Российская Академия Наук

РОССИЙСКАЯ АКАДЕМИЯ НАУК

ИНСТИТУТ ПРОБЛЕМ БЕЗОПАСНОГО РАЗВИТИЯ АТОМНОЙ ЭНЕРГЕТИКИ

RUSSIAN ACADEMY OF SCIENCES

NUCLEAR SAFETY INSTITUTE

Препринт ИБРАЭ № IBRAE-2001-18

Preprint IBRAE-2001-18

R. Parkin, M. Kanevski, M. Maignan, G. Raspa, T. Hakamata, E. Savelieva, E. Kalantarov

PRINCIPAL COMPONENT ANALYSIS OF SPATIAL DATA

Москва
2001

Moscow
2001

УДК 502.3

Паркин Р., Каневский М., Майгнан М., Распа Дж., Хакамата Т., Савельева Е., Калантаров Е. АНАЛИЗ ПРИНЦИПИАЛЬНЫХ КОМПОНЕНТ ДЛЯ ПРОСТРАНСТВЕННЫХ ДАННЫХ. Препринт № IBRAE-2001-18. Москва: Институт проблем безопасного развития атомной энергетики РАН, 2001. 21 с. — Библиогр.: 41 назв.

Аннотация

В работе представлен обзор различных методов анализа принципиальных компонент и применение одного из этих методов к пространственным данным с использованием программного пакета “Multigeo”. Анализ принципиальных компонент является одним из самых распространенных методов обработки, сжатия и визуализации данных большой размерности, хотя эффективность данного метода ограничена его линейностью. Данный подход был применен к данным по загрязнению донных отложений Женевского озера и территории Японии тяжелыми металлами.

Parkin R., Kanevski M., Maignan M., Raspa G., Hakamata T., Savelieva E. Kalantarov E. PRINCIPAL COMPONENT ANALYSIS OF SPATIAL DATA. Preprint IBRAE-2001-18. Moscow: Nuclear Safety Institute RAS, 2001. 21 p. — Refs.: 41 items.

Abstract

This work presents review of various Principal Component Analysis (PCA) methods and application PCA for the spatial prediction of concentration of metals using software “Multigeo”. Principal component analysis is one of the most popular techniques for processing, compressing and visualizing data, although its effectiveness is limited by its global linearity. The technique is illustrated using the real data on Geneva Lake sediments contamination and Japanese soil contamination by heavy metals.

Principal Component Analysis of Spatial Data

R. Parkin, M. Kanevski, M. Maignan (1), G. Raspa (2), T. Hakamata (3), E. Savelieva, E. Kalantarov

Institute of Nuclear Safety (IBRAE)
B. Tulskaya 52, 113191 Moscow, Russia
phone: (095) 955-22-31, fax: (095) 958-11-51, e-mail: park@ibrae.ac.ru

(1) University of Lausanne, Switzerland

(2) University La Sapienza, Roma, Italy

(3) NIAES Nippon Agro-Environmental Institute of Environmental Studies,Tsukuba, Japan,

Contents

1 Introduction

High dimensional data analysis is becoming increasingly common. With high dimensional data, it is difficult to understand the underlying structure: it is difficult to "see the wood for the trees." Additionally, the storage, transmission and processing of high dimensional data place great demands on systems. Hence, it is desirable to reduce the dimensionality of the data, whilst maintaining as much of its original structure [4].

Since the beginning of last century, several researchers (see, for example, [5, 6, 7, 8]) have developed dimensionality reduction techniques. Principal component analysis (PCA) is one of these important techniques.

In mathematical terms, n correlated random variables are transformed into a set of $d \leq n$ uncorrelated variables. These uncorrelated variables are linear combinations of the original variables and can be used to express the data in a reduced form. Data modeling and pattern recognition are better able to work on this reduced form, and the form is efficient for storage and transmission. PCA is also sometimes used as a data visualization technique since high dimensional datasets can be reduced to a low dimension and then plotted [4].

Thus, Principal Component Analysis is widely used for several different applications, below are some examples:

- Noise reduction
- Data Compression
- Visualization oh high dimensional data

Principal component analysis (PCA) is one of the most popular techniques for processing, compressing and visualizing data, although its effectiveness is limited by its global linearity.

In this paper Principal Component Analysis (PCA) was carried out and obtained results were discussed. Case studies are based on the real data on Geneva Lake sediments contamination and Japanese soil contamination by heavy metals.

2 Principal Component Analysis review

The various types of methods have been used for PCA. There are the more conventional *matrix method*s. In these methods all the data are used to calculate the variance-covariance structure and express it in a matrix. Most multivariate analysis textbooks (for example, [2, 3, 13, 14, 15 and 16]) describe matrix methods for performing PCA. The goal is to find the eigenvectors of the covariance matrix. These eigenvectors correspond to the directions of the principal components of the original data, their statistical significance is given by their corresponding eigenvalues [4,11]. In recent years, the QR algorithm has been the most widely used algorithm for calculating the complete set of eigenvalues of a matrix [17, 18]. Cyclic Jacobi methods are particularly suited for implementation in a parallel computer [17, 18]. The divide and conquer method of Cuppen is a relatively new method for calculating the complete eigensystem of a symmetric, tridiagonal matrix [17]. The singular value decomposition (SVD) of a real, symmetric, positive semidefinite matrix is equivalent to the orthogonal decomposition in terms of eigenvalues/eigenvectors [19]. Therefore, algorithms for computing the SVD can also be used for PCA. The power method and its variants [20, 21] are some of the simplest techniques for finding a few of the dominant eigenvalue/eigenvector [12].

PCA's can be also neurally realized (for example, [24, 25, 26, 27, 28, 29]). The PCA network used in [22] is a one layer feedforward neural network which is able to extract the principal components of the stream of input vectors. Typically, Hebbian type learning rules are used, based on the one unit learning algorithm originally proposed by Oja [26]. Many different versions and extensions of this basic algorithm have been proposed during the recent years (see [28, 29, 30, 31]). The structure of the PCA NN can be summarized as follows: there is one input layer, and one forward layer of neurons totally connected to the inputs; during the learning phase there are feedback links among neurons, that classify the network structure as either hierarchical or symmetric. After the learning phase the network becomes purely feedforward. The hierarchical case leads to the well known GHA algorithm [28, 30]; in the symmetric case we have the Oja's subspace network [26]. PCA neural algorithms can be derived from optimization problems, such as variance maximization and representation error minimization. We can generalize these problems to nonlinear problems, getting nonlinear algorithms (and relative networks) [22]. These have the same structure of the linear ones: either hierarchical or symmetric. These learning algorithms can be further classified in: robust PCA algorithms and nonlinear PCA algorithms [29, 30]. We define robust PCA so that the objective function grows less than quadratically [22]. The non linear learning function appears at selected places only. In nonlinear PCA algorithms all the outputs of the neurons are nonlinear function of the responses [22].

Principal Component Analysis has found wide applications in various disciplines: psychology [32, 33], genetics [34], pattern recognition [35], remote sensing [36] and seismic data analysis [37, 38]. PCA has been used for data mining and detecting linear associative rules [39]. Merz and Pazzani [40] have used a PCA-based technique for combining regression estimates [22]. A maximum-likelihood-based framework for constructing mixture models of PCA is proposed by Tipping and Bishop [41], [22].

3 Theory

3.1 Principal Component Analysis

Principal component analysis is the most widely used method of multivariate data analysis owing to the simplicity of its algebra and to its straightforward interpretation.

A linear transformation is defined, which transforms a set of correlated variables into uncorrelated factors. These orthogonal factors can be shown to extract successively a maximal part of the total variance of the variables. A graphical display can be produced which shows the position of the variables in the plane spanned by two factors [2].

3.1.1 Transformation into factors

The basic problem solved by principal component analysis is to transform a set of correlated variables into uncorrelated quantities, which could be interpreted in an ideal (multi-Gaussian) context as independent factors underlying the phenomenon. That is why the uncorrelated quantities are called *factors*, although such an interpretation is not always perfectly adequate [2].

Z is the $n \times N$ matrix of data from which the means of the variables have already been subtracted. The corresponding $N \times N$ variance-covariance matrix **V** then is

$$\mathbf{V} = [\sigma_{ij}] = \frac{1}{n}\mathbf{Z}^{\mathrm{T}}\mathbf{Z}$$

Let **Y** be an $n \times N$ matrix containing in its rows the *n* samples of factors Y_p (p=1,…, N), which are uncorrelated and of zero mean.

The variance-covariance matrix of the factors is diagonal, owing to the fact that the covariances between factors are nil by definition

$$\mathbf{D} = \frac{1}{n}\mathbf{Y}^{\mathrm{T}}\mathbf{Y} = \begin{vmatrix} d_{11} & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & d_{NN} \end{vmatrix}$$

and the diagonal elements d_{pp} are the variances of the factors.

