabrera. Grand tour and projection pursuit. *Journal of Computational and Graphical Statistics*, 4(3):155–172, 1995.

[22] W. W. Cooley and P. R. Lohnes. *Multivariate Data Analysis*. John Wiley and Sons, New York, 1971.

[23] T. F. Cootes, G. J. Edwards, and C. J. Taylor. Active appearance models. In *Proc. European Conf. on Computer Vision*, volume 2,

pages 484–498. Springer, 1998.

[24] T. F. Cootes, G. J. Edwards, and C. J. Taylor. Active appearance models. *IEEE Trans. on Pattern Recognition and Machine Intelligence*, 23(6):681–685, 2001.

[25] T. F. Cootes, C. J. Taylor, D. H. Cooper, and J. Graham. Active shape models - their training and application. *Computer Vision and Image Understanding*, 61(1):38–59, 1995.

[26] R. Dave and R. Krishnapuram. Robust clustering methods: A unified view. *IEEE Transactions on Fuzzy Systems*, 5(2):270–293, May 1997.

[27] R. N. Dave. Characterization and detection of noise in clustering. *Pattern recognition letters*, 12(11):657–664, 1991.

[28] I. L. Dryden and K. V. Mardia. *Statistical Shape Analysis*. John Wiley, Chichester, 1998.

[29] W. M. Elsasser. *Atom and Organism: A new approach to theoretical biology*. Princeton University Press, Princeton, 1966.

[30] Bjarne K. Ersbøll. *Transformations and Classifications of Remotely Sensed Data: Theory and Geological Cases*. PhD thesis, Institute of Mathematical Statistics and Operations Research, Technical University of Denmark, Lyngby, 1989. 297 pp.

[31] M. Barni et al. A robust fuzzy clustering algortihm for the classification of remote sensing images. *IEEE (IGARSS) International Geoscience and Remote Sensing Symposium*, 2000.

[32] Nasa Facts. El Niño, the earth science enterprise series, nf.211, 1999.

[33] E. B. Fowlkes and J. R. Kettenring. The ACE method of optimal transformation (discussion of Breiman and Friedman, Estimating optimal transformations for multiple regression and correlation). *Journal of the American Statistical Association*, 80:607–613, 1985.

[34] C. Fraley and A. Raftery. How many clusters? which clustering method? answers via model-based cluster analysis. Technical Report 329, Department of Statistics, University of Washington, Seattle, WA, 1998.

[35] I. E. Frank and J. H. Friedman. A statistical view of some chemometrics regression tools. *Technometrics*, 35(2):109–135, 1993.

[36] Y. Freund and R. Schapire. A short introduction to boosting. *Society for Artificial Intelligence*, 14(5):771–780, 1999.

[37] J. Friedman and W. Stuetzle. Smoothing of scatterplots. Technical Report Orion 3, Dept. of Statistics, Stanford University., 1982.

[38] J. H. Friedman. Exploratory projection pursuit. *Journal of the American Statistical Association*, 82:249–266, 1987.

[39] J H Friedman and W Stuetzle. Projection pursuit regression. *Journal of the American Statistical Association*, 76:817–823, 1981.

[40] J. H. Friedman and R. Tibshirani. The monotone smoothing of scatterplots. *Technometrics*, 26:243–250, 1984.

[41] I. Gath and A. B. Geva. Unsupervised optimal fuzzy clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 11(7):773–781, July 1989.

[42] P. Geladi and B. R. Kowalski. Partial least-squares regression: a tutorial. *Analytica Chimica Acta*, 185:1–17, 1986.

[43] S. Geman and D. Geman. Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6:721–741, 1984.

[44] S. Geman and C. Graffigne. Markov random fields and their applications to computer vision. In A.M. Gleason, editor, *Proceedings of the International Congress of Mathematicians*, pages 1496–1517, Berkeley, California, 1987.

[45] Stuart Geman and Chii-Ruey Hwang. Diffusions for global optimization. *SIAM Journal on Control and Optimization*, 24(5):1031–1043, 1986.

