TECHNISCHE UNIVERSITÄT WIEN

DISSERTATION

# Non-Gaussian Feature Extraction for Complex Data

ausgeführt zum Zwecke der Erlangung des akademischen Grades
eines Doktors der Naturwissenschaften unter der Leitung von

**Klaus Nordhausen**
Institute of Statistics and Mathematical Methods in Economics, Vienna University of Technology
Department of Mathematics and Statistics, University of Jyväskylä

eingereicht an der Technischen Universität Wien
Fakultät für Mathematik und Geoinformation

von

**Una Radojičić**
Matrikelnummer: 11937135

Diese Dissertation haben begutachtet:

1. **Dipl.-Stat. University Lecturer Klaus Nordhausen, PhD**
   Institute of Statistics and Mathematical Methods in Economics, Vienna University of Technology
   Department of Mathematics and Statistics, University of Jyväskylä
2. **Prof. Dr. Uwe Schmock**
   Institute of Statistics and Mathematical Methods in Economics, Vienna University of Technology
3. **Associate Professor Pauliina Ilmonen, PhD**
   Department of Mathematics and Systems Analysis, Aalto University

Wien, September 14, 2021

# Abstract

In recent years, the size and complexity of the available data have grown rapidly, making visualization and exploratory data analysis very difficult. Therefore, researchers have proposed many methods to reduce the dimensionality of the data, by finding suitable data transformations that map the observed (measured) covariates into a smaller set of *features*, which hopefully then contain all the relevant or discriminatory information while simultaneously eliminating redundancies, and noise. Due to their simplicity, linear data transformations have been of special interest, and are very often obtained using the projection pursuit, a family of methods searching for, mostly univariate, projections of the data which maximize a predefined objective function, also known as projection index. Given the assumption that the data are a mixture of low-dimensional, non-Gaussian signal and independent high-dimensional Gaussian noise, the appropriate statistical framework is the non-Gaussian components analysis (NGCA) model. Finding the non-Gaussian components of the data is often considered as an important preprocessing step for efficient data analysis, thus making the aim of the feature extraction within the NGCA model to project the data onto a subspace that contains only the signal. Therefore, the aim of the thesis is to first study the feature extraction, as well as the estimation of the dimension of the feature subspace, under the multivariate non-Gaussian component model, with a special focus on homoscedastic Gaussian mixture model with two classes and the projection pursuit, and then to further extend the obtained methods to accommodate for the data which naturally allows a matrix representation such as e.g. grayscale images.

# Acknowledgement

First and foremost I wish to thank my supervisor Klaus Nordhausen, PhD for introducing me to the world of independent component analysis, for his continuous support during my study, and for never lacking patience and motivation, even though I myself many times did. His guidance helped deeply me in both the research as well as in writing this thesis, which without his help would have lacked a considerable amount of articles, among others. He never lacked a kind word or a recommendation for a nice place to visit in moments I was adjusting to life in Vienna. I would also like to express my gratitude to Joni Virta, PhD for introducing me to the interesting topic of unsupervised estimation of the linear discriminant, his patience and guidance in answering an infinite number of questions, and helping me fix just as many mistakes. I would like to thank both Klaus Nordhausen and Joni Virta for their willingness to host me at the Universities of Jyväskylä and Turku; I hope I will finally come to visit this year. I have been more than fortunate to have amazing co-authors for all my publications. Therefore, I would also like to deeply thank Professor Hannu Oja and Niko Lietzén, PhD. It was a great pleasure and experience to work with you.

I would like to express my gratitude to my referees Professor Uwe Schmock and Pauliina Ilmonen, PhD for their willingness to read and review this thesis. Furthermore, I would like to thank my dear friend, Iva Klaric for finding time in an insanely busy schedule to help to improve the language aspects of the thesis.

The research was financially supported by the interdisciplinary project *Blind Source Separation for Tensor-Valued Mass Spectral Data (GIP105000BSS)* and was carried during my work as a university assistant in the computational statistics group at the Institute of Statistics & Mathematical Methods in Economics, Vienna University of Technology. Therefore, I wish to thank all my colleagues from CSTAT group for making my time spent there a memorable experience.

I wish to warmly thank all my friends and family, my father Slavoljub as well as my brother Leon, sister Sara, and aunt Slavica whose doors have always been open to my visits and phones ready to take my call. Their constant encouragement persuaded me to continue my education in Vienna. Furthermore, I would like to thank my "roommates" Nikolina and Snježana, for all the wonderful time we spent together, as well as for giving me a sense of home and family in a foreign country. Especially, I would like to thank my fiancé Davorin, who put up with me all these years and supported and encouraged me to pursue all my dreams.

I would not have made it without all of you! Thank you!

# Eidesstattliche Erklärung

Ich erkläre an Eides statt, dass ich die vorliegende Dissertation selbstständig und ohne fremde Hilfe verfasst, andere als die angegebenen Quellen und Hilfsmittel nicht benutzt bzw. die wörtlich oder sinngemäß entnommenen Stellen als solche kenntlich gemacht habe.

Wien, am September 14, 2021

Una Radojičić

# Original Publications

This thesis consists of an introductory part and the following publications.

I. U. Radojicic and K. Nordhausen. Non-Gaussian component analysis: testing the dimension of the non-Gaussian subspace. In *M. Maciak, M. Pestas and M. Schindler, editors, Analytical Methods in Statistics, AMISTAT 2019*, pages 101–123, Springer Cham, 2019.

II. U. Radojicic, K. Nordhausen and J. Virta. Large-sample properties of blind estimation of the linear discriminant using projection pursuit. *arXiv preprint arXiv: 2103.04678*, 2021.

III. U. Radojicic, N. Lietzen, K. Nordhausen and J. Virta. Dimension estimation in two-dimensional PCA. In *Proceedings of the 12th International Symposium on Image and Signal Processing and Analysis*, pages 16–22, 2021.

IV. U. Radojicic, K. Nordhausen and J. Virta. Kurtosis-based projection pursuit for matrix-valued data. *arXiv preprint arXiv:2109.04167*, 2021.

V. U. Radojicic, K. Nordhausen and H. Oja. Notion of information and independent component analysis. *Applications of Mathematics*, 65:311–330, 2020.

# Contents

# Part I.

# Summary

# 1. Introduction

Nowadays, due to technological advancements, the size and complexity of the available data sets have grown rapidly, making visualization and exploratory data analysis very difficult. Therefore, in recent years, researchers have proposed many methods to reduce the dimensionality of the data and extract important data features. Simply stated, the goal of the feature extraction is to find a transformation of the data which maps the observed (measured) covariates into a smaller set of *features*, which hopefully then contain all the relevant or discriminatory information about the data while eliminating irrelevant data, redundancies, and noise (Foley and Sammon, 1975). Therefore, feature extraction is closely related to dimension reduction, and the two are often considered synonyms. Due to their simplicity, linear data transformations, i.e. those which map the original vector of covariates into the feature subspace through a projection matrix have been of special interest. Principal component analysis (PCA) (Jolliffe, 2002) and linear discriminant analysis (LDA) (Fisher, 1936) are perhaps the two most well-known linear feature extraction methods. In PCA, a projection matrix is chosen such that the obtained subspace accounts for as much of the variability in the data as possible, under the orthogonality of extracted directions. On the other hand, LDA projects the data onto the feature subspace which accounts for the maximum separability in data. In general, to direct the search for relevant low-dimensional features, while eliminating noise, statistical models may be specified. If one assumes the data are a mixture of low-dimensional, non-Gaussian signal and independent high-dimensional Gaussian noise, then the appropriate framework is the non-Gaussian components analysis (NGCA) model (Bickel and Levina, 2004). Accordingly, the aim of the feature extraction within the NGCA model is to project the data onto a subspace that contains only the signal (Blanchard et al., 2006; Kawanabe et al., 2007; Theis et al., 2011; Bean, 2014; Sasaki et al., 2016; Virta et al., 2016). A special case of the wide NGCA model is the non-Gaussian independent component model (NGICA), in which signal components are assumed to be independent, thus making the aim of the feature extraction in NGICA the recovery of the independent and non-Gaussian signals (Nordhausen et al., 2017; Risk et al., 2019). If data originates from a heterogeneous population, where the underlying populations (classes) have different means, then Everitt and Hand (1981) suggest a more parametric approach, thus proposing a mixture model as a suitable statistical framework, where perhaps the most famous among all mixture models are Gaussian mixture models (GMM). Nordhausen et al. (2017) argue how the $p$-variate homoscedastic GMM with $p+1$ classes is an NGCA model, while in Paper II it is shown that any homoscedastic GMM with 2 classes is in fact an NGICA model. Feature extraction within GMM is mostly done with the aim of optimal group separation. Even though LDA is essentially a model-free method, it bears a close connection to the homoscedastic Gaussian mixture model with two classes, where it is shown to be optimal in the sense that it generates a projection direction that is used to construct the optimal Bayes classifier (Rao, 1948).

Feature extraction is intimately connected to the projection pursuit (PP) (Huber, 1985), where PP denotes a family of methods searching for, mostly univariate, projections of the

data, which maximize a predefined objective function (projection index). Thus, extracted features usually maximize some criterion of high-variability, non-Gaussianity, or independence specified by a projection index.

Besides high-dimensional multivariate data, the availability and the need for analysis of matrix-variate data, such as e.g. image data (also known as *image processing*) has also increased rapidly over the years, thus naturally raising the question of dimension reduction and feature extraction in the matrix-variate setting. Examples of such data are abundant nowadays, the most notable example being image data where the elements of the matrix represent the gray-scale intensities of the individual pixels of an image. For example, in a coronary tissue (CT) image each pixel represents a tissue and it has an assigned value of grayscale level between 0 and 255. This grayscale value represents the X-ray beam attenuation to the tissue. Pixels with values close to 0 (darker pixels) represent structures having less attenuation to the beam, i.e. soft tissue, while pixels close to 255 (light pixels) represent structures having high attenuation, i.e., calcifications (Athanasiou et al., 2017). Thus, extracting features that would characterize the existence and the type of calcification in the soft tissue could be used for diagnostic purposes, by allowing classification of CT images w.r.t presence and the type of the calcification. In general, image classification and recognition, as well as image compression are classical problems in image processing.

When the observations are matrix-valued, traditional data analysis techniques can be implemented to tackle presented problems by first vectorizing the observed matrices into a long vector. However, this approach is often suboptimal, since it ignores the underlying data structure while at the same time producing a high-dimensional vector, making further data analysis difficult. Thus, the aim of the thesis is to first study the feature extraction, as well as the estimation of the dimension of the feature subspace, under the multivariate NGCA model, with a special focus on homoscedastic GMM with two classes and the projection pursuit, and then to further extend the obtained methods to accommodate for the matrix structure of the data.

The thesis is structured as follows. The first part of the thesis considers feature extraction and dimension reduction for vector-valued data under various multivariate models. In the first section, we define location and scatter functionals and focus on feature extraction, noise, and dimension reduction in the scope of elliptical models. The following two sections consider dimension reduction in the scope of non-Gaussian component models and independent component models, using two-scatter matrices and the projection pursuit approach. We conclude the discussion on the feature extraction in the vector-valued models with Gaussian mixture models, with a special focus on two-class models and linear discriminant analysis. In cases where data are matrix-valued, as discussed, ignoring the matrix structure loses information about the data, thus implying that simply applying methods for vector-valued data to vectorized matrices yields sub-optimal procedures. Thus, the second part of the thesis focuses on feature extraction and dimension reduction for matrix-valued data. Section 3.2 considers dimension reduction in the context of the noisy second-order matrix model, while in Section 3.4 we discuss the use of orthogonal rank-1 tensor projections in the context of LDA with aim of data clustering using various projection indices with emphasis on matrix Gaussian mixtures. Throughout the thesis, various properties of data were discussed as projection pursuit indices in the quest for meaningful directions in multivariate data; cumulants, entropy, Gaussianity, independence, to name a few. Thus, we conclude with a discussion on some characteristics of univariate distributions as well as

their application to the ordering of random variables. The importance of Chapter 4 lays in the connections between information, non-Gaussianity, and statistical independence in the context of independent component analysis, which is often taken for granted. Chapter 5 then concludes the thesis with a discussion on future work.

# 2. Feature extraction for multivariate data

## 2.1. Location-scatter model and its generalizations

Before the main discussion, we introduce the general framework under which we will be working in this chapter. Namely, all models discussed in this section can be derived from the location-scatter model or its extensions (Oja, 2010). The location-scatter model we consider is

$$\mathbf{x} = \mathbf{A}\mathbf{z} + \boldsymbol{\mu}, \tag{2.1}$$

where $\mathbf{x}$ is an observable $p$-variate random vector, $\mathbf{z}$ is a latent $p$-variate random vector whose distribution we will discuss in each model separately, $\boldsymbol{\mu}$ is a $p$-variate location vector and $\mathbf{A}$ is a full-rank, $p \times p$-mixing matrix. Usually, the latent vector $\mathbf{z}$ is assumed to be standardized in a way specified later in the thesis, and imposing various assumptions on the distribution of $\mathbf{z}$ yields a large variety of multivariate models. For example, assuming that $\mathbf{z}$ has a standard normal distribution yields the multivariate normal model for $\mathbf{x}$.

### 2.1.1. Multivariate normal model

The multivariate normal model is probably the most studied model in the multivariate analysis. In the context of the location-scatter Model (2.1), $\mathbf{z}$ is required to have a multivariate standard normal distribution, $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_p)$, yielding $\mathbb{E}(\mathbf{x}) = \boldsymbol{\mu}$ and $\mathrm{Cov}(\mathbf{x}) = \mathbf{A}\mathbf{A}'$. The probability density function of the random vector $\mathbf{x}$ from a multivariate normal distribution is given by

$$\phi(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{p/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\{-\tfrac{1}{2}\|\boldsymbol{\Sigma}^{-1/2}(\mathbf{x} - \boldsymbol{\mu})\|_2^2\}.$$

Note that the multivariate normal model is fully characterized by its mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. The multivariate normal model is also fully symmetric around the location $\boldsymbol{\mu}$ and all of its marginal distributions are normally distributed. Thus, the kurtosis of each component is the same and is equal to 3. The standard normal distribution $\mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I}_p)$, $\sigma^2 > 0$ is the only spherical distribution with independent components, thus guaranteeing principal component analysis to find independent projections (Bilodeau and Brenner, 1999).

However, the multivariate normal model has a drawback in not being able to model heavy-tailed phenomena. Hence, a natural generalization of the multivariate normal model is the elliptical model.

### 2.1.2. Elliptical model

Before considering the elliptical model, let us define the spherical model.

**Definition 1.** *A $p$-variate random vector $\boldsymbol{x}$ has a spherical distribution if $\boldsymbol{x} - \boldsymbol{\mu} \sim \boldsymbol{U}(\boldsymbol{x} - \boldsymbol{\mu})$, for all orthogonal matrices $\boldsymbol{U} \in \mathcal{O}^{p\times p}$. The $p$-variate vector $\boldsymbol{\mu} \in \mathbf{R}^p$ is the location vector of $\boldsymbol{x}$.*

From the definition of the spherical model, it is clear that it is symmetric around its location and that all of its marginals are equally distributed. This further implies that, under the assumption that the first two moments exist, $\mathrm{Cov}(\mathbf{x}) = \sigma^2\mathbf{I}_p$, for $\sigma^2 > 0$. Since it is invariant to rotations, one can argue that all directions are equally (un)interesting. The elliptical model is then a special case of location-scatter model

$$\mathbf{x} = \mathbf{A}\mathbf{z} + \boldsymbol{\mu}, \tag{2.2}$$

where the latent vector $\mathbf{z} \in \mathbb{R}^p$ has spherical distribution around the origin. The density function $f$ of a random vector $\mathbf{x}$ which has an elliptical distribution is given by

$$f(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = |\boldsymbol{\Sigma}|^{-1/2}\exp\{-g(\|\boldsymbol{\Sigma}^{-1/2}(\mathbf{x} - \boldsymbol{\mu})\|_2^2)\},$$

where $\boldsymbol{\mu}$ is a location vector as above, $\boldsymbol{\Sigma}$ is a symmetric semi-positive definite matrix, which determines the scale and the correlation structure of $\mathbf{x}$ and $g$ is a real-valued function (Frahm et al., 2003). Assuming the existence of first two moments, $\boldsymbol{\mu} = \mathbf{E}(\mathbf{x})$ and $\mathbf{Cov}(\mathbf{x}) \propto \boldsymbol{\Sigma}$. The multivariate normal distribution is a member of the elliptical model, with $g(t) = 1/2t + p/2\ln 2/\pi$. Another well-known member of the elliptical model is the family of the multivariate $t$-distributions (Kotz and Nadarajah, 2004), as well as the uniform distribution on an ellipsoid. Thus, the elliptical model is an extension of the normal model which allows for both lighter and heavier tails, while still requiring that all the marginals are alike in shape. Owen and Rabinovitch (1983) argue how for full rank matrix $\mathbf{B} \in \mathbb{R}^{p\times d}$, the projection $\mathbf{B}'\mathbf{x}$ is again a member of an elliptical model. Especially, all the univariate projections of the random vectors from an elliptical model are symmetric. It is important to note that the elliptical model is identifiable only up to post-multiplication by an orthogonal matrix, in sense that if $p \times p$ matrix $\mathbf{W}$ recovers the latent vector $\mathbf{z} = \mathbf{W}(\mathbf{x} - \boldsymbol{\mu})$, then due to spherical symmetry of $\mathbf{z}$, so does $\mathbf{WU}$, for any orthogonal matrix $\mathbf{U} \in \mathcal{O}^{p\times p}$.

In general, feature extraction in the elliptical models is closely related to the spread, and interesting data projections are usually found using the first two moments, assuming these exist. For that purpose, we next discuss general location and scatter functionals.

## 2.2. Location and scatter functionals and their usage

Let again $\mathbf{x}$ be a $p$-variate random vector with distribution function $F_\mathbf{x}$. Then a $p$-vector-valued functional $\mathbf{T}(F_\mathbf{x}) = \mathbf{T}(\mathbf{x})$ is called a location functional if it is affine equivariant in the sense that

$$\mathbf{T}(\mathbf{A}\mathbf{x} + \mathbf{b}) = \mathbf{A}\mathbf{T}(\mathbf{x}) + \mathbf{b},$$

for all full rank $p \times p$ matrices $\mathbf{A}$ and all $p$-variate vectors $\mathbf{b}$.

A $p \times p$ matrix-valued functional $\mathbf{S}(F_\mathbf{x}) = \mathbf{S}(\mathbf{x})$ is called a scatter functional if it is symmetric, positive semi-definite and affine equivariant in the sense that

$$\mathbf{S}(\mathbf{A}\mathbf{x} + \mathbf{b}) = \mathbf{A}\mathbf{S}(\mathbf{x})\mathbf{A}',$$

for all full rank $p \times p$ matrices $\mathbf{A}$ and all $p$-variate vectors $\mathbf{b}$ (Oja, 2010). A broader class of scatter functionals are so-called orthogonally equivariant scatters, which are required to satisfy affine equivariance only for orthogonal matrices. Thus, location and scatter functionals are a way to describe the centrality and spread of the data and are estimated by

replacing $F_{\mathbf{x}}$ with the empirical distribution. Furthermore, location and scatter functionals are often used to derive various skewness and kurtosis measures. More precisely, for two location functionals $\mathbf{T}_1$, $\mathbf{T}_2$ and two scatter functionals $\mathbf{S}_1$, $\mathbf{S}_2$, $\mathbf{T}_1(\mathbf{x}) - \mathbf{T}_2(\mathbf{x})$ is a measure of skewness of $\mathbf{x}$, while the eigenvalues of $\mathbf{S}_1^{-1}(\mathbf{x})\mathbf{S}_2(\mathbf{x})$ can be seen as a kurtosis measures in the direction of the corresponding eigenvectors (Tyler et al., 2009).

Probably the most widely used pair of a location and scatter functionals are the expected value $\mathbb{E}(\mathbf{x})$ and the covariance matrix

$$\mathrm{Cov}(\mathbf{x}) = \mathbb{E}((\mathbf{x} - \mathbb{E}(\mathbf{x})(\mathbf{x} - \mathbb{E}(\mathbf{x})').$$

However, estimation of some location and scatter functionals, including mean and covariance, is heavily influenced by the presence of outliers, and not efficient for heavy-tailed distributions. Under the elliptical model, location and scatter functionals have interesting properties. If $\mathbf{x}$ has an elliptical distribution with $p$-variate location vector $\boldsymbol{\mu}$, then $\mathbf{T}(\mathbf{x}) = \boldsymbol{\mu}$, for all location functionals $\mathbf{T}$, provided that they exist. Furthermore, assuming the covariance matrix $\mathrm{Cov}(\mathbf{x})$ exists, then $\mathbf{S}(\mathbf{x}) \propto \mathrm{Cov}(\mathbf{x})$, for all scatter matrices $\mathbf{S}$. This specifically means that in the elliptical model, all location functionals are equal and correspond to the center of symmetry, while all scatter matrices are proportional to each other, and especially to the covariance matrix if it exists (Oja et al., 2006), thus estimating the symmetry center and the scatter.

The literature is full of alternatives for the mean vector and the covariance matrix, which have different desirable properties, like robustness or efficiency, at specific models. A large family of functionals which we will discuss in the following in more detail are the $M$-estimators of location and scatter (Maronna, 1976). Some additional classes of robust location and scatter functionals are $S$-functionals (Davies, 1987) and $\tau$-functionals (Lopuhaä, 1991), just to name a few, which are all constructed for inference only in the scope of elliptical models. For a general review of robust estimators of multivariate location and scatter functionals see Maronna and Yohai (2016).

### M-estimators of location and scatter

$M$-functionals of location and scatter were first introduced by Maronna (1976) and are usually jointly estimated as the solutions of the two following implicit equations:

$$\mathbf{T}(\mathbf{x}) = \mathbb{E}\left(w_1(r)\right)^{-1}\mathbb{E}(w_1(r)\mathbf{x})$$

and

$$\mathbf{S}(\mathbf{x}) = \mathbb{E}\left(w_2(r)\left(\mathbf{x} - \mathbf{T}(\mathbf{x})\right)\left(\mathbf{x} - \mathbf{T}(\mathbf{x})\right)'\right),$$

where $w_1(r)$ and $w_2(r)$ are nonnegative continuous functions of the (pseudo) Mahalanobis distance $r = ||\mathbf{S}(\mathbf{x})^{-1/2}(\mathbf{x} - \mathbf{T}(\mathbf{x}))||$. Thus, one can think of $M$-functionals of location and scatter as weighted variants of the mean and the covariance matrix yielding them as special cases when choosing $w_1(r) = w_2(r) = 1$. Usually the weight functions are chosen to be non-increasing to obtain estimators that may be robust. Some popular members of the family of $M$-estimators are Huber's $M$-estimators (Huber, 1964) which have the weight functions

$$w_1(r) = \begin{cases} 1 & r \le c \\ c/r & r > c \end{cases} \quad \text{and} \quad w_2(r) = \begin{cases} 1/\sigma^2 & r \le c \\ c/(r^2\sigma^2) & r > c \end{cases}.$$

The scaling factor $\sigma^2$ is chosen so that $\mathbb{E}(Qw_2(\sqrt{Q})) = p$ and $c$ is a tuning constant chosen to satisfy $\mathbb{P}(Q \leq c^2) = q$, where $Q \sim \chi^2_p$, and $q \in (0, 1)$. Another class of $M$-estimators are those based on the likelihood of a $t$-distribution having $\nu \geq 1$ degrees of freedom (Kent and Tyler, 1991), which yields weight functions

$$w_1(r) = w_2(r) = \frac{p + \nu}{r^2 + \nu}.$$

A special case of $M$-estimators based on the $t$-distribution are Cauchy $M$-estimators, which correspond to $t$-estimators with $\nu = 1$. Traditionally, $M$-estimators of location and scatter are computed via fixed-point algorithms which are iterated from an initial starting point until the difference in successive functional values is less than some predetermined threshold (Huber, 1964; Kent and Tyler, 1991). Depending on the weight functions there are however also other algorithms available. For example, a gradient descent method and a partial Newton-Raphson method are discussed in Dümbgen et al. (2016). For a recent general review of M-estimators see Düembgen et al. (2013).

