WORKING PAPERS

N° 1268

November 2021

"On the usage of joint diagonalization in multivariate statistics"

Klaus Nordhausen and Anne Ruiz-Gazen

Toulouse
School of
Economics

On the usage of joint diagonalization in multivariate statistics

Klaus Nordhausen[1] and Anne Ruiz-Gazen[2]
Department of Mathematics and Statistics, University of Jyväskylä, Finland
Toulouse School of Economics, Université de Toulouse Capitole, France

November 23, 2021

Abstract

Scatter matrices generalize the covariance matrix and are useful in many multivariate data analysis methods, including well-known principal component analysis (PCA), which is based on the diagonalization of the covariance matrix. The simultaneous diagonalization of two or more scatter matrices goes beyond PCA and is used more and more often. In this paper, we offer an overview of many methods that are based on a joint diagonalization. These methods range from the unsupervised context with invariant coordinate selection and blind source separation, which includes independent component analysis, to the supervised context with discriminant analysis and sliced inverse regression. They also encompass methods that handle dependent data such as time series or spatial data.

Keywords: Blind Source Separation, Dimension Reduction, Independent Component Analysis, Invariant Component Selection, Scatter Matrices, Supervised Dimension Reduction.

1 Introduction

Classical multivariate analysis, such as that presented in [Anderson, 2003] assumes that the data at hand follow a multivariate normal model. This is very convenient as the multivariate normal distribution is fully specified by its mean vector and covariance matrix and these two statistics suffice to develop tractable and optimal inference tools for this model. Early on, it was known that statistical methods based on the mean vector and covariance matrix are very sensitive to atypical observations in the data and are not very efficient for observations coming from a heavy-tailed distribution. To alleviate these problems, the normal model is commonly broadened to the elliptical model, which keeps the shape of the probability contours but allows for one additional kurtosis parameter, thus allowing heavier and lighter tails than the normal model. For robustness and optimality reasons, alternative location measures for the mean vector and alternative dispersion measures for the covariance matrix were developed in the elliptical framework. These measures are often expected to have certain properties under affine transformations of the data, in which case they are called location functionals $\boldsymbol{T}$ and scatter functionals $\boldsymbol{S}$. It can then be shown that in an elliptical model, all location functionals, including the mean vector, correspond to the symmetry center and that all scatter functionals are proportional to the covariance matrix, if they exist Oja [2010], Nordhausen and Tyler [2015]. Thus, scatter functionals measure the same population quantity in the elliptical model, and it is sufficient to use one location functional and one scatter functional for inference purposes (A slightly larger model where location and scatter functionals measure the same population quantities is also discussed in Oja [2010], Nordhausen and Tyler [2015].) Since approximately the turn of the last century, interest in the simultaneous use of two or more scatter matrices, which is of course most interesting when these functionals do not measure the same population quantities, has increased.

In the present paper we will show how two or more scatter functionals are jointly used in multivariate statistics and in which models this is of interest. For this purpose we recall first in Section 2 the concept of scatter functionals in detail and discuss some of their properties. Section 3 gives details on the simultaneous and joint diagonalization of scatter functionals which is the main tool used in our context. Invariant coordinate selection (ICS) is discussed in Section 4, blind source separation (BSS) is discussed in Section 5 and the use of joint diagonalization in the context of supervised dimension reduction (SDR) methods is discussed in Section 6. Finally, the paper is concluded in Section 7.

2 Scatter matrices

Joint diagonalization has been used in unsupervised and supervised contexts and for independent and dependent data.

In the unsupervised case with independent data, the definition of a scatter matrix, sometimes also called pseudo-covariance, is a generalization of the covariance matrix definition [see Huber and Ronchetti, 2011, Maronna and Yohai, 2016, Croux and Haesbroeck, 2000, Tyler et al., 2009, Nordhausen and Tyler, 2015, among others].