[46] N. A. Gershenfeld. *The Nature of Mathematical Modeling*. University Press, Cambridge, 1999.

[47] B. Gidas. The langevin equation as a global minimization algorithm. In E. Bienenstock, F. Fogelman Souli'e, and G. Weisbuch, editors, *Disordered Systems and Biological Organization*, pages 321–326. Springer-Verlag, Berlin, 1986.

[48] E. D. Goldberg, editor. *North Sea Science*. The Massachusetts Institute of Technology, MIT, 1973.

[49] Colin Goodall. Procrustes methods in the statistical analysis of shape. *Journal of the Royal Statistical Society, Series B*, 53(2):285–339, 1991.

[50] J. C. Gower. Generalized Procrustes analysis. *Psychometrika*, 40:33–50, 1975.

[51] A. A. Green, M. Berman, P. Switzer, and M. D. Craig. Transformation for ordering multispectral data in terms of image quality with implications for noise removal. *IEEE Transactions on Geoscience and Remote Sensing*, 26(1):65–74, 1988.

[52] P. Green and B. Silverman. *Nonparametric Regression and Generalized Linear Models*. Chapman and Hall., 1994.

[53] Joseph C. Harsanyi and Chein-I Chang. Hyperspectral image classifi-

cation and dimensionality reduction: An orthogonal subspace projection approach. *IEEE Transactions on Geoscience and Remote Sensing*, 32(4):779–785, 1994.

[54] T. Hastie and R. Tibshirani. *Generalized additive models*. Chapman and Hall, 1990.

[55] T. Hastie, R. Tibshirani, and J. H. Friedman. *The Elements of Statistical Learning*. Springer, 2001.

[56] K. B. Hilger, A. A. Nielsen, O. B. Andersen, and P. Knudsen. An ACE-based nonlinear extension to traditional empirical orthogonal function analysis. In *MultiTemp, Venice*, 2001.

[57] K. B. Hilger, A. A. Nielsen, and R. Larsen. A scheme for initial exploratory data analysis of multivariate image data. In *Scandinavian Image Analysis Conference (SCIA)'01*, 2001.

[58] K. B. Hilger and M. B. Stegmann. Madcam - the multispectral active decomposition camera. In *Proc. 10th Danish Conference on Pattern Recognition and Image Analysis*, Copenhagen, Denmark, 2001.

[59] Klaus Baggesen Hilger and Allan Aasbjerg Nielsen. Targeting input data for change detection studies by suppression of undesired spectra. In *Proceedings of Seminar on Remote sensing and image analysis techniques for revision of topographic databases*, KMS, The National Survey and Cadastre, Copenhagen, Denmark, February 2000.

[60] Klaus Baggesen Hilger, Allan Aasbjerg Nielsen, Per Knudsen, and Ole B. Andersen. Enhancement of ocean related signal by suppression of undesired spectra in remotely sensed multivariate SeaWiFS images in the GEOSONAR project. In *American Geophysical Union (AGU) Fall Meeting*, San Francisco, California, USA, December 2000.

[61] P. M. Holligan, T. Aarup, and S. B. Groom. The north sea: Satellite colour atlas. *Continental Shelf Research*, 9(8):667–765, 1989.

[62] P. Horst. Relations among m sets of measures. *Psychometrika*, 26:129–149, 1961.

[63] A. Höskuldsson. PLS regression methods. *Journal of Chemometrics*, 2:211–228, 1986.

[64] Harold Hotelling. Analysis of a complex of statistical variables into principal components. *J. Educ. Psych.*, 24:417–441, 1933.

[65] Harold Hotelling. Relations between two sets of variates. *Biometrika*, XXVIII:321–377, 1936.

[66] SAS Institute Inc. *SAS/STAT User's Guide, ver. 6*. SAS Institute, 1990.

[67] Internet http://neptune.gsfc.nasa.gov/ocean.html. Goddard Space Flight Center, National Aeronautics and Space Administration, Greenbelt, Maryland, USA.

[68] Internet http://pao.gsfc.gov/. Goddard Space Flight Center, serviced by NASA.