Since it can sometimes be computationally demanding running the whole iterative process until convergence, a compromise in the iterative process are the so-called one-step $M$-estimators of location and scatter. One step $M$-estimators start with a pair of location and scatter functionals $(\mathbf{T}_1, \mathbf{S}_1)$ and two weight functions $w_1 = w_1(r)$ and $w_2 = w_2(r)$ and then use just one updating step to obtain weighted new functionals

$$\mathbf{T}_2(\mathbf{x}) = \mathbb{E}\left(w_1(r_1)\right)^{-1}\mathbb{E}(w_1(r_1)\mathbf{x}), \quad \mathbf{S}_2(\mathbf{x}) = \mathbb{E}\left(w_2(r_1)\left(\mathbf{x} - \mathbf{T}_1(\mathbf{x})\right)\left(\mathbf{x} - \mathbf{T}_1(\mathbf{x})\right)'\right),$$

where $r_1 = \|\mathbf{S}_1(\mathbf{x})^{-1/2}(\mathbf{x} - \mathbf{T}_1(\mathbf{x}))\|$. A scatter functional from this family which we will consider later is the scatter matrix of fourth moments which starts with the pair of location and scatter functionals $(\mathbf{T}_1, \mathbf{S}_1)$=$(\mathbb{E}, \mathrm{Cov})$ and weight functions $w_1(r) = r^2$, $w_2(r) = r^2/(p+2)$. The resulting location-scatter pair $(\mathbf{T}_2, \mathbf{S}_2)$ are then the vector of the third moments $\mathbf{T}_2(\mathbf{x}) = \mathbb{E}_3(\mathbf{x})$ and the matrix of the fourth moments $\mathbf{S}_2(\mathbf{x}) = \mathrm{Cov}_4(\mathbf{x})$ defined as

$$\mathbb{E}_3(\mathbf{x}) = \frac{1}{p}\mathbb{E}(r^2\mathbf{x}), \quad \mathrm{Cov}_4(\mathbf{x}) = \frac{1}{p+2}\mathbb{E}\left(r^2\left(\mathbf{x} - \mathbb{E}(\mathbf{x})\right)\left(\mathbf{x} - \mathbb{E}(\mathbf{x})\right)'\right),$$

where $r = ||\mathrm{Cov}(\mathbf{x})^{-1/2}(\mathbf{x} - \mathbb{E}(\mathbf{x}))||$.

### Independence properties of scatter functionals

Even though scatter and location functionals are discussed mainly in the context of elliptical models, they have an important role in analyzing independent component models as well. However, as the Gaussian distribution is the only elliptical distribution with independent components, additional properties of scatter functionals are of interest when exploring meaningful directions of multivariate vectors from a non-elliptical model. One of such properties is independence. A scatter functional $\mathbf{S}(\mathbf{x})$ is said to have the (full) independence property if

$$\mathbf{S}(\mathbf{x}) = \mathbf{D}(\mathbf{x})$$

for all $\mathbf{x}$ having independent components, where $\mathbf{D}(\mathbf{x})$ denotes a diagonal matrix. Moreover, if the $p$-variate vector $\mathbf{x} = (\mathbf{x}_1', \ldots, \mathbf{x}_k')'$ has $k$ independent sub-vectors with corresponding

block dimensions $p_1, \ldots, p_k$, then a scatter functional $\mathbf{S}(\mathbf{x})$ is said to have the block independence property if

$$\mathbf{S}(\mathbf{x}) = \mathbf{B}(\mathbf{x}),$$

where $\mathbf{B}(\mathbf{x})$ is the block diagonal matrix with block dimensions $p_1, \ldots, p_k$. If $\mathbf{x}$ has independent components, we can consider each component to be a univariate block. Therefore, block independence is the stronger assumption and it implies full independence.

Most scatter functionals do not possess the full or block independence property, however Cov and $\text{Cov}_4$ do. On the other hand, Nordhausen and Tyler (2015) argue how all scatter functionals $\mathbf{S}(\mathbf{x})$ are diagonal and block diagonal if all but one of the independent parts of $\mathbf{x}$ are symmetric. Thus, exploiting the concept of symmetry, we can define symmetrized scatter functionals. More precisely, let $\mathbf{S}$ denote any scatter functional. Then its symmetrized version is defined as

$$\mathbf{S}_{sym}(\mathbf{x}) := \mathbf{S}(\mathbf{x}^1 - \mathbf{x}^2),$$

where $\mathbf{x}^1$ and $\mathbf{x}^2$ are independent copies of $\mathbf{x}$. For example, Nordhausen and Tyler (2015) show that every symmetrized scatter functional possess both full and block independence properties. Observe that symmetrized scatter functionals do not require a location functional. In fact, they are usually computed using all pairwise differences and then computing the original scatter with respect to the origin. Interestingly, both Cov and $\text{Cov}_4$ can be expressed as functions of pairwise differences. Symmetrized $M$-estimators of scatter are investigated in Sirkiä et al. (2007), while the computational issues are especially discussed in Dümbgen et al. (2016) and Miettinen et al. (2016).

### 2.2.1. Principal component analysis

Principal component analysis (PCA) (Jolliffe, 2002) is a data analysis method defined as an orthogonal linear transformation, that creates a new coordinate system (rotates the data set), so that new coordinates are uncorrelated. More precisely, the analysis aims to create a new set of axes along which the variation of the data is maximized. If we write

$$\text{Cov}(\mathbf{x}) = \mathbf{U}\mathbf{D}\mathbf{U}'$$

to be the eigendecomposition of $\text{Cov}(\mathbf{x})$, where $\mathbf{D} = \text{diag}(d_1, \ldots, d_p)$, $d_1 \geq d_2 \geq \cdots \geq d_p \geq 0$ is a diagonal matrix containing the ordered eigenvalues of $\text{Cov}(\mathbf{x})$, and $\mathbf{U}$ is an orthogonal matrix containing the eigenvectors of $\text{Cov}(\mathbf{x})$ as columns, then, projecting the observations along the eigenvectors in $\mathbf{U}$, one obtains the principal component scores (features) $\mathbf{y} = \mathbf{U}'\mathbf{x}$. The covariance matrix of such obtained features is now diagonal, implying that the features are uncorrelated, while the eigenvalues correspond to variances of the corresponding components.

In general, provided that the eigenvalues of $\text{Cov}(\mathbf{x})$ are distinct, ordering of the eigenvalues in $\mathbf{D}$ implies ordering of corresponding features in $\mathbf{y}$. In the context of PCA, features obtained by projecting data onto eigenvectors of $\text{Cov}(\mathbf{x})$ along which the variance is larger (that correspond to larger eigenvalues) are considered more important. Then, one usually discards a certain number of features with the lowest variance, thus obtaining a data transformation of a lower dimension. The question is now, how many featuers to keep? Before we tackle this question, let us discuss the special connection between PCA and elliptical models.

In the case where $\mathbf{x}$ is a $p$-variate random vector from elliptical model (2.2), the covariance matrix of $\mathbf{x}$ is, as discussed,

$$\mathrm{Cov}(\mathbf{x}) = \sigma^2 \mathbf{A}\mathbf{A}',$$

for $\sigma^2 > 0$, which corresponds to variance of each individual component in $\mathbf{z}$. In that case, $\mathbf{U}$ corresponds to left-singular matrix of $\mathbf{A}$, making $\mathbf{y} = \mathbf{D}^{1/2}\mathbf{V}'\mathbf{z} + \mathbf{U}'\boldsymbol{\mu}$, the solution to the elliptical model (which is identifiable up to post-multiplication by an orthogonal matrix), thus implying the connection between the PCA and elliptical models. Another reason we stress the connection of PCA and elliptical models is that the most common use of the elliptical models in practice is to craft distributions that share some key properties of the normal distribution while at the same time having heavier tails, making estimation of standard covariance matrix in these models potentially not efficient. However, the previous derivation of PCA in elliptical models holds not just for the covariance matrix but for any affine equivariant scatter matrix (remember that these are all proportional), thus potentially obtaining a more robust procedure (Marden, 1999; Visuri et al., 2000).

### Dimension reduction in PCA

The question of dimension reduction becomes more straightforward if we assume that $\mathrm{Cov}(\mathbf{x})$ is a singular matrix, i.e. if some of the measured features in $\mathbf{x}$ are mutually linearly dependent and thus obsolete. This raises the question of estimation of rank of $\mathrm{Cov}(\mathbf{x})$. However, it is more often in practice for data to be measured with a certain error, which is usually independent of the signal part. Thus, Jolliffe (2002) proposes the so-called principal component model

$$\mathbf{x} = \mathbf{s} + \mathbf{n}, \tag{2.3}$$

where the signal component $\mathbf{s}$ and noise component $\mathbf{n}$ are independent and are such that $\mathrm{Cov}(\mathbf{s}) = \mathbf{M}$ is singular with rank $d$ and $\mathrm{Cov}(\mathbf{n}) = \sigma_{\mathbf{n}}^2 \mathbf{I}_p$. The problem now is the estimation of the rank of the matrix $\mathbf{M}$. The inference on the dimension of the signal subspace in PCA is generally done using eigenvalues of the sample covariance matrix, see e.g. Jolliffe (2002) and Schott (2006) and references therein. A popular graphical technique for selecting the number of significant components is the plot of eigenvalues of the covariance matrix in the decreasing order, also known as the *scree plot*. One then looks for the "elbow" in the plot, i.e. the point where the eigenvalues seem to equalize, thus selecting the components prior to this point as significant.

However, variation in the eigenvectors of $\mathrm{Cov}(\mathbf{x})$ also carries information on the dimension of the signal subspace. Simply stated, when the eigenvalues of a random matrix are close together, their eigenvectors tend to vary greatly, while when the eigenvalues are far apart, their variability tends to be small (Luo and Li, 2016). Methods based on detecting changes in the variation of eigenvectors mostly use bootstrap resampling techniques to approximate the variation of the span of the first $k$ sample eigenvectors, where high variation indicates that the chosen eigenvectors belong to the same eigenspace, i.e. the difference between the corresponding eigenvalues is small, as it is the case for eigenvectors corresponding to the negligible eigenvalues. For more details see Ye and Weiss (2003). The work in Luo and Li (2016) combines the previous two methodologies, where the negligible eigenvalues are identified by using both the information on the magnitude of the eigenvalue as well as the variability in the corresponding eigenspace. Luo and Li (2021) suggested an alternative approach in estimating the variation in eigenvectors, by employing data augmentation

and argues that, if a suitable random component $\mathbf{x}_S$ of dimension $r > 0$, mimicking the first and the second order behaviour of $\mathbf{n}$ is added to the observed $\mathbf{x} \in \mathbb{R}^p$, thus obtaining $\mathbf{x}^* = (\mathbf{x}', \mathbf{x}_S')' \in \mathbb{R}^{p+r}$, then the augmented parts of the first $d$ eigenvectors of the matrix $\mathrm{Cov}(\mathbf{x}^*)$ are smaller in magnitude than the augmented parts of the latter eigenvectors of the corresponding matrix. When combining this information with the information from the eigenvalues, one obtains a curve whose minimum inclines to occur at the true dimension. For more insight see Figure 2.1, where one can see how information obtained from the norm of augmented subvectors greatly helps in determining how many components should be kept. Interesting to mention is the connection between the multivariate normal distribution

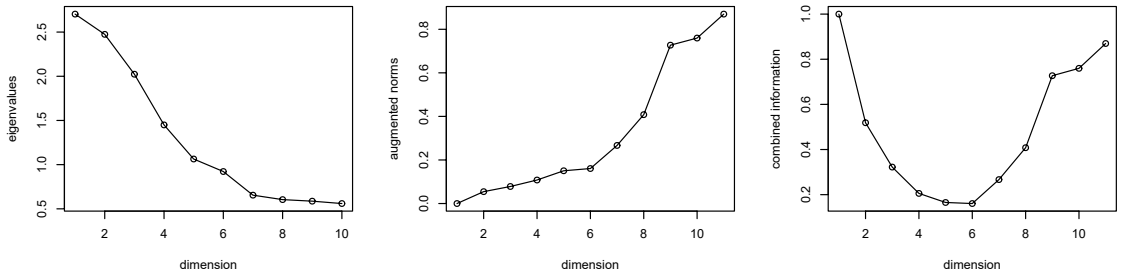



Figure 2.1.: Left to right, scree plot, augmented norms and the objective function for the augmented estimator using $r = 5$ as a combination of the augmented norms and the scaled eigenvalues, calculated for the data from Model (2.3) with the signal $\mathbf{s}$ from the multivariate normal distribution with mean $\mathbb{E}(\mathbf{s}) = \mathbf{0}$ and the covariance matrix $\mathrm{Cov}(\mathbf{s}) = \mathrm{diag}(2.1, 1.6, 1.1, 0.6, 0.1, 0.1, 0, 0, 0, 0)$, and the additive noise from the multivariate normal distribution with $\mathbb{E}(\mathbf{n}) = \mathbf{0}$ and $\mathrm{Cov}(\mathbf{n}) = 0.9\,\mathbf{I}_{10}$. Signal dimension is $d = 6$ and coincides with the minimum in the right panel.

and the dimension reduction using PCA. Namely, Linsker (1988) showed that if $\mathbf{x}$ obeys the PCA model (2.3) in which additionally both the signal and the noise have a multivariate normal distribution, then PCA maximizes the mutual information between the signal and the reduced transformation of $\mathbf{x}$. Tipping and Bishop (1999) demonstrate how the principal axes may be determined through maximum-likelihood estimation of parameters in a latent variable model, where both the distribution of latent vector and the conditional distribution of the observable vector given the latent vector are Gaussian. The approach is known as the probabilistic PCA.

### 2.2.2. Whitening

Whitening is a basic data transformation that subtracts location to shift the data to the origin and then rotates data so that new components are uncorrelated. Finally, it rescales rotated components to have unit variance. Formally, the transformation is

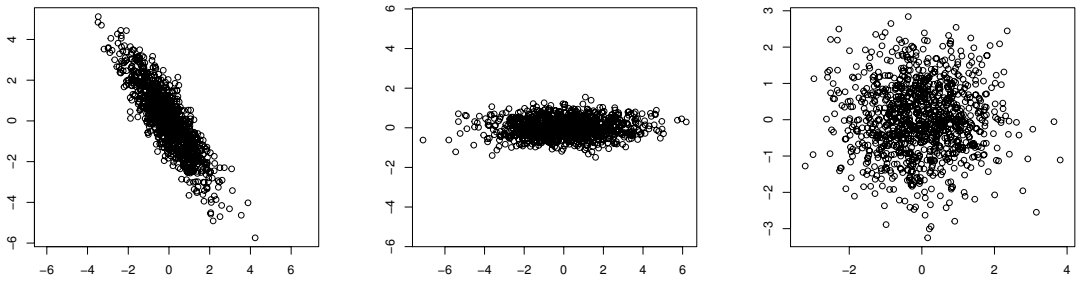
$$\mathbf{x}_{st} = \mathrm{Cov}(\mathbf{x})^{-1/2}(\mathbf{x} - \mathbb{E}(\mathbf{x})),$$

and satisfies

$$\mathbb{E}(\mathbf{x}_{st}) = \mathbf{0}, \quad \mathrm{Cov}(\mathbf{x}_{st}) = \mathbf{I}_p.$$

Observe that if $\mathbf{x}$ has a multivariate normal distribution, then $\mathbf{x}_{st}$ has independent and identically distributed margins, while in the case where $\mathbf{x}$ follows an elliptical model, $\mathbf{x}_{st}$ has a spherical distribution. The difference between PCA and whitening is that in PCA, one does not rescale the components to have the unit variance, but uses this information to order obtained features. For more insight see Figure 2.2.

No direction in whitened $\mathbf{x}_{st}$ is now more interesting than the other, with respect to the scale. Whitening is often referred to as standardization, and can be done using any pair of a location and scatter functionals, not necessarily only $(\mathbb{E}, \mathrm{Cov})$. For more details on whitening see for example Ilmonen et al. (2012).



(a) Original data (b) Principal components (c) Whitened data

Figure 2.2.: Left to right, a sample of size 500 from the centered multivariate normal distribution, corresponding principal components, and whitened data.

## 2.3. Non-Gaussian component model

It is often that the observed data is a mixture between low-dimensional signal and noise which originates from various sources, making the non-Gaussian component (NGCA) model (Blanchard et al., 2005) the right statistical framework to guide the feature extraction in that case. The NGCA model is a location-scatter model

$$\mathbf{x} = \mathbf{A}\mathbf{z} + \boldsymbol{\mu} = \mathbf{A}_1\mathbf{s} + \mathbf{A}_2\mathbf{n} + \boldsymbol{\mu},$$

where $\mathbf{z} = (\mathbf{s}', \mathbf{n}')'$ is a latent $p$-variate vector consisting of the $q$-variate non-Gaussian signal vector $\mathbf{s}$ and the $(p-q)$-variate Gaussian noise vector $\mathbf{n}$. The signal and noise vectors are independent, and both are assumed to be standardized. The full-rank matrices $\mathbf{A}_1 \in \mathbb{R}^{p\times q}$ and $\mathbf{A}_2 \in \mathbb{R}^{p\times(p-q)}$ specify the signal and noise parts of $\mathbf{x}$ respectively, while $p$-variate vector $\boldsymbol{\mu}$ is a location vector.

NGCA model can be seen as a special case of independent subspace analysis (ISA) (Cardoso, 1998) with two independent blocks, where ISA assumes that the latent vector $\mathbf{z}$ consists of $k$ independent sub-vectors. The aim is then to find an estimate of a transformation matrix to recover the independent sub-vectors. For details about ISA see for example Theis (2007); Nordhausen and Oja (2016).

The aim of the feature extraction within the NGCA model is to identify the signal part of $\mathbf{x}$, i.e., to project the data onto the signal subspace. Thus, one seeks a $p \times p$ full rank unmixing block matrix $\mathbf{W} = (\mathbf{W}_1' \mathbf{W}_2')'$ with submatrices $\mathbf{W}_1$ and $\mathbf{W}_2$, such that $\mathbf{W}_1\mathbf{x}$ recovers the non-Gaussian signal subspace and $\mathbf{W}_2\mathbf{x}$ the Gaussian noise subspace. Observe that matrices $\mathbf{A}_1$ and $\mathbf{A}_2$ are identifiable only up to post-multiplication with $q \times q$ and $(p-q)\times(p-q)$ dimensional orthogonal matrices, respectively. Thus, $\mathbf{A}$, and consequentially $\mathbf{W}$ are not identifiable either, making the goal of NGCA analysis to estimate subspace spanned by columns of $\mathbf{W}_1$. The identifiability issue is closely related to the lack of the assumptions posed to signal component $\mathbf{s}$, thus resulting that the individual signals can not be recovered.

A special case of the NGCA model in which all components of signal $\mathbf{s}$ are independent (remember that components of $\mathbf{s}$ are only needed to be uncorrelated in the NGCA model) is a non-Gaussian independent component (NGICA) model. The NGICA model has the advantage over the general NGCA model in that the signal components of $\mathbf{s}$ are identifiable up to their order and signs, thus making the aim of the feature extraction within the NGICA model to estimate the individual signals. NGICA was for example considered in Nordhausen et al. (2017); Risk et al. (2019).

If we further assume that the data originates from the various independent signal sources and at most a single noise source, we find ourselves in the domain of the independent component (ICA) model.

### 2.3.1. Independent component model

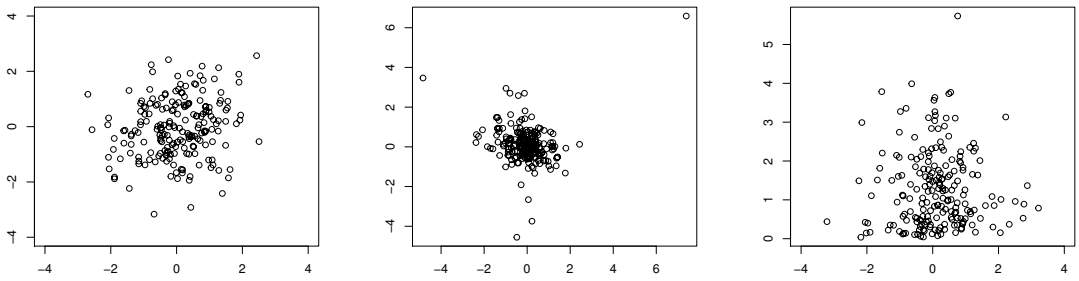
The ICA model is an NGICA model in which the noise component is either univariate or does not exist at all. More precisely, the independent component model is a location-scatter model

$$\mathbf{x} = \mathbf{A}\mathbf{z} + \boldsymbol{\mu}, \tag{2.4}$$

for $p$-variate, standardized, latent vector $\mathbf{z}$ with independent components and at most one normally distributed component, a non-singular matrix $\mathbf{A}$ and the location vector $\boldsymbol{\mu}$. Restrictions on the number of the Gaussian components of $\mathbf{z}$ are due to the identifiability of the model. Remember that the standard normal distribution is spherical, and is thus invariant to orthogonal transformations. The aim of the feature extraction in the ICA model is then to recover independent components. The reason we make a distinction from the wider, NGICA model is that in the NGICA model, the dimension of the signal subspace is usually unknown. Thus, the problem of feature extraction is intertwined with the estimation of the dimension of the signal subspace.

Even though the ICA model is significantly narrower than the wide NGCA model, it is still fitting to model e.g. skewed or even clustered data. Figure 2.3 shows scatter plots of random samples from the standard normal model, heavy-tailed elliptical model, and ICA model, showcasing how these can differ.

For more details on the independent component model see for example Hyvärinen et al. (2001); Hyvärinen (1999); Comon and Jutten (2010); Nordhausen and Oja (2018).



(a) Multivariate normal model (b) Elliptical model (c) ICA model

Figure 2.3.: From left to right, random samples of size 200 from the multivariate normal distribution, multivariate t-distribution with 3 degrees of freedom, and an independent component model with $t_3$ and $\chi^2_3$ components. Each sample is standardized.

### 2.3.2. Independent component analysis

As discussed, feature extraction in the ICA model is done with the aim of extraction of the independent components. Hyvärinen and Oja (1997) make a heuristic argument on how a sum of two independent random variables, as a consequence of central limit theorem, usually has a distribution that is closer to Gaussian than any of the two original random variables, thus justifying one of the principles of ICA given in Hyvärinen and Oja (1997): "Non-Gaussian is independent". Therefore, exploring various measures of non-Gaussianity, like e.g., kurtosis and negentropy, one obtains a wide range of algorithms for ICA, most of which are projection pursuit based. Moreover Hyvärinen and Oja (1997) make an explicit connection between the ICA and PP, emphasizing that if the ICA model holds, optimization of the non-Gaussianity measures produces independent components, while if the model does not hold, then what is obtained are the projection pursuit directions.

#### Projection pursuit based ICA

One of the most popular ICA methods is the so-called fastICA algorithm (Hyvärinen and Oja, 1997). It is a PP based method, where the projection index is tailored such that the extracted components maximize negentropy, which is mostly taken as a non-Gaussianity measure. For a standardized continuous random variable $x$ with the probability density function $f$, negentropy is defined as

$$\mathrm{NH}(\mathbf{x}) = \mathrm{H}(z) - \mathrm{H}(x) \geq 0,$$

where $z \sim \mathcal{N}(0,1)$ and $\mathrm{H}(x) = -\mathbb{E}(\log f(x))$ is a differential entropy. Negentropy is always non-negative since the standard normal distribution is the one with the largest entropy among all distributions of unit variance (Cover and Thomas, 2006). However, it is difficult to directly apply negentropy, since the knowledge of the density $f$ is usually lacking. This makes it necessary to approximate it. One possible approximation is a cumulant based approximation proposed in Jones and Sibson (1987), where NH of the random variable $x$ is

approximated by

$$\mathrm{NH}(x) \approx \frac{1}{12}(\mathbb{E}(x^3))^2 + \frac{1}{48}(\mathbb{E}(x^4) - 3)^2.$$

The first part of the approximation is skewness-, while the latter is kurtosis-based. We will consider a similar projection pursuit index when blindly estimating linear discriminant direction in data from the Gaussian mixture model. However, this approximation is lacking robustness. An alternative approach proposed in Hyvärinen (1997) is

$$\mathrm{NH}(x) \approx (\mathbb{E}(G(x)) - \mathbb{E}(G(z)))^2,$$

where $z \sim N(0,1)$ and $G$ is usually taken to be $G(x) = \log(\cosh(\alpha x))/\alpha$ or $G(x) = -\exp(-x^2/2)$, for tuning parameter $1 \leq \alpha \leq 2$. In case the components are sequentially extracted, the method is known as deflation-based fastICA. For more details fastICA see for example Hyvärinen and Oja (1997); Miettinen et al. (2017) and references therein.