A matrix **A** is sought, $N \times N$ orthogonal, which linearly transforms the measured variables into synthetic factors

$$\mathbf{Y} = \mathbf{ZA} \qquad \text{with } \mathbf{A}^{\mathrm{T}}\mathbf{A} = \mathbf{I}$$

Multiplying this equation from the left by 1/*n* and $\mathbf{Y}^{\mathrm{T}}$, we have

$$\frac{1}{n}\mathbf{Y}^{\mathrm{T}}\mathbf{Y} = \frac{1}{n}\mathbf{Y}^{\mathrm{T}}\mathbf{ZA}$$

and replacing **Y** by **ZA** on the right hand side, it follows

$$\frac{1}{n}(\mathbf{ZA})^{\mathrm{T}}(\mathbf{ZA}) = \frac{1}{n}\mathbf{A}^{\mathrm{T}}\mathbf{Z}^{\mathrm{T}}\mathbf{ZA} = \mathbf{A}^{\mathrm{T}}\frac{1}{n}(\mathbf{Z}^{\mathrm{T}}\mathbf{Z})\mathbf{A}$$

Finally

$$\mathbf{D} = \mathbf{A}^{\mathrm{T}}\mathbf{VA}$$

that is

$$\mathbf{VA} = \mathbf{AD}$$

It can immediately be seen that the matrix **Q** of orthonormal eigenvectors of **V** offers a solution to the problem and that the eigenvalues λ_p are then simply the variances of the factors Y_p. Principal component analysis is nothing else than a statistical interpretation of the eigenvalue problem

$$\mathbf{VQ} = \mathbf{Q\Lambda} \qquad \text{with } \mathbf{Q}^{\mathrm{T}}\mathbf{Q} = \mathbf{I}$$

defining the factors as

$$\mathbf{Y} = \mathbf{ZQ}$$

3.1.2 Maximization of the variance of a factor

Another important aspect of principal component analysis is that it allows to define a sequence of orthogonal factors, which successively absorb a maximal amount of the variance of the data [2].

Take a vector $\mathbf{y}_1$ corresponding to the first factor obtained by transforming the centered data matrix **Z** with a vector $\mathbf{a}_1$ calibrated to unit length

$$\mathbf{y}_1 = \mathbf{Z}\mathbf{a}_1 \qquad \text{with } \mathbf{a}_1^{\mathrm{T}}\mathbf{a}_1 = 1$$

The variance of $\mathbf{y}_1$ is

$$\mathrm{var}(\mathbf{y}_1) = \frac{1}{n}\mathbf{y}_1^{\mathrm{T}}\mathbf{y}_1 = \frac{1}{n}\mathbf{a}_1^{\mathrm{T}}\mathbf{Z}^{\mathrm{T}}\mathbf{Z}\mathbf{a}_1 = \mathbf{a}_1^{\mathrm{T}}\mathbf{V}\mathbf{a}_1$$

To attribute a maximal part of the variance of the data to $\mathbf{y}_1$, we define an objective function ϕ_1 with a Lagrange parameter λ_1, which multiplies the constraint that the transformation vector $\mathbf{a}_1$ should be of unit norm

$$\phi_1 = \mathbf{a}_1^{\mathrm{T}}\mathbf{V}\mathbf{a}_1 - \lambda_1(\mathbf{a}_1^{\mathrm{T}}\mathbf{a}_1 - 1)$$

Setting the derivative with respect to $\mathbf{a}_1$ to zero

$$\frac{\partial \phi_1}{\partial \mathbf{a}_1} = 0 \quad \Leftrightarrow \quad 2\mathbf{V}\mathbf{a}_1 - 2\lambda_1\mathbf{a}_1 = 0$$

we see that λ_1 is an eigenvalue of the variance-covariance matrix and that $\mathbf{a}_1$ is equal to the eigenvector $\mathbf{q}_1$ associated with this eigenvalue

$$\mathbf{V}\mathbf{q}_1 = \lambda_1\mathbf{q}_1$$

We are interested in a second vector $\mathbf{y}_2$ orthogonal to the first

$$\mathrm{cov}(\mathbf{y}_2, \mathbf{y}_1) = \mathrm{cov}(\mathbf{Z}\mathbf{a}_2, \mathbf{Z}\mathbf{a}_1) = \mathbf{a}_2^{\mathrm{T}}\mathbf{V}\mathbf{a}_1 = \mathbf{a}_2^{\mathrm{T}}\lambda_1\mathbf{a}_1 = 0$$

The function ϕ_2 to maximize incorporates two constraints: the fact that $\mathbf{a}_2$ should be unit norm and the orthogonality between $\mathbf{a}_2$ and $\mathbf{a}_1$. These constraints bring up two new Lagrange multipliers λ_2 and μ

$$\phi_2 = \mathbf{a}_2^{\mathrm{T}}\mathbf{V}\mathbf{a}_2 - \lambda_2(\mathbf{a}_2^{\mathrm{T}}\mathbf{a}_2 - 1) + \mu\mathbf{a}_2^{\mathrm{T}}\mathbf{a}_1$$

Setting the derivative with respect to $\mathbf{a}_2$ to zero

$$\frac{\partial \phi_2}{\partial \mathbf{a}_2} = 0 \quad \Leftrightarrow \quad 2\mathbf{V}\mathbf{a}_2 - 2\lambda_2\mathbf{a}_2 + \mu\mathbf{a}_1 = 0$$

What is the value of μ? Multiplying the equation by $\mathbf{a}_1^{\mathrm{T}}$ from the left

$$2\underbrace{\mathbf{a}_1^{\mathrm{T}}\mathbf{V}\mathbf{a}_2}_{0} - 2\lambda_2\underbrace{\mathbf{a}_1^{\mathrm{T}}\mathbf{a}_2}_{0} + \mu\underbrace{\mathbf{a}_1^{\mathrm{T}}\mathbf{a}_1}_{1} = 0$$

we see that μ is nil (the constraint is not active) and thus

$$\mathbf{V}\mathbf{a}_2 = \lambda_2\mathbf{a}_2$$

Again λ_2 turns out to be an eigenvalue of the variance-covariance matrix and $\mathbf{a}_2$ is the corresponding eigenvector $\mathbf{q}_2$. Continuing in the same way we find the rest of the *N* eigenvalues and eigenvectors of **V** as an answer to our maximization problem [2].

3.1.3 Interpretation of the factor variances

Numbering the eigenvalues of **V** from the largest to the lowest, we obtain a sequence of *N* uncorrelated factors, which provide an optimal decomposition (in the least squares sense) of the total variance as

$$\mathrm{tr}(\mathbf{V}) = \sum_{i=1}^{N}\sigma_{ii} = \sum_{p=1}^{N}\lambda_p$$

The eigenvalues indicate the amount of the total variance associated with each factor and the ratio

$$\frac{\text{variance of the factor}}{\text{total variance}} = \frac{\lambda_p}{\mathrm{tr}(\mathbf{V})}$$

gives a numerical indication, usually expressed in %, of the importance of the factor.

Generally it is preferable to standardize the variables (subtracting the means and dividing by the standard deviations), so that the principal component analysis is performed on the correlation matrix **R**. In this framework, when an eigenvalue is lower than 1, we may consider that the associated factor has less explanatory value than any single variable, as its variance is inferior to the unit variance of each variable [2].

3.1.4 Correlation of the variables with the factors

In general it is preferable to work with standardized variables $\tilde{z}_{\alpha i}$ to set them on a common scale and make them comparable

$$\tilde{z}_{\alpha i} = \frac{z_{\alpha i} - m_i}{\sqrt{\sigma_{ii}}}$$

where m_i and $\sqrt{\sigma_{ii}}$ are the mean and the standard deviation of the variable $z_{\alpha i}$ [2].

The variance-covariance matrix associated to standardized data is the correlation matrix

$$\mathbf{R} = \frac{1}{n}\tilde{\mathbf{Z}}^{\mathrm{T}}\tilde{\mathbf{Z}}$$

which can be decomposed using its eigensystem as

$$\mathbf{R} = \tilde{\mathbf{Q}}\tilde{\mathbf{\Lambda}}\tilde{\mathbf{Q}}^{\mathrm{T}} = \tilde{\mathbf{Q}}\sqrt{\tilde{\mathbf{\Lambda}}}\left(\tilde{\mathbf{Q}}\sqrt{\tilde{\mathbf{\Lambda}}}\right)^{\mathrm{T}} = \tilde{\mathbf{A}}\tilde{\mathbf{A}}^{\mathrm{T}}$$

The vectors $\tilde{\mathbf{a}}_i$, columns of $\tilde{\mathbf{A}}^{\mathrm{T}}$, are remarkable in the sense that they indicate the correlations between a variable $\tilde{\mathbf{z}}_i$ and the factors $\mathbf{y}_p$ because

$$\mathrm{corr}(\tilde{\mathbf{z}}_i, \mathbf{y}_p) = \sqrt{\tilde{\lambda}_p}\,\tilde{q}_{ip} = \tilde{a}_{ip}$$

The vectors $\tilde{\mathbf{a}}_i$ are of unit length and their cross product is equal to the correlation coefficient

$$\tilde{\mathbf{a}}_i^{\mathrm{T}}\tilde{\mathbf{a}}_j = \rho_{ij}$$

Owing to their geometry the vectors $\tilde{\mathbf{a}}_i$ can be used to represent the position of the variables on the surface of the unit hypersphere centered at the origin. The correlation coefficients ρ_{ij} are the cosines of the angles between the vectors referring to two different variables.

The projection of the position of the variables on the surface of the hypersphere towards a plane defined by a pair of axes of factors yields a graphical representation called the *circle of correlations.* The circle of correlations shows the proximity of the variables inside a unit circle and is useful to evaluate the affinities and the antagonisms between the variables. Statements can easily be made about variables, which are located near the circumference of the unit circle because the proximities in 2-dimensional space then correspond to proximities in *N*-dimensional space. For the variables located near the center of a unit circle it is necessary to check whether the proximities really correspond to proximities on the hypersphere by looking at the correlation circles generated by other factor pairs [2].