[69] Internet http://podaac.jpl.nasa.gov. Jet Propulsion Laboratory, National Aeronautics and Space Administration, Pasadena, California, USA.

[70] Internet http://www.teknologisk.dk/251. COMB, the Industrial Centre for Surfacemicroscopy, Microanalysis and Image Analysis.

[71] A. Jain, P. Duin, and J. Mao. Statistical pattern recognition: A review. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(1):pp. 4–37, 2000.

[72] A.K. Jain and R.C. Dubes. *Algorithms For Clustering Data*. Prentice-Hall, Englewood Cliffs, New Jersy, 1988.

[73] L. Kaufman. Finding groups in data: an introduction to cluster analysis. In *Finding Groups in Data: An Introduction to Cluster Analysis*. Wiley, New York, 1990.

[74] J. R. Kettenring. Canonical analysis of several sets of variables. *Biometrika*, 58:433–451, 1971.

[75] P. Knudsen, O. B. Andersen, and T. Knudsen. ATSR sea surface temperature data in a global analysis with TOPEX/POSEIDON altimetry. *Geophysical Research Letters*, 23(8):821–824, 1996.

[76] T. Kohonen. The self-organising map. *Proceedings of the IEEE*, 78(9):1464–1480, 1990.

[77] T. Kohonen. *Self-Organizing Maps*. Springer-Verlag, 1995.

[78] R. Krishnapuram and J. M. Keller. The possibilistic c-means algorithm: Insights and recommendations. *IEEE Transactions on Fuzzy Systems*, 4(3), 1996.

[79] J. B. Kruskal. Non-metric multidimensional scaling: a numerical method. *Psychometrika*, 29:115–129, 1964.

[80] Rasmus Larsen. MAF and other transformations applied in remote sensing. Master's thesis, Institute of Mathematical Statistics and Operations Research, Technical University of Denmark, Lyngby, 1991. In Danish. 130 pp.

[81] Rasmus Larsen, Hrafnkell Eiriksson, and Mikkel B. Stegmann. Q-MAF shape decomposition. In *Medical Image Computing and Computer-Assisted Intervention - MICCAI 2001, 4th International Conference, Utrecht, The Netherlands*, volume 2208 of *Lecture Notes in Computer Science*, pages 837–844. Springer, 2001.

[82] T. Lindeberg. Scale-space: A framework for handling image structures at multiple scales. In *Proceedings for CERN School of Computing*, Egmond aan Zee, The Netherlands, September 1996.

[83] J. MacQueen. Some methods of classification and analysis of multivariate observations. In *The Fifth Berkeley Symposium on Mathemtical Statistics and Probability*, pages 281–297, 1967.

[84] S. Mika, G. Rätsch, J. Weston, B. Schölkopf, and K.-R. Müller. Fisher discriminant analysis with kernels. In Y.-H. Hu, J. Larsen, E. Wilson, and S. Douglas, editors, *Neural Networks for Signal Processing IX*, pages 41–48. IEEE, 1999.

[85] John W. V. Miller, James B. Farison, and Youngin Shin. Spatially invariant image sequences. *IEEE Transactions on Image Processing*, 1(2):148–161, 1992.

[86] L. Molgedey and H. G. Schuster. Separation of a mixture of independent signals using time delayed correlations, 1994.

[87] Bart De Moor. Generalizations of the OSVD: Structure, properties and applications. In R. J. Vaccaro, editor, *SVD and Signal Processing II*, pages 83–98. Elsevier Science Publisher, 1991.

[88] A. A. Nielsen. Stream sediment geochemistry in South Greenland: Data report. Technical report, Department of Mathematical Modelling, DTU, October 1992.

[89] A. A. Nielsen. *Analysis of Regularly and Irregularly Sampled Spatial, Multivariate, and Multi-temporal Data*. PhD thesis, Department of Mathematical Modelling, Technical University of Denmark, Lyngby, 1994. Internet http://www.imm.dtu.dk/∼aa/phd/.