Following Nordhausen and Tyler [2015], let us first define the functional version of a scatter estimator. For a p-dimensional vector $\boldsymbol{X}$ with distribution function $F_{\boldsymbol{X}}$, a functional $\boldsymbol{S}(F_{\boldsymbol{X}})$ also denoted by $\boldsymbol{S}(\boldsymbol{X})$ is called a scatter functional if it is a $p \times p$ symmetric positive semidefinite and affine equivariant matrix. Note that in Tyler et al. [2009], the definition is more stringent than that in Nordhausen and Tyler [2015], and assumes that a scatter matrix is positive definite. We recall that an affine equivariant matrix $\boldsymbol{S}(\boldsymbol{X})$ is such that

$$\boldsymbol{S}(\boldsymbol{A}\boldsymbol{X}+\boldsymbol{b}) = \boldsymbol{A}\boldsymbol{S}(\boldsymbol{X})\boldsymbol{A}^\top,$$

where $^\top$ denotes the transpose operator, $\boldsymbol{A}$ is a full rank $p \times p$ matrix and $\boldsymbol{b}$ a p-vector.

For distributions $F_{\boldsymbol{X}}$ with finite second moments, the covariance functional is defined by:

$$\operatorname{Cov}(\boldsymbol{X}) = E\left[(\boldsymbol{X}-E(\boldsymbol{X}))(\boldsymbol{X}-E(\boldsymbol{X}))^\top\right]$$

and is affine equivariant.

Let us now consider the empirical version of a scatter functional. This means that we have a p-variate dataset $\boldsymbol{X}_n = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)^\top$ and the scatter functional $\boldsymbol{S}(F_n)$ for the empirical distribution F_n. A scatter matrix statistic or estimator is thus a $p \times p$ symmetric positive semidefinite and affine equivariant matrix. In this framework, an affine equivariant matrix $\boldsymbol{S}(\boldsymbol{X}_n)$ is such that

$$\boldsymbol{S}(\boldsymbol{X}_n\boldsymbol{A}+\boldsymbol{1}_n\boldsymbol{b}^\top) = \boldsymbol{A}^\top\boldsymbol{S}(\boldsymbol{X}_n)\boldsymbol{A},$$

where $\boldsymbol{A}$ is a full rank $p \times p$ matrix, $\boldsymbol{b}$ a p-vector and $\boldsymbol{1}_n$ an n-vector full of ones.

The empirical covariance matrix is defined by:

$$\operatorname{Cov}(\boldsymbol{X}_n) = \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{x}_i - \bar{\boldsymbol{x}}_n)(\boldsymbol{x}_i - \bar{\boldsymbol{x}}_n)^\top$$

where $\bar{\boldsymbol{x}}_n = 1/n \sum_{i=1}^n \boldsymbol{x}_i$ is the empirical mean. The mean is an affine equivariant location estimator $\boldsymbol{T}$ such that:

$$\boldsymbol{T}(\boldsymbol{A}\boldsymbol{X} + \boldsymbol{b}) = \boldsymbol{A}\boldsymbol{T}(\boldsymbol{X}) + \boldsymbol{b},$$

for the functional version and

$$\boldsymbol{T}(\boldsymbol{X}_n\boldsymbol{A}^\top + \mathbf{1}_n\boldsymbol{b}^\top) = \boldsymbol{A}\boldsymbol{T}(\boldsymbol{X}_n) + \boldsymbol{b},$$

for the empirical version where $\boldsymbol{A}$ is a full rank $p \times p$ matrix and $\boldsymbol{b}$ a p-vector.

For elliptical distributions with second moments, scatter functionals are proportional to the covariance matrix [see, e.g., Bilodeau and Brenner, 2008].

Many scatter matrices have been defined with the objective of making the covariance matrix estimator more robust [see, e.g., Huber and Ronchetti, 2011, Maronna et al., 2019]. Tyler et al. [2009] divide the scatter matrices in three classes depending on their robustness properties. The first class includes scatter estimators with a zero breakdown point such as the usual covariance matrix but also the one-step M-estimators with a functional defined by:

$$\mathrm{Cov}_w(\boldsymbol{X}) = E\left[w(D^2(\boldsymbol{X}))(\boldsymbol{X} - E(\boldsymbol{X}))(\boldsymbol{X} - E(\boldsymbol{X}))^\top\right],$$

where w is a non-negative and continuous weight function and $D^2(\boldsymbol{X}) = (\boldsymbol{X} - E(\boldsymbol{X}))^\top \mathrm{Cov}(\boldsymbol{X})^{-1}(\boldsymbol{X} - E(\boldsymbol{X}))$ is the Mahalanobis distance. The sample version of the one-step M-estimator is:

$$\mathrm{Cov}_w(\boldsymbol{X}_n) = \frac{1}{n}\sum_{i=1}^n w(D^2(\boldsymbol{x}_i))(\boldsymbol{x}_i - \bar{\boldsymbol{x}}_n)(\boldsymbol{x}_i - \bar{\boldsymbol{x}}_n)^\top,$$

where $D^2(\boldsymbol{x}_i) = (\boldsymbol{x}_i - \bar{\boldsymbol{x}}_n)^\top \mathrm{Cov}(\boldsymbol{X}_n)^{-1}(\boldsymbol{x}_i - \bar{\boldsymbol{x}}_n)$.