### ICA based on two scatter matrices

As mentioned, there are many methods for extraction of independent components within the ICA model, and many of these are based on projection pursuit ideas. The approach of interest in this section is based however on the simultaneous use of two scatter functionals $\mathbf{S}_1$ and $\mathbf{S}_2$.

One of earliest contributions to ICA includes the fourth order blind identification method (FOBI), originally suggested as an ICA method in Cardoso (1989) and considered in an exploratory data analysis context in Tyler et al. (2009). It starts by choosing two scatters $\mathbf{S}_1 = \mathrm{Cov}$ and $\mathbf{S}_2 = \mathrm{Cov}_4$ and defining the fourth-order-blind-identification (FOBI) functional $\mathbf{W}$, which jointly diagonalizes $\mathbf{S}_1$ and $\mathbf{S}_2$. More precisely, let $\mathbf{x}$ be a $p$-variate random vector with finite fourth moments and set $\mathbf{S}_1 = \mathrm{Cov}$ and $\mathbf{S}_2 = \mathrm{Cov}_4$. Then the FOBI functional is defined as the $p \times p$ matrix-valued functional $\mathbf{W}$ for which

$$\mathbf{W}(\mathbf{x})\mathbf{S}_1(\mathbf{x})\mathbf{W}(\mathbf{x})^\top = \mathbf{I}_p \quad \text{and} \quad \mathbf{W}(\mathbf{x})\mathbf{S}_2(\mathbf{x})\mathbf{W}(\mathbf{x})^\top = \mathbf{D}(\mathbf{x}), \tag{2.5}$$

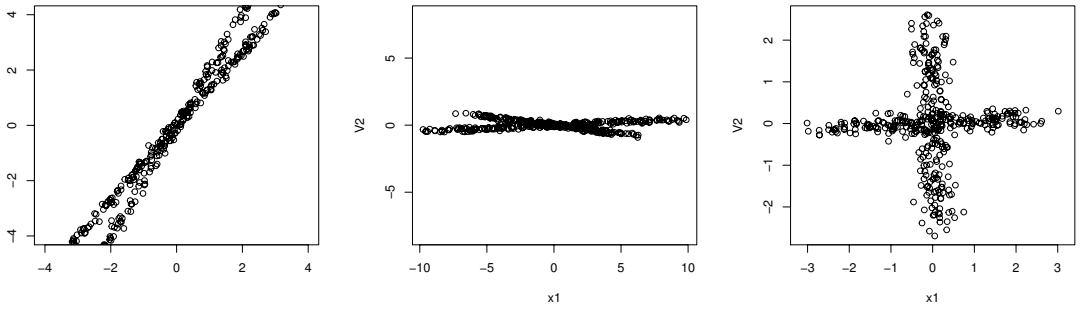
where $\mathbf{D}(\mathbf{x})$ is a diagonal matrix with decreasing diagonal elements. For convenience and when the context is clear, the dependence on $\mathbf{x}$ of $\mathbf{S}_1$, $\mathbf{S}_2$, $\mathbf{W}$ and $\mathbf{D}$ will be omitted. In Miettinen et al. (2015) it is shown that in the ICA model the diagonal elements $d_1, \ldots, d_p$ of $\mathbf{D}$, correspond to kurtosis measures of latent variables $\mathbf{z}$, yielding $d_i = 1$ if and only if $\mathbb{E}(z_i^4) = 3$. Thus, in ICA, the FOBI functional is well-defined (up to signs) if all independent components have distinct kurtoses and in that case, $\mathbf{z}$ corresponds to the original independent components up to signs and order.

The FOBI functional $\mathbf{W}$ is usually obtained by first whitening $\mathbf{x} \mapsto \mathbf{x}^{st} = \mathbf{S}_1(\mathbf{x})^{-1/2}(\mathbf{x} - \mathbb{E}(\mathbf{x}))$ and then performing an eigendecomposition of $\mathbf{S}_2(\mathbf{x}^{st}) = \mathbf{U}\mathbf{D}\mathbf{U}'$. It can then be shown that $\mathbf{W} = \mathbf{U}\mathbf{S}_1^{-1/2}$, and that $\mathbf{D}$ in the eigendecomposition of $\mathbf{S}_2(\mathbf{x}^{st})$ is equal to $\mathbf{D}$ from Definition 2.5 of the FOBI functional. The latent components $z_1, \ldots, z_p$ are then obtained as

$$\mathbf{z} = \mathbf{W}(\mathbf{x} - \boldsymbol{\mu}).$$

The intuition behind this transformation is that $\mathbf{W} = \mathbf{U}\mathbf{S}_1^{-1/2}$ gives latent components $\mathbf{z} = \mathbf{W}(\mathbf{x} - \boldsymbol{\mu})$ obtained by first whitening $\mathbf{x}$ with respect to $\mathbf{S}_1$ and then choosing $\mathbf{z}$ to be

the principal components, with respect to $\mathbf{S}_2$ of the whitened $\mathbf{x}$. In that sense, one can think of FOBI based ICA as an extension of PCA. Figure 2.4 illustrates the benefit of an additional rotation when moving from (scaled) PCA, i.e. whitening, to ICA. Virta (2018) argues




(a) Original data (b) Principal components (c) Independent components

Figure 2.4.: Bivariate scatters of a sample of size 400 from ICA model; left to right, observed data, corresponding principal components, and independent components obtained using FOBI algorithm, respectively.

how in general, ICA is considered superior to PCA (PCA finds (only) uncorrelated, while ICA finds independent components). However, one can also see them as parallel methods, since both solve a generalization of the multivariate normal model. Especially, under the multivariate normal model PCA actually recovers independent components, further making PCA and ICA, the signature methods of the elliptical and independent component model, equivalent in the intersection of the two models.

It is important to mention how the joint usage of two general scatter functionals (not just Cov and $\mathrm{Cov}_4$) is of interest in ICA but is however not limited to it. Tyler et al. (2009) give then a general method for exploring multivariate data using two scatter matrices, based on the eigenvalue–eigenvector decomposition of one scatter matrix w.r.t. to another. The method is called invariant co-ordinate selection (ICS) and FOBI can be considered to be a special case of it.

Let now $\mathbf{S}_1$ and $\mathbf{S}_2$ be two scatter functionals. Then, in general, $p \times p$ variate functional $\mathbf{W}$, which satisfies

$$\mathbf{W}(\mathbf{x})\mathbf{S}_1(\mathbf{x})\mathbf{W}(\mathbf{x})^\top = \mathbf{I}_p \quad \text{and} \quad \mathbf{W}(\mathbf{x})\mathbf{S}_2(\mathbf{x})\mathbf{W}(\mathbf{x})^\top = \mathbf{D}(\mathbf{x}),$$

is of interest especially outside of an elliptical model, where $\mathbf{S}_1$ and $\mathbf{S}_2$ are sometimes required to satisfy certain properties. The reason why the combination $\mathbf{S}_1 - \mathbf{S}_2$ is considered especially outside an elliptical model is that if $\mathbf{x}$ has an elliptical distribution all scatters calculated at $\mathbf{x}$, provided that they exist, are proportional to each other. In Oja et al. (2006) it is shown that any two scatter functionals that have the full independence property can be used in such a way as an ICA method, provided that the diagonal elements in $\mathbf{D}$ are distinct. The independence property can be neglected if $p - 1$ latent components have symmetric distributions (Tyler et al., 2009). Namely, as discussed, the eigenvalues of $\mathbf{S}_1^{-1}(\mathbf{x})\mathbf{S}_2(\mathbf{x})$ can be seen as a kurtosis measures in the direction of the corresponding

eigenvectors. Thus, diagonal elements $d_1, \ldots, d_p$ in $\mathbf{D}$ correspond to $\mathbf{S}_1 - \mathbf{S}_2$-kurtoses of the marginal latent components in $\mathbf{z}$. Observe that the matrix $\mathbf{D}$ depends on the scaling of scatter functionals $\mathbf{S}_1$ and $\mathbf{S}_2$. Often, $\mathbf{S}_1$ and $\mathbf{S}_2$ are standardized under the multivariate normal model implying that $\mathbf{D}(\mathbf{x}) = \mathbf{I}_p$, for $\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. That is for example the case for $\mathrm{Cov} - \mathrm{Cov}_4$ combination of scatters.

Therefore, in order to identify individual independent components using the $\mathbf{S}_1 - \mathbf{S}_2$ combination of scatters, it is important for $\mathbf{S}_1 - \mathbf{S}_2$ kurtoses of the components to differ. For example, in a model where one component has Gaussian distribution and the other homoscedastic Gaussian mixture distribution with two classes and mixing proportion of $1/2 - 1/\sqrt{12}$, both components have the classical $(\mathrm{Cov} - \mathrm{Cov}_4)$ kurtosis value of 1, making them indistinguishable w.r.t. classical kurtosis. However, these could still be separated using the different combination of scatters.

For the exploratory use, there are also some guidelines provided by Tyler et al. (2009) on how to choose the two scatters while arguing that there is no general best combination. The joint use of two scatters is of interest in the mixture models as well. Namely, Tyler et al. (2009) show that in an elliptical mixture model, ICS can be seen as an unsupervised Fisher's linear discriminant method.

### 2.3.3. Dimension estimation in NGCA

Estimation of the signal subspace dimension in the NGCA model is crucial to the successful separation of the signal subspace from the noise. While there are meanwhile many suggestions, like in Blanchard et al. (2006); Kawanabe et al. (2007); Theis et al. (2011); Bean (2014); Sasaki et al. (2016); Virta et al. (2016) to name a few, on how to perform NGCA there is not much research yet on how to estimate the dimensions of the two subspaces.

In the NGCA model, the FOBI functional has the advantage that the eigenvalues in $\mathbf{D}$ of Gaussian components are known to be one. Nordhausen et al. (2021) and Nordhausen et al. (2017) provide then asymptotic results and bootstrapping strategies to obtain p-values when testing the hypothesis

$$H_{0k} : \text{There are exactly } k \text{ non-Gaussian components,} \tag{2.6}$$

by testing that there are $p - k$ eigenvalues in $\mathbf{D}$ equal to 1, and using the mean squared deviation of the $p-k$ eigenvalues in $\mathbf{D}$ closest to 1, from the theoretical value of 1, as a test statistic. Successive application of the tests can then be used to estimate the dimension.

On the other side, the idea of the bootrstrap strategy is to first sample with replacement an $n$-dimensional sample $\tilde{\mathbf{X}} = (\tilde{\mathbf{x}}_1, \ldots, \tilde{\mathbf{x}}_n) \in \mathbb{R}^{p \times n}$ from a random sample $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n) \in \mathbb{R}^{p \times n}$ and estimate its signal component as $\tilde{\mathbf{S}} = (\tilde{\mathbf{s}}_1, \ldots, \tilde{\mathbf{s}}_n) = \hat{\mathbf{W}}_1 \tilde{\mathbf{X}} \in \mathbb{R}^{k \times n}$. Then, in order to make the noise space Gaussian, transform $\tilde{\mathbf{X}} \leftarrow \hat{\mathbf{W}}^{-1}(\tilde{\mathbf{S}}' \mathbf{N}')'$, where $\mathbf{N} \in \mathbb{R}^{(p-k) \times n}$ is an $n$-dimensional random sample from $\mathcal{N}(\mathbf{0}, \mathbf{I}_{p-k})$. The value of the test statistics is then calculated in each of the bootstrap samples and the estimated p-values are obtained by comparing the value of the test statistics calculated in sample $\mathbf{X}$ to those of bootstrap samples.

In Paper I, we show that any two scatter matrix combination can be used in NGCA to separate the Gaussian from the non-Gaussian subspace, provided that $\mathbf{S}_1$–$\mathbf{S}_2$- kurtoses of the signals differ from the corresponding Gaussian value. However, it is then usually

not known what is the noise eigenvalue. Therefore, when testing the hypothesis (2.6) we suggest to identify the set of the noise components as the one corresponding to the set of eigenvalues having a minimal variance, under the assumption that the Gaussian subspace is larger than any set of the signal components which would share the same eigenvalue. Asymptotic results for the test statistic depend on the specific choice of $\mathbf{S}_1$ and $\mathbf{S}_2$, and are in most cases not tractable. However, as discussed, the bootstrap strategy described above for FOBI can be adapted for a general combination of scatters, where the used test statistics is the variance of the $p-k$- subset of eigenvalues that has the smallest variance. For more details see Paper I.

Even though the NGCA model is successfully used to model clustered data, a natural framework when working with data from multiple underlying populations is finite mixture models. Among those, we place a special focus on finite mixtures of multivariate Gaussian distributions.

## 2.4. Gaussian mixture model

A Gaussian mixture model (GMM) is often referred to as parametric probability density function, which is represented as a convex combination of Gaussian densities, with not necessarily equal location and scale parameters (McLachlan and Peel, 2000). For a $p$-dimensional random vector $\mathbf{x}$ from Gaussian mixture model with $k$-classes, $k \in \mathbb{N}$, the probability density function is given by

$$f(\mathbf{x}; \alpha_i, \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) = \sum_{i=1}^{k} \alpha_i \phi(\mathbf{x}; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i), \tag{2.7}$$

where $\alpha_1, \ldots, \alpha_k \geq 0$, $\sum_{i=1}^{k} \alpha_i = 1$ are mixing proportions, $\boldsymbol{\mu}_i$ and $\boldsymbol{\Sigma}_i$ are $p$-variate mean vector and $p \times p$ covariance matrix of $i$-th population (class), respectively, for $i = 1, \ldots, k$. $\phi(\cdot\,; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ is the probability density function of $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ distribution. The mean vector and the covariance matrix of $\mathbf{x}$ are

$$\boldsymbol{\mu} = \sum_{i=1}^{k} \alpha_i \boldsymbol{\mu}_i, \quad \mathrm{Cov}(\mathbf{x}) = \sum_{i=1}^{k} \alpha_i \boldsymbol{\Sigma}_i + \sum_{i=1}^{k} \alpha_i (\boldsymbol{\mu}_i - \boldsymbol{\mu})(\boldsymbol{\mu}_i - \boldsymbol{\mu})',$$

respectively, where

$$\boldsymbol{\Sigma}_W = \sum_{i=1}^{k} \alpha_i \boldsymbol{\Sigma}_i \quad \text{and} \quad \boldsymbol{\Sigma}_B = \sum_{i=1}^{k} \alpha_i (\boldsymbol{\mu}_i - \boldsymbol{\mu})(\boldsymbol{\mu}_i - \boldsymbol{\mu})',$$

are so-called the *within-class* and the *between-class* scatters, respectively. In case where all the class covariances are equal, the corresponding GMM is called *homoscedastic* GMM. Nordhausen et al. (2017) argues how homoscedastic Gaussian mixture models with $p+1$ classes are included in the NGCA model. Moreover, in Paper II it is shown that homoscedastic GMM with 2 classes is in fact an NGICA model. How wide the class of GMM is, is for example discussed in McLachlan and Peel (2000), where it is stated how any continuous distribution can be approximated arbitrarily well by a finite mixture of normal densities

with common covariance. To our particular interest in following section is the homoscedastic GMM with two classes.

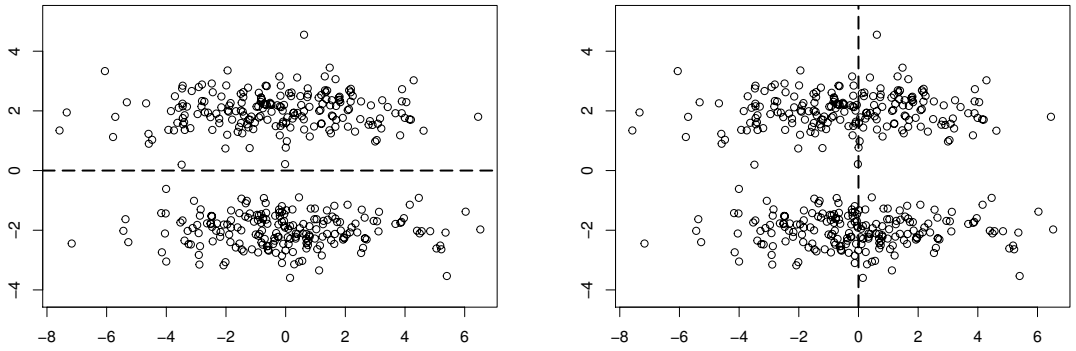
As discussed, GMM is particularly convenient for modeling heterogeneous data, originating from multiple underlying populations, wherein the absence of class labels one wishes to correctly classify the observed data. In general, parameter estimation and clustering in the scope of GMM are usually considered jointly. When it comes to parameter estimation, maximum likelihood approach is considered a gold standard and one of the most popular algorithms for that purpose is known as Expectation-Maximization (EM) algorithm. For more details on the EM algorithm see for example Friedman et al. (2001), while for more details on the implementation of the EM algorithm see documentation on R-package *mclust* (Scrucca et al., 2016). Other well-known clustering methods for Gaussian mixture modeling are e.g. $k$-means clustering or spectral clustering, see Friedman et al. (2001); Von Luxburg (2007). However, when dealing with high-dimensional data, the likelihood function is usually highly multi-modal, thus making high-dimensional maximization a rather difficult and often unreachable task. Moreover, the space over which we are optimizing grows exponentially with the dimension of the data, making a blind search for global optimum a wild goose chase. Thus, one is advised to perform exploratory data analysis to initialize model parameters more accurately, which is again rather difficult if the dimension $p$ of $\mathbf{x}$ is large. A partial solution to the problem lies in first doing dimension reduction, i.e. projecting data to an appropriate lower-dimensional space, which hopefully then contains all the discriminatory features of the original data set. One of the most common unsupervised methods for revealing clusters is still arguably PCA. However, it is also well known that PCA does not, in general, yield a consistent estimator of the optimal linear discriminant direction. A standard example demonstrating this is the extreme case where the within-class covariance matrix $\mathbf{\Sigma}$ is heavily concentrated on a direction orthogonal to the difference of the group means $\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1$. In such a case, the projections of the two group means onto the first principal component direction overlap, making clustering based on the direction impossible. Figure 2.5 illustrates how PCA fails in extracting feature which contains discriminatory information on the data. In case the class membership of the data is known, linear discriminant analysis (LDA) (Fisher, 1936) can be used to extract such directions.

### Linear discriminant analysis

Linear discriminant analysis was originally derived to discriminate between the two classes, under the assumption of homoscedasticity and normality, i.e. in the scope of homoscedastic Gaussian mixture model with two classes (Fisher, 1936). For the direction $\mathbf{w} \in \mathbb{R}^p$ and the $p$-variate random vector $\mathbf{x}$ from homoscedastic two-class GMM with common variance $\mathbf{\Sigma}$ and class means $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$, the within-class and between-class variances of the projection $\mathbf{w}'\mathbf{x}$ are

$$\sigma_W^2 = \mathbf{w}'\mathbf{\Sigma}\mathbf{w}, \quad \sigma_B^2 = (\mathbf{w}'\boldsymbol{\mu}_1 - \mathbf{w}'\boldsymbol{\mu}_2)^2,$$

respectively. The optimal linear discriminant direction is obtained following Fisher's linear discriminant rule, which searches for projection $\mathbf{w}'\mathbf{x}$ of the data $\mathbf{x}$, for which the ratio of the between-class variance to the within-class variance is maximal. The solution $\mathbf{w}$ to the given maximization problem is the leading eigenvector of $\Sigma^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)'$, thus implying that $\mathbf{w} \propto \Sigma^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$. Fisher's linear discriminant and LDA are often used as synonyms. In the homoscedastic, two-class GMM, LDA is optimal in the sense that the




(a) Optimal PCA direction

(b) Optimal LDA direction

Figure 2.5.: The dashed line gives the optimal PCA direction (left) and LDA direction (right) for random samples of size 400 from balanced, two-component Gaussian mixture model.

optimal Bayes classifier (the one having the minimal misclassification rate out of all classifiers), depends on the data only through the projection $(\Sigma^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2))'\mathbf{x}$, see e.g. Mardia et al. (1995). In Paper II it is shown that GMM with 2 classes with equal covariances is in fact an NGICA model. Moreover, inspecting the proof of the statement yields that the $p-1$ of latent independent components are Gaussian, while the remaining non-Gaussian component has homoscedastic Gaussian distribution and is up to a scale equal to optimal discriminant feature, extracted by the LDA.

In the case of more than two classes, one can extend the analysis used in the derivation of Fisher's discriminant to find a subspace that contains most of the class variability. Again, using Fisher's discriminant rule, the class separation direction $\mathbf{w}$ maximizes the ratio of the variance between- to the variance within the classes

$$\frac{\mathbf{w}'\boldsymbol{\Sigma}_W\mathbf{w}}{\mathbf{w}'\boldsymbol{\Sigma}_B\mathbf{w}},$$

where in the homoscedastic setting with the common variance $\boldsymbol{\Sigma}$, within-class variance $\boldsymbol{\Sigma}_W = \boldsymbol{\Sigma}$. In the case where $\boldsymbol{\Sigma}_W^{-1}\boldsymbol{\Sigma}_B$ is diagonalizable, the class variability is contained in the subspace spanned by the eigenvectors of $\boldsymbol{\Sigma}_W^{-1}\boldsymbol{\Sigma}_B$, which later serve primarily in the dimension reduction, as it is in the PCA. Another approach in multiclass LDA is to partition the classes, and then to use standard, two-class LDA to classify each partition separately. A common way of partitioning is the extraction of one group at a time, meaning to combine all but one group. The procedure is then applied $k-1$ times, resulting in $k-1$ linear discriminant directions. An alternative is a pairwise classification, where one considers all of $k(k-1)/2$ pairs of groups, thus obtaining $k(k-1)/2$ discriminant directions. As mentioned in Section 2.3, Tyler et al. (2009) show that in an elliptical mixture model, ICS can be seen as an LDA method without knowing the class labels. For more details on LDA see e.g. Hastie et al. (2004); Rao (1948) and references therein.

However, in practice, it is rarely the case that the data directly follow a GMM. It is more often the case that data contain a certain amount of redundant information, i.e. noise, which then interferes with clustering. Especially, in the setting where the magnitude of the noise is comparable to the one of the signal which reveals the cluster structure, discussed classification methods usually do not apply directly. To illustrate the problem, consider a random sample from the NGCA model

$$\mathbf{x} = \mathbf{A}(\mathbf{s}', \mathbf{n}')', \tag{2.8}$$

where signal $\mathbf{s}$ follows GMM with density $f(\mathbf{s}) = 0.5\phi(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}) + 0.5\phi(\boldsymbol{\mu}_2, \boldsymbol{\Sigma})$, while the noise $\mathbf{n}$ has two independent, standard normal components, where $\boldsymbol{\mu}_1 = (-4, -4)$, $\boldsymbol{\mu}_2 = (2, 1)$ and the covariance matrix $\boldsymbol{\Sigma}$ and the mixing matrix $\mathbf{A}$ are

$$\boldsymbol{\Sigma} = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}, \qquad \mathbf{A} = \begin{pmatrix} 1 & 0 & 3 & 0 \\ 0 & -1 & 0 & 3 \\ -0.2 & 0 & 1 & 1 \\ 0 & -0.3 & 1 & 1 \end{pmatrix},$$

respectively. Figure 2.6 and 2.7 illustrate how scatter plots of the observed sample and the corresponding principal components reveal no clear cluster structure. However, if one extracts only a signal subspace, then the underlying cluster structure becomes obvious, as it is shown in Figure 2.8.