[90] A. A. Nielsen. An extension to a filter implementation of local quadratic surface for image noise estimation. In *Proceedings of the 10th International Conference on Image Analysis and Processing (ICIAP'99)*, pages 119–124, Venice, Italy, 27-29 September 1999.

[91] A. A. Nielsen, K. Conradsen, J. L. Pedersen, and A. Steenfelt. Maximum autocorrelation factorial kriging. In W. J. Kleingeld and D. G. Krige, editors, *Geostats 2000 Capetown*, 2000.

[92] A. A. Nielsen, K. Conradsen, and J. J. Simpson. Multivariate alteration detection (MAD) and MAF post-processing in multispectral, bi-temporal image data: New approaches to change detection studies. *Remote Sensing of Environment*, 64:1–19, 1998.

[93] A. A. Nielsen, K. B. Hilger, O. B. Andersen, and P. Knudsen. A bivariate extension to traditional empirical orthogonal function analysis. In *MultiTemp, Venice*, 2001.

[94] A. A. Nielsen, K. B. Hilger, O. B. Andersen, and P. Knudsen. A temporal extension to traditional empirical orthogonal function analysis. In *MultiTemp, Venice*, 2001.

[95] Allan Aasbjerg Nielsen. Linear mixture models, full and partial unmixing in multi- and hyperspectral image data. In B. E. Ersbøll and P. Johansen, editors, *Proceedings of the Scandinavian Image Analysis Conference (SCIA'99), vol. 2*, pages 898–902, Kangerlussuaq, Greenland, 7-11 June 1999.

[96] Allan Aasbjerg Nielsen. Partial unmixing in hyperspectral image data. In ERIM, editor, *Proceedings from the Fourth International Airborne Remote Sensing Conference and Exhibition, Vol. II*, pages 535–542, Ottawa, Ontario, Canada, 21-24 June 1999.

[97] Allan Aasbjerg Nielsen. Multiset canonical correlations analysis and multispectral, truly multi-temporal remote sensing data. *IEEE Transactions on Image Processing*, November 2001.

[98] Allan Aasbjerg Nielsen. Spectral mixture analysis: Linear and semi-parametric full and iterated partial unmixing in multi- and hyperspectral image data. *Journal of Mathematical Imaging and Vision*, 15:17–37, 2001.

[99] Allan Aasbjerg Nielsen and Knut Conradsen. Multivariate alteration detection (MAD) in multispectral, bi-temporal image data: A new approach to change detection studies. Technical Report 1997-11, Department of Mathematical Modelling, Technical University of Denmark, 1997. Internet http://www.imm.dtu.dk/∼aa/tech-rep-1997-11/.

[100] Allan Aasbjerg Nielsen and Klaus Baggesen Hilger. Spectral-spatial decomposition of multivariate SeaWiFS images with suppression of undesired spectra and noise. In *Oceans from Space*, Venice, October 2000.

[101] Søren Ingvor Olsen. Estimation of noise in images: An evaluation. *Graphical Models and Image Processing*, 55(4):319–323, 1993.

[102] P. Perona and J. Malik. Scale-space and edge detection using anisotropic diffusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12(7):629–639, 1990.

[103] D. Pregibon and Y. Vardi. Comment (discussion of Breiman and Friedman, Estimating optimal transformations for multiple regression and correlation). *Journal of the American Statistical Association*, 80:598–601, 1985.

[104] R. W. Preisendorfer. *Principal Component Analysis in Meteorology*

and Oceanography. posthumously compiled and edited by C. D. Morley. Developments in Atmospheric Science, 17, Elsevier, 1988.

[105] W. Press, B. Flannery, S. Teukolsky, W. Vetterling, and C. Recipes. *Numerical Recipes in C*. Cambridge University Press, 1988.

[106] M. Razaee, P. van der Zwet, B. Lelieveldt, R. van der Geest, and J. Reiber. A multiresolution image segmentation technique based on pyr amidal segmentation and fuzzy clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 9(7):1238–1248, July 2000.