The covariance matrix is obtained with $w(d) = 1$ while we get the Cov_{-1} matrix defined by Critchley et al. [2006] when $w(d) = 1/d$. As noticed by Nordhausen [2014], when $w(d) = d^\alpha$ with $\alpha < 0$, such estimators down-weight values with large Mahalanobis distance and so have a robust flavour even if they have a zero breakdown point. From the same class, the fourth-moment based estimator Cov_4 obtained with $w(d) = d$ is widely used in the blind source separation literature [see, e.g., Theis and Inouye, 2006, Nordhausen and Virta, 2019]. It is highly nonrobust since it up-weights values with large Mahalanobis distances but it proves to be useful in particular situations.

The second class of estimators contains scatter matrices with a moderate breakdown point such as the class $(\boldsymbol{T}, \boldsymbol{S})$ of M-estimators that are defined [see, e.g., Maronna, 1976] as solutions of systems of equations of the following form:

$$\begin{aligned} E\left[u_1\left[(\boldsymbol{X} - \boldsymbol{T}(\boldsymbol{X}))^\top \boldsymbol{S}(\boldsymbol{X})^{-1}(\boldsymbol{X} - \boldsymbol{T}(\boldsymbol{X}))\right](\boldsymbol{X} - \boldsymbol{T}(\boldsymbol{X}))\right] &= \mathbf{0} && (1) \\ E\left[u_2\left[(\boldsymbol{X} - \boldsymbol{T}(\boldsymbol{X}))^\top \boldsymbol{S}(\boldsymbol{X})^{-1}(\boldsymbol{X} - \boldsymbol{T}(\boldsymbol{X}))\right](\boldsymbol{X} - \boldsymbol{T}(\boldsymbol{X}))(\boldsymbol{X} - \boldsymbol{T}(\boldsymbol{X}))^\top\right] &= \boldsymbol{S}(\boldsymbol{X}) && (2) \end{aligned}$$

where u_1 and u_2 are appropriate weight functions. The sampling version of M-estimators is easily derived from the previous equations.

The third class contains high breakdown point estimators such as S-estimators or minimum covariance determinant estimators (see Maronna et al. [2019] for details and other scatter estimators from the same class).

In some statistical methods, such as independent component analysis, a scatter matrix is also expected to verify the joint independence property or the block independence property. A scatter functional $\boldsymbol{S}(\boldsymbol{X})$ has the joint independence property if for a vector with mutually independent components $\boldsymbol{S}(\boldsymbol{X})$ is diagonal. In the class of one-step M-estimators with nonnegative and continuous weight function, Virta [2016] proves that the only scatter functionals with the independence property are the Cov and Cov_4 estimators and their nonnegative linear combinations. In the case where all components of $\boldsymbol{X}$ are not necessarily independent but consist of independent blocks of components, the block dependence property states that the mutually independent subvectors of $\boldsymbol{X}$ correspond to diagonal submatrices leading to a block diagonal scatter matrix (see Nordhausen and Tyler [2015], Tyler et al. [2009] for more details).

Note that it is also possible to define some symmetrized version of the previous scatter matrices by considering $\boldsymbol{S}(\boldsymbol{U}-\boldsymbol{V})$, where $\boldsymbol{U}$ and $\boldsymbol{V}$ are independent copies of $\boldsymbol{X}$ [see Tyler et al., 2009, Nordhausen and Tyler, 2015]. In other words, the symmetrisation is obtained by applying the scatter functional to pairwise differences. The symmetrized scatter matrices possess the joint and block independence property [see Oja et al., 2006, Nordhausen and Tyler, 2015].