In the general case of non-standardized data it is possible to build a graph showing the correlations between the set of variables and a pair of factors. The variance-covariance matrix **V** is multiplied from the left and the right with the matrix $\mathbf{D}_{\sigma^{-1}}$ of the inverses of the standard deviations

$$\mathbf{D}_{\sigma^{-1}}\mathbf{V}\mathbf{D}_{\sigma^{-1}} = \mathbf{D}_{\sigma^{-1}}\mathbf{Q}\Lambda\mathbf{Q}^{\mathrm{T}}\mathbf{D}_{\sigma^{-1}} = \mathbf{D}_{\sigma^{-1}}\mathbf{Q}\sqrt{\mathbf{\Lambda}}\left(\mathbf{D}_{\sigma^{-1}}\mathbf{Q}\sqrt{\mathbf{\Lambda}}\right)^{\mathrm{T}}$$

from what the general formula to calculate the correlation between a variable and a factor can be deduced

$$\mathrm{corr}(\mathbf{z}_i, \mathbf{y}_p) = \sqrt{\frac{\lambda_p}{\sigma_{ii}}}\,q_{ip}$$

3.2 Multivariate Nested Variogram

The multivariate regionalization of a set of random functions can be represented with a spatial multivariate linear model. The associated multivariate nested variogram model is easily fitted to the multivariate data. Several coregionalization matrices describing the multivariate correlation structure at different scales of a phenomenon

result from the variogram fit. The relation between the coregionalization matrices and the classical variance-covariance matrix is examined [2].

3.2.1 Linear model of coregionalization

A set of real second-order stationary random functions $\{Z_i(\mathbf{x});\ i=1,...,N\}$ can be decomposed into sets $\{Z_u^i(\mathbf{x});\ u=0,...,S\}$ of spatially uncorrelated components

$$Z_i(\mathbf{x}) = \sum_{u=0}^{S} Z_u^i(\mathbf{x}) + m_i$$

where for all values of the indices i, j, u and v

$$\mathrm{E}[Z_i(\mathbf{x})] = m_i$$
$$\mathrm{E}[Z_u^i(\mathbf{x})] = 0$$

and

$$\mathrm{cov}(Z_u^i(\mathbf{x}), Z_u^j(\mathbf{x}+\mathbf{h})) = \mathrm{E}[Z_u^i(\mathbf{x}) Z_u^j(\mathbf{x}+\mathbf{h})] = C_{ij}^u(\mathbf{h})$$
$$\mathrm{cov}(Z_u^i(\mathbf{x}), Z_v^j(\mathbf{x}+\mathbf{h})) = 0 \qquad \text{when} \quad u \neq v$$

The cross covariance functions $C_{ij}^u(\mathbf{h})$ associated with the spatial components are composed of real coefficients b_{ij}^u and are proportional to real correlation functions $\rho_u(\mathbf{h})$

$$C_{ij}(\mathbf{h}) = \sum_{u=o}^{S} C_{ij}^u(\mathbf{h}) = \sum_{u=0}^{S} b_{ij}^u \rho_u(\mathbf{h})$$

which implies that the cross covariance functions are even in this model [2].

Coregionalization matrices $\mathbf{B}_u$ of order $N \times N$ can be set up and we have a multivariate nested covariance function model

$$\mathbf{C}(\mathbf{h}) = \sum_{u=0}^{S} \mathbf{B}_u \rho_u(\mathbf{h})$$

with positive semi-definite coregionalization matrices $\mathbf{B}_u$.

Each spatial component $Z_u^i(\mathbf{x})$ can itself be represented as a set of uncorrelated factors $Y_u^p(\mathbf{x})$ with transformation coefficients a_{up}^i

$$Z_u^i(\mathbf{x}) = \sum_{p=1}^{N} a_{pu}^i Y_u^p(\mathbf{x})$$

where for all values of the indices i, j, u, v, p and q

$$\mathrm{E}[Y_u^p(\mathbf{x})] = 0$$

and

$$\mathrm{cov}(Y_u^p(\mathbf{x}), Y_u^p(\mathbf{x}+\mathbf{h})) = \rho_u(\mathbf{h})$$
$$\mathrm{cov}(Y_u^p(\mathbf{x}), Y_v^q(\mathbf{x}+\mathbf{h})) = 0 \qquad \text{when} \quad u \neq v \quad \text{or} \quad p \neq q$$

Combining the spatial with the multivariate decomposition, we obtain the *linear model of coregionalization*

$$Z_i(\mathbf{x}) = \sum_{u=0}^{S} \sum_{p=1}^{N} a_{pu}^i Y_u^p(\mathbf{x})$$

In practice first a set of correlation functions $\rho_u(\mathbf{h})$ (i.e. normalized variograms $g_u(\mathbf{h})$) is selected, taking care to keep S reasonably small. Then the coregionalization matrices are fitted using a weighted least squares algorithm (described below). The weighting coefficients are chosen by the practitioner so as to provide a graphically satisfactory fit which downweights arbitrarily distance classes, which do not comply with the shape suggested by the experimental variograms. Finally the coregionalization matrices are decomposed, yielding the transformation coefficients a_{up}^i, which specify the linear coregionalization model

$$\mathbf{B}_u = \mathbf{A}_u \mathbf{A}_u^{\mathrm{T}} \qquad \text{where} \quad \mathbf{A}_u = [a_{pu}^i]$$

The decomposition of the $\mathbf{B}_u$ into the product of $\mathbf{A}_u$ with its transpose is usually based on the eigenvalue decomposition of each coregionalization matrix [2].

3.2.2 Bivariate fit of the experimental variograms

The multivariate nested variogram model associated with a linear model of intrinsically stationary random functions is

$$\boldsymbol{\Gamma}(\mathbf{h}) = \sum_{u=0}^{S} \mathbf{B}_u g_u(\mathbf{h})$$

where the $g_u(\mathbf{h})$ are normalized variograms and the $\mathbf{B}_u$ are positive semi-definite matrices [2].

In the case of two variables it is simple to design a procedure for fitting the variogram model to the experimental variograms. We start by fitting the two direct variograms using a nested model. At least one structure $g_u(\mathbf{h})$ should be common to both variograms to obtain a non-trivial coregionalization model. Then we are able to fit the sills b_{ij}^u of the cross variogram, using the sills of the direct variograms to set bounds within which the coregionalization model is authorized

$$\left| b_{ij}^u \right| \leq \sqrt{b_{ii}^u b_{jj}^u}$$

because the second order principal minors of $\mathbf{B}_u$ are positive.

Constrained weighted least squares routines exist, which allow integrating these constraints into an automated fit for a set of predefined structures [2].

The extension of this bivariate procedure to more than two variables does not guarantee a priori an authorized model, because higher order principal minors of the coregionalization matrices are not constrained to be positive.

3.2.3The need for an analysis of the coregionalization

We can use classical principal component analysis to define the values of factors at sample locations and then krige the factors over the whole region to make maps. What would be the benefit of a spatial multivariate analysis based on the linear model of coregionalization and the corresponding multivariate nested variogram? [2]

To answer this question we restrict the discussion to a second order stationary context, to make sure that the variance-covariance matrix $\mathbf{V}$, on which classical principal component analysis is based, exists from the point of view of the model.

If we let the lag $\mathbf{h}$ go to infinity in a second order stationary context with structures $g_u(\mathbf{h})$ having unit sills, we notice that the multivariate variogram model is equal to the variance-covariance matrix for large $\mathbf{h}$

$$\boldsymbol{\Gamma}(\mathbf{h}) \to \mathbf{V} \qquad \text{for} \quad \mathbf{h} \to \infty$$

In this setting the variance-covariance matrix is simply a mixture of coregionalization matrices

$$\mathbf{V} = \sum_{u=0}^{S} \mathbf{B}_u$$

We realize that when the variables are not intrinsically correlated, it is necessary to analyze separately each coregionalization matrix $\mathbf{B}_u$. The variance-covariance matrix $\mathbf{V}$ is a blend of different correlation structures stemming from all scales covered by the sampling grid and is likely to be meaningless. Furthermore, coregionalization matrices $\mathbf{B}_u$ can be obtained under any type of stationarity hypothesis, while the variance-covariance matrix $\mathbf{V}$ is only meaningful with data fitting into a framework of second-order stationarity [2].

3.3 Coregionalization Analysis

The geostatistical analysis of multivariate spatial data can be subdivided into two steps

- The analysis of the coregionalization of a set of variables leading to the definition of a linear model of coregionalization;
- The cokriging of specific factors at characteristic scales.

These techniques have originally been called *factorial kriging analysis* (from the French *analyse krigeante* [1]). They allow to isolate and to display sources of variation acting at different spatial scales with a different correlation structure [2].

3.3.1 Regionalized principal component analysis

Principal component analysis can be applied to coregionalization matrices, which are the variance-covariance matrices describing the correlation structure of a set of variables at characteristic spatial scales [2].

Regionalized principal component analysis consists in decomposing each matrix $\mathbf{B}_u$ into eigenvalues and eigenvectors

$$\mathbf{B}_u = \mathbf{A}_u \mathbf{A}_u^{\mathrm{T}} \qquad \text{with} \quad \mathbf{A}_u = \mathbf{Q}_u \sqrt{\mathbf{\Lambda}_u} \quad \text{and} \quad \mathbf{Q}_u \mathbf{Q}_u^{\mathrm{T}} = \mathbf{I}$$

The matrices $\mathbf{A}_u$ specify the coefficients of the linear model of coregionalization. The transformation coefficients

$$a_{pu}^{i} = \sqrt{\lambda_u^p}\, q_p^i$$

are the covariances between the original variables $Z_i(\mathbf{x})$ and the factors $Y_u^p(\mathbf{x})$. They can be used to plot the position of the variables on correlation circles for each characteristic spatial scale of interest. These plots are helpful to compare the correlation structure of the variables at the different spatial scales.