[107] B. D. Ripley. *Pattern Recognition and Neural Networks*. Cambridge University Press, 1996.

[108] J. W. Sammon, Jr. A non-linear mapping for data structure analysis. *IEEE Transactions on Computers*, 18:401–409, 1969.

[109] Bernhard Schölkopf, Alexander Smola, and Klaus-Robert Muller. Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation*, 10(5):1299–1319, 1998.

[110] Bernhart Schölkopf. Statistical learning and kernel methods. Technical Report MSR-TR-2000-23, Microsoft Research Limited, February 2000.

[111] A. Scott. *Stairway to the Mind: The Controversial New Science of Consciousness*. Springer-Verlag (Copernicus), 1995.

[112] SeaWiFS. http://seawifs.gsfc.nasa.gov/SEAWIFS.html. Goddard Space Flight Center, National Aeronautics and Space Administration, Greenbelt, Maryland, USA.

[113] M. B. Stegmann. Active appearance models: Theory, extensions and cases. Master's thesis, Informatics and Mathematical Modelling, Technical University of Denmark, Lyngby, 2000. http://www.imm.dtu.dk/∼aam/.

[114] P. Switzer and A. A. Green. Min/max autocorrelation factors for multivariate spatial imagery. Technical Report 6, Dept. of statistics, Stanford University, 1984.

[115] P. Switzer and S. E. Ingebritsen. Ordering of time-difference data from multispectral imagery. *Remote Sensing of Environment*, 20:85–94, 1986.

[116] R. Tibshirani. Estimating transformations for regression via additivity and variance stabilization. *Journal of the American Statistical Association*, 83(402):394–405, June 1988.

[117] W. Torgerson. Multidimensional scaling: I. Theory and method. *Psychometrika*, 17:401–419, 1952.

[118] D. Tran and M. Wagner. Fuzzy entropy clustering. In *Proceedings of Fuzzy Systems Conf, IEEE*, 2000.

[119] J. Tukey. *Exploratory Data Analysis*. AddisonWesley Publishing Co, 1977.

[120] Internet http://www.forskraad.dk/saerforsk/jordobs/joobs0.html. The Earth Observation Programmes funded by the joint Danish Research Councils.

[121] Vidal V. V. Applied simulated annealing. In *Lect. Notes in Econom. and Math.Syst.*, volume 396. Springer Verlag, Berlin, 1993.

[122] H. van der Vorst. Subspace iteration for eigenproblems. *CWI Quarterly*, 9:151–160, 1996.

[123] Vladimir N. Vapnik. *The nature of statistical learning theory*. Springer Verlag, Heidelberg, 1995.

[124] W. N. Venables and B. D. Ripley. *Modern Applied Statistics with S-Plus*. Springer, third edition, 2000.

[125] Joachim Weickert. A review of nonlinear diffusion filtering. In *Scale-Space Theories in Computer Vision*, pages 3–28, 1997.

[126] R. Wiemker. Unsupervised fuzzy classification of multispectral imagery using spatial-spectral features. In I. Balderjahn, R. Mathar, and M. Schader, editors, *Data Highways and Information Flooding, A Challenge for Classification and Data Analysis*. Springer, 1997.

[127] R. Wilson and C.-T. Li. Hidden multiresolution random fields and their application to image segmentation. In *Proceedings of the 10th International Conference on Image Analysis and Processing (ICIAP'99)*, pages 346–351, Venice, Italy, 27-29 September 1999.

[128] K. Windfeld. *Application of Computer Intensive Data Analysis: Methods to the Analysis of Digital Images and Spatial Data*. PhD thesis, Department of Mathematical Modelling, Technical University of Denmark, Lyngby, 1992.

[129] S. Wold, A. Ruhe, H. Wold, and W. J. Dunn. The collinearity problem in linear regression. The partial least squares (PLS) approach to generalized inverses. *SIAM Journal of Scientific Statistical Computing*, 5(3):735–743, 1984.

[130] W. Pitts W.S. McCulloch. A logical calculus of ideas immanent in nervous activity. *Bulletin of Mathematical Biophysics*, 5:115–133, 1943.