In Liski et al. [2014], the definition of a scatter matrix is extended to the context of supervised methods. In this context, in addition to the p-vector $\boldsymbol{X}$, a response variable Y is available. Following Liski et al. [2014], a supervised scatter functional $\boldsymbol{S}$ is a function of the joint distribution $F_{\boldsymbol{X},Y}$ of $(\boldsymbol{X},Y)$ which is affine equivariant in the sense that

$$\boldsymbol{S}(F_{\boldsymbol{AX}+\boldsymbol{b},Y}) = \boldsymbol{A}\boldsymbol{S}(F_{\boldsymbol{X},Y})\boldsymbol{A}^\top,$$

for all full rank matrices $\boldsymbol{A}$ and all p-vectors $\boldsymbol{b}$. One example of such a supervised scatter functional is:

$$\boldsymbol{S}_{SIR}(F_{\boldsymbol{X},Y}) = \mathrm{Cov}(E(\boldsymbol{X}|Y)).$$

Note that in the case of a discrete response variable, $\boldsymbol{S}_{SIR}$ corresponds to the between covariance matrix $\mathrm{Cov}_{\boldsymbol{B}}$.

Thus far, we have focused on samples of independent data but as will be detailed below, joint diagonalization is also widely used in the context of time series and spatial data. In such contexts, affine equivariant estimators that are not necessarily positive semidefinite are considered and go beyond the scatter matrix definition above.

In the context of p-variate time series, let us consider a stochastic process $\boldsymbol{X}_T = (\boldsymbol{x}_1,\ldots,\boldsymbol{x}_T)$ measured at time $t \in \{1,\ldots,T\}$. For a given lag $\tau \in \{0,1,\ldots\}$, the sample version of the cross-autocovariance matrix $\mathrm{ACov}_\tau(\boldsymbol{X}_T)$ is given by

$$\mathrm{ACov}_\tau(\boldsymbol{X}_T) = \frac{1}{T-\tau}\sum_{t=1}^{T-\tau}(\boldsymbol{x}_t - \bar{\boldsymbol{x}}_T)(\boldsymbol{x}_{t+\tau} - \bar{\boldsymbol{x}}_T)^\top$$

where $\bar{\boldsymbol{x}}_T = 1/T\sum_{t=1}^T \boldsymbol{x}_t$. Note that $\mathrm{ACov}_0(\boldsymbol{X}_T) = \mathrm{Cov}(\boldsymbol{X}_T)$.

The ACov_τ matrix is not necessarily symmetric and is sometimes symmetrized when it is expected to be symmetric for the model under consideration [see Tong et al., 1990, Miettinen et al., 2012]. A symmetrized version of ACov_τ is defined by

$$\mathrm{ACov}_\tau^S = \frac{1}{2}(\mathrm{ACov}_\tau + \mathrm{ACov}_\tau^\top).$$

Let us now consider multivariate data measured at spatial locations $\boldsymbol{s}_1, \dots, \boldsymbol{s}_n$ in a domain $\mathcal{S} \subseteq \mathbb{R}^d$. $\boldsymbol{X}_n = (\boldsymbol{x}(\boldsymbol{s}_1), \dots, \boldsymbol{x}(\boldsymbol{s}_n))$, Bachoc et al. [2020] define local covariance, or scatter, matrices, by:

$$\mathrm{LCov}_f(\boldsymbol{X}_n) = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{n} f(\boldsymbol{s}_i - \boldsymbol{s}_j)(x(\boldsymbol{s}_i) - \bar{\boldsymbol{x}}_n)(\boldsymbol{x}(\boldsymbol{s}_j) - \bar{\boldsymbol{x}}_n)^\top, \tag{3}$$

where $\bar{\boldsymbol{x}}_n = 1/n \sum_{i=1}^n \boldsymbol{x}(\boldsymbol{s}_i)$ and $f : \mathbb{R}^d \to \mathbb{R}$ is called the kernel function. Examples of kernels f are the ball and ring kernels $B(h)(\boldsymbol{s}) = I(\|\boldsymbol{s}\| \leq h)$ with fixed $h \geq 0$ and $R(h_1, h_2)(\boldsymbol{s}) = I(h_1 \leq \|\boldsymbol{s}\| \leq h_2)$ with fixed $h_2 \geq h_1 \geq 0$ where $\|.\|$ denotes the euclidian norm and $I(\cdot)$ denotes the indicator function.

Note that the ACov and LCov are considered as scatter matrices but may not be semipositive definite.