The correlation circle plots can be used to identify intrinsic correlation: if the plots show the same patters of correlation, this means that the eigenvectors of the coregionalization matrices are similar and the matrices only differ by their eigenvalues. Thus the matrices $\mathbf{B}_u$ are all proportional to a matrix $\mathbf{B}$, the coregionalization matrix of the intrinsic correlation model [2].

3.3.2 Generalizing the analysis

The analysis can be generalized by choosing eigenvectors, which are orthogonal with respect to a symmetric matrix $\mathbf{M}_u$ representing a metric

$$\mathbf{A}_u = \mathbf{Q}_u \mathbf{M}_u \sqrt{\mathbf{\Lambda}_u} \qquad \text{with} \quad \mathbf{Q}_u \mathbf{M}_u \mathbf{Q}_u^{\mathrm{T}} = \mathbf{I}$$

A possibility is to use the metric

$$\mathbf{M}_u = \sum_{v=0}^{S} \mathbf{B}_v$$

which is equivalent to the variance-covariance matrix $\mathbf{V}$ in a second order stationary model. This metric generates a contrast between the global variation and the variation at a specific scale described by a coregionalization matrix $\mathbf{B}_u$ [2].

3.3.3 Cokriging regionalized factors

The linear model of coregionalization defines factors at particular spatial scales. We wish to estimate a regionalized factor from data in a local neighborhood around each estimation location $\mathbf{x}_0$ [2].

The estimator of a specific factor $Y_{u_0}^{p_0}(\mathbf{x})$ at a location $\mathbf{x}_0$ is a weighted average of data from variables in the neighborhood with unknown weights ω_α^i

$$Y_{p_0 u_0}^{*}(\mathbf{x}_0) = \sum_{i=1}^{N} \sum_{\alpha=1}^{n_i} \omega_\alpha^i Z_i(\mathbf{x}_\alpha)$$

In the framework of local second-order stationarity, in which local means m_l^i for the neighborhood around $\mathbf{x}_0$ are meaningful, an unbiased estimator is built for the factor (of zero mean, by construction) by using weights summing up to zero for each variable

$$\mathrm{E}\left[Y_{p_0 u_0}^{*}(\mathbf{x}) - Y_{u_0}^{p_0}(\mathbf{x})\right] = \sum_{i=1}^{N} m_l^i \underbrace{\sum_{\alpha=1}^{n} \omega_\alpha^i}_{0} = 0$$

The effect of the constraints on the weights is to filter out the local means of the variables $Z_i(\mathbf{x})$.

The estimation variance σ_E^2 is

$$\sigma_E^2 = \mathrm{E}\left[\left(Y^*_{p_0 u_0}(\mathbf{x}) - Y^{p_0}_{u_0}(\mathbf{x})\right)^2\right]$$
$$= 1 + \sum_{i=1}^{N}\sum_{j=1}^{N}\sum_{\alpha=1}^{n}\sum_{\beta=1}^{n}\omega^i_\alpha\omega^j_\beta C_{ij}(\mathbf{x}_\alpha - \mathbf{x}_\beta) - 2\sum_{i=1}^{N}\sum_{\alpha=1}^{n}\omega^i_\alpha a^i_{u_0 p_0}\rho_{u_0}(\mathbf{x}_\alpha - \mathbf{x}_0)$$

The minimal estimation variance is realized by the cokriging system

$$\begin{cases} \sum_{j=1}^{N}\sum_{\beta=1}^{n_j}\omega^j_\beta C_{ij}(\mathbf{x}_\alpha - \mathbf{x}_\beta) - \mu_i = a^i_{u_0 p_0}\rho_{u_0}(\mathbf{x}_\alpha - \mathbf{x}_0) & \text{for } i = 1,\dots N; \\ & \alpha = 1,\dots n \\ \sum_{\beta=1}^{n_i}\omega^i_\beta = 0 & \text{for } i = 1,\dots N \end{cases}$$

We find in the right hand side of this system the transformation coefficients $a^i_{p_0 u_0}$ of the factor of interest. These coefficients are multiplied by values of the spatial correlation function $\rho_{u_0}(\mathbf{h})$, which describes the correlation at the scale of interest.

The factor cokriging is used to estimate a regionalized factor at the nodes of a regular grid, which serves to draw a map [2].

3.3.4 Regionalized multivariate analysis

Cokriging a factor is more cumbersome and computationally more intensive than kriging it. Coregionalization analysis is more lengthy than a traditional analysis, which ignores spatial scale. When is all this effort necessary and worthwhile? When can it be avoided? The answer is based on the notion of intrinsic correlation [2].

The key question to investigate is whether the correlation between variables is dependent on spatial scale. Three ways to test for scale-dependent correlation have been described

1. codispersion coefficients $\mathrm{cc}_{ij}(\mathbf{h})$ can be computed and plotted: if they are not constant for each variable pair, the correlation structure of the variable set is affected by spatial scale;
2. cross variograms between principal components of the variables can be computed: if they are not zero for each principal component pair at any lag $\mathbf{h}$, the classical principal components are meaningless because the variance-covariance matrix of the variable set is merely a mixture of different variance-covariance structures at various spatial scales;
3. plots of correlation circles in a regionalized principal component analysis can be examined: if the patterns of association between the variables are not identical for the coregionalization matrices, the intrinsic correlation model is not appropriate for this data set. With only few variables it is possible to look directly at a table of regionalized correlation coefficients instead of the regionalized principal components.

If the data appears to be intrinsically correlated, we can apply any classical method of multivariate analysis, calculate the direct variograms of the factors, krige them on a grid and represent them as maps. But if correlation is affected by spatial scale, we need to fit a linear model of coregionalization and to cokrige the factors [2].

4 Case studies

4.1 Data

The methods described above were applied for the analysis of the spatial data on contamination by heavy metals both in Geneva Lake (Leman) sediments and of Japanese soils.

Leman is situated in the southwest part of Switzerland. Dataset of Leman contains 9 metals (Pb, Cu, Cd, Zn, Cr, Ni, Be, B, Mn). Analogously Japanese dataset contains 8 metals (Cu, Zn, Cr, Ni, Be, B, Mn). Leman and Japan have the complex monitoring network (almost linear), which complicates analysis. The values of

contamination for different metals (both Leman and Japan) are correlated among themselves. The metals are both linear correlated with some metals and poorly correlated with other [10].

The technique, named Principal Component Analysis (PCA), has allowed to reduce the input dimension of data.

4.2 Analysis

In the given work the Principal Component Analysis was carried out for:

1. To select the most significant few principal components which will describe all of the real variations in the data while the rest of the components will contain mostly uncorrelated noise;
2. To carry out factorial kriging and factors mapping and using an inversion to initial variables to carry out the variables mapping.

To carry out the Principal Component Analysis (PCA) with the help of Multigeo research software program [3], it is necessary:

1) To ascertain the variables correlation for dataset;
2) To carry out the spatial correlation analysis of initial dataset (to construct experimental variograms and cross-variograms, to select theoretical models [2]);
3) To carry out the principal component analysis for each model structure;

The considered variables are correlated and their correlations are presented in Table 1 and 2 for each model structure.

Table 1. The correlation matrix for model structures: nugget – above the diagonal, spherical – below the diagonal (Japan)

	Cu	*Zn*	*Cd*	*Pb*	*Cr*	*Mn*	*Ni*	*As*
Cu	**1.0000**	0.2805	0.2567	0.4359	0.0224	0.1876	0.0142	0.2608
Zn	0.6143	**1.0000**	0.2848	0.326	0.0915	0.2613	0.0571	0.1447
Cd	0.47	0.3356	**1.0000**	0.2718	0.0237	0.0648	0.0318	0.0556
Pb	0.4566	0.1075	0.0378	**1.0000**	0.0339	0.1953	0.0368	0.5669
Cr	0.2537	0.2201	0.0651	0.0078	**1.0000**	0.1383	0.8773	-0.0351
Mn	0.7086	0.7411	0.6234	-0.0136	0.5815	**1.0000**	0.0815	0.2321
Ni	0.1656	0.1799	0.0175	-0.0494	0.9196	0.4871	**1.0000**	-0.0395
As	0.4482	0.107	0.4883	0.7774	-0.0355	0.2559	-0.0855	**1.0000**

Table 2. The correlation matrix for model structures: nugget – above the diagonal, spherical – below the diagonal (Leman)

	Cd	*Zn*	*Cu*	*Mn*	*Cr*	*B*	*Be*	*Pb*	*Ni*
Cd	**1.0000**	0.4750	0.3397	0.1820	0.2507	0.1611	0.2402	0.3565	0.2516
Zn	0.8227	**1.0000**	0.6386	0.1364	0.7091	0.3783	0.6039	0.5209	0.6810
Cu	0.5025	0.7269	**1.0000**	0.1098	0.6792	0.2881	0.5496	0.3981	0.6008
Mn	0.1805	0.2167	0.0236	**1.0000**	0.1067	0.1417	0.0345	0.1203	0.2151
Cr	0.3020	0.5647	0.3827	0.3255	**1.0000**	0.6816	0.6554	0.1840	0.9610
B	0.2570	0.2562	0.1588	0.2397	-0.0415	**1.0000**	0.6675	0.1023	0.6740
Be	0.0505	0.2463	0.0152	0.3333	0.9214	-0.1363	**1.0000**	0.0055	0.7207
Pb	0.6416	0.6073	0.3730	0.2175	0.6117	-0.2194	0.4875	**1.0000**	0.1688
Ni	0.3361	0.6270	0.4366	0.4229	0.9436	0.1348	0.8360	0.4413	**1.0000**

For Leman dataset both variograms and cross-variograms in all directions and in the direction of 22,5° were constructed . Modeled variograms and cross-variograms, obtained for Leman, consist of nugget and spherical (with radius 30 km) model structures [10]:

$$\gamma(h) = \begin{cases} c_0 + (c - c_0)\left(\dfrac{1.5h}{a} - \dfrac{0.5h^3}{a^3}\right), & h \le a \\ c, & h > a \end{cases}, \quad (1)$$

where c_0 is nugget, c is sill, a is range.