[131] F.Y. Wu. The Potts model. *Reviews of Modern Physics*, 54(1):235–268, 1982.

Ph.D. theses from IMM

1. **Larsen, Rasmus.** (1994). *Estimation of visual motion in image sequences. xiv* + 143 pp.
2. **Rygard, Jens Moberg.** (1994). *Design and optimization of flexible manufacturing systems. xiii* + 232 pp.
3. **Lassen, Niels Christian Krieger.** (1994). *Automated determination of crystal orientations from electron backscattering patterns. xv* + 136 pp.
4. **Melgaard, Henrik.** (1994). *Identification of physical models. xvii* + 246 pp.
5. **Wang, Chunyan.** (1994). *Stochastic differential equations and a biological system. xxii* + 153 pp.
6. **Nielsen, Allan Aasbjerg.** (1994). *Analysis of regularly and irregularly sampled spatial, multivariate, and multi-temporal data. xxiv* + 213 pp.
7. **Ersbøll, Annette Kjær.** (1994). *On the spatial and temporal correlations in experimentation with agricultural applications. xviii* + 345 pp.
8. **Møller, Dorte.** (1994). *Methods for analysis and design of heterogeneous telecommunication networks.* Volume 1-2, *xxxviii* + 282 pp., 283-569 pp.
9. **Jensen, Jens Christian.** (1995). *Teoretiske og eksperimentelle dynamiske undersøgelser af jernbanekøretøjer. viii* + 174 pp.
10. **Kuhlmann, Lionel.** (1995). *On automatic visual inspection of reflective surfaces.* Volume 1, *xviii* + 220 pp., (Volume 2, *vi* + 54 pp., fortrolig).
11. **Lazarides, Nikolaos.** (1995). *Nonlinearity in superconductivity and Josephson Junctions. iv* + 154 pp.
12. **Rostgaard, Morten.** (1995). *Modelling, estimation and control of fast sampled dynamical systems. xiv* + 348 pp.
13. **Schultz, Nette.** (1995). *Segmentation and classification of biological objects. xiv* + 194 pp.
14. **Jørgensen, Michael Finn.** (1995). *Nonlinear Hamiltonian systems. xiv* + 120 pp.
15. **Balle, Susanne M.** (1995). *Distributed-memory matrix computations. iii* + 101 pp.
16. **Kohl, Niklas.** (1995). *Exact methods for time constrained routing and related scheduling problems. xviii* + 234 pp.
17. **Rogon, Thomas.** (1995). *Porous media: Analysis, reconstruction and percolation. xiv* + 165 pp.
18. **Andersen, Allan Theodor.** (1995). *Modelling of packet traffic with matrix analytic methods. xvi* + 242 pp.
19. **Hesthaven, Jan.** (1995). *Numerical studies of unsteady coherent structures and transport in two-dimensional flows.* Risø-R-835(EN) 203 pp.
20. **Slivsgaard, Eva Charlotte.** (l995). *On the interaction between wheels and rails in railway dynamics. viii* + 196 pp.
21. **Hartelius, Karsten.** (1996). *Analysis of irregularly distributed points. xvi* + 260 pp.
22. **Hansen, Anca Daniela.** (1996). *Predictive control and identification - Applications to steering dynamics. xviii* + 307 pp.
23. **Sadegh, Payman.** (1996). *Experiment design and optimization in complex systems. xiv* + 162 pp.
24. **Skands, Ulrik.** (1996). *Quantitative methods for the analysis of electron microscope images. xvi* + 198 pp.
25. **Bro-Nielsen, Morten.** (1996). *Medical image registration and surgery simulation. xxvii* + 274 pp.
26. **Bendtsen, Claus.** (1996). *Parallel numerical algorithms for the solution of systems of ordinary differential equations. viii* + 79 pp.
27. **Lauritsen, Morten Bach.** (1997). *Delta-domain predictive control and identification for control. xxii* + 292 pp.
28. **Bischoff, Svend.** (1997). *Modelling colliding-pulse mode-locked semiconductor lasers. xxii* + 217 pp.
29. **Arnbjerg-Nielsen, Karsten.** (1997). *Statistical analysis of urban hydrology with special emphasis on rainfall modelling.* Institut for Miljøteknik, DTU. *xiv* + 161 pp.
30. **Jacobsen, Judith L.** (1997). *Dynamic modelling of processes in rivers affected by precipitation runoff. xix* + 213 pp.