3 Simultaneous and joint diagonalization

In PCA [see, e.g., Jolliffe, 2002], the covariance $\mathrm{Cov}(\boldsymbol{X}_n)$ (or correlation) matrix, which is a symmetric real valued matrix, is diagonalized. It means that the following transformation is calculated:

$$\boldsymbol{U}(\boldsymbol{X}_n)\,\mathrm{Cov}(\boldsymbol{X}_n)\boldsymbol{U}(\boldsymbol{X}_n)^\top = \boldsymbol{\Lambda}(\boldsymbol{X}_n)$$

where $\boldsymbol{\Lambda}(\boldsymbol{X}_n) = \mathrm{diag}(\lambda_1 \geq \dots \geq \lambda_p)$ is the diagonal matrix containing the ordered eigenvalues of $\mathrm{Cov}(\boldsymbol{X}_n)$ and $\boldsymbol{U}(\boldsymbol{X}_n) = (\boldsymbol{u}_1, \dots, \boldsymbol{u}_p)^\top$ contains its corresponding orthonormal eigenvectors as rows. Because the matrix $\mathrm{Cov}(\boldsymbol{X}_n)$ is symmetric, the matrix $\boldsymbol{U}(\boldsymbol{X}_n)$ is orthogonal with respect to the usual inner-product and can be chosen such that $\boldsymbol{U}(\boldsymbol{X}_n)\boldsymbol{U}(\boldsymbol{X}_n)^\top = \boldsymbol{U}(\boldsymbol{X}_n)^\top\boldsymbol{U}(\boldsymbol{X}_n) = \boldsymbol{I}_p$. This procedure is also called the spectral or the eigenvalue-eigenvector decomposition or eigendecomposition of $\mathrm{Cov}(\boldsymbol{X}_n)$.

When considering two scatter matrices $\boldsymbol{S}_1(\boldsymbol{X}_n)$ and $\boldsymbol{S}_2(\boldsymbol{X}_n)$, it is possible to find a matrix $\boldsymbol{W}(\boldsymbol{X}_n)$ such that both matrices are transformed into diagonal matrices:

$$\boldsymbol{W}(\boldsymbol{X}_n)\boldsymbol{S}_1(\boldsymbol{X}_n)\boldsymbol{W}(\boldsymbol{X}_n)^\top = \boldsymbol{\Lambda}_1(\boldsymbol{X}_n) \text{ and } \boldsymbol{W}(\boldsymbol{X}_n)\boldsymbol{S}_2(\boldsymbol{X}_n)\boldsymbol{W}(\boldsymbol{X}_n)^\top = \boldsymbol{\Lambda}_2(\boldsymbol{X}_n) \tag{4}$$

where $\boldsymbol{\Lambda}_1(\boldsymbol{X}_n)$ and $\boldsymbol{\Lambda}_2(\boldsymbol{X}_n)$ are diagonal matrices (see e.g., Tyler et al. [2009]).

This procedure is called simultaneous diagonalization [see, e.g., Schott, 2005]. It leads to the diagonalization of $\boldsymbol{S}_1(\boldsymbol{X}_n)^{-1}\boldsymbol{S}_2(\boldsymbol{X}_n)$ which is not necessarily symmetric:

$$\boldsymbol{S}_1(\boldsymbol{X}_n)^{-1}\boldsymbol{S}_2(\boldsymbol{X}_n) = \boldsymbol{W}(\boldsymbol{X}_n)^\top\boldsymbol{\Lambda}(\boldsymbol{X}_n)\boldsymbol{W}(\boldsymbol{X}_n)^{-1}$$

with $\boldsymbol{\Lambda}(\boldsymbol{X}_n) = \boldsymbol{\Lambda}_1^{-1}(\boldsymbol{X}_n)\boldsymbol{\Lambda}_2(\boldsymbol{X}_n)$.

Generally, $\boldsymbol{\Lambda}_1(\boldsymbol{X}_n)$ is taken as the identity matrix and we focus on this particular case from now on. In this case, Problem (4) is equivalent to the diagonalization of $\boldsymbol{S}_2(\boldsymbol{X}_n)$ with a matrix of eigenvectors $\boldsymbol{W}(\boldsymbol{X}_n)$ that is orthogonal with respect to the inner product induced by $\boldsymbol{S}_1(\boldsymbol{X}_n)$ instead of the canonical inner product.