The share of accident (noncorrelatedness) in spatial distribution of the data is determined by nugget.

Theoretical models of variograms and cross-variograms variograms in all directions and in the direction of 22,5° and also their values are discussed in work [10]. Selection parameters of models were carried out method with the help of least square. The models in the direction of 22,5° were used for PCA, because these models modeled experimental variograms and cross-variograms better than the models in all directions. The variability percentage of the nugget and the spherical model structure for each considered variable is shown in Table 3.

Table 3. The variability percentage of the nugget and the spherical structure for each variable (Leman)

Spatial component	***Variance***	
	Nugget	*Spherical*
Cd	43.18	56.82
Zn	26.67	73.33
Cu	16.69	83.31
Mn	16.82	83.18
Cr	22.18	77.82
B	15.86	84.14
Be	17.11	82.89
Pb	34.68	65.32
Ni	32.40	67.60

Similar analysis was carried out for Japanese data: experimental variograms and cross-variograms in all directions were constructed. The models (1), consisting the nugget and the spherical (with radius 400 km) model structures, have been used for modeling as well as for Leman. Some models are presented in Figure 1.

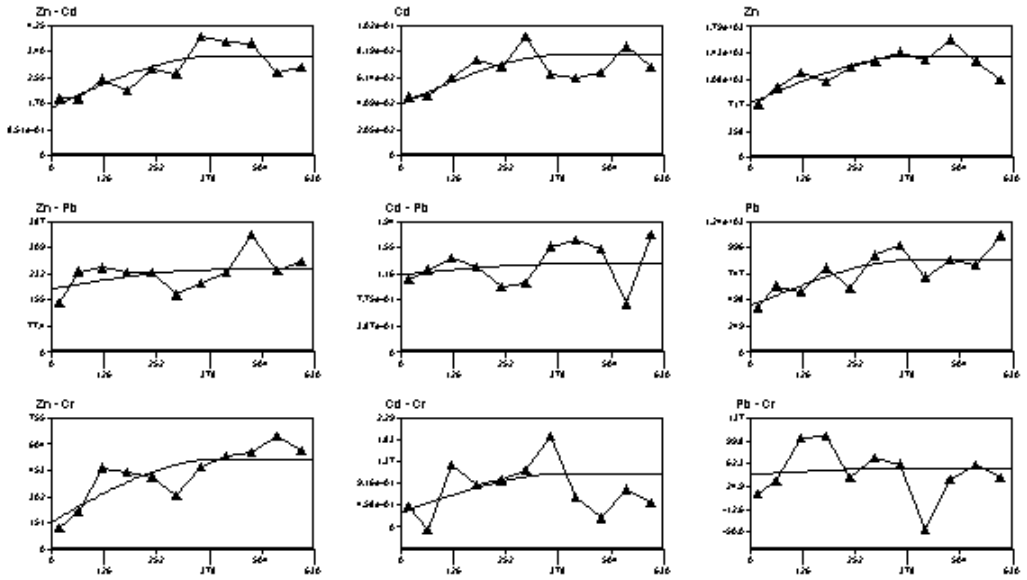

Figure 1. The some variograms and cross-variograms for Japanese data

The correlations of metals in Japanese data and the variability percentage of model structures are shown in Table 1 and 4 respectively.

Table 4: The variability percentage of the nugget and the spherical structure for each variable (Japan)

Spatial component	*Variance*	
	Nugget	*Spherical*
Cu	74.12	25.88
Zn	54.05	45.95
Cd	51.34	48.66
Pb	49.94	50.06
Cr	48.55	51.45
Mn	72.38	27.62
Ni	41.28	58.72
As	73.38	26.62

These dataset variables (the metals concentrations) are transformed, through the PCA, into a new set of orthogonal axes. This set contains 8 components for Leman. For Japanese data another result is obtained: 7 components for the spherical structure and 8 – for the nugget. Also the ninth new orthogonal axis for Leman dataset and the eighth new axis for the spherical structure of Japanese dataset are obtained, but their eigenvalues and correlations with the variables are equal zero. Therefore these axes were not examined.

The correlations of principal components with initial variables for Leman and Japan datasets are presented in Table 5, 7 and Table 6, 8 for the nugget and the spherical structures respectively.

Table 5. Correlations of principal components with initial variables for nugget structure model (Leman)

	PC1	**PC2**	**PC3**	**PC4**	**PC5**	**PC6**	**PC7**	**PC8**
Cd	0.4576	-0.557	-0.0588	-0.6723	0.0229	-0.111	-0.1077	0.0229
Zn	0.8442	-0.2865	0.1471	0.0278	0.0380	-0.0140	0.3899	-0.1708
Cu	0.7664	-0.2019	0.1966	0.2018	0.3962	0.1656	-0.3213	-0.0692
Mn	0.2143	-0.241	-0.9243	0.1407	0.1189	0.0805	0.0341	-0.0104
Cr	0.9162	0.2195	0.0369	0.1244	0.0465	-0.2989	-0.0555	-0.0314
B	0.7017	0.4072	-0.1726	-0.08962	-0.497	0.08131	-0.1838	-0.129
Be	0.7944	0.3551	0.0709	-0.1991	0.0638	0.4004	0.1385	0.1209
Pb	0.3812	-0.7463	0.1555	0.3145	-0.3925	0.0882	-0.0299	0.1087
Ni	0.9118	0.2399	-0.0858	0.0949	0.0367	-0.2373	0.0441	0.1873

Table 6. Correlations of principal components with initial variables for spherical structure model (Leman)

	PC1	**PC2**	**PC3**	**PC4**	**PC5**	**PC6**	**PC7**	**PC8**
Cd	0.6571	-0.6002	-0.1117	0.3028	0.2267	0.1597	-0.1637	-0.0115
Zn	0.8388	-0.4695	-0.054	-0.0415	0.0058	0.1746	0.2019	-0.0108
Cu	0.5987	-0.5213	-0.1556	-0.3787	-0.4159	-0.1534	-0.0745	-0.0045
Mn	0.4228	0.1765	0.6029	0.5303	-0.3812	0.0097	-0.0004	-0.0088
Cr	0.8941	0.4008	-0.016	-0.1806	0.0477	-0.0289	-0.0066	-0.0628
B	0.1214	-0.4655	0.7847	-0.1633	0.2993	-0.1908	0.0146	0.0017
Be	0.694	0.6923	0.0402	-0.0781	0.1701	-0.0365	-0.0322	-0.0097
Pb	0.7512	-0.0174	-0.434	0.392	0.1056	-0.2802	0.052	0.03
Ni	0.8899	0.2883	0.1956	-0.248	-0.0225	0.1362	-0.0332	0.0709

Table 7. Correlations of principal components with initial variables for nugget structure model (Japan)

	PC1	**PC2**	**PC3**	**PC4**	**PC5**	**PC6**	**PC7**	**PC8**
Cu	0.6526	-0.1788	0.1241	-0.1537	-0.7008	-0.0051	-0.1098	0.0007
Zn	0.6058	-0.0397	0.3714	0.3645	0.1781	-0.5663	-0.0905	0.0088
Cd	0.4645	-0.0882	0.6933	-0.169	0.2582	0.4403	-0.0816	-0.0045
Pb	0.7816	-0.2202	-0.1829	-0.2935	0.1241	-0.0658	0.4486	-0.0061
Cr	0.2769	0.9248	-0.0497	-0.0674	-0.0008	-0.0068	-0.0214	-0.246
Mn	0.4867	0.0541	-0.24	0.7535	-0.0503	0.3554	0.0754	0.0151
Ni	0.2544	0.9246	-0.0417	-0.1381	0.0061	-0.0019	-0.0134	0.2437
As	0.6025	-0.272	-0.5707	-0.2084	0.2583	0.0376	-0.3546	-0.0003

Table 8. Correlations of principal components with initial variables for spherical structure model (Japan)

	PC1	PC2	PC3	PC4	PC5	PC6	PC7
Cu	0.8326	-0.2585	0.0643	-0.2567	0.4096	-0.0106	0.0454
Zn	0.7136	0.0382	0.3968	-0.4891	-0.2969	-0.067	-0.0018
Cd	0.6284	-0.2678	0.4152	0.573	-0.0041	-0.1742	-0.0477
Pb	0.3483	-0.657	-0.615	-0.2192	-0.0435	-0.1165	-0.0733
Cr	0.5784	0.6685	-0.4199	0.0938	0.0109	0.041	-0.1779
Mn	0.9216	0.2193	0.2622	0.0433	-0.0139	0.1778	-0.004
Ni	0.499	0.7117	-0.4321	0.1087	-0.0472	-0.1137	0.1757
As	0.4996	-0.7181	-0.3334	0.2615	-0.1627	0.1516	0.0758

Similarly the eigenvalues of principal components are shown in Table 9 (Leman) and Table 10 (Japan).