31. **Sommer, Helle Mølgaard.** (1997). *Variability in microbiological degradation experiments - Analysis and case study. xiv* + 211 pp.
32. **Ma, Xin.** (1997). *Adaptive extremum control and wind turbine control. xix* + 293 pp.
33. **Rasmussen, Kim Ørskov.** (1997). *Nonlinear and stochastic dynamics of coherent structures. x* + 215 pp.
34. **Hansen, Lars Henrik.** (1997). *Stochastic modelling of central heating systems. xxii* + 301 pp.
35. **Jørgensen, Claus.** (1997). *Driftsoptimering på kraftvarmesystemer.* 290 pp.
36. **Stauning, Ole.** (1997). *Automatic validation of numerical solutions. viii* + 116 pp.
37. **Pedersen, Morten With.** (1997). *Optimization of recurrent neural networks for time series modeling. x* + 322 pp.
38. **Thorsen, Rune.** (1997). *Restoration of hand function in tetraplegics using myoelectrically controlled functional electrical stimulation of the controlling muscle. x* + 154 pp. + Appendix.
39. **Rosholm, Anders.** (1997). *Statistical methods for segmentation and classification of images. xvi* + 183 pp.
40. **Petersen, Kim Tilgaard.** (1997). *Estimation of speech quality in telecommunication systems. x* + 259 pp.
41. **Jensen, Carsten Nordstrøm.** (1997). *Nonlinear systems with discrete and continuous elements.* 195 pp.
42. **Hansen, Peter S.K.** (1997). *Signal subspace methods for speech enhancement. x* + 226 pp.
43. **Nielsen, Ole Møller.** (1998). *Wavelets in scientific computing. xiv* + 232 pp.
44. **Kjems, Ulrik.** (1998). *Bayesian signal processing and interpretation of brain scans. iv* + 129 pp.
45. **Hansen, Michael Pilegaard.** (1998). *Metaheuristics for multiple objective combinatorial optimization. x* + 163 pp.
46. **Riis, Søren Kamaric.** (1998). *Hidden markov models and neural networks for speech recognition. x* + 223 pp.
47. **Mørch, Niels Jacob Sand.** (1998). *A multivariate approach to functional neuro modeling. xvi* + 147 pp.
48. **Frydendal, Ib.** (1998.) *Quality inspection of sugar beets using vision. iv* + 97 pp. + app.
49. **Lundin, Lars Kristian.** (1998). *Parallel computation of rotating flows. viii* + 106 pp.