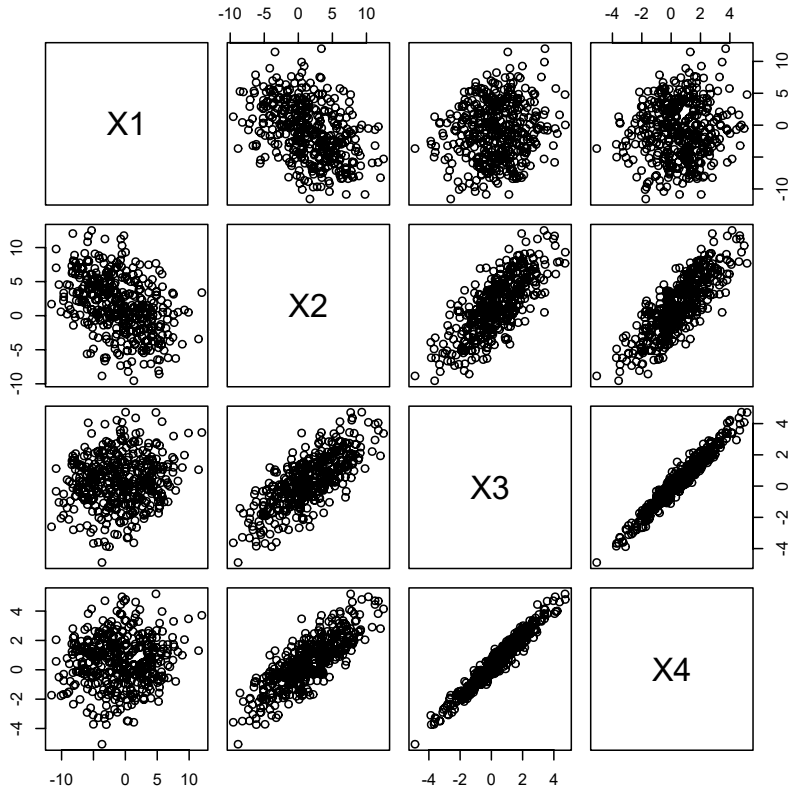



Figure 2.6.: Scatter plot of the sample of size $n = 400$ from Model (2.8).

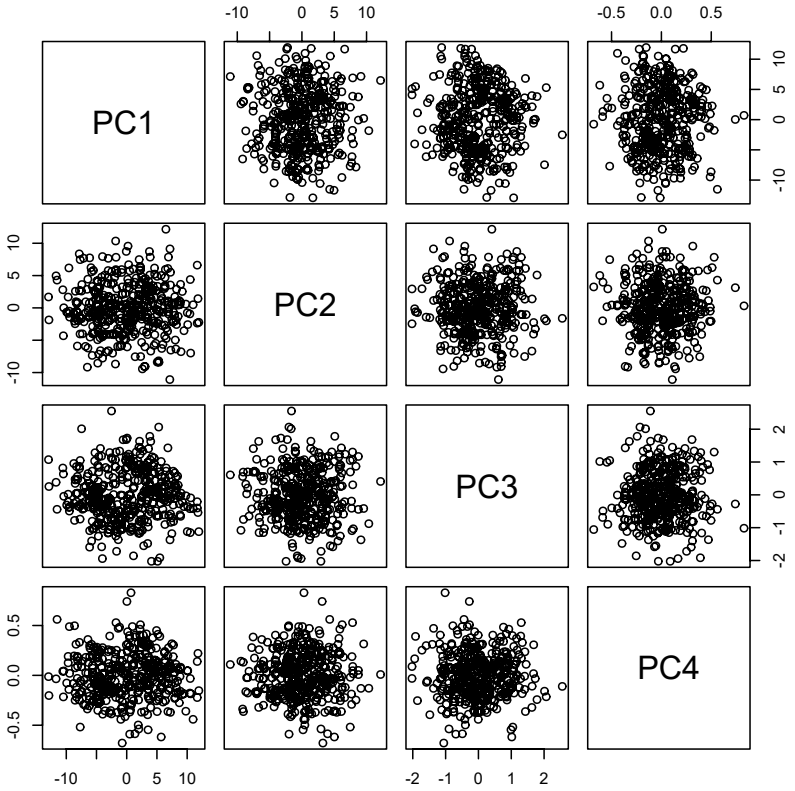
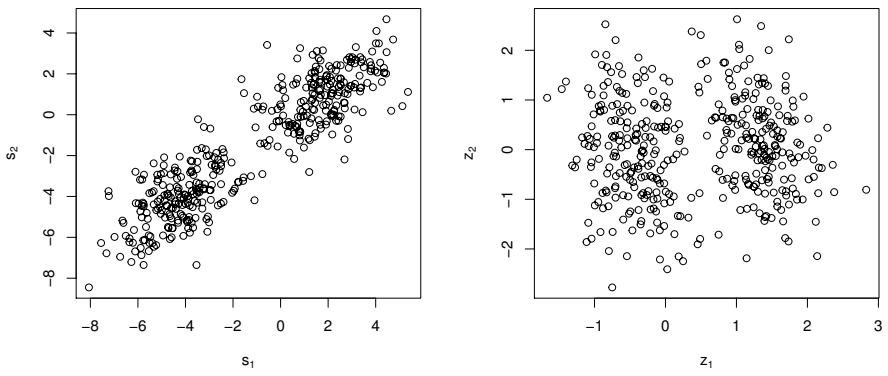



Figure 2.7.: Scatter plot of the principal components of the sample of size $n = 400$ from Model (2.8).




(a) Original signal (b) Estimated signal

Figure 2.8.: Signal part of the sample of size $n = 400$ from Model (2.8) (left), and corresponding estimated signal using FOBI method.

### 2.4.1. A supervised estimation of the linear discriminant

When it comes to the estimation of the linear discriminant direction, in case of the supervised estimation, the approach is straightforward. We give the estimator in the case of homoscedastic GMM with two classes, as it will be relevant for further discussion. Let $\mathbf{x}$ be a random vector from homoscedastic GMM with two classes, and let the Bernoulli random variable $y \sim Bernoulli(\alpha_1)$ describe the class membership, where $\alpha_1 > 0$ is the mixing proportion. If we have a sample $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)$ from the distribution of the full pair $(\mathbf{x}, y)$ available, the standard estimator of $\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)$ is the plug-in estimator (which is also its MLE, up to the scaling of the pooled covariance matrix). That is, using the notation,

$$\bar{\mathbf{x}}_{n1} := \frac{1}{\sum_{i=1}^n y_i} \sum_{i=1}^n y_i \mathbf{x}_i, \quad \bar{\mathbf{x}}_{n2} := \frac{1}{\sum_{i=1}^n (1-y_i)} \sum_{i=1}^n (1-y_i)\mathbf{x}_i,$$

$$\mathbf{S}_n := \frac{1}{n-2} \left\{ \sum_{i=1}^n y_i (\mathbf{x}_i - \bar{\mathbf{x}}_1)(\mathbf{x}_i - \bar{\mathbf{x}}_1)' + \sum_{i=1}^n (1-y_i)(\mathbf{x}_i - \bar{\mathbf{x}}_2)(\mathbf{x}_i - \bar{\mathbf{x}}_2)' \right\},$$

the plug-in estimator of $\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)$ is defined as,

$$\mathbf{w}_n := \mathbf{S}_n^{-1}(\bar{\mathbf{x}}_{n2} - \bar{\mathbf{x}}_{n1}).$$

Asymptotic results for LDA are very standard in the literature, see e.g. Anderson (2003) and are usually given in the case of fixed group sizes, whereas in our model the group sizes are determined by the indicator variables $y_1, \ldots, y_n$ and are, as such, random. Paper II gives the limiting distribution of plug-in estimator for the presented model and shows that the magnitude of the variance of the estimator is larger the more imbalanced the groups are, as well as the larger amount of variation in the direction of the optimal discriminant direction is.

However, it is often the case that the knowledge of group membership is lacking and one wishes to cluster data into meaningful clusters. Furthermore, in the situations where the class label is given, supervised methods usually first use a so-called *training* sample with known labels to estimate the parameters of the method, and then treat new, unlabeled data, based on the estimated parameters. This way supervised methods, in general, gain no additional information from the new data.

### 2.4.2. Projection pursuit based estimation of the linear discriminant

Projection pursuit is a general family of methods searching for a projection direction that maximizes the value of the so-called *projection index*, and can be seen as an alternative to cluster analysis (Huber, 1985; Bolton and Krzanowski, 2003; Bickel et al., 2018; Fischer et al., 2019). Huber (1985), for example, suggested that interesting projections are those that produce minimum entropy (non-normal) distributions, thus implying that any test statistic used for testing normality could potentially be used as a projection index. Especially, Huber (1985) suggested the use of standardized absolute cumulants as projection indices for cluster detection. The approach was later followed by Jones and Sibson (1987) who proposed the use of a linear combination of squared third and fourth cumulants of

projection of the standardized data, as a projection index for cluster identification. As discussed in Section 2.3, the proposed projection index also serves as the approximation of the negentropy in fastICA. Asymptotic results for general projection indices have been derived earlier in the context of ICA, see, e.g., Ollila (2009); Dermoune and Wei (2013); Miettinen et al. (2015); Virta et al. (2016).

Peña and Prieto (2001) studied the use of kurtosis $\kappa : \mathbb{S}^{p-1} \to \mathbb{R}$

$$\kappa(\mathbf{u}) = \frac{\mathrm{E}\{(\mathbf{u}'(\mathbf{X} - \mathrm{E}(\mathbf{X})))^4\}}{[\mathrm{E}\{(\mathbf{u}'(\mathbf{X} - \mathrm{E}(\mathbf{X})))^2\}]^2},$$

as projection index in this setting, where $\mathbb{S}^{p-1} = \{\mathbf{x} \in \mathbb{R}^p : \|\mathbf{x}\| = 1\}$ is the centered unit sphere in $\mathbf{R}^p$. Namely, denoting the common covariance matrix by $\boldsymbol{\Sigma}$ and the two group means by $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$, Peña and Prieto (2001) showed that using kurtosis as the projection index in the projection pursuit allows the unsupervised estimation of the projection direction $\boldsymbol{\theta} := \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)$ used in the linear discriminant analysis to construct the optimal Bayes classifier, in the absence of information on group membership, except in some special cases. Namely, for mixing proportion $\alpha_1 \in \{\delta_1, \delta_2\}$, where $\delta_1 = 1/2 - 1/\sqrt{12}$ and $\delta_2 = 1/2 + 1/\sqrt{12}$, corresponding GMM is indistinguishable from normal model, w.r.t. kurtosis. The result of Peña and Prieto (2001) raises a natural question regarding the efficiency of the procedure. The main question we address in Paper II is how much do we lose by not knowing the group membership and relying solely on the projection pursuit to recover direction $\boldsymbol{\theta}$, compared to using the supervised LDA estimator to recover the same direction, working, for simplicity, under the assumption of homoscedastic GMM with two classes.

Use of the kurtosis as a projection index as described in Peña and Prieto (2001) in practice requires information about the mixing proportion $\alpha_1$. Namely, if $\alpha_1 \in (\delta_1, \delta_2)$ then the linear discriminant $\boldsymbol{\theta}/\|\boldsymbol{\theta}\|$ is found as the minimizer of $\kappa$, whereas if $\alpha_1 \in (0, \delta_1) \cup (\delta_2, 1)$ then $\boldsymbol{\theta}/\|\boldsymbol{\theta}\|$ is found as the maximizer of $\kappa$. Thus, to obtain a truly blind estimator, in Paper II we propose using the squared *excess kurtosis* $(\kappa(\mathbf{u}) - 3)^2$ as an objective function and show that it yields a Fisher consistent estimate of the linear discriminant, apart from the degenerate cases $\alpha_1 \in \{\delta_1, \delta_2\}$, where excess kurtosis vanishes. Interestingly, the limiting covariance is shown to be proportional to the one of the supervised linear discriminant estimator. For more details on the limiting distribution and the proportionality constant see Paper II.

While kurtosis is arguably the most popular choice for the projection index in PP, skewness $\gamma : \mathbb{S}^{p-1} \to \mathbb{R}$

$$\gamma(\mathbf{u}) = \frac{\mathrm{E}\{(\mathbf{u}'(\mathbf{X} - \mathrm{E}(\mathbf{X})))^3\}}{[\mathrm{E}\{(\mathbf{u}'(\mathbf{X} - \mathrm{E}(\mathbf{X})))^2\}]^{3/2}},$$

is a common alternative. Loperfido (2013) shows that skewness has the same ability to find the optimal projection direction in the absence of the group membership information, except in the symmetric setting ($\alpha_1 = 0.5$), where due to symmetry, skewness of all projections is zero. Therefore, in Paper II we discuss also skewness-based projection pursuit in this context, showing that the limiting covariance of the unsupervised skewness-based estimator of the linear discriminant is proportional to those of kurtosis-based- and supervised estimator. As discussed, a drawback of both kurtosis and skewness is that the two indices are unable to recover the optimal projection direction for some particular values of the mixing proportion, that however can be mitigated by combining both cumulants into a

single projection index, in a form of a convex combination $\eta : \mathbb{S}^{p-1} \to \mathbb{R}$

$$\eta(\mathbf{u}) = \eta(\mathbf{u}; w_1) := w_1 \gamma(\mathbf{u})^2 + w_2 \{\kappa(\mathbf{u}) - 3\}^2,$$

where $w_1, w_2 \geq 0$, $w_1 + w_2 = 1$. Virta et al. (2016) e.g. studied some asymptotic properties of the hybrid index $\eta$ in the context of independent component analysis, while in Paper II we derive the limiting distribution of the hybrid-estimator of the linear discriminant giving also a full covariance matrix, which again turns out to be proportional to those of kurtosis and skewness based estimators. Studying the asymptotic relative efficiency of the hybrid estimator w.r.t. the supervised one, we discuss optimal weighting if information on mixing proportion is available. As a further interesting observation, when the mixture model approaches the multivariate normal model, the limit of the optimal weight seems to approach the value 0.8, which is the exact weighting used in the Jarque-Bera test statistic for testing normality, $(n/6)\gamma_n^2 + (n/24)(\kappa_n - 3)^2$ (Jarque and Bera, 1980). It is concluded that in the case of moderately balanced and infinitely well-separated groups, projection pursuit is able to reach asymptotic efficiency equal to LDA with an optimal choice of weighting.

It is important to mention that from an inferential point of view, skewness and kurtosis as test statistics were first introduced by Malkovich and Afifi (1973). Machado (1983) discusses asymptotic distribution under the null hypothesis of normality, of statistics proposed by Malkovich and Afifi (1973). Friedman (1987) and Posse (1995) propose to assess the significance of results by comparing the observed value of the projection index with the sampling distribution of the same index, obtained by simulating many random samples from a Gaussian distribution of the same dimension and cardinality as the data. Kuriki and Takemura (2008) use a geometric approach to derive exact formulae for the tail probabilities of Malkovich and Afifi (1973) statistics. Loperfido (2018) conjectures that the asymptotic distribution of the maximal skewness attainable by a linear combination of Gaussian random variables is skew-normal. In that manner, work presented in Paper II can be considered in the direction of an inferential PP, where the role of inferential procedures is in deciding whether the directed structure is real or just the artifact of the noise.

# 3. Feature extraction for matrix-variate data

We next move to an extension of concepts presented in Chapter 1 to the setting where the observations are assumed to be matrix-valued. Examples of such data are abundant nowadays, the most notable example being image data where the elements of the matrix represent the gray-scale intensities of the individual pixels of an image. For example, images of hand-written signs (digits, letters...) are represented this way, where one often aims to classify observed signs to e.g. transfer hand-written documents into the digital format. Figure 3.1 shows a sample of handwritten digits 1 and 2, from data set *digits*, available freely in the R package tensorBSS (Virta et al., 2021a) and consisting of $16 \times 16$ grayscale images of normalized handwritten digits $1-9$, automatically scanned from envelopes by the U.S. Postal Service. Furthermore, magnetic resonance imaging (MRI) and coronary tissue

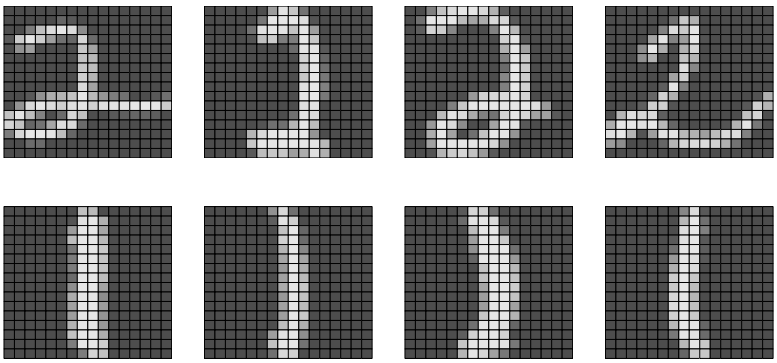


Figure 3.1.: A collection of images of digits 1 and 2 (up) from the *digits* data set.

(CT) data are naturally represented in that way. Other examples include e.g. biological abundance data where each matrix contains the abundances of a single species in $p$ regions (the rows) over $q$ time points (the columns).

Methods for analysis of matrix-valued data can be roughly divided into two groups. The first one are vectorization-based methods which constitute the simplest approach to matrix modeling. Namely, one defines vectorization operator $\text{vec} : \mathbb{R}^{p\times q} \to \mathbb{R}^{pq}$, where for observed $p \times q$ matrices $\mathbf{X}_i$, $i = 1, \ldots, n$, $\text{vec}(\mathbf{X}_i) = \mathbf{x}_i \in \mathbb{R}^{pq}$ obtained by stacking columns of $\mathbf{X}_i$ on top of each other into a large vector. Such vector $\mathbf{x}_i$ is then subjected to the desired multivariate method. Although probably a natural way to tackle the problem, the simplicity of analysis introduced by vectorization, comes with a price. The vectorization loses the spatial structure one had in original matrices, while at the same time produces vector $\mathbf{x}_i$ of potentially high dimension $pq$. The alternative way of dealing with the matrix-variate data is to keep the matrix structure and develop methods to accommodate it. We will focus on

the latter approach. In general, analyzing raw, high-dimensional matrix-variate data can be computationally expensive and often intractable. Consequentially, feature extraction is an important and often occurring problem in image processing, where observed matrix-variate (image) data is transformed into a lower-dimensional space in a way that the obtained representation contains important properties of the original data.

As in the multivariate setting, we introduce the statistical frameworks that will guide the feature extraction of the matrix-variate data.

## 3.1. Location-scatter model

We motivate the following models again through a general location-scatter model. We say that a $(p \times q)$-variate random matrix $\mathbf{X}$ obeys the matrix location-scatter model if

$$\mathbf{X} = \mathbf{A}\mathbf{Z}\mathbf{B}' + \mathbf{T},$$

where location matrix $\mathbf{T} \in \mathbb{R}^{p\times q}$ and non-singular mixing matrices $\mathbf{A} \in \mathbb{R}^{p\times p}$, $\mathbf{B} \in \mathbb{R}^{q\times q}$ are unknown parameters and latent random matrix $\mathbf{Z} \in \mathbb{R}^{p\times q}$ satisfies

$$\mathbb{E}(\mathrm{vec}(\mathbf{Z})) = \mathbf{0}, \quad \mathrm{Cov}(\mathrm{vec}(\mathbf{Z})) = \mathbf{I}_{pq}.$$

When vectorizing the location-scatter matrix-variate model we obtain a structured vector-valued location-scatter model, where the mixing matrix has the Kronecker structure $\mathbf{B}\otimes\mathbf{A}$. For matrices $\mathbf{A} \in \mathbb{R}^{p\times p}$ and $\mathbf{B} \in \mathbb{R}^{q\times q}$, the Kronecker product is $pq \times pq$ block matrix, such that $(i, j)$-block of $\mathbf{A} \otimes \mathbf{B}$ is $\mathbf{A}_{i,j}\mathbf{B}$, for $i, j = 1 \ldots, p$ (Henderson and Searle, 1981). Observe now if one would simply vectorize the location-scatter model with aim of estimation of mixing matrix $\mathbf{B} \otimes \mathbf{A}$, one would estimate $pq(pq + 1)/2$ unknown parameters. However, if the underlying structure of the model is considered, then the number of parameters one estimates is $p(p + 1)/2 + q(q + 1)/2$ and is considerably smaller than in the previous case. The approach in which we keep the structure of the vectorized data is known as the structured covariance estimation (Srivastava et al., 2008), and to some extent shows how not all methods starting with vectorization lose all the spatial structure, what is important is how one proceeds post vectorization.

As is the case for vector-valued models, posing various additional assumptions on the latent matrix $\mathbf{Z}$ leads into various families of models, the simplest of which being the matrix-variate normal model.

### 3.1.1. Matrix-variate normal model

**Definition 2.** *The $p\times q$ random matrix $\boldsymbol{X}$ is said to have matrix-variate normal distribution with $p \times q$ mean matrix $\boldsymbol{T}$ and covariance matrix $\boldsymbol{\Omega} \otimes \boldsymbol{\Sigma}$, if $\mathrm{vec}(\boldsymbol{X})$ follows $pq$-variate normal distribution $\mathcal{N}(\mathrm{vec}(\boldsymbol{T}), \boldsymbol{\Omega} \otimes \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma} \in \mathbb{R}^{p\times p}$ and $\boldsymbol{\Omega} \in \mathbb{R}^{q\times q}$ are positive-definite symmetric matrices. We write $\boldsymbol{X} \sim \mathcal{MN}_{p\times q}(\boldsymbol{T}, \boldsymbol{\Sigma}, \boldsymbol{\Omega})$.*

Definition 2 implies that if $\mathbf{X} \sim \mathcal{MN}_{p\times q}(\mathbf{T}, \boldsymbol{\Sigma}, \boldsymbol{\Omega})$, then $\mathbf{X}' \sim \mathcal{MN}_{p\times q}(\mathbf{T}', \boldsymbol{\Omega}, \boldsymbol{\Sigma})$. Let now $\mathbf{X} = (\mathbf{x}_1', \ldots, \mathbf{x}_q')' \sim \mathcal{MN}_{p,q}(\mathbf{M}, \boldsymbol{\Sigma}, \boldsymbol{\Omega})$, where $\mathbf{x}_i \in \mathbb{R}^p$ is $i$-th column of $\mathbf{X}$, $i = 1, \ldots, q$. The structured covariance matrix of $\mathrm{vec}(\mathbf{X})$ implies that $\mathrm{Cov}(\mathbf{x}_i, \mathbf{x}_j) = \boldsymbol{\Omega}_{i,j}\boldsymbol{\Sigma}$, $i, j = 1, \ldots, q$. Therefore, $\mathbb{E}((\mathbf{X} - \mathbf{T})(\mathbf{X} - \mathbf{T})') = \sum_{i=1}^{q} \mathrm{Cov}(\mathbf{x}_i) = \mathrm{tr}(\boldsymbol{\Omega})\boldsymbol{\Sigma}$. Similarly, $\mathrm{Cov}(\tilde{\mathbf{x}}_i, \tilde{\mathbf{x}}_j) =$

$\mathbf{\Sigma}_{i,j}\mathbf{\Omega}$, $i,j = 1,\dots,p$, where $\tilde{\mathbf{x}}_i \in \mathbb{R}^q$ is $i$-th row of $\mathbf{X}$, $i = 1,\dots,p$, further implying that $\mathbb{E}((\mathbf{X}-\mathbf{T})'(\mathbf{X}-\mathbf{T})) = \mathrm{tr}(\mathbf{\Sigma})\mathbf{\Omega}$. Observe that, if e.g. $\mathbf{\Sigma} = \mathbf{I}_p$, then $\mathrm{Cov}(\tilde{\mathbf{x}}_i, \tilde{\mathbf{x}}_j) = 0$, for $i \neq j$, implying that rows of $\mathbf{X}$ are mutually independent and up to a location, identically distributed. Furthermore, in that case $\mathbb{E}((\mathbf{X}-\mathbf{T})'(\mathbf{X}-\mathbf{T}))/p = \sum_{i=1}^{p} \mathrm{Cov}(\tilde{\mathbf{x}}_i)/p = \mathbf{\Omega}$, thus implying that in case where the rows are independent copies from multivariate Gaussian distribution, $\mathbf{\Omega}$ is the corresponding (column) covariance. Similar is true $\mathbf{\Sigma}$ as well. That is the reason why we often refer to $\mathbf{\Sigma}$ and $\mathbf{\Omega}$ as to row and column covariance matrices, respectively.