Table 9. The eigenvalue and variability percentage of structures for each factor (Leman)

PC	*Eigenvalue*		*Percentile of variance*		*Cumulative percentile*	
	Nugget	*Spherical*	*Nugget*	*Spherical*	*Nugget*	*Spherical*
1	4.50	4.32	49.94	48.05	49.94	48.05
2	1.45	1.82	16.06	20.26	66.01	68.31
3	0.99	1.25	10.95	13.86	76.96	82.17
4	0.68	0.80	7.60	8.87	84.57	91.05
5	0.58	0.50	6.46	5.58	91.03	96.62
6	0.37	0.22	4.08	2.39	95.10	99.02
7	0.33	0.08	3.63	0.87	98.74	99.89
8	0.11	0.01	1.26	0.11	100.00	100.00

Table 10. The eigenvalue and variability percentage of structures for each factor (Japan)

PC	*Eigenvalue*		*Percentile of variance*		*Cumulative percentile*	
	Nugget	*Spherical*	*Nugget*	*Spherical*	*Nugget*	*Spherical*
1	2.36	3.40	29.51	42.52	29.51	42.52
2	1.88	2.09	23.46	26.11	52.97	68.62
3	1.05	1.26	13.19	15.69	66.16	84.31
4	0.91	0.77	11.32	9.65	77.48	93.97
5	0.67	0.29	8.43	3.59	85.91	97.55
6	0.65	0.12	8.08	1.47	93.99	99.03
7	0.36	0.08	4.50	0.97	98.50	100.00
8	0.12	0.0	1.50	0.0	100.00	100.00

4.3 Discussion

After the results of the PCA have been obtained it is possible to reply to the following question: “How many parameters are required to adequately describe the metals concentrations?”

Because of the data compression property of PCA, it is an ideal method for investigating this question. Although a physical description of the new axes will not often be immediately apparent, the *number* of principal components that contain most of the information in the metals concentrations tells how many different parameters are needed to fully describe the data set [9]. In this case the data are the metals concentrations. Therefore to answer this question, it is necessary:

a) To examine circles of variables correlations in the plane defined by a pair of axes of factors;
b) To analyze eigenvalues of principal components.

If to look at Figures 2, 3, 4 and 5 it is possible to see that each of them presents *circle of correlations.* This is the projection of the position of the variables on the surface of the hypersphere towards a plane defined by a pair of axes of factors yielding a graphical representation. The circle of correlations shows the proximity of the variables inside a unit circle and is useful to evaluate the affinities and the antagonisms between the variables [2].

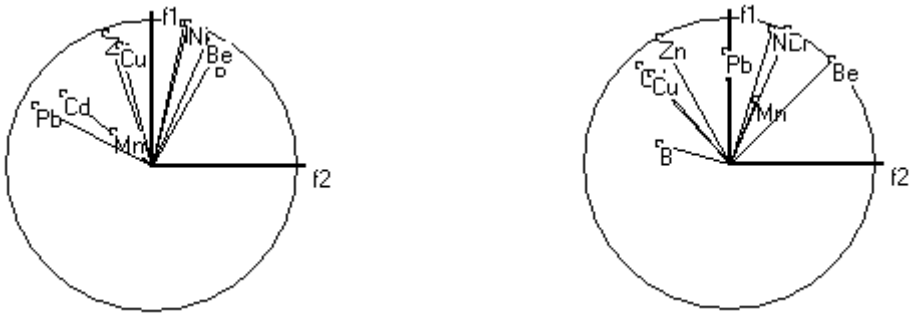

Figure 2. Circle correlation on the plane of the first two principal components PC1 and PC2 (factors f1, f2) for nugget structure (left) and spherical structure (right) (Leman)

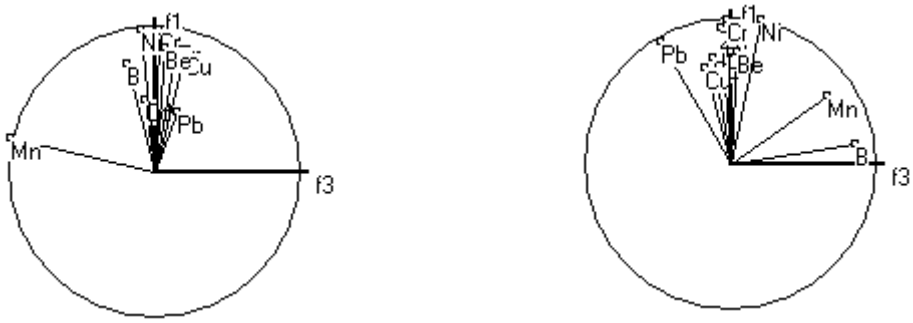

Figure 3. Circle correlation on the plane of the first two principal components PC1 and PC3 (factors f1, f3) for nugget structure (left) and spherical structure (right) (Leman)

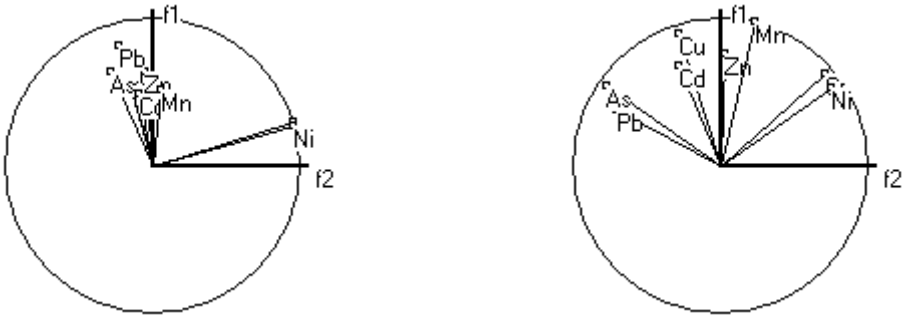

Figure 4. Circle correlation on the plane of the first two principal components PC1 and PC2 (factors f1, f2) for nugget structure (left) and spherical structure (right) (Japan)

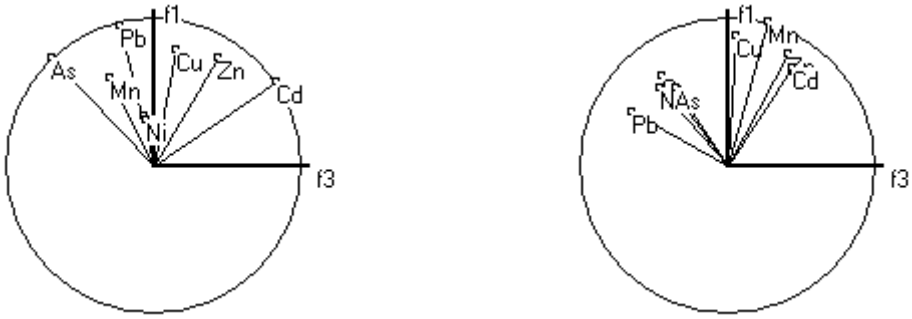

Figure 5. Circle correlation on the plane of the first two principal components PC1 and PC3 (factors f1, f3) for nugget structure (left) and spherical structure (right) (Japan)

These circles of correlations are shown for various principal components and model structures of Leman and Japanese datasets. One can see that variables (for Leman dataset – Pb, Be, Ni, Cu, Zn, Cr, Cd; for Japan dataset – Cu, Mn, Zn, Cd) begin to group around the axe, defined first factor, at successive substitution of second axe on subsequent principal component, thus the strong correlation between the first principal component and initial variables is observed.

Figures 5 and 6 show the eigenvalues for the two different model structures of Leman and Japanese datasets: filled circles for the nugget model structure and filled squares for the spherical model structure. Remembering that the eigenvalues show the amount of the total sample variance accounted for in that principal component, each of these diagrams displays the distribution of variances in the new axes for each of the two structures.