50. **Borges, Pedro.** (1998). *Multicriteria planning and optimization. - Heuristic approaches. xiv* + 219 pp.
51. **Nielsen, Jakob Birkedal.** (1998). *New developments in the theory of wheel/rail contact mechanics. xviii* + 223 pp.
52. **Fog, Torben.** (1998). *Condition monitoring and fault diagnosis in marine diesel engines. xii* + 178 pp.
53. **Knudsen, Ole.** (1998). *Industrial vision. xii* + 129 pp.
54. **Andersen, Jens Strodl.** (1998). *Statistical analysis of biotests. - Applied to complex polluted samples. xx* + 207 pp.
55. **Philipsen, Peter Alshede.** (1998). *Reconstruction and restoration of PET images. vi* + 132 pp.
56. **Thygesen, Uffe Høgsbro.** (1998). *Robust performance and dissipation of stochastic control systems.* 185 pp.
57. **Hintz-Madsen, Mads.** (1998). *A probabilistic framework for classification of dermatoscopic images. xi* + 153 pp.
58. **Schramm-Nielsen, Karina.** (1998). *Environmental reference materials methods and case studies. xxvi* + 261 pp.
59. **Skyggebjerg, Ole.** (1999). *Acquisition and analysis of complex dynamic intra- and intercellular signaling events.* 83 pp.
60. **Jensen, Kåre Jean.** (1999). *Signal processing for distribution network monitoring. xv* + 199 pp.
61. **Folm-Hansen, Jørgen.** (1999). *On chromatic and geometrical calibration. xiv* + 238 pp.
62. **Larsen, Jesper.** (1999). *Parallelization of the vehicle routing problem with time windows.xx* + 266 pp.
63. **Clausen, Carl Balslev.** (1999). *Spatial solitons in quasi-phase matched struktures. vi* + (flere pag.)
64. **Kvist, Trine.** (1999). *Statistical modelling of fish stocks. xiv* + 173 pp.
65. **Andresen, Per Rønsholt.** (1999). *Surface-bounded growth modeling applied to human mandibles. xxii* + 125 pp.
66. **Sørensen, Per Settergren.** (1999). *Spatial distribution maps for benthic communities.*
67. **Andersen, Helle.** (1999). *Statistical models for standardized toxicity studies. viii* + (flere pag.)
68. **Andersen, Lars Nonboe.** (1999). *Signal processing in the dolphin sonar system. xii* + 214 pp.
69. **Bechmann, Henrik.** (1999). *Modelling of wastewater systems. xviii* + 161 pp.

70. **Nielsen, Henrik Aalborg.** (1999). *Parametric and non-parametric system modelling. xviii* + 209 pp.
71. **Gramkow, Claus.** (1999). *2D and 3D object measurement for control and quality assurance in the industry. xxvi* + 236 pp.
72. **Nielsen, Jan Nygaard.** (1999). *Stochastic modelling of dynamic systems. xvi* + 225 pp.
73. **Larsen, Allan.** (2000). *The dynamic vehicle routing problem. xvi* + 185 pp.
74. **Halkjær, Søren.** (2000). *Elastic wave propagation in anisotropic inhomogeneous materials. xiv* + 133 pp.
75. **Larsen, Theis Leth.** (2000). *Phosphorus diffusion in float zone silicon crystal growth. viii* + 119 pp.
76. **Dirscherl, Kai.** (2000). *Online correction of scanning probe microscopes with pixel accuracy.* 146 pp.
77. **Fisker, Rune.** (2000). *Making deformable template models operational. xx* + 217 pp.
78. **Hultberg, Tim Helge.** (2000).*Topics in computational linear optimization. xiv* + 194 pp.
79. **Andersen, Klaus Kaae.** (2000).*Stochastic modelling of energy systems. xiv* + 191 pp.
80. **Thyregod, Peter.** (2001). *Modelling and monitoring in injection molding. xvi* + 132 pp.
81. **Schjødt-Eriksen, Jens.** (2001). *Arresting of collapse in inhomogeneous and ultrafast Kerr media.*
82. **Bennetsen, Jens Christian.** (2000). *Numerical simulation of turbulent airflow in livestock buildings. xi* + 205 pp + Appendix.
83. **Højen-Sørensen, Pedro A.d.F.R.** (2001). *Approximating methods for intractable probabilistic models: - Applications in neuroscience. xi* + 104 pp + Appendix.
84. **Nielsen, Torben Skov.** (2001). *On-line prediction and control in non-linear stochastic systems. xviii* + 242 pp.
85. **Öjelund, Henrik.** (2001). *Multivariate calibration of chemical sensors. xviii* + 184 pp.
86. **Adeler, Pernille Thorup.** (2001). *Hemodynamic simulation of the heart using a 2D model and MR data. xv* + 180 pp.
87. **Nielsen, Finn Årup.** (2001). *Neuroinformatics in functional neuroimaging.* 330 pp.
88. **Kidmose, Preben.** (2001). *Blind separation of heavy tail signals. viii* + 136 pp.

89. **Hilger, Klaus Baggesen.** (2001). *Exploratory analysis of multivariate data. xxiv* + 186 pp.