It is easy to see that Problem (4) is equivalent to the usual diagonalization of $\boldsymbol{S}_1(\boldsymbol{X}_n)^{-1/2}\boldsymbol{S}_2(\boldsymbol{X}_n)\boldsymbol{S}_1(\boldsymbol{X}_n)^{-1/2}$ which is a symmetric matrix with ordered eigenvalues given by $\boldsymbol{\Lambda}_2(\boldsymbol{X}_n)$ and orthonormal eigenvectors given by $\boldsymbol{S}_1(\boldsymbol{X}_n)^{1/2}\boldsymbol{W}(\boldsymbol{X}_n)$.

Finally, Problem (4) is also equivalent to the problem of finding values $\lambda_i(\boldsymbol{X})$ and vectors $\boldsymbol{w}_i(\boldsymbol{X})$, $i \in \{1, \dots, p\}$, such that

$$\boldsymbol{S}_2(\boldsymbol{X}_n)\boldsymbol{w}_i(\boldsymbol{X}) = \lambda_i(\boldsymbol{X})\boldsymbol{S}_1(\boldsymbol{X}_n)\boldsymbol{w}_i(\boldsymbol{X})$$

which is called the generalized eigendecomposition problem.

This simultaneous diagonalisation procedure is used in different contexts and takes different names depending on the context. For instance, when using the scatter matrices Cov and Cov_4, the method is called FOBI in the signal processing literature (see e.g., Nordhausen and Virta [2019] and Subsection 5.1 below). When considering the usual covariance matrix and one autocovariance matrix ACov, it is called AMUSE in the time series context (see e.g., Pan et al. [2021] and Subsection 5.2 below).

Going beyond two scatter matrices is more challenging but has also been studied in the literature. We will call the procedure "joint diagonalization" as soon as the number of scatter matrices is larger than two. Let us consider $\boldsymbol{S}_0(\boldsymbol{X}), \boldsymbol{S}_1(\boldsymbol{X}), \ldots, \boldsymbol{S}_K(\boldsymbol{X})$, i.e., $K+1$ scatter matrices associated with a random vector $\boldsymbol{X}$. It is known that, for such a collection of symmetric matrices, there exists a matrix $\boldsymbol{P}(\boldsymbol{X})$, such that $\boldsymbol{P}(\boldsymbol{X})\boldsymbol{S}_k(\boldsymbol{X})\boldsymbol{P}(\boldsymbol{X})^\top$ is diagonal, for each $k \in \{0, \ldots, K\}$, if and only if all pairs of scatter matrices commute [see Schott, 2005].

In the blind source separation model (see Section 5), the assumption that the scatter matrices commute is true, and the joint diagonalization is possible. However, when considering the sampling versions of the scatter matrices, the property is lost. In such a situation, we can try to make the matrices jointly "as diagonal as possible" [see, e.g., Flury and Gautschi, 1986, Clarkson, 1988, Miettinen, 2015].

One idea is to take one of the scatter matrices, let us say $\boldsymbol{S}_0(\boldsymbol{X}_n)$, as a reference and to find a transformation $\boldsymbol{W}(\boldsymbol{X}_n)$ such that $\boldsymbol{S}_0(\boldsymbol{X}_n)$ is diagonalized while the other scatter matrices, $\boldsymbol{S}_1(\boldsymbol{X}_n), \ldots, \boldsymbol{S}_K(\boldsymbol{X}_n)$ are only approximately diagonalized. A popular criterion is based on a least squares approach. More precisely, the criterion consists in looking for $\boldsymbol{W}(\boldsymbol{X}_n)$ which minimises the sum of squares of the off-diagonal elements of all possible scatter matrices after transformation, and under the constraint that $\boldsymbol{W}(\boldsymbol{X}_n)\boldsymbol{S}_0(\boldsymbol{X}_n)\boldsymbol{W}(\boldsymbol{X}_n)^\top$ is the identity

$$\min \sum_{k=1}^{K} ||\ \mathrm{off}(\boldsymbol{W}(\boldsymbol{X}_n)\boldsymbol{S}_k(\boldsymbol{X}_n)\boldsymbol{W}(\boldsymbol{X}_n)^\top)||^2, \text{ subject to } \boldsymbol{W}(\boldsymbol{X}_n)\boldsymbol{S}_0(\boldsymbol{X}_n)\boldsymbol{W}(\boldsymbol{X}_n)^\top = \boldsymbol{I}_p$$

where $\mathrm{off}(\boldsymbol{A}) = \boldsymbol{A} - \mathrm{diag}(\boldsymbol{A})$, for a square matrix $\boldsymbol{A}$, and $||\cdot||$ denotes the matrix Frobenius norm.