It is now clear that not all random matrices with normally distributed entries are from matrix-variate normal model. As it is the case for multivariate normal distribution, the family of matrix-variate normal distributions is also closed under linear transformations. More precisely, if $\mathbf{X} \sim \mathcal{MN}_{p\times q}(\mathbf{T}, \mathbf{\Sigma}, \mathbf{\Omega})$ then for full rank matrices $\mathbf{A} \in \mathbb{R}^{p\times r}$, $r \leq p$ and $\mathbf{B} \in \mathbb{R}^{q\times s}$, $s \leq q$ is

$$\mathbf{A}'\mathbf{X}\mathbf{B} \sim \mathcal{MN}_{r\times s}(\mathbf{A}'\mathbf{T}\mathbf{B}, \mathbf{A}'\mathbf{\Sigma}\mathbf{A}, \mathbf{B}'\mathbf{\Omega}\mathbf{B}).$$

In the context of location-scatter model, if we assume that $\mathbf{Z} \sim \mathcal{MN}_{p\times q}(\mathbf{0}, \mathbf{I}_p, \mathbf{I}_q)$, then $\mathbf{X} = \mathbf{A}\mathbf{Z}\mathbf{B}' + \mathbf{T} \sim \mathcal{MN}_{p\times q}(\mathbf{T}, \mathbf{A}\mathbf{A}', \mathbf{B}\mathbf{B}')$, which gives a mean of sampling from matrix-variate distribution. In the following, we refer to $\mathcal{MN}_{p\times q}(\mathbf{0}, \mathbf{I}_p, \mathbf{I}_q)$ as to matrix-variate standard normal distribution, that is, as its vector counterpart, a spherical distribution, while $\mathbf{Z} \sim \mathcal{MN}_{p\times q}(\mathbf{0}, \mathbf{I}_p, \mathbf{I}_q)$ has independent entries from univariate standard normal distribution. Consequentially, matrix-variate normal distribution belongs to the class of matrix-variate elliptical distributions.

The probability density function of $\mathbf{X} \sim \mathcal{MN}_{p\times q}(\mathbf{T}, \mathbf{\Sigma}, \mathbf{\Omega})$ is

$$\phi^{(p\times q)}(\mathbf{X}; \mathbf{T}, \mathbf{\Sigma}, \mathbf{\Omega}) = \frac{\exp(-\frac{1}{2}\mathrm{tr}\left[\mathbf{\Omega}^{-1}(\mathbf{X}-\mathbf{T})'\mathbf{\Sigma}^{-1}(\mathbf{X}-\mathbf{T})\right])}{(2\pi)^{pq/2}\det(\mathbf{\Omega})^{p/2}\det(\mathbf{\Sigma})^{q/2}},$$

thus showing that the matrix-variate normal model is fully characterized by the mean matrix $\mathbf{T}$ and row and column covariance matrices $\mathbf{\Sigma}$ and $\mathbf{\Omega}$. For more details and properties of matrix-variate normal distribution see Gupta and Nagar (1999), and references therein.

### 3.1.2. Elliptical models

As is the case in the vector setting, matrix-valued elliptical models are derived from the location-scatter model assuming that latent matrix $\mathbf{Z}$ has matrix-variate spherical distribution.

**Definition 3.** *A $p \times q$-variate random matrix $\boldsymbol{X}$ is said to have*

*a) right spherical distribution if $(\boldsymbol{X}-\boldsymbol{T}) \sim (\boldsymbol{X}-\boldsymbol{T})\boldsymbol{V}$, for all orthogonal matrices $\boldsymbol{V} \in \mathcal{O}^{q\times q}$,*

*b) left spherical distribution if $(\boldsymbol{X}-\boldsymbol{T}) \sim \boldsymbol{U}'(\boldsymbol{X}-\boldsymbol{T})$, for all orthogonal matrices $\boldsymbol{U} \in \mathcal{O}^{p\times p}$,*

*c) spherical distribution if $(\boldsymbol{X}-\boldsymbol{T}) \sim \boldsymbol{U}'(\boldsymbol{X}-\boldsymbol{T})\boldsymbol{V}$, for all orthogonal matrices $\boldsymbol{U} \in \mathcal{O}^{p\times p}$ and $\boldsymbol{V} \in \mathcal{O}^{q\times q}$,*

*where* $\boldsymbol{T} \in \mathbb{R}^{p\times q}$ *is a location matrix of* $\boldsymbol{X}$.

An alternative way of defining matrix spherical distributions is using vectorization operator and requiring that vec($\mathbf{Z}$) has vector-valued spherical distribution. Such definition gives a somewhat narrower class of distributions (Arashi, 2017). However, we define matrix spherical distribution according to the first definition, as it is given in Gupta and Nagar (1999).

If the second moments of spherically distributed $\mathbf{X}$ exist, then $\mathrm{Cov}(\mathrm{vec}(\mathbf{X})) \propto \mathbf{I}_p \otimes \mathbf{I}_q = \mathbf{I}_{pq}$. The class of $p \times q$ matrix-valued elliptical distributions is now defined as the set of all distributions of the form

$$\mathbf{X} = \mathbf{A}\mathbf{Z}\mathbf{B}' + \mathbf{T},$$

where $\mathbf{A} \in \mathbb{R}^{p\times p}$, $\mathbf{B} \in \mathbb{R}^{q\times q}$, $\mathbf{T} \in \mathbb{R}^{p\times q}$ and latent matrix $\mathbf{Z} \in \mathbb{R}^{p\times q}$ has matrix spherical distribution around the origin. As discussed, a matrix-variate normal distribution is a member of the elliptical family. Another well-known member of the matrix-elliptical model is matrix-variate $t$-distribution.

**Definition 4.** *The random matrix* $\boldsymbol{X} \in \mathbb{R}^{p\times q}$ *is said to have matrix-variate* $t$*-distribution with* $\nu > 0$ *degrees of freedom, and parameters* $\boldsymbol{T} \in \mathbb{R}^{p\times q}$, $\boldsymbol{\Sigma} \in \mathbb{R}^{p\times p}$ *and* $\boldsymbol{\Omega} \in \mathbb{R}^{q\times q}$, $\boldsymbol{X} \sim \boldsymbol{T}_{p\times q}(\nu, \boldsymbol{T}, \boldsymbol{\Sigma}, \boldsymbol{\Omega})$, *if its probability density function is given by*

$$\begin{aligned} f_{\boldsymbol{X}}(\boldsymbol{X}) = &\frac{\Gamma_p\left(\frac{\nu+p+q-1}{2}\right)}{\pi^{\frac{pq}{2}}\Gamma_p\left(\frac{\nu+p-1}{2}\right)} \det(\boldsymbol{\Omega})^{-p/2}\det(\boldsymbol{\Sigma})^{-q/2} \\ &\times \det(\boldsymbol{I}_p + \boldsymbol{\Sigma}^{-1}(\boldsymbol{X}-\boldsymbol{T})\boldsymbol{\Omega}^{-1}(\boldsymbol{X}-\boldsymbol{T})')^{-\frac{\nu+p+q-1}{2}}, \end{aligned}$$

*where* $\boldsymbol{\Sigma} \in \mathbb{R}^{p\times p}$ *and* $\boldsymbol{\Omega} \in \mathbb{R}^{q\times q}$ *are positive definite symmetric matrices.*

For $\nu = 1$, matrix-variate $t$-distribution with $\nu$ degrees of freedom is often referred to as Cauchy distribution. In particular, if $\mathbf{T} = \mathbf{0}$, (centered) matrix-variate $t$ distribution belongs to class of left spherical distributions for $\boldsymbol{\Sigma} = \mathbf{I}_p$, right spherical distributions for $\boldsymbol{\Omega} = \mathbf{I}_q$ and the class of spherical distributions if $\boldsymbol{\Sigma} = \mathbf{I}_p$ and $\boldsymbol{\Omega} = \mathbf{I}_q$. For more details about the matrix-variate $t$-distribution see Gupta and Nagar (1999).

In general, in the case where the elliptically distributed random matrix $\mathbf{X} = \mathbf{A}\mathbf{Z}\mathbf{B}' + \mathbf{T}$ has a probability density function, it is of the form

$$f_{\mathbf{X}}(\mathbf{X}) = \frac{g(\mathrm{tr}\left[\boldsymbol{\Omega}^{-1}(\mathbf{X}-\mathbf{T})'\boldsymbol{\Sigma}^{-1}(\mathbf{X}-\mathbf{T})\right])}{\det(\boldsymbol{\Omega})^{p/2}\det(\boldsymbol{\Sigma})^{q/2}},$$

for symmetric positive definite matrices $\boldsymbol{\Sigma} = \mathbf{A}\mathbf{A}'$ and $\boldsymbol{\Omega} = \mathbf{B}\mathbf{B}'$ and real valued function $g$. In case the second moments of $\mathbf{X}$ exits, $\mathbb{E}(\mathrm{vec}(\mathbf{X})) = \mathrm{vec}(\mathbf{T})$, while $\mathrm{Cov}(\mathrm{vec}(\mathbf{X})) \propto \boldsymbol{\Omega} \otimes \boldsymbol{\Sigma}$, where the proportionality constant depends on the specific distribution. For further results on matrix-variate elliptical distributions see e.g. Gupta and Nagar (1999); Fang and Ting (1990). As in the vector case, analysis of matrix-variate elliptical models is mostly done by inspecting second-order moment behaviour. Thus, in the next section we discuss extensions of PCA for matrix-variate data, with the note that in general, matrix-variate extensions of PCA are model-free procedures, just as multivariate PCA is.

## 3.2. Matrix-variate principal component analysis

As discussed in Section 2.2.1, PCA (Jolliffe, 2002) is often used with aim of dimension reduction in the high-dimensional data analysis, by searching for the transformation of the data onto a low-dimensional space that retains maximal variation. When the data is matrix-valued, traditional analysis vectorizes each observation into a long vector, thus producing a model with a large number of parameters. For example, in the PCA-based face recognition methods (Turk and Pentland, 1991; Zhao et al., 2003), the face image matrices must be pre-vectorizes into large vectors, which often leads to high-dimensional vector spaces, making it rather difficult to accurately estimate the covariance matrix (Yang et al., 2004). Thus, in cases where the sample size is relatively small, many existing vector-valued statistical methods fail to work satisfactorily. A possible strategy for overcoming the presented difficulty is to take advantage of the matrix structure of the data. Probably most straightforward generalization of the standard (vector) PCA to matrices searches for low-dimensional projections of matrix objects which again capture maximal data variation, where methods differ based on how one measures variation in higher-order data. E.g. in two-component principal component analysis (Zhang and Zhou, 2005) it is characterized by the trace of the covariance of the projection while Lu et al. (2008) use the squared Frobenius norm of the covariance of the transformation.

Probably the biggest difference between the vectorial and matrix PCA is that in the latter, connection with matrix-variate elliptical models is not as clear. For it to hold, one would need to define affine equivariant scatter functionals $\mathbf{S}_i : \mathbb{R}^{p_1 \times p_2} \to \mathbb{R}^{p_i \times p_i}$, $i = 1, 2$, such that

$$\mathbf{S}_1(\mathbf{AXB}') = \mathbf{A}\mathbf{S}_1(\mathbf{X})\mathbf{A}', \quad \mathbf{S}_2(\mathbf{AXB}') = \mathbf{B}\mathbf{S}_2(\mathbf{X})\mathbf{B}',$$

for all regular matrices $\mathbf{A} \in \mathbb{R}^{p \times p}$, $\mathbf{B} \in \mathbb{R}^{q \times q}$. Virta et al. (2017) conjectures that such functionals in general do not exist. However, orthogonally equivariant analogs to presented scatters do exist with $\mathbb{E}((\mathbf{X} - \mathbb{E}(\mathbf{X})(\mathbf{X} - \mathbb{E}(\mathbf{X}))')$ and $\mathbf{E}((\mathbf{X} - \mathbb{E}(\mathbf{X}))'(\mathbf{X} - \mathbb{E}(\mathbf{X}))$ being one of them. Thus, we continue discussion on matrix PCA methods in scope of Model (3.1).

**Definition 5.** *The random matrix $\boldsymbol{X} \in \mathbb{R}^{p_1 \times p_2}$ follows the noisy second-order (NS) model if it allows a representation*

$$\boldsymbol{X} = \boldsymbol{T} + \boldsymbol{U}_1 \boldsymbol{Z} \boldsymbol{U}_2' + \boldsymbol{\varepsilon}, \tag{3.1}$$

*where $\boldsymbol{T} \in \mathbb{R}^{p_1 \times p_2}$ is the mean matrix, $\boldsymbol{U}_1 \in \mathbb{R}^{p_1 \times d_1}$, $\boldsymbol{U}_2 \in \mathbb{R}^{p_2 \times d_2}$ are unknown matrices with orthonormal columns and $\boldsymbol{Z}$ is a $d_1 \times d_2$ core matrix with zero mean and dimensions $d_1 \leq p_1$, $d_2 \leq p_2$.*

Additionally, one poses the technical assumptions that $\mathbb{E}\|\mathbf{Z}\|^2 < \infty$ and that $\mathbb{E}(\mathbf{ZZ}')$ and $\mathbb{E}(\mathbf{Z}'\mathbf{Z})$ are positive definite matrices. The additive $p_1 \times p_2$ noise matrix $\boldsymbol{\varepsilon}$ is taken to be independent from the core $\mathbf{Z}$ and has a matrix spherical distribution, implying that $\mathbb{E}(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}') = \sigma^2 \mathbf{I}_{p_1}$ for some $\sigma^2 \geq 0$. Furthermore, for sake of identifibility of the parameters of Model (3.1), one assumes that multiplicity of eigenvalues of $\mathbb{E}(\mathbf{ZZ}')$ and $\mathbb{E}(\mathbf{Z}'\mathbf{Z})$ is 1, respectively.

To overcome the problems obtained by applying vector PCA to matrix-valued observations, Zhang and Zhou (2005) propose an image projection technique, called two-dimensional

principal component analysis (2DPCA). For the $p_1 \times p_2$ random matrix $\mathbf{X}$ from Model (3.1), 2DPCA searches for mutually orthogonal directions $\mathbf{v}_k \in \mathbb{R}^{p_2}$, $k = 1, \ldots, d_2$ such that the total variation of the projection $\mathbf{X}\mathbf{v}_k$ of $\mathbf{X}$ onto $\mathbf{v}_k$, which is being characterised by the trace of the covariance matrix of the projection, is maximal. More precisely, $\mathbf{v}_1 \in \mathbb{R}^{p_2}$ maximizes

$$\operatorname{tr}\left(\mathbb{E}((\mathbf{X}\mathbf{v} - \mathbb{E}(\mathbf{X}\mathbf{v}))(\mathbf{X}\mathbf{v} - \mathbb{E}(\mathbf{X}\mathbf{v}))')\right) = \mathbf{v}'\mathbb{E}((\mathbf{X} - \mathbb{E}(\mathbf{X}))(\mathbf{X} - \mathbb{E}(\mathbf{X}))')\mathbf{v}.$$

The matrix $\mathbb{E}((\mathbf{X}-\mathbb{E}(\mathbf{X}))(\mathbf{X}-\mathbb{E}(\mathbf{X}))')$ is called image covariance or scatter matrix. Observe that it is not a covariance matrix in a standard definition of it. However, due to the resemblance to the covariance matrix of a random vector, we refer to it as such. Second direction $\mathbf{v}_2$ then maximizes $\mathbf{v}'\mathbb{E}((\mathbf{X} - \mathbb{E}(\mathbf{X}))(\mathbf{X} - \mathbb{E}(\mathbf{X}))')\mathbf{v}$, subject to being orthogonal to $\mathbf{v}_1$, and so on. It is now clear that $\mathbf{v}_1, \ldots, \mathbf{v}_{d_2}$ can be found as first $d_2$ eigenvectors of positive, semi-definite symmetric matrix $\mathbb{E}((\mathbf{X}-\mathbb{E}(\mathbf{X}))(\mathbf{X}-\mathbb{E}(\mathbf{X}))')$. The obtained result of 2DPCA is now $\mathbf{X}\mathbf{V}_2$, where $\mathbf{V}_2 = (\mathbf{v}_1, \ldots, \mathbf{v}_{d_2})$. Observe that 2DPCA is essentially working in the row-direction of the matrix $\mathbf{X}$. One extracts information contained in columns of $\mathbf{X}$ by applying the analogous procedure to $\mathbf{X}'$, thus projecting $\mathbf{X}'$ onto orthogonal matrix $\mathbf{V}_1 \in \mathbb{R}^{p_1 \times d_1}$. Within the scope of Model (3.1) and under the additional assumption on multiplicity of eigenvalues of $\mathbb{E}(\mathbf{Z}\mathbf{Z}')$ and $\mathbb{E}(\mathbf{Z}'\mathbf{Z})$, 2DPCA recovers mixing matrices $\mathbf{U}_1$ and $\mathbf{U}_2$, up to ordering and sign changes of the columns.

Zhang and Zhou (2005) propose simultaneous use of matrices $\mathbf{V}_1$ and $\mathbf{V}_2$ obtained by 2DPCA through the projection $\mathbf{V}_1'\mathbf{X}\mathbf{V}_2$. The method is known as two-component-two-direction principal component analysis (2D$^2$PCA). Zhang and Zhou (2005) further compare the accuracy of classification of gray-scale images of frontal faces pre-transformed using PCA, 2DPCA, and 2D$^2$PCA, concluding that 2D$^2$PCA is superior to the former two methods both in accuracy and computational time.

Alternatively to the presented procedure, one can estimate mixing matrices simultaneously. The approach called multilinear principal component analysis (MPCA) was given in Lu et al. (2008) for general order $m$ tensors. For sake of simplicity, we present the idea of the method for order-2 tensors, i.e. matrices. The MPCA solution to Model (3.1) is $\mathbf{V}_1'\mathbf{X}\mathbf{V}_2$, where the unmixing matrices $\mathbf{V}_1 \in \mathcal{O}^{p_1 \times d_1}$ and $\mathbf{V}_2 \in \mathcal{O}^{p_2 \times d_2}$ with orthogonal columns are found such that total variation captured by the projection is maximal. More precisely, $\mathbf{V}_1$ and $\mathbf{V}_2$ maximize $\mathbb{E}\|\mathbf{V}_1'(\mathbf{X} - \mathbb{E}(\mathbf{X}))\mathbf{V}_2\|_F^2$, where $\|\cdot\|_F$ denotes Frobenius matrix norm. The solution to the presented maximization problem can be found by the higher-order orthogonal iteration algorithm (HOOI) (Sheehan and Saad, 2007), which relies on the estimating equations

$$\mathbb{E}\left((\tilde{\mathbf{X}}\mathbf{V}_2)((\tilde{\mathbf{X}}\mathbf{V}_2)')\right)\mathbf{V}_1 = \mathbf{V}_1\mathbf{D}_1, \quad \mathbb{E}\left(\tilde{\mathbf{X}}'\mathbf{V}_1)(\tilde{\mathbf{X}}'\mathbf{V}_1)'\right)\mathbf{V}_2 = \mathbf{V}_2\mathbf{D}_2,$$

where $\tilde{\mathbf{X}} = \mathbf{X} - \mathbb{E}(\mathbf{X})$, and $\mathbf{D}_1 \in \mathbb{R}^{d_1 \times d_1}$, $\mathbf{D}_2 \in \mathbb{R}^{d_2 \times d_2}$ are diagonal matrices containing the eigenvalues of $\mathbb{E}((\tilde{\mathbf{X}}\mathbf{V}_2)((\tilde{\mathbf{X}}\mathbf{V}_2)')$ and $\mathbb{E}((\tilde{\mathbf{X}}'\mathbf{V}_1)((\tilde{\mathbf{X}}'\mathbf{V}_1)')$, respectively. As it is the case for 2DPCA, MPCA also recovers the latent mixing matrices $\mathbf{U}_1$ and $\mathbf{U}_2$ under Model (3.1) and additional assumption posed on multiplicity of eigenvalues of $\mathbb{E}(\mathbf{Z}\mathbf{Z}')$ and $\mathbb{E}(\mathbf{Z}'\mathbf{Z})$, up to the sign and order changes of their columns. Hung et al. (2012) argues that MPCA is asymptotically more efficient than 2D$^2$PCA in estimating the target dimension reduction subspace. However, the price to pay is increased computational cost.

Regardless of the estimation procedure used for estimation of latent mixing matrices $\mathbf{U}_1 \in \mathbb{R}^{p_1 \times d_1}$, $\mathbf{U}_2^{p_2 \times d_2}$, knowledge of the dimension of the core matrix $\mathbf{Z} \in \mathbb{R}^{d_1 \times d_2}$ is needed.

### Dimension estimation in two-dimensional PCA

As discussed, Model (3.1) itself can be thought of as a form of dimension reduction for the images where, for each original image $\mathbf{X}_i$, there exists a low-rank latent core image $\mathbf{Z}_i$ that contains the signal/information content of the image. This signal is then contaminated by the noise $\boldsymbol{\varepsilon}_i$ to produce the observed image. Thus the "true" row and column dimensions of the images are $d_1$ and $d_2$, respectively, and the objective is to estimate them based solely on the sample $\mathbf{X}_1, \ldots, \mathbf{X}_n$ from Model (3.1). Standard dimension estimation techniques rely solely on the magnitude of the eigenvalues of $\mathbb{E}(\tilde{\mathbf{X}}'\tilde{\mathbf{X}})$ and $\mathbb{E}(\tilde{\mathbf{X}}\tilde{\mathbf{X}}')$ and use mostly a rule of thumb where enough components (features) are selected to reach a pre-determined amount of "explained variation" (Yang et al., 2004; Lu et al., 2008). A naive "automated" way of estimating the dimensions would be to plot the eigenvalues of $\mathrm{E}(\tilde{\mathbf{X}}\tilde{\mathbf{X}}')$ and $\mathrm{E}(\tilde{\mathbf{X}}'\tilde{\mathbf{X}})$ as a scree plot and search for an "elbow". However, this is often difficult to locate (see middle Figure 3.2), and additional information are usually needed in order to estimate dimensions accurately.