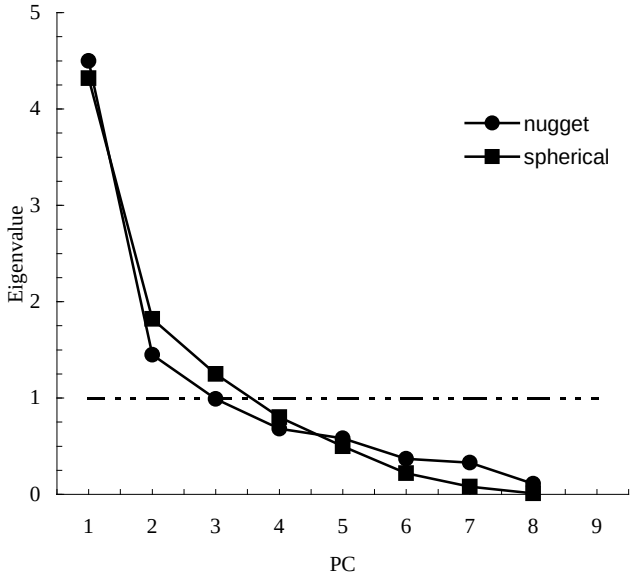

Figure 6. The distribution in metal concentration variations contained in each principal component for the nugget and the spherical model structures for Leman

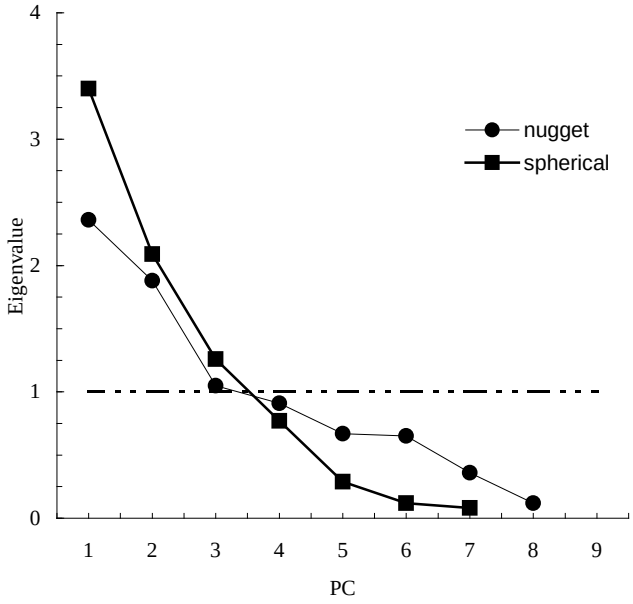

Figure 7. The distribution in metal concentration variations contained in each principal component for the nugget and the spherical model structures for Japan

One feature of the diagram for Leman that is clear at a first look is the likeness in the shape of the variances distribution in the nugget structure relative to the spherical structure. For Japanese diagram the distinction in the shape of the variances distribution is observed. But this distinction is observed at the eigenvalues, which less than one, i.e. at the components containing nearly no concentrations information. The first component in the structures contains 2.4 to 3.1 times (nugget and spherical, respectively) the variance of the second component for Leman data and 1.3 to 1.6 times – for Japanese data.

The dot-dashed line presents the value of the eigenvalues for a sample of 100% uncorrelated variances and can act as a different type of benchmark [9]. Taking the shallow trend of the structures eigenvalues as the benchmark of components containing nearly no concentrations information, it follows that, at most, three components in the datasets are required to describe most of the metal concentration variation there.

The Leman data have been used also in the work "Multivariate geostatistical mapping of contamination in Geneva lake sediments. Case study with Multigeo" [10]. This work has been drawn the following conclusion: three different additional variables of the eight (Zn, Cu, B) are enough to obtain better estimation and estimation variation by cokriging. The results of the Principal Component Analysis carried out in the present work confirm this conclusion also.

Other aim of the PCA – factorial kriging and factors mapping in order to carry out the variables mapping using an inversion to initial variables.

For Leman and Japanese data factorial kriging was carried out using Multigeo. Some maps of estimations and the variances factors estimations are presented in Figures 8, 9, 10, 11, 12 and 13.

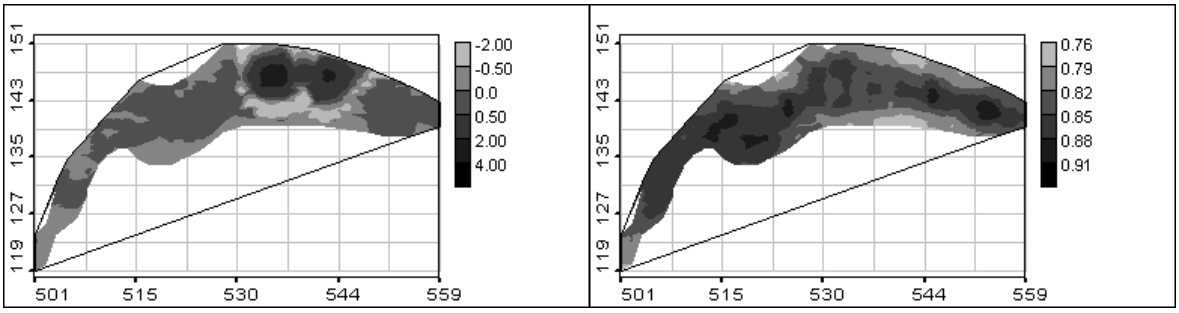

Figure 8. The kriging estimations (left) and variations (right) of the principal component PC1 (Leman)

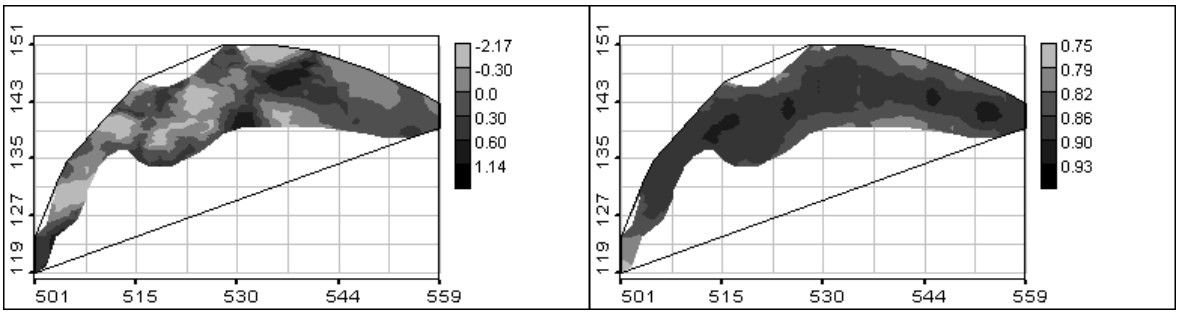

Figure 9. The kriging estimations (left) and variations (right) of the principal component PC2 (Leman)

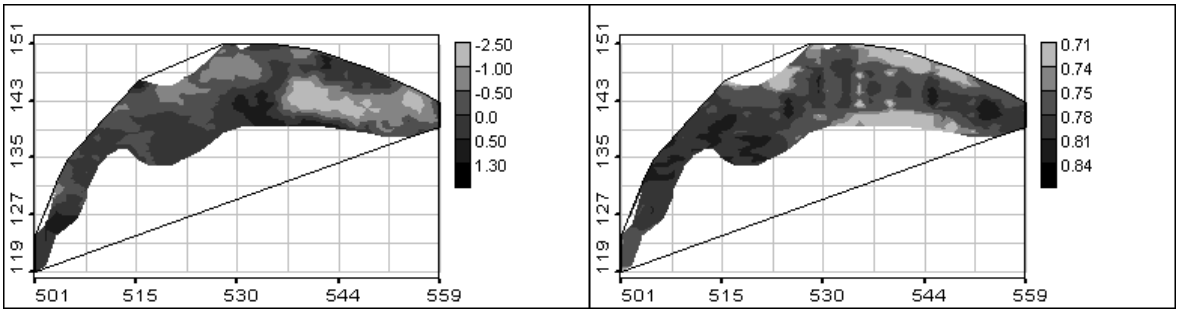

Figure 10. The kriging estimations (left) and variations (right) of the principal component PC3 (Leman)

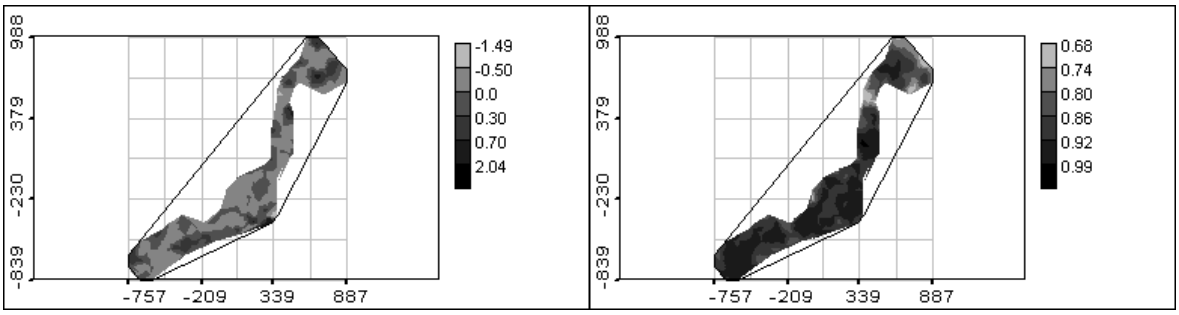

Figure 11. The kriging estimations (left) and variations (right) of the principal component PC1 (Japan)

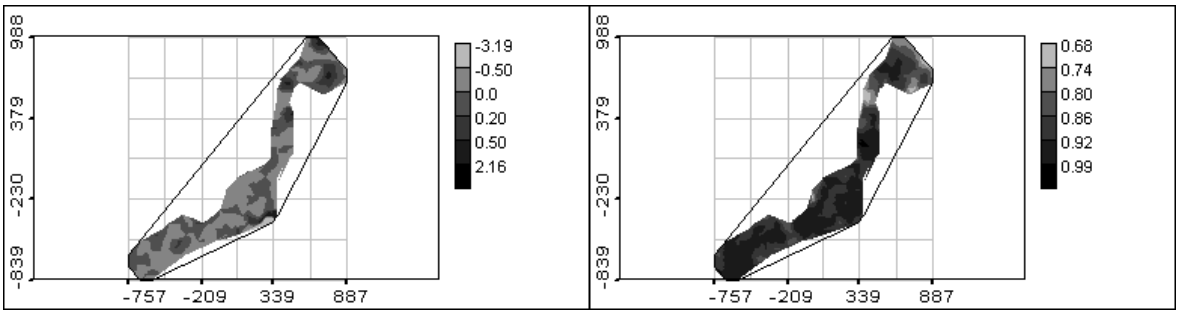

Figure 12. The kriging estimations (left) and variations (right) of the principal component PC2 (Japan)