In Paper III, we propose an automatic tool for determining the optimal number of components, in the context of Model (3.1) and 2D$^2$PCA, by extending the approach presented in Luo and Li (2021) for vector-valued observations. Since the Model (3.1) is fully symmetric, we further clarify the proposed augmentation estimator of the row-dimension $d_1 > 0$ only, where for the simplicity of the notation we assume that $\mathbb{E}(\mathbf{X}) = \mathbf{0}$. That being said, the augmentation estimator extends the idea presented in Luo and Li (2021) for vector-valued observations and concatenates the observed $\mathbf{X}$ with additional artificial normally distributed rows $\mathbf{X}_S \in \mathbb{R}^{r \times p_2}$ that mimic the first and second-moment behavior of the error $\boldsymbol{\varepsilon}$ in Model (3.1) to produce the augmented observation $\mathbf{X}^* = (\mathbf{X}', \mathbf{X}_S')'$, where number of rows $r \geq 1$ of $\mathbf{X}_S$ is a tuning parameter. The augmented (artificially added) part of the first $d_1$ eigenvectors of $\mathrm{E}\{\mathbf{X}^*(\mathbf{X}^*)'\}$ turns out to be negligible when compared to the augmented parts of the latter eigenvectors, allowing us to distinguish between the eigenvectors belonging to the first $d_1$, significant, eigenvalues, and the remaining ones. For illustration see Figure 3.2. More precisely, in Model (3.1), $\mathbb{E}(\mathbf{X}\mathbf{X}') = \mathbf{U}_1\mathbb{E}(\mathbf{Z}\mathbf{Z}')\mathbf{U}_1' + \mathbb{E}(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}')$ where $\mathbb{E}(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}') = \sigma_1^2\mathbf{I}_{p_1}$ for some $\sigma_1^2 \geq 0$. Consequently, the rank of $\mathbb{E}(\mathbf{X}\mathbf{X}') - \sigma_1^2\mathbf{I}_{p_1}$ is precisely the dimension $d_1$ we aim to estimate. Then, for $r > 0$, and $\mathbf{X}^* = (\mathbf{X}', \mathbf{X}_S')'$ as above,

$$\mathbf{M}^* := \mathbb{E}\{\mathbf{X}^*(\mathbf{X}^*)'\} - \sigma_1^2\mathbf{I}_{p_1+r} = \begin{pmatrix} \mathbf{U}_1\mathbb{E}(\mathbf{Z}\mathbf{Z}')\mathbf{U}_1' & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}$$

and $\mathbf{U}_1\mathbb{E}(\mathbf{Z}\mathbf{Z}')\mathbf{U}_1'$ are of the same rank and also have the same positive eigenvalues. Denote next the eigenvalues of $\mathbb{E}(\mathbf{Z}\mathbf{Z}')$ by $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_{d_1} > 0$ and let the $(p_1 + r)$-variate vector $\boldsymbol{\beta}_i^* = (\boldsymbol{\beta}_i', \boldsymbol{\beta}_{i,S}')'$, $i = 1, \ldots, p_1 + r$, be any eigenvector of $\mathbf{M}^*$ corresponding to its $i$th eigenvalue. We call the $r$-dimensional subvector $\boldsymbol{\beta}_{i,S}$ the *augmented part* (subvector) of the $i$th eigenvector. Then, for $i \leq d_1$, $\mathbf{M}^*\boldsymbol{\beta}_i^* = (\mathbf{U}_1\mathbb{E}(\mathbf{Z}\mathbf{Z}')\mathbf{U}_1'\boldsymbol{\beta}_i', \mathbf{0}')' = \lambda_i(\boldsymbol{\beta}_i', \boldsymbol{\beta}_{i,S}')'$, implying that $\boldsymbol{\beta}_{i,S} = \mathbf{0}$ for $i = 1, \ldots d_1$. Observe also that the same does not hold for the later eigenvectors.

A similar procedure is then applied to $\mathbf{X}'$ with aim of estimation of $d_2$. The estimation of unknown parameters $\sigma_1^2 = \mathbb{E}(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}')$ and $\sigma_2^2 = \mathbb{E}(\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon})$ is crucial for successful implementation of presented method. As both row and column dimensions are usually unknown and

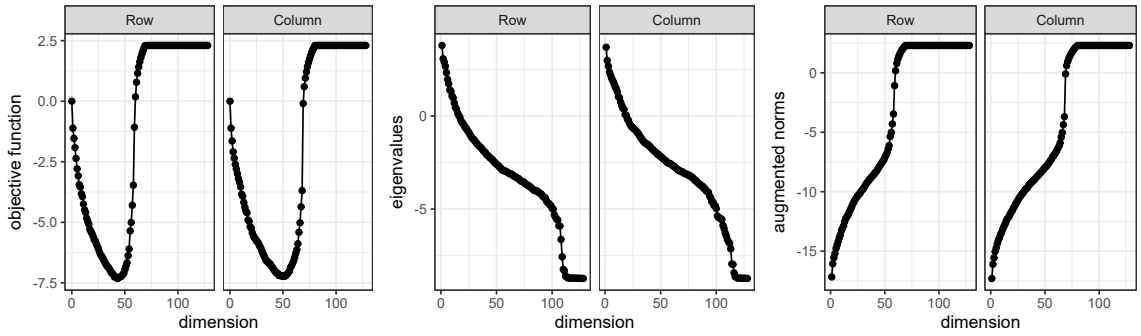



Figure 3.2.: Left to right, the logarithmized objective function for the augmented ladle estimator using $r = 10$ augmented components as a combination of the augmented norms and the scaled eigenvalues, logarithmized scree plot for row and column PCs, and logarithmized augmented norms, calculated for the subset of *fingers* data set, available freely in `https://www.kaggle.com/koryakinp/fingers` and consisting of $128 \times 128$ grayscale images of hands with 0 and 5 fingers extended.

estimated, several consistent pooled estimators of unknown quantities are given in Paper III. The automated procedure is implemented in R-package *tensorBSS* (Virta et al., 2021a) and is shown to perform very well in both simulated and real data.

It is worth mentioning that to our best knowledge, automated dimension selection in this context has been developed earlier only by Tu et al. (2019) who use Stein's unbiased risk estimation (SURE) for the task, which is however computational very expansive and often unfeasible, as it iterates through all possible combinations of $p$ and $q$.

## 3.3. Matrix-variate independent component model

Matrix-variate elliptical models inherit the symmetry properties from the multivariate elliptical models. Thus, to model e.g. skewed data, one could use the matrix-variate independent component model as a working framework.

**Definition 6.** *A random matrix $\boldsymbol{X} \in \mathbb{R}^{p \times q}$ is said to follow matrix-variate independent component model if it allows representation*

$$\boldsymbol{X} = \boldsymbol{A}\boldsymbol{Z}\boldsymbol{B}' + \boldsymbol{T}, \tag{3.2}$$

*where $\boldsymbol{T} \in \mathbb{R}^{p \times q}$ is the mean matrix, the invertible $\boldsymbol{A} \in \boldsymbol{R}^{p \times p}$ and $\boldsymbol{B} \in \mathbb{R}^{q \times q}$ are unknown mixing matrices and the latent matrix $\boldsymbol{Z}$ is assumed to have mutually independent, marginally standardized components.*

Thus, in the context of the location-scatter model, an additional assumption of independence is posed to latent matrix $\mathbf{Z}$. As in the multivariate setting, ICA aims to estimate unknown mixing matrices, i.e. to identify latent matrix $\mathbf{Z}$. Under the additional assumption that at most one entire row, and most one entire column of $\mathbf{Z}$ have a multivariate normal distribution, latent matrix $\mathbf{Z}$ is identifiable up to the order, joint scaling, and signs of its rows and columns. On the contrary, in Model (3.2) $\mathbf{Z}$ is allowed to contain up to

$pq - \max(p, q)$ normal components, given that all the non-normal ones are suitably located. As discussed for the general location-scatter model, vectorizing Model (3.2) yields a Kronecker-structured multivariate independent component model

$$\mathrm{vec}(\mathbf{X}) = (\mathbf{B} \otimes \mathbf{A})\mathrm{vec}(\mathbf{Z}) + \mathrm{vec}(\mathbf{T}).$$

Naturally, any reasonable multivariate ICA method applied to the vectorized model should take into account the special form of the mixing matrix. However, to our knowledge, such methods were not developed. On the other hand, the problem of estimation of unknown mixing matrices under the matrix representation of Model (3.2) is considered in e.g. Virta et al. (2017, 2018, 2021b), where two classical multivariate ICA procedures, the fourth-order blind identification (FOBI) (Cardoso, 1989) and joint diagonalization of eigenmatrices (JADE) (Cardoso and Souloumiac, 1993), were extended to TFOBI and TJADE, respectively, to estimate unknown parameters in Model (3.2). It is important to emphasize that Model (3.2) and the discussed method for its solving are derived in more general form, for general $r$th order tensors.

To the best of our knowledge there are not yet matrix-variate extensions of multivariate NGCA and NGICA models, thus making it difficult to put it into a framework for feature extraction as e.g. a preprocessing step for clustering as in the vector case.

## 3.4. Matrix-variate Gaussian mixture model

As it is the case in a multivariate setting, matrix mixture models are particularly convenient for modeling data originating from heterogeneous populations (the data naturally groups into several classes), where the aim of feature extraction in those models is usually classification or clustering, depending on whether the class membership is known or not. Multivariate Gaussian mixture models then naturally generalize to the matrix-variate setting as mixtures of matrix-variate normal distributions.

**Definition 7.** *Random matrix $\boldsymbol{X} \in \mathbb{R}^{p\times q}$ is said to follow matrix-variate Gaussian mixture model (MGMM) with $k$-classes if its probability density function is given by*

$$f(\boldsymbol{X}; \alpha_1, \ldots, \alpha_k, \boldsymbol{\Theta}_1, \ldots, \boldsymbol{\Theta}_k) = \sum_{i=1}^{k} \alpha_i \phi_i^{(p\times q)}(\boldsymbol{X}; \boldsymbol{T}_i, \boldsymbol{\Sigma}_i, \boldsymbol{\Omega}_i),$$

*where $\boldsymbol{\Theta}_i = (\boldsymbol{T}_i, \boldsymbol{\Sigma}_i, \boldsymbol{\Omega}_i)$ and $\phi_i^{(p\times q)}(\cdot; \boldsymbol{T}_i, \boldsymbol{\Sigma}_i, \boldsymbol{\Omega}_i)$, $i = 1, \ldots, k$ denote the set of parameters and probability density function of $i$th matrix-variate normally distributed class, respectively.*

We say that MGMM is homoscedastic if all classes have common covariances, i.e. when $\boldsymbol{\Sigma}_i = \boldsymbol{\Sigma}$ and $\boldsymbol{\Omega}_i = \boldsymbol{\Omega}$, for all $i = 1, \ldots, k$, where $\boldsymbol{\Sigma} \in \mathbb{R}^{p\times p}$ and $\boldsymbol{\Omega} \in \mathbb{R}^{q\times q}$ are positive definite symmetric matrices. As a consequence of the fact that vectorization transfers MGMM to multivariate GMM, all MGMM belong to location-scatter models. Moreover, homoscedastic MGMM have a location-scatter representation in which the latent matrix $\mathbf{Z}$ is a mixture of spherical, matrix-variate normal distributions (Viroli, 2011). More precisely, a $p \times q$ random matrix $\mathbf{X}$, from homoscedastic MGMM with density $f(\mathbf{X}) = \sum_{i=1}^{k} \alpha_i \phi_i^{(p\times q)}(\mathbf{X}; \mathbf{T}_i, \boldsymbol{\Sigma}, \boldsymbol{\Omega})$,

allows a representation as $\mathbf{X} = \mathbf{\Sigma}^{1/2}\mathbf{Z}\mathbf{\Omega}^{1/2} + \mathbf{T}$, where $\mathbf{Z}$ follows homoscedastic MGMM with $k$-classes and probability density function

$$f(\mathbf{Z}) = \sum_{i=1}^{k} \alpha_i \phi_i^{(p\times q)}(\mathbf{Z}; \mathbf{\Sigma}^{-1/2}(\mathbf{T}_i - \mathbf{T})\mathbf{\Omega}^{-1/2}, \mathbf{I}_p, \mathbf{I}_q),$$

where $\mathbf{T} = \sum_{i=1}^{k} \alpha_i \mathbf{T}_i$. The mean and the covariance matrix of the random matrix $\mathbf{X}$ from MGMM (7) are

$$\begin{aligned}
\mathbb{E}(\mathrm{vec}(\mathbf{X})) &= \sum_{i=1}^{k} \alpha_i \mathrm{vec}(\mathbf{T}_i) \quad \text{and} \\
\mathrm{Cov}(\mathrm{vec}(\mathbf{X})) &= \sum_{i=1}^{k} \alpha_i \left(\mathrm{vec}(\mathbf{T}_i)\mathrm{vec}(\mathbf{T}_i)' + \mathbf{\Omega}_i \otimes \mathbf{\Sigma}_i\right) \\
&\quad - \left(\sum_{i=1}^{k} \alpha_i \mathrm{vec}(\mathbf{T}_i)\right)\left(\sum_{i=1}^{k} \alpha_i \mathrm{vec}(\mathbf{T}_i)\right)',
\end{aligned}$$

respectively. Mixtures of matrix-variate Gaussian distribution inherit many properties from matrix-variate normal distributions. Especially, for random matrix $\mathbf{X}$ from MGMM (7) and full rank matrices $\mathbf{A} \in \mathbb{R}^{p\times r}$, $r \leq p$ and $\mathbf{B} \in \mathbb{R}^{q\times s}$, $s \leq q$, probability density function of the linear transformation $\mathbf{A}'\mathbf{X}\mathbf{B} \in \mathbb{R}^{r\times s}$ is given by

$$f(\mathbf{A}'\mathbf{X}\mathbf{B}; \alpha_1, \ldots, \alpha_k, \mathbf{\Theta}_1, \ldots, \mathbf{\Theta}_k) = \sum_{i=1}^{k} \alpha_i \phi_i^{(r\times s)}(\mathbf{X}; \mathbf{A}'\mathbf{T}_i\mathbf{B}, \mathbf{A}'\mathbf{\Sigma}_i\mathbf{A}, \mathbf{B}'\mathbf{\Omega}_i\mathbf{B}).$$

The proof of the statement can be found in Viroli (2011), and its importance lies in the fact that it shows how low-rank projections of MGMM are again MGMM. Thus, appropriate linear transformations can be applied for visualizations, dimension reduction, and consequentially classification. As a special case, when $\mathbf{A}$ or (and) $\mathbf{B}$ are vectors, the above consideration shows the connection between matrix-variate and multivariate (univariate) GMM. Especially relevant for our purposes are rank-1 projections obtained for $r = s = 1$, i.e., when both $\mathbf{A}$ and $\mathbf{B}$ are vectors.

Parameter estimation in matrix-variate GMM is done similarly as in the multivariate setting. With maximum likelihood estimation as a gold standard, parameters in MGMM can be estimated through EM algorithm (Dempster et al., 1977; Friedman et al., 2001), where the concrete adaptation of multivariate EM algorithm to MGMM can be found in Viroli (2011) and references therein. However, as the problem of multimodality of the likelihood function in high-dimensional multivariate settings is even more troubling in the matrix-variate case as the dimensions increase, so does the need for performing appropriate dimension reduction. As discussed in Section 3.2, probably the most widely used class of methods for this purpose are various extensions of PCA. However, in e.g. both 2D$^2$PCA and MPCA, the criterion by which the feature extraction is done is not tailored in a way to ensure maximal separation between the classes in the lower-dimensional space. Thus, we next discuss the extension of multivariate LDA to the matrix-variate setting.

### 3.4.1. Matrix-variate linear discriminant analysis

Based on the Fisher discriminant rule Foley and Sammon (1975) presented a method for the extraction of features in two-class image data. To use the matrix structure of $(p \times q)$ random matrix $\mathbf{X}$, Foley and Sammon (1975) seek $0 < d \leq q$ mutually orthogonal linear discriminant directions $\mathbf{u}_1, \ldots, \mathbf{u}_d$ that maximize the trace of the between-class scatter over the trace of the within-class scatter of the projection $\mathbf{x}_k = \mathbf{X}\mathbf{u}_k$, $k = 1, \ldots d$ . The method naturally involves the inverse of the within-class scatter, which can, due to the small sample size when compared to data dimensionality, be singular. Various algorithms which overcome this problem and also generalize to multiple classes have been developed over the years. One of the most popular is perhaps 2D-LDA (Li and Yuan, 2005). However, 2D-LDA ignores the between-row correlations in the matrix observations which can lead to substantially higher missclassification error than applying multivariate Fisher's LDA to vectorized observations, when the rows are correlated (Zheng et al., 2008). Thus, to discriminate the matrix-variate observations Zhong et al. (2015) propose one-step method to find $d$ *rank-1-projections* $\mathbf{u}_i'\mathbf{X}\mathbf{v}$, $i = 1, \ldots, d$ which exhibit the maximum ratio of between-class- to the within-class variance, where $\mathbf{u}_i \in \mathbb{R}^d$ are mutually orthogonal and $\mathbf{v} \in \mathbb{R}^q$. Zhong et al. (2015) also discusses a two-step generalization of the initially proposed method that accommodates for multiple $\mathbf{v}$ as well. The term *rank-1 projection* stems from the representation $\mathbf{u}'\mathbf{X}\mathbf{v} = \operatorname{tr}(\mathbf{v}\mathbf{u}'\mathbf{X}) = \langle \mathbf{u}\mathbf{v}', \mathbf{X} \rangle$ of projecting $\mathbf{X}$ onto the rank-1 matrix $\mathbf{u}\mathbf{v}'$ (w.r.t. the standard Euclidean geometry on matrices), and has a long use in the machine learning literature, most often in the context of classification using projection indices based on second moments, see, e.g., Hua et al. (2007); Liu et al. (2011); Wu et al. (2011a,b).

On the other hand, probably the most straightforward extension of multivariate LDA to matrix-variate setting is to vectorize matrix-valued observations and proceed by applying well-studied multivariate LDA. In recent years, some progress has been made on developing sparse multivariate LDA using $l_1$-regularization (Tibshirani, 1996), including Shao et al. (2011); Fan et al. (2012); Mai et al. (2012). However, all these methods, as well as the classical multivariate LDA deal with vector-valued covariates. Also, $l_1$-regularization, which has shown success in a high-dimensional vector setting, does not necessarily work well in the matrix-variate context because the underlying matrix-variate signals are usually approximately low-rank rather than $l_0$-sparse (Hu et al., 2020). Thus, in Paper IV we propose a special kind of regularization to LDA optimal direction, where we assume that it allows a natural matrix representation, and then proceed by unsupervised estimation of such matrix, one *rank-1-block* at the time, where the individual blocks are estimated using projection pursuit, with a special focus on estimation in homoscedastic MGMM with two classes.

### 3.4.2. Projection pursuit based estimation of the linear discriminant

As discussed, when it comes to the feature extraction, one of the most common approaches is projection pursuit, and in the recent years, due to the growing complexity of the available data sets, the need for projection pursuit has only increased. However, applying projection pursuit to large, high-dimensional data, where the large sample size $n$ is not necessarily larger then the large number of observed covariates $p$, is not as straightforward. Namely, PP faces several issues in very high-dimensional setting. Diaconis and Freedman (1984)

show how most univariate projections of the high-dimensional point clouds are approximately normal. Huber (1985) thus argues that in high-dimensional settings there is little information to be found by projection pursuit, since Gaussianity is usually considered as noise. On the other hand, Pires and Branco (2019) show that when the dimension of the data $p$ is larger then the available sample size $n$, one can always find a two-dimensional projection equal to (up to affine transformation) arbitrary configuration of points in $\mathbb{R}^2$. Thus, applying PP in the high-dimensional setting may seem like a rather futile exercise. However, these issues can be bypassed by imposing a suitable structure on the data. In Paper IV we therefore assume that the observable $pq$-variate random vector $\mathbf{x}$ allows a natural representation as a random $p \times q$ matrix $\mathbf{X}$, such that $\mathrm{vec}(\mathbf{X}) = \mathbf{x}$, thus further working with projections of the form $\mathbf{u}'\mathbf{X}\mathbf{v}$ where $\mathbf{u} \in \mathbb{S}^{p-1}$, $\mathbf{v} \in \mathbb{S}^{q-1}$. Such projections have a total of $p+q-2$ degrees of freedom, where the standard approach of vectorizing $\mathbf{X}$ and working with the usual projections $\mathbf{w}'\mathrm{vec}(\mathbf{X})$ involves $pq-1$ degrees of freedom. For illustration, take a horizontal slice of an fMRI image at typical resolution of $64 \times 64$ pixels (Lindquist, 2008). The rank-1 projection on $\mathbf{u} \in \mathbb{S}^{63}$, $\mathbf{v} \in \mathbb{S}^{63}$ involves estimation of 126 parameters, whereas a projection after vectorization has in total $64^2 - 1 = 4095$ parameters, with a difference of more than one order of magnitude. Observe that the structure-ignoring projection $\mathbf{w}'\mathrm{vec}(\mathbf{X})$ can be represented as a *rank-d projection* since $\mathbf{w}'\mathrm{vec}(\mathbf{X}) = \langle \mathbf{W}, \mathbf{X} \rangle$, where $\mathrm{vec}(\mathbf{W}) = \mathbf{w}$, and $d = \mathrm{rank}(\mathbf{W})$. This shows that rank-1 projections and projection after vectorization, are two ends of a range of projections of different rank.

Given a $p \times q$ random matrix $\mathbf{X}$ (with finite fourth moments) and the projection directions $(\mathbf{u}, \mathbf{v}) \in \mathbb{S}^{p-1} \times \mathbb{S}^{q-1}$, in Paper IV we propose the kurtosis of the projection $\mathbf{u}'\mathbf{X}\mathbf{v}$,

$$\kappa_{\mathbf{X}}(\mathbf{u}, \mathbf{v}) = \frac{\mathrm{E}\left([\mathbf{u}'\{\mathbf{X} - \mathrm{E}(\mathbf{X})\}\mathbf{v}]^4\right)}{\left\{\mathrm{E}\left([\mathbf{u}'\{\mathbf{X} - \mathrm{E}(\mathbf{X})\}\mathbf{v}]^2\right)\right\}^2},$$

as the PP index. As an alternative to $\kappa_{\mathbf{X}}$, we provide an index which measures the importance of the "one-sided" projection $\mathbf{u}'\mathbf{X}$ using the Mardia's measure of multivariate kurtosis (Mardia, 1970), defined for a $q$-dimensional random vector $\mathbf{x}$ as

$$\psi(\mathbf{x}) = \mathrm{E}\left[\{\mathbf{x} - \mathrm{E}(\mathbf{x})\}'\mathrm{Cov}(\mathbf{x})^{-1}\{\mathbf{x} - \mathrm{E}(\mathbf{x})\}\right]^2.$$

Hence, given a $p \times q$ random matrix $\mathbf{X}$ (with finite fourth moments) and the projection direction $\mathbf{u} \in \mathbb{S}^{p-1}$, Mardia's kurtosis of the projection $\mathbf{u}'\mathbf{X}$ is,

$$\psi_{\mathbf{X}}(\mathbf{u}) = \mathrm{E}\left[\mathbf{u}'(\mathbf{X} - \mathrm{E}(\mathbf{X}))\left[\mathrm{E}\left((\mathbf{X} - \mathrm{E}(\mathbf{X}))'\mathbf{u}\mathbf{u}'(\mathbf{X} - \mathrm{E}(\mathbf{X}))\right)\right]^{-1}(\mathbf{X} - \mathrm{E}(\mathbf{X}))'\mathbf{u}\right]^2.$$

To estimate the projection direction $\mathbf{v}$, the right-hand side analogue of $\psi_{\mathbf{X}}$ is naturally needed. We show that under mild assumptions about moments of $\mathbf{X}$, both $\kappa_{\mathbf{X}}$ and $\psi_{\mathbf{X}}$ have both a minimizer and a maximizers in $\mathbb{S}^{p-1} \times \mathbb{S}^{q-1}$ and $\mathbb{S}^{p-1}$, respectively. However, we consider those indices especially in the scope of homoscedastic MGMM with two classes, i.e. we assume that the $p \times q$ random matrix $\mathbf{X}$ has a probability density function

$$f(\mathbf{X}; \alpha, \mathbf{T}_1, \mathbf{T}_2, \mathbf{\Sigma}, \mathbf{\Omega}) = \alpha_1 \phi_1^{(p\times q)}(\mathbf{X}; \mathbf{T}_1, \mathbf{\Sigma}, \mathbf{\Omega}) + (1-\alpha_1)\phi_2^{(p\times q)}(\mathbf{X}; \mathbf{T}_2, \mathbf{\Sigma}, \mathbf{\Omega}), \qquad (3.3)$$

for class means $\mathbf{T}_1 \in \mathbb{R}^{p\times q}$, $\mathbf{T}_2 \in \mathbb{R}^{p\times q}$, $\mathbf{T}_1 \neq \mathbf{T}_2$, positive definite common covariances $\mathbf{\Sigma} \in \mathbb{R}^{p\times p}$ and $\mathbf{\Omega} \in \mathbb{R}^{q\times q}$, and mixing proportion $\alpha_1 > 0$. Under Model (3.3), the optimal

projection for separating the parts of the mixture in the sense of LDA is $\langle \mathbf{W}_{\text{LDA}}, \mathbf{X} \rangle$, where the projection direction is

$$\mathbf{W}_{\text{LDA}} := \mathbf{A}^{-1}(\mathbf{T}_2 - \mathbf{T}_1)\mathbf{B}^{-1}.$$

In Paper IV we show that under mild conditions, sequential optimization under right linear constraints of both projection indices allows the reconstruction of $\mathbf{W}_{\text{LDA}}$, one rank-1 block at a time. The rank-1 projection of the data onto each of the $d$ extracted rank-1 blocks is then considered an interesting data feature, which enables discriminating between the two classes, where $d = \text{rank}(\mathbf{W}_{\text{LDA}})$. Furthermore, the value of the PP index then serves as an ordering of the extracted features, allowing us to further reduce the dimension of the feature subspace.

On the other side, we also show that the same is not possible for the second-order counterparts. That is, we give necessary and sufficient conditions for which the leading projections extracted by the second-order methods MPCA and $(2\text{D})^2$PCA to be able to recover optimal LDA projection, in the case where the mean difference $\mathbf{T}_1 - \mathbf{T}_2$ is of rank 1.

Furthermore, we establish strong consistency results for optimizers of both $\kappa_{\mathbf{X}}$ and $\psi_{\mathbf{X}}$, which further imply strong consistency of estimated $\mathbf{W}_{\text{LDA}}$ and conjecture that the $\kappa_{\mathbf{X}}$-based estimate of $\mathbf{W}_{\text{LDA}}$ is asymptotically normal, with standard, $\sqrt{n}$ convergence rate. It is important to mention that, unlike the second-order methods, sequential optimizers of both projection indices possess an affine equivariance property.

# 4. The notion of information, Gaussianity, and independence

In the engineering literature independent component analysis (Hyvärinen, 1999; Nordhausen and Oja, 2018) is often described as a search for the uncorrelated linear combinations of the original variables that maximize non-Gaussianity. As discussed in Section 2.3, in case of estimation by FOBI method, first the vector of principal components is found and the components are standardized to have zero means and unit variances, and second, the vector is further rotated so that the new components maximize a selected measure of non-Gaussianity. It is then argued that the features extracted this way are as independent as possible or that they display the maximal information. The aim of this chapter is to discuss and clarify to some extent the somewhat vague connections between non-Gaussianity, independence and notions of information of univariate random variables in the context of the independent component analysis.

## 4.1. Orderings of random variables

Let us start by defining some elementary characteristics of univariate random variables, and arguably some of the most classical ones are the location and dispersion. Location and dispersion of a random variable $x$ are often considered by defining the corresponding measures (functionals) for these properties, as functions of the distribution of $x$. More precisely, we say that a functional

i) $T = T(x) \in \mathbb{R}$ is a location measure if $T(ax + b) = aT(x) + b$, for all $a, b \in \mathbb{R}$,

ii) $S = S(x) \in \mathbb{R}_+$ is a dispersion measure if $S(ax + b) = |a|S(x)$, for all $a, b \in \mathbb{R}$.

Consequentially, a functional $S^2 = S^2(x) \in \mathbb{R}_+$ is a squared dispersion measure if $S^2(ax + b) = a^2 S^2(x)$, for all $a, b \in \mathbb{R}$. Observe that location measures and squared dispersion measures are in fact univariate counterparts of location and scatter functionals, and thus share all their properties. For squared dispersion measures Huber (1985) considered the concepts of additivity, subadditivity and superadditivity, which are in Paper V shown to be crucial when considering projection indices for ICA. We say that a squared dispersion measure is (sub)[super]additive if $S^2(x + y)(\leq)[\geq] = S^2(x) + S^2(y)$, for all independent $x$ and $y$. For example, Huber (1985) shows that the cumulants $\kappa_k^{2/k}(x)$, $k \geq 2$, when calculated for standardized distributions, and exponential negentropy $\exp\{\mathrm{NH}(x)\}$, among many others notions of information, provide subadditive squared dispersion measures.

Comparing different location and dispersion measures yields measures of skewness and kurtosis

$$\eta(x) = \frac{T_1(x) - T_2(x)}{S(x)} \quad \text{and} \quad \kappa(x) = \frac{S_1^2(x)}{S_2^2(x)},$$

respectively, and we have studied their use as projection indices for traditional choice of location and scatter. However, one can raise a question if it is possible to order random variables with respect to discussed measures, thus answering if one random variable "posses" e.g. kurtosis, more strongly then the other one. One can answer the question by defining the partial orderings using the, so-called *shift function* $\Delta$, where for continuous $x$ and $y$ with cumulative distribution functions (cdf) $F$ and $G$, $\Delta(x) = G^{-1}(F(x)) - x$ (Bickel and Lehmann, 1975, 1976; Zwet, 1964; Oja, 1981). The transformation $x \mapsto x + \Delta(x)$ is the Monge-Kantorovich optimal transport map, when transporting $x$ to $y$ (Rachev, 1998).

Oja (1981) for example argues how $F$ and $G$ are comparable in the location sense, if the corresponding shift function $\Delta$ is positive. Observe that $\Delta(x) \geq 0 \iff F(X) \geq G(x)$, so the partial ordering coincides with well-known stochastic order. Similarly, Oja (1981) argues how $F$ and $G$ are comparable in the dispersion sense, for $\Delta$ increasing, while these are comparable in the skewness sense, if $\Delta$ is convex. For symmetrical distributions Oja (1981) proposes kurtosis ordering by stating that $F$ and $G$ are comparable in the kurtosis sense if $\Delta$ is concave-convex around the center of the symmetry of $F$, while for the non-symmetrical distributions the mode of $F$ is taken instead of the symmetry center. Bickel and Lehmann (1975, 1976); Oja (1981) argue how in addition to properties like affine invariance and equivariance, measures of location, dispersion, skewness and kurtosis should be monotone w.r.t. the corresponding orderings.

However, when considering orderings of random variables, it seems only natural to order them by the amount of information they carry.

### 4.1.1. Information orderings for discrete distributions

In general, for a discrete distribution with $k$ possible values and probabilities listed in $p = (p_1, ..., p_k)$, it is often presumed that it is more informative if the result of the experiment involving $p$ is known with a high probability (Cover and Thomas, 2006), or that $p$ contains only a very small portion of the very high probabilities $p_i$. These somewhat naive characterizations suggest the following well-known partial ordering for discrete distributions (Marshall et al., 2011).

**Definition 8.** *For two discrete distributions $p$ and $q$, we say that $p$ is majorized by $q$, and write $p \prec q$ if*

$$\sum_{i=1}^{j} p_{(i)} \geq \sum_{i=1}^{j} q_{(i)}, \ j = 1, ..., k.$$

Pecaric et al. (1992) then gives a characterization of the majorization, which directly implies that for all discrete $p$ with up to $k$ values, $(1/k, ..., 1/k) \prec p \prec (0, ..., 0, 1)$ and, for simple mixtures, $p \prec q \quad \Rightarrow \quad p \prec \lambda p + (1 - \lambda)q \prec q, \ 0 \leq \lambda \leq 1$.

### 4.1.2. Information orderings for continuous distributions

When exploring concepts of an information ordering for the continuous distributions, one starts by mimicking those for the discrete variables.

**Definition 9.** *For a continuous random variable $x$ with probability density function (pdf) $f$ on $(0,1)$, $f_{\downarrow}(u) = \sup\{y : m(y) > u\}$, $u \in (0,1)$, provides the decreasing rearrangement of $f$, where $m(y) = \mu\{u : f(u) > y\}$ and $\mu$ is Lebesgue measure.*

Observe that decreasing rearrangements to some extent generalize the concept of ordering probabilities in the discrete distribution. For more details and examples of decreasing rearrangements, see e.g. Kristiansson (2002). Then, using the decreasing rearrangement, and mimicking majorization of discrete distributions, one constructs a partial ordering of continuous random variables with support on $(0,1)$.

**Definition 10.** *Let $f$ and $g$ be density functions on the interval $(0,1)$. Then $g$ has more information than $f$, write $f \prec g$, if*

$$\int_0^u f_{\downarrow}(v)\mathrm{d}v \leq \int_0^u g_{\downarrow}(v)\mathrm{d}v, \quad \textit{for all } u \in (0,1)$$

Ryff (1963) then gives a characterization of the partial orderings for functions with the support on $(0,1)$, as $f \prec g$ if and only if

$$\int_0^1 C(f(u))\mathrm{d}u \leq \int_0^1 C(g(u))\mathrm{d}u \quad \text{for all continuous convex functions } C.$$

Jensen inequality then implies that if $f$ is probability density function of $\mathcal{U}(0,1)$, then $f \prec g$, for all $g$.

When it comes to extending partial orderings from Definition 10, the approach we take is mapping the distribution of the random variable with support on $\mathbb{R}$, into the one with bounded, $(0,1)$ support. Furthermore, the transformation should be location and scale invariant. To find a location and scale-free version of the density, Staudte (2017) proposed the transformation

$$f(x),\ x \in \mathbb{R} \quad \mapsto \quad f^*(u) = \frac{f(F^{-1}(u))}{\mathbb{E}[f(x)]},\ u \in (0,1).$$

Function $f^*$, called *the probability density quantile* (pdQ), is a probability density function on $(0,1)$ which is invariant under linear transformations of the original variable $x$ (Staudte, 2017). Furthermore, for given $f^*$, the original $f$ is known up location and scale (Staudte, 2017).

As an alternative to pdQ, to find a location and scale-free version of the density, in Paper V we propose the transformation

$$f(x),\ x \in \mathbb{R} \quad \mapsto \quad f : \varphi(u) = \frac{f(\Phi^{-1}(u))}{\varphi(\Phi^{-1}(u))}, \quad u \in (0,1),$$

where $\varphi$ and $\Phi$ are the pdf and the cdf of a normal distribution with mean $\mathbb{E}(x)$ and variance $\mathrm{Var}(x)$, respectively. It is clear that if $f$ is a pdf of a normal distribution, then $f : \varphi$ is a pdf of $\mathcal{U}(0,1)$. Observe that the negative differential entropy of the transformation $f : \varphi(u)$, $-\mathrm{H}(f : \varphi(u)) = \mathrm{KL}(f, \varphi) \geq 0$, where $\mathrm{KL}(f, \varphi) = \mathbb{E}_f(\log(f(x)/g(x)))$ is the Kullback-Leibler (KL) divergence (Cover and Thomas, 2006) between $f$ and $g$, that, in Lehman's terms, measures how much the probability distribution $g$ differs from the reference distribution $f$. Figure 4.1 shows comparison of $f$, $f^*$ and $f : \varphi$ for normal and uniform distribution.

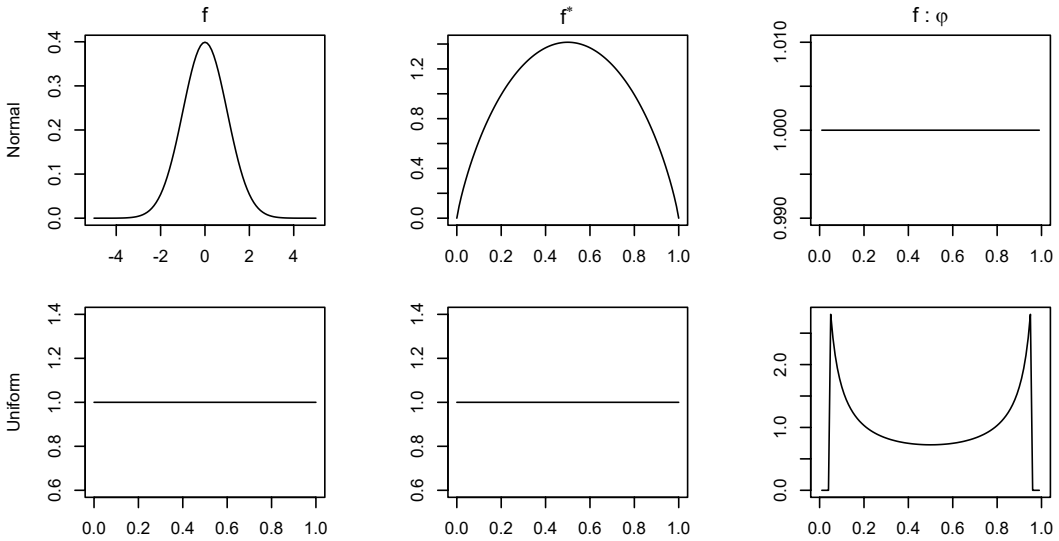



Figure 4.1.: Comparison of $f$, $f^*$ and $f : \varphi$ for normal and uniform distribution.

## 4.2. Independent component analysis, projection pursuit and information measures

Let $p$-variate random vector $\mathbf{x} = \mathbf{A}\mathbf{z} + \boldsymbol{\mu}$ follow ICA model (2.4), where $\mathbf{z} = (z_1, ..., z_p)'$ is the vector of standardized independent components. When extracting features in ICA model, one often uses projection pursuit with projection index tailored as some measure of non-Gaussianity. The question is, what justification one has for the projection index $D(\mathbf{u}) = D(\mathbf{u}'\mathbf{x})$ to find the independent components via projection pursuit? Moreover, are the extracted features provided by the most informative directions as it is often stated in the literature? We partially answer these questions in Paper V, where we argue how any subadditive squared dispersion measure, as well as its monotone transformations, can justly be used in order to recover independents components. Furthermore, the same is true for monotonic decreasing transformations of a superadditive squared dispersion measure.

Due to Huber (1985), many notions of information, exponential negentropy being one of them, are subadditive squared dispersion measures. Therefore, their monotone transformations, like e.g. negentropy, can be used as projection indices when searching for independent components, making the components extracted that way to indeed provide the most informative features. On the other side, 3rd and 4th cumulants, among others, when calculated in a standardized distribution can also be justly used as projection indices while at the same time measuring deviation from Gaussianity.

In Paper V, we further define monotone information measures, as measures of distribution that are monotone w.r.t. corresponding partial orderings, and show that many famous notions of information are in fact information measures, with negative differential entropy (and negentropy as its monotone transformation) being one of them. Thus, if the ordering is defined by pdQ $f^*$ as the transformation of pdf $f$, measures of information will attain their minimum at uniform distribution. If however the corresponding ordering is defined by $f : \varphi$ transformation, then monotone information measures will attain their minimum at normal distribution. Furthermore, in that case negentropy of such location and scale-free

version of $f$ then measures how much $f$ differs from the normal distribution, thus further connecting concepts of non-Gaussianity and information measure.

Recall that information as stated for discrete distributions is invariant under the permutations of the probabilities in $(p_1, ..., p_k)$, while all permutations consist of successive pairwise exchanges of two probabilities. In Paper V we extend that concept to continuous distributions and show that many notions of information are invariant under such elemental probability transformation. For more details and examples of monotone information measures and their properties, see Paper V.

# 5. Final remarks

In the thesis we considered dimension reduction and feature extraction for vector- and matrix-valued data, using projection pursuit in the scope of the non-Gaussian component model. Starting with the vector setting, we studied simultaneous use of two scatter functionals with aim of the dimension reduction in the non-Gaussian component model, and discussed under which conditions two different scatters can be used to estimate the subspaces. Based on this consideration we suggest bootstrap techniques to test for a specific subspace dimension and also show how successive applications of the presented tests can be used to obtain an estimate of the dimensions of interest. As illustrated in Figures 2.6 and 2.7, the successful estimation of the latent dimension is of great need for feature extraction prior to classification or clustering. We then focused on feature extraction in the homoscedastic GMM with two classes, that we showed is in fact an NGICA model, using projection pursuit. We conduct an asymptotic comparison of two popular estimators of the linear discriminant direction, supervised plug-in LDA estimator and projection pursuit estimator based on skewness and kurtosis. For the latter, we proposed using the convex combination of squared excess kurtosis and squared skewness as the projection index (giving the individual cumulants as special cases). Both the theoretical results and simulations indicate that with a suitable choice of weighting, such projection pursuit achieves good performance compared to supervised LDA, considering it operates in absence of group membership information. Moreover, in the case of moderately balanced and infinitely well-separated groups, projection pursuit is able to reach asymptotic efficiency equal to LDA with an optimal choice of weighting. We further studied the usage of various information criteria in vector-variate ICA, as well as the connections between notions of information and statistical independence, and the special role of the Gaussian distribution, while giving the result which justified their use as projection indices in a projection pursuit governed ICA (and homoscedastic GMM with two classes for that matter).

Furthermore, we generalized the ideas presented for the vector-valued observations to the matrix-variate setting. However, in absence of the matrix-variate extension of the vector-valued NGCA model, we move our focus to the matrix-variate PCA and estimation of the latent dimension in Model (3.1). Thus, we extended the augmentation-based estimator introduced in Luo and Li (2021) for vector-valued observations, to the matrix-variate setting and demonstrated its excellent performance for both simulated and real data. As the part of the future work, we will also derive the theoretical properties of the estimator and extend it to the general tensorial PCA case to also cover, for example, color images and video data. We further extended the ideas presented in Paper II to the homoscedastic MGMM with two classes, thus developing projection pursuit for the data that admit a natural representation in the matrix form. The projection indices we propose are extensions of the classical kurtosis and Mardia's multivariate kurtosis and we show that both are able to recover the optimally separating projection in the full absence of any label information, while also establishing the strong consistency of the corresponding sample estimators. As the part of the future work, we will derive limiting distribution of the estimators, which we conjecture are Gaussian with

$\sqrt{n}$ convergence rate, and further extend results from Paper II to matrix variate setting, in sense of using skewness as well as the combination of skewness and the kurtosis as projection indices.

One of the particular difficulties regarding the presented model is that it is highly affected by the outliers. Thus, we will also consider more robust alternatives to the presented projection indices. Furthermore, we will consider extending the NGCA and NGICA models to matrix variate setting while also investigating if the matrix-variate PP can also be used for the matrix-variate NGICA, as it is successfully done in the multivariate setting.

# References


T. W. Anderson. *An Introduction to Multivariate Statistical Analysis*. Wiley, New York, 3 edition, 2003.

M. Arashi. Some theoretical results on tensor elliptical distribution. *arXiv preprint at arXiv:1709.00801*, 2017.

L. S. Athanasiou, D. I. Fotiadis, and L. K. Michalis. Principles of coronary imaging techniques. In L. S. Athanasiou, D. I. Fotiadis, and L. K. Michalis, editors, *Atherosclerotic Plaque Characterization Methods Based on Coronary Imaging*, pages 23–47. Academic Press, Oxford, 2017.

D. M. Bean. *Non-Gaussian Component Analysis*. PhD thesis, University of California, Berkeley, 2014.

P. J. Bickel and E. L. Lehmann. Descriptive statistics for nonparametric models II: Location. *The Annals of Statistics*, 3:1045–1069, 1975.

P. J. Bickel and E. L. Lehmann. Descriptive statistics for nonparametric models III: Dispersion. *The Annals of Statistics*, 4:1139–1158, 1976.

P. J. Bickel and E. Levina. Some theory for Fisher's linear discriminant function, 'naive Bayes', and some alternatives when there are many more variables than observations. *Bernoulli*, 10:989–1010, 2004.

P. J. Bickel, G. Kur, and B. Nadler. Projection pursuit in high dimensions. *Proceedings of the National Academy of Sciences*, 115:9151–9156, 2018.

M. Bilodeau and D. Brenner. *Theory of Multivariate Statistics*. Springer, New York, 1999.

G. Blanchard, M. Sugiyama, M. Kawanabe, V. Spokoiny, and K.-R. Müller. Non-Gaussian component analysis: a semi-parametric framework for linear dimension reduction. In *Advances in Neural Information Processing Systems*, pages 131–138, 2005.

G. Blanchard, M. Kawanabe, M. Sugiyama, V. Spokoiny, and K.-R. Müller. In search of non-Gaussian components of a high-dimensional distribution. *Journal of Machine Learning Research*, 7:247–282, 2006.

R. J. Bolton and W. J. Krzanowski. Projection pursuit clustering for exploratory data analysis. *Journal of Computational and Graphical Statistics*, 12:121–142, 2003.

J.-F. Cardoso. Source separation using higher order moments. In *International Conference on Acoustics, Speech, and Signal Processing*, pages 2109–2112, 1989.

J.-F. Cardoso. Multidimensional independent component analysis. In *Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP '98*, volume 4, pages 1941–1944, 1998.

J.-F. Cardoso and A. Souloumiac. Blind beamforming for non-Gaussian signals. In *IEE Proceedings F (Radar and Signal Processing)*, volume 140, pages 362–370. IET, 1993.

P. Comon and C. Jutten. *Handbook of Blind Source Separation: Independent Component Analysis and Applications*. Academic Press, Amsterdam, 2010.

T. M. Cover and J. A. Thomas. *Elements of Information Theory*. Wiley-Interscience, New York, 2006.

P. L. Davies. Asymptotic behaviour of S-estimates of multivariate location parameters and dispersion matrices. *The Annals of Statistics*, 15:1269–1292, 1987.

A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B*, 39:1–38, 1977.

A. Dermoune and T. Wei. FastICA algorithm: five criteria for the optimal choice of the nonlinearity function. *IEEE Transactions on Signal Processing*, 61:2078–2087, 2013.

P. Diaconis and D. Freedman. Asymptotics of graphical projection pursuit. *The Annals of Statistics*, 12:793–815, 1984.

L. Dümbgen, K. Nordhausen, and H. Schuhmacher. New algorithms for M-estimation of multivariate scatter and location. *Journal of Multivariate Analysis*, 144:200–217, 2016.

L. Düembgen, M. Pauly, and T. Schweizer. A survey of M-functionals of multivariate location and scatter. *Statistics Surveys*, 9:32–105, 2013.

B. Everitt and D. J. Hand. *Finite Mixture Distributions*. Chapman and Hall, London, 1981.

J. Fan, Y. Feng, and X. Tong. A road to classification in high dimensional space: The regularized optimal affine discriminant. *Journal of the Royal Statistical Society, Series B.*, 74:745–771, 2012.

K. T. Fang and Z. T. Ting. *Generalized Multivariate Analysis*. Science Press, Beijing, 1990.

D. Fischer, A. Berro, K. Nordhausen, and A. Ruiz-Gazen. REPPlab: An R package for detecting clusters and outliers using exploratory projection pursuit. *Communications in Statistics-Simulation and Computation*, pages 1–23, 2019.

R. A. Fisher. The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7:179–188, 1936.

D. H. Foley and J. Sammon. An optimal set of discriminant vectors. *IEEE Transactions on Computers*, C-24:281–289, 1975.

G. Frahm, M. Junker, and A. Szimayer. Elliptical copulas: applicability and limitations. *Statistics & Probability Letters*, 63:275–286, 2003.

J. Friedman, T. Hastie, and R. Tibshirani. *The Elements of Statistical Learning*. Springer, New York, 2001.

J. H. Friedman. Exploratory projection pursuit. *Journal of the American Statistical Association*, 82:249–266, 1987.

A. K. Gupta and D. K. Nagar. *Matrix Variate Distributions*. Chapman and Hall/CRC, 1st edition, 1999.

T. Hastie, R. Tibshirani, J. Friedman, and J. Franklin. The elements of statistical learning: Data mining, inference, and prediction. *Mathematical Intelligencer*, 27:83–85, 2004.

H. V. Henderson and S. R. Searle. The vec-permutation matrix, the vec operator and Kronecker products: a review. *Linear and Multilinear Algebra*, 9:271–288, 1981.

W. Hu, W. Shen, H. Zhou, and D. Kong. Matrix linear discriminant analysis. *Technometrics*, 62:196–205, 2020.

G. Hua, P. A. Viola, and S. M. Drucker. Face recognition using discriminatively trained orthogonal rank one tensor projections. In *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE, 2007.

P. J. Huber. Robust estimation of a location parameter. *The Annals of Mathematical Statistics*, 35:73–101, 1964.

P. J. Huber. Projection pursuit. *The Annals of Statistics*, 13:435–475, 1985.

H. Hung, P. Wu, I. Tu, and S. Huang. On multilinear principal component analysis of order-two tensors. *Biometrika*, 99:569–583, 2012.

A. Hyvärinen. New approximations of differential entropy for independent component analysis and projection pursuit. In *Proceedings of the 10th International Conference on Neural Information Processing Systems*, page 273–279. MIT Press, 1997.

A. Hyvärinen. Gaussian moments for noisy independent component analysis. *IEEE Signal Processing Letters*, 6:145–147, 1999.

A. Hyvärinen and E. Oja. A fast fixed-point algorithm for independent component analysis. *Neural computation*, 9:1483–1492, 1997.