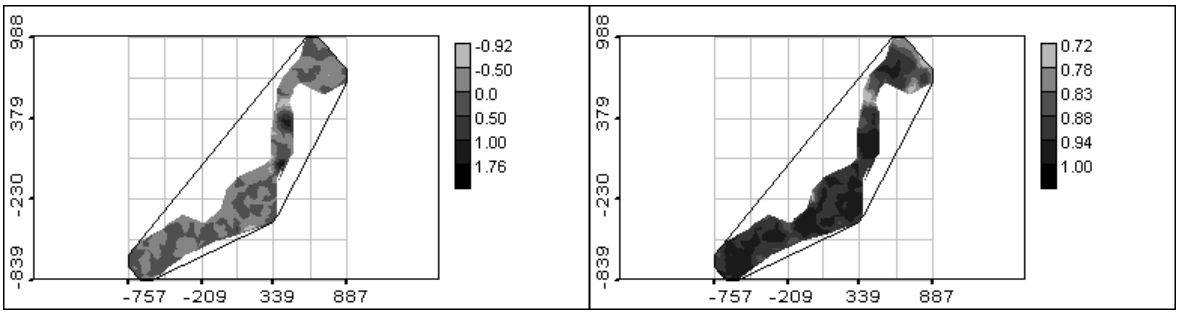

Figure 13. The kriging estimations (left) and variations (right) of the principal component PC3 (Japan)

But there is a series of problems with the research software program Multigeo. Multigeo allows to carry out the factorial kriging and mapping factors, however it is impossible to make the inversion to initial variables, i.e. one cannot tell which variables describe the greatest part of the real variations in the data. This is the defect of the Multigeo for the Principal Component Analysis.

5 Conclusions

In the present work the Principal Component Analysis (PCA) has been described and applied for geostatistical data. This analysis has been carried out using the data on the contamination of Geneva Lake and Japanese soils by heavy metals. The values of contamination for different metals are correlated among themselves. According to the results of Principal Component Analysis it is possible to draw the following conclusions:

1. The Principal Component Analysis allows reducing the dimensionality of the data, whilst maintaining as much of its original structure.
2. The three independent parameters are required to account for the metal concentration variations in the data on the contamination of Geneva Lake and Japanese soils with metals.
3. The results obtained in the present work with regard to Leman data confirm other studies on the same data well.

To analyze and visualize raw data and the results Multigeo software was used.

6 Acknowledgements

The work was supported in part by INTAS grant 99-00099.

7 References

[1] Matheron G (1982) *Pour un*e *Analyse Krigeante des Données Régionalisées.* Publication N-732, Centre de Géostatistique, Fontainebleau, France, 22p.

[2] H. Wackernagel. Multivariate Geostatistics. Springler-Verlag, Berlin, 1995

[3] Guiseppe Raspa, Massimiliano Ticci, Ismail Hoxha, Multigeo, User's Guide, Università La Sapienza Dipartimento ICMMPM Materie Prime, September 1999.

[4] R. A. Calvo, M. G. Partridge, and M. A. Jabri. A comparative study of principal component analysis techniques. In Proc. Ninth Australian Conf. on Neural Networks, Brisbane, QLD, Feb. 1998

[5] A. Bravais. Analyse mathematique sur les probabilite's des erreurs de situation d'un point. *Sci. Math Phys*., 9:255–332, 1846.

[6] H. Hotelling. Analysis of a complex of statistical variables into principal components. *J. Educ. Psych*., 24:417–441, 498–520, 1933.

[7] R. Frisch. Correlation and scatter in statistical variables. *Nordic Stat.* J, 8:36–102, 1929.

[8] K. Pearson. On lines and planes of closest fit to a system of points in space. *Philosophical Magazin*e, 2:79–156, 1901.

[9] Lewis A. Jones, Warrick J. Couch. A Statistical Comparison of Line Strength Variations in Coma and Cluster Galaxies at z~0.3. PASA, 15 (3), 309

[10] R. Parkin, M. Kanevski, M. Maignan, G. Raspa, E. Savelieva. Multivariate geostatistical mapping of contamination in Geneva lake sediments. Case study with Multigeo. Preprint IBRAE-2001-04. Moscow: Nuclear Safety Institute RAS, 2001, p.20

[11] M. Partridge and R. A. Calvo. Fast dimensionality reduction and simple pca. *Intelligent Data Analysis*, 2(3), 1998 (to appear).

[12] H. Kargupta, W. Huang, S. Krishnamoorthy, E. Johnson. Distributed Clustering Using Collective Principal Component Analysis. Accepted for publication in Knowledge and Information Systems Journal Special Issue on Distributed and Parallel Knowledge Discovery. 2000 (In press).

[13] A. Basilevsky. Statistical factor analysis and related methods: theory and applications. Wiley, New York, 1994.

[14] C. Chatfield. Introduction to multivariate analysis. Chapman and Hall, London; New York, 1980.

[15] I. T. Jolliffe. *Principal component analysi*s. Springer-Verlag, New York, 1986.

[16] A. C. Rencher. *Methods of multivariate analysi*s. Wiley, New York, 1995.

[17] D. S. Watkins. *Fundamentals of Matrix Computations.* John Wiley, New York, 1991.

[18] G. H. Golub and C. F. V. Loan. *Matrix Computations.* The Johns Hopkins University Press, Baltimore, 1989.

[19] R. A. Horn and C. R. Johnson. *Matrix Analysis.* Cambridge University Press, Cambridge, 1985.

[20] M. L. Maron and R. J. Lopez. *Numerical Analysis: A Practical Approach.* Wadsworth Publishing Company, Belmont, California, 1992.

[21] J. H. Mathews. *Numerical Methods for Mathematics, Science, and Engineering.* Prentice Hall, 1992.

[22] R. Tagliaferri, G. Longo, S. Andreon, S. Zaggia, N. Capuano, G. Gargiulo. Astronomical Object Recognition by means of Neural Networks. In Proceedings of "Neural Nets WIRN VIETRI-98", Springer, 1998.

[23] E. Oja, H. Ogawa, J. Wangviwattana. Learning in nonlinear constrained Hebbian network. In T. Kohonen et al. (Eds.), Artificial neural networks, 385-390, Amsterdam: North-Holland, 1991.

[24] P. Baldi, K. Hornik. Neural networks for principal component analysis: learning from examples without local minima. Neural Networks, 2, (7), 53-58, 1989.

[25] C. Jutten, J. Herault. Blind separation of sources, part I: an adaptive algorithm based on neuromimetic architecture. Signal Processing, 24, (1), 1-10, 1991.

[26] E. Oja. A simplified neuron model as a principal component analyzer. Journal of Mathematical Biology, 15, 267-273, 1982.

[27] M. Plumbley. A Hebbian/anti Hebbian network which optimizes information capacity by orthonormalizing the principal subspace. In Proc. IEE Conf. on Artificial Neural Networks, Brighton, UK, 86-90, 1993.

[28] T. D. Sanger. Optimal unsupervised learning in a single-layer linear feedforward network. Neural Networks, 2, 459-473, 1989.

[29] J. Karhunen, J. Joutsensalo. Representation and separation of signals using nonlinear PCA type learning. Neural Networks, 7, 113-127, 1994.

[30] J. Karhunen, J. Joutsensalo. Generalization of principal component analysis, optimization problems, and neural networks. Neural Networks, 8, 549-562, 1995.

[31] E. Oja, J. Karhunen, L. Wang, R. Vigario. Principal and independent components in neural networks – recent developments. Seventh Italian Workshop on Neural Networks, Vietri 1995, M.Marinaro & R.Tagliaferri Ed.s, World Scientific Pu. Singapore, 16-35, 1996.

[32] D. F. Morrison. *Multivariate statistical methods.* McGraw Hill, New York, 1976.

[33] J. E. Birren and D. F. Morrison. Analysis of WAIS subtests in relation to age and education. *Journal of Gerontology*, 16:363-369, 1961.

[34] S. Hashiguchi and H. Morishima. Estimation of genetic contribution of principal components to individual variates concerned. *Biometrics*, 25:9-15, 1969.

[35] S. Wold. Pattern recognition by means of disjoint principal components models. *Pattern Recognition*, 8:127-139, 1976.

[36] J. B. Lee, A. S. Woodyatt, M. Berman. Enhancement of high spectral resolution remote sending data by a noise-adjusted principal component transform. *IEEE Transactions on Geoscience and Remote Sensing*, 28(3):295-304, May 1990.

[37] C. H. Hermon and D. Mace. Use of Karhunen-Loeve transformation in seismic data processing. *Geophys. Prosp.*, 26:600-626, 1978.

[38] I. F. Jones and S. Levy. Signal-to-noise ratio enhancement in multichannel seismic data via the Karhunen-Loeve transform. *Geophys. Prosp.*, 35:12-32, 1987.

[39] C. Faloutsos, F. Korn, A. Labrinidis, Y. Kotidis, A. Kaplunovich, and D. Perkovic. Quantifiable data mining using principal component analysis. Technical report, 1997. Institute for Systems Research, University of Maryland technical Report TR 97-25.

[40] C. Merz and M. Pazzani. A principal components approach to combining regression estimates. *Machine Learning*, 36:9-32, 1999.

[41] M Tipping. Probabilistic visualisation of high-dimensional binary data. To appear in Advances in Neural Information Processing Systems 11. Editors: Michael S. Kearns, Sara A. Solla and David A. Cohn., 1999.