A. Hyvärinen, J. Karhunen, and E. Oja. *Independent Component Analysis*. John Wiley & Sons, 2001.

P. Ilmonen, H. Oja, and R. Serfling. On invariant coordinate system (ICS) functionals. *International Statistical Review*, 80:93–110, 2012.

C. M. Jarque and A. K. Bera. Efficient tests for normality, homoscedasticity and serial independence of regression residuals. *Economics Letters*, 6:255–259, 1980.

I. Jolliffe. *Principal Component Analysis*. Springer, New York, 2 edition, 2002.

M. C. Jones and R. Sibson. What is projection pursuit? *Journal of the Royal Statistical Society. Series A*, 150:1–18, 1987.

M. Kawanabe, M. Sugiyama, G. Blanchard, and K.-R. Müller. A new algorithm of non-Gaussian component analysis with radial kernel functions. *Annals of the Institute of Statistical Mathematics*, 59:57–75, 2007.

J. T. Kent and D. E. Tyler. Redescending M-estimates of multivariate location and scatter. *The Annals of Statistics*, 19:2102–2119, 1991.

S. Kotz and S. Nadarajah. *Multivariate t-Distributions and Their Applications*. Cambridge University Press, 2004.

E. Kristiansson. Decreasing rearrangement and Lorentz L(p,q) spaces. Master's thesis, Department of Mathematics of the Lulea University of Technology, 2002.

S. Kuriki and A. Takemura. The tube method for the moment index in projection pursuit. *Journal of Statistical Planning and Inference*, 138:2749–2762, 2008.

M. Li and B. Yuan. 2D-LDA: a statistical linear discriminant analysis for image matrix. *Pattern Recognition Letters*, 26:527–532, 2005.

M. Lindquist. The statistical analysis of fMRI data. *Statistical Science*, 23:439–464, 2008.

R. Linsker. Self-organization in a perceptual network. *IEEE Computer*, 21:105–117, 1988.

C. Liu, K. He, J.-l. Zhou, and C.-B. Gao. Discriminant orthogonal rank-one tensor projections for face recognition. In *Asian Conference on Intelligent Information and Database Systems*, pages 203–211. Springer, 2011.

N. Loperfido. Skewness and the linear discriminant function. *Statistics & Probability Letters*, 83:93–99, 2013.

N. Loperfido. Skewness-based projection pursuit: A computational approach. *Computational Statistics & Data Analysis*, 120:42–57, 2018.

H. P. Lopuhaä. Multivariate $\tau$-estimators for location and scatter. *The Canadian Journal of Statistics / La Revue Canadienne de Statistique*, 19:307–321, 1991.

H. Lu, K. N. Plataniotis, and A. N. Venetsanopoulos. MPCA: Multilinear principal component analysis of tensor objects. *IEEE Transactions on Neural Networks and Learning Systems*, 19:18–39, 2008.

W. Luo and B. Li. Combining eigenvalues and variation of eigenvectors for order determination. *Biometrika*, 103:875–887, 2016.

W. Luo and B. Li. On order determination by predictor augmentation. *Biometrika*, 108: 557–574, 2021.

S. G. Machado. Two statistics for testing for multivariate normality. *Biometrika*, 70:713–718, 1983.

Q. Mai, H. Zou, and M. Yuan. A direct approach to sparse discriminant analysis in ultra-high dimensions. *Biometrika*, 99:29–42, 2012.

J. F. Malkovich and A. A. Afifi. On tests for multivariate normality. *Journal of the American Statistical Association*, 68:176–179, 1973.

J. I. Marden. Some robust estimates of principal components. *Statistics & Probability Letters*, 43:349–359, 1999.

K. Mardia, J. Kent, and J. Bibby. *Multivariate Analysis*. Academic Press, 1995.

K. V. Mardia. Measures of multivariate skewness and kurtosis with applications. *Biometrika*, 57:519–530, 1970.

R. A. Maronna. Robust M-estimators of multivariate location and scatter. *The Annals of Statistics*, 4:51–67, 1976.

R. A. Maronna and V. J. Yohai. *Robust Estimation of Multivariate Location and Scatter*, pages 1–12. American Cancer Society, 2016.

A. W. Marshall, I. Olkin, and B. C. Arnold. *Inequalities: Theory of Majorization and Its Application*. Springer, New York, 2 edition, 2011.

G. J. McLachlan and D. Peel. *Finite Mixture Models*. John Wiley & Sons, New York, 2000.

J. Miettinen, S. Taskinen, K. Nordhausen, and H. Oja. Fourth moments and independent component analysis. *Statistical Science*, 30:372–390, 2015.

J. Miettinen, K. Nordhausen, S. Taskinen, and D. Tyler. On the computation of symmetrized M-estimators of scatter. In C. Agostinelli, A. Basu, P. Filzmoser, and D. Mukherjee, editors, *Recent Advances in Robust Statistics: Theory and Applications*, pages 151–167, New Delhi, 2016. Springer India.

J. Miettinen, K. Nordhausen, and S. Taskinen. Blind source separation based on joint diagonalization in R: The packages JADE and BSSasymp. *Journal of Statistical Software*, 76:1–31, 2017.

K. Nordhausen and H. Oja. Independent subspace analysis using three scatter matrices. *Austrian Journal of Statistics*, 40:93–101, 2016.

K. Nordhausen and H. Oja. Independent component analysis: A statistical perspective. *Wiley Interdisciplinary Reviews: Computational Statistics*, 10:e1440, 2018.

K. Nordhausen and D. E. Tyler. A cautionary note on robust covariance plug-in methods. *Biometrika*, 102:573–588, 2015.

K. Nordhausen, H. Oja, D. E. Tyler, and J. Virta. Asymptotic and bootstrap tests for the dimension of the non-Gaussian subspace. *IEEE Signal Processing Letters*, 24:887–891, 2017.

K. Nordhausen, H. Oja, and D. E. Tyler. Asymptotic and bootstrap tests for subspace dimension. *Journal of Multivariate Analysis*, online first:104830, 2021.

H. Oja. On location, scale, skewness and kurtosis of univariate distributions. *Scandinavian Journal of Statistics*, 8:154–168, 1981.

H. Oja. *Multivariate Nonparametric Methods with R: An approach based on spatial signs and ranks*. Springer, New York, 2010.

H. Oja, S. Sirkiä, and J. Eriksson. Scatter matrices and independent component analysis. *Austrian Journal of Statistics*, 35:175–189, 2006.

E. Ollila. The deflation-based FastICA estimator: Statistical analysis revisited. *IEEE transactions on Signal Processing*, 58:1527–1541, 2009.

J. Owen and R. Rabinovitch. On the class of elliptical distributions and their applications to the theory of portfolio choice. *The Journal of Finance*, 38:745–752, 1983.

J. E. Pecaric, F. Proschan, and J. L. Tong. *Convex Functions, Partial Orderings, and Statistical Applications*. Academic Press, 1992.

D. Peña and F. J. Prieto. Cluster identification using projections. *Journal of the American Statistical Association*, 96:1433–1445, 2001.

A. Pires and J. Branco. High dimensionality: The latest challenge to data analysis. *arXiv preprint arXiv:1902.04679*, 2019.

C. Posse. Projection pursuit exploratory data analysis. *Computational Statistics & Data Analysis*, 20:669–687, 1995.

S. T. Rachev. *Mass Transportation Problems*. Springer, New York, 1 edition, 1998.

C. R. Rao. The utilization of multiple measurements in problems of biological classification. *Journal of the Royal Statistical Society. Series B*, 10:159–203, 1948.

B. B. Risk, D. S. Matteson, and D. Ruppert. Linear non-Gaussian component analysis via maximum likelihood. *Journal of the American Statistical Association*, 114:332–343, 2019.

J. V. Ryff. On the representation of doubly stochastic operators. *Pacific Journal of Mathematics*, 13:1379–1386, 1963.

H. Sasaki, G. Niu, and M. Sugiyama. Non-Gaussian component analysis with log-density gradient estimation. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, pages 1177–1185, 2016.

J. R. Schott. A high-dimensional test for the equality of the smallest eigenvalues of a covariance matrix. *Journal of Multivariate Analysis*, 97:827–843, 2006.

L. Scrucca, M. Fop, T. B. Murphy, and A. E. Raftery. mclust 5: clustering, classification and density estimation using Gaussian finite mixture models. *The R Journal*, 8:289–317, 2016.

J. Shao, Y. Wang, X. Deng, and S. Wang. Sparse linear discriminant analysis by thresholding for high dimensional data. *The Annals of Statistics*, 39:1241 – 1265, 2011.

B. N. Sheehan and Y. Saad. Higher order orthogonal iteration of tensors (HOOI) and its relation to PCA and GLRAM. In *Proceedings of the 2007 SIAM International Conference on Data Mining (SDM)*, pages 355–365, 2007.

S. Sirkiä, S. Taskinen, and H. Oja. Symmetrised M-estimators of multivariate scatter. *Journal of Multivariate Analysis*, 98:1611–1629, 2007.

M. S. Srivastava, T. von Rosen, and D. Von Rosen. Models with a Kronecker product covariance structure: estimation and testing. *Mathematical Methods of Statistics*, 17: 357–370, 2008.

R. G. Staudte. The shapes of things to come: probability density quantiles. *Statistics*, 51: 782–800, 2017.

F. J. Theis. Towards a general independent subspace analysis. In B. Schölkopf, J. C. Platt, and T. Hoffman, editors, *Advances in Neural Information Processing Systems 19*, pages 1361–1368. MIT Press, Cambridge, MA, 2007.

F. J. Theis, M. Kawanabe, and K. R. Müller. Uniqueness of non-Gaussianity-based dimension reduction. *IEEE Transactions on Signal Processing*, 59:4478–4482, 2011.

R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B*, 58:267–288, 1996.

M. E. Tipping and C. M. Bishop. Probabilistic principal component analysis. *Journal of the Royal Statistical Society. Series B*, 61:611–622, 1999.

I.-P. Tu, S.-Y. Huang, and D.-N. Hsieh. The generalized degrees of freedom of multilinear principal component analysis. *Journal of Multivariate Analysis*, 173:26–37, 2019.

M. A. Turk and A. P. Pentland. Face recognition using eigenfaces. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 586–587, 1991.

D. E. Tyler, F. Critchley, L. Dümbgen, and H. Oja. Invariant co-ordinate selection. *Journal of the Royal Statistical Society. Series B*, 71:549–592, 2009.

C. Viroli. Finite mixtures of matrix normal distributions for classifying three-way data. *Statistics and Computing*, 21:511–522, 2011.

J. Virta. *Independent component analysis for non-standard data structures*. PhD thesis, University of Turku – Annales Universitatis Turkuensis, 2018.

J. Virta, K. Nordhausen, and H. Oja. Projection pursuit for non-Gaussian independent components. *arXiv preprint arXiv:1612.05445*, 2016.

J. Virta, B. Li, K. Nordhausen, and H. Oja. Independent component analysis for tensor-valued data. *Journal of Multivariate Analysis*, 162:172–192, 2017.

J. Virta, B. Li, K. Nordhausen, and H. Oja. JADE for tensor-valued observations. *Journal of Computational and Graphical Statistics*, 27:628–637, 2018.

J. Virta, C. L. Koesner, K. Nordhausen, H. Oja, B. Li, and U. Radojičić. *tensorBSS: Blind Source Separation Methods for Tensor-Valued Observations*, 2021a. URL `https://CRAN.R-project.org/package=tensorBSS`. R package version 0.3.8.

J. Virta, N. Lietzén, P. Ilmonen, and K. Nordhausen. Fast tensorial JADE. *Scandinavian Journal of Statistics*, 48:164–187, 2021b.

S. Visuri, V. Koivunen, and H. Oja. Sign and rank covariance matrices. *Journal of Statistical Planning and Inference*, 91:557–575, 2000.

U. Von Luxburg. A tutorial on spectral clustering. *Statistics and computing*, 17:395–416, 2007.

S. Wu, W. Li, Z. Wei, and J. Yang. Local discriminative orthogonal rank-one tensor projection for image feature extraction. In *The First Asian Conference on Pattern Recognition*, pages 367–371. IEEE, 2011a.

X. Wu, J. Lai, and X. Chen. Rank-1 tensor projection via regularized regression for action classification. *International Journal of Wavelets, Multiresolution and Information Processing*, 9:1025–1041, 2011b.

J. Yang, D. Zhang, A. Frangi, and J.-Y. Yang. Two-dimensional PCA: A new approach to appearance-based face representation and recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26:131–137, 2004.

Z. Ye and R. E. Weiss. Using the bootstrap to select one of a new class of dimension reduction methods. *Journal of the American Statistical Association*, 98:968–979, 2003.

D. Zhang and Z.-H. Zhou. $(2D)^2$PCA: Two-directional two-dimensional PCA for efficient face representation and recognition. *Neurocomputing*, 69:224–231, 2005.

W. Zhao, R. Chellappa, P. Phillips, and A. Rosenfeld. Face recognition: A literature survey. *ACM Computing Surveys*, 35:399–458, 2003.

W.-S. Zheng, J. H. Lai, and S. Z. Li. 1D-LDA vs. 2D-LDA: When is vector-based linear discriminant analysis better than matrix-based? *Pattern Recognition*, 41:2156–2172, 2008.

W. Zhong, X. Xing, and K. Suslick. Tensor sufficient dimension reduction. *WIREs Computational Statistics*, 7:178–184, 2015.

W. R. v. Zwet. *Convex transformations of random variables*, volume 10. Mathematisch Centrum, Amsterdam, 1964.

# Curriculum vitae

## Personal data


| | |
|---|---|
| Name | **Una Radojičić** |
| Contact address | CSTAT - Computational Statistics<br>Institute of Statistics & Mathematical Methods in Economics<br>Vienna University of Technology<br>Wiedner Hauptstrasse 8-10<br>A-1040 Vienna |
| Phone | +43 1 58801 10566 |
| E-Mail | una.radojicic@tuwien.ac.at |


---

## Work experience

| | |
|---|---|
| 10/2019 - present | University Assistant at the CSTAT group, Institute of Statistics and Mathematical Methods in Economics, Vienna University of Technology, Austria |
| 10/2017 - 09/2019 | University Assistant at the Department of Mathematics, Josip Juraj Strossmayer University of Osijek, Croatia |

---

## Education

| | |
|---|---|
| 10/2019 - present | **Doctoral program** in Natural Sciences Technical Mathematics, Vienna University of Technology, Austria |
| 10/2015 - 10/2017 | **Mag.Math.**, Master program in mathematics - Financial Mathematics and Statistics, Josip Juraj Strossmayer University of Osijek, Croatia<br>*Summa cum laude* |
| 10/2012 - 10/2015 | **BSc**, Bachelor program in mathematics, Josip Juraj Strossmayer University of Osijek, Croatia<br>*Summa cum laude* |

---

## List of presentations

**Talk**, Dimension estimation in two-dimensional PCA (12th Int'l Symposium on Image and Signal Processing and Analysis, Zagreb, Croatia (ISPA 2021))

**Talk**, Large-sample properties of blind estimation of the linear discriminant using projection pursuit (Data Science, Statistics & Visualisation (DSSV 2021))

**Invited talk**, Non-Gaussian component analysis: testing the dimension of the signal subspace (13th International Conference of the ERCIM WG on Computational and Methodological Statistics (CMStatistics 2020))

**Invited talk**, Notion of information and independent component analysis (Statistical Seminar of the Department of Mathematics, Josip Juraj Strossmayer University of Osijek, Croatia)

**Poster**, Notion of information and independent component analysis (Data Science, Statistics, and Visualisation (DSSV 2020))

**Invited talk**, Algorithms for initialization of Gaussian mixture models (24th Young Statisticians Meetings, Basovizza, Italy 2019 (YSM 2019))

**Invited talk**, Algorithms for initialization of Gaussian mixture models (21th European Young Statisticians Meetings, Belgrade, Serbia 2019 (EYSM 2019))

**Talk**, Fast and efficient method for solving multiple closed curve detection problem (International Conference on Pattern Recognition Applications and Methods, Prague, Czech Republic (ICPRAM 2019))

**Talk**, Application of the adaptive annealing method to the generalized incremental algorithm (International statistical conference in Croatia, Opatija, Croatia 2018 (ISCCRO'18))

**Talk**, Mathematical model of AIDS disease (Regional Primatijada, Čanj, Montenegro, 2017)

**Talk**, Motivational problems and elementariness of calculus of variations (Primatijada, Poreč, Croatia, 2015)

---

## List of software

J. Virta, C. Koesner, B. Li, K. Nordhausen, H. Oja and **U. Radojicic**. tensorBSS: Blind source separation methods for tensor-valued observations. R package version 0.3.8. `https://CRAN.R-project.org/package=tensorBSS`. 2021.

K. Riemer, G. Frahm, K. Nordhausen and **U. Radojicic**. shapeNA: M-estimation of shape for data with missing values. R package version 0.0.2. `https://CRAN.R-project.org/package=shapeNA`. 2021.

---

## List of publications

**U. Radojicic**, K. Nordhausen and J. Virta. Kurtosis-based projection pursuit for matrix-valued data. *arXiv preprint arXiv:2109.04167*, 2021.

**U. Radojicic**, K. Nordhausen and J. Virta. Large-sample properties of blind estimation of the linear discriminant using projection pursuit. *arXiv preprint arXiv:2103.04678*, 2021.

**U. Radojicic**, N. Lietzen, K. Nordhausen and J. Virta. Dimension estimation in two-dimensional PCA. In *Proceedings of the 12th International Symposium on Image and Signal Processing and Analysis*, pages 16-22, 2021.

K. Nordhausen, **U. Radojicic**. Least absolute value. To appear in Encyclopedia of Mathematical Geosciences. Encyclopedia of Earth Sciences Series. Springer, Cham.

**U. Radojicic** and K. Nordhausen. Non-Gaussian component analysis: testing the dimension of the non-Gaussian subspace. In *M. Maciak, M. Pestas and M. Schindler, editors, Analytical Methods in Statistics, AMISTAT 2019*, pages 101-123, Springer Cham, 2019.

**U. Radojicic**, K. Nordhausen and H. Oja. Notion of information and independent component analysis. *Applications of Mathematics*, 65:311-330, 2020.

**U. Radojicic**, R. Scitovski and K. Sabo. A fast and efficient method for solving the multiple closed curve detection problem. In *Proceedings of the 8th International Conference on Pattern Recognition Applications and Methods*, pages 269-276. 2019.

**U. Radojicic**, K. Sabo and R. Scitovski. A fast and efficient method for solving the multiple line detection problem. *Rad Hrvatske akademije znanosti i umjetnosti*, 23:123-140, 2019.

K Burazin, **U. Radojicic**. Introduction to the calculus of variations and its history. *Osječki matematički list*, 16:111-133, 2016.

Vienna, September 14, 2021

Una Radojičić

# Part II.

# Publications

# Non-Gaussian component analysis: testing the dimension of the signal subspace

## Summary

Publication I considers the simultaneous use of two scatter functionals in the context of the dimension reduction in the non-Gaussian component model, where the aim is to divide the data into a non-Gaussian part, the signal, and a Gaussian part, the noise. For this purpose bootstrap test strategies are introduced to test the dimension of the non-Gaussian subspace, and an extensive simulation study was conducted to estimate the power of the test for various combinations of scatters. Sequential application of the test can then be used to estimate the signal dimension.

## Bibliographic information

U. Radojicic and K. Nordhausen. Non-Gaussian component analysis: testing the dimension of the non-Gaussian subspace. In *M. Maciak, M. Pestas and M. Schindler, editors, Analytical Methods in Statistics, AMISTAT 2019*, pages 101-123, Springer Cham, 2019.

## Author's contribution

U. Radojicic participated in several discussions with the coauthor to develop the idea and the methodology. Furthermore, U. Radojicic conducted the simulation study and implemented the R code for the examples. U. Radojicic contributed in deriving main results as well as the proofs, and also contributed in overall writing and editing of the paper as well as the review of the paper based on joint discussion with coauthor and the suggestions of the reviewers.

# Large-sample properties of blind estimation of the linear discriminant using projection pursuit



## Summary

Publication II considers the estimation of the linear discriminant with projection pursuit, in an unsupervised manner. The viewpoint we take is asymptotic and, as our main contribution, we derive central limit theorems for estimators based on three different projection indices, skewness, kurtosis and their convex combination. The results show that in each case the limiting covariance matrix is proportional to that of LDA, a supervised estimator of the discriminant. An extensive comparative study between the asymptotic variances reveals that projection pursuit is able to achieve efficiency equal to LDA when the groups are arbitrarily well-separated and their sizes are reasonably balanced. Simulations reveal very good performance of the methods, while also confirming the validity of the obtained asymptotic results.

## Bibliographic information

U. Radojicic, K. Nordhausen and J. Virta. Large-sample properties of blind estimation of the linear discriminant using projection pursuit. *arXiv preprint arXiv:2103.04678*, 2021.

## Author's contribution

U. Radojicic participated in several discussions with the coauthors to develop the idea and the methodology. Furthermore, U. Radojicic carried out the simulation study and contributed in deriving the proofs of the presented results. U. Radojicic also contributed in overall writing and editing of the paper.

# Dimension estimation in two-dimensional PCA

## Summary

Publication III considers the estimation of the optimal number of low-rank components in the dimension reduction of image data. For that purpose we develop an automated method based on the combination of two-dimensional principal component analysis and an augmentation estimator proposed in Luo and Li (2021) for vector-valued observations. Simply stated, the method combines a scree plot with information extracted from the eigenvectors of the covariance matrix. Simulation studies show that the method performs well and gives the accurate estimates of the latent dimensions, while the real data example showcasts good performance in practice.

## Bibliographic information

U. Radojicic, N. Lietzen, K. Nordhausen and J. Virta. Dimension estimation in two-dimensional PCA. In *Proceedings of the 12th International Symposium on Image and Signal Processing and Analysis*, pages 16-22, 2021.

## Author's contribution

U. Radojicic participated in several discussions with the coauthors to develop the idea and the methodology. Furthermore, U. Radojicic contributed in deriving the results presented in the paper, as well as in implementing the method to tensorBSS package. U. Radojicic wrote the first version of the draft and reviewed the manuscript based on joint discussions with the coauthors and suggestions of the reviewers.

# Kurtosis-based projection pursuit for matrix-valued data

## Summary

Publication IV considers projection pursuit for data that admit a natural representation in matrix form. The projection indices we propose are extensions of the classical kurtosis and Mardia's multivariate kurtosis. The first index estimates projections for both sides of the matrices simultaneously, while the second one finds the two projections separately. Both indices are shown to recover the optimally separating projection for two-group Gaussian mixtures in the full absence of any label information. We further establish the strong consistency of the corresponding sample estimators. Simulations and a real data example on hand-written postal code data demonstrate good performance of the proposed method in the simulated as well the real data.

## Bibliographic information

U. Radojicic, K. Nordhausen and J. Virta. Kurtosis-based projection pursuit for matrix-valued data. *arXiv preprint arXiv:2109.04167*, 2021.

## Author's contribution

U. Radojicic participated in several discussions with the coauthors to develop the idea and the methodology. Furthermore, U. Radojicic contributed in deriving the asymptotical results and those considering second-order methods, as well as the corresponding proofs. U. Radojicic implemented the methods in R and carried out the simulations and the real data example, while also contributing in overall writing and editing of the paper.

# Notion of information and independent component analysis

## Summary

Publication V considers partial orderings and measures of information for continuous univariate random variables, while discussing the special roles of the Gaussian and uniform distribution. As discussed in Huber (1985), the information measures and measures of non-Gaussianity including the third and fourth cumulants are generally used as projection indices in the projection pursuit approach. We derive a result which justifies their use in the independent component analysis. Furthermore, we discuss in detail the connections between information, non-Gaussianity and statistical independence in the context of independent component analysis.

## Bibliographic information

U. Radojicic, K. Nordhausen and H. Oja. Notion of information and independent component analysis. *Applications of Mathematics*, 65:311-330, 2020.

## Author's contribution

U. Radojicic participated in several discussions with the coauthors to develop the idea and the methodology. Furthermore, U. Radojicic contributed in deriving the information orders for continuous distributions as well as the examples and figures. U. Radojcic contributed in overall writing and editing of the paper as well as the review of the paper based on joint discussion with coauthors and the suggestions of the reviewers.