This criterion is equivalent to maximizing the following sum

$$\sum_{k=1}^{K} ||\mathrm{diag}(\boldsymbol{W}(\boldsymbol{X}_n)\boldsymbol{S}_k(\boldsymbol{X}_n)\boldsymbol{W}(\boldsymbol{X}_n)^\top)||^2, \text{ subject to } \boldsymbol{W}(\boldsymbol{X}_n)\boldsymbol{S}_0(\boldsymbol{X}_n)\boldsymbol{W}(\boldsymbol{X}_n)^\top = \boldsymbol{I}_p. \quad (5)$$

Popular algorithms for this approximate joint diagonalization are based on Jacobi rotations [see, e.g., Clarkson, 1988, Cardoso and Souloumiac, 1996, Miettinen, 2015]. In the multivariate time series context, with $\boldsymbol{S}_0(\boldsymbol{X}_n) = \mathrm{Cov}(\boldsymbol{X}_n)$ and, for $\boldsymbol{S}_1(\boldsymbol{X}_n), \ldots, \boldsymbol{S}_K(\boldsymbol{X}_n)$, autocovariance matrices with different lags, this algorithm is called SOBI [Miettinen et al., 2016]. Other possible algorithms are also discussed in Illner et al. [2015]. In particular, Miettinen et al. [2014] introduced a deflation-based algorithm such that the single rows of the matrix $\boldsymbol{W}(\boldsymbol{X}_n)$ are calculated one after the other by looking at a maximization problem similar to (5) but for each row of $\boldsymbol{W}(\boldsymbol{X}_n)$. The existence and the uniqueness of the solution are discussed.

There exist also several other proposals [see, e.g., Miettinen, 2015, Illner et al., 2015]. If there is no reason that $\boldsymbol{S}_0(\boldsymbol{X}_n)$ plays a special role, it is possible to replace the constraint $\boldsymbol{W}(\boldsymbol{X}_n)\boldsymbol{S}_0(\boldsymbol{X}_n)\boldsymbol{W}(\boldsymbol{X}_n)^\top = \boldsymbol{I}_p$ by other constraints [see, e.g., Ziehe et al., 2004, Yeredor, 2002]. Moreover, the square in the maximisation criterion can be replaced by other functions allowing to weight differently the involved scatter matrices [see Miettinen et al., 2014, Miettinen, 2015, for details].

In the rest of this article, we present many existing multivariate data analysis methods that use a simultaneous or approximate joint diagonalisation of scatter matrices.

4 Invariant coordinate selection

There are two popular data transformations based on a single scatter matrix. The first one is principal component analysis, which uses the transformation matrix as the orthogonal matrix $\boldsymbol{U}$ obtained via the eigenvalue-eigenvector decomposition of $\boldsymbol{S}(\boldsymbol{X}_n) = \boldsymbol{U}(\boldsymbol{X}_n)^\top \boldsymbol{D}(\boldsymbol{X}_n)\boldsymbol{U}(\boldsymbol{X}_n)$ where the ith row of $\boldsymbol{U}(\boldsymbol{X}_n)$ contains the ith eigenvector of $\boldsymbol{S}(\boldsymbol{X}_n)$ and the diagonal matrix $\boldsymbol{D}(\boldsymbol{X}_n)$ contains on its diagonal the corresponding eigenvalues for which we assume that they are ordered in descending order. The principal components are the observations projected along the principal vectors, i.e. $\boldsymbol{z}_i = \boldsymbol{U}(\boldsymbol{X}_n)\boldsymbol{x}_i$, $i \in \{1, \dots, n\}$, where it is often assumed that the observations are centered. The principal components then have the property that they are uncorrelated with respect to $\boldsymbol{S}(\boldsymbol{X}_n)$, i.e., $\boldsymbol{S}(\boldsymbol{Z}_n) = \boldsymbol{D}(\boldsymbol{X}_n)$. Traditional PCA is based on the regular covariance matrix [see, e.g., Jolliffe, 2002, for details]; however, within an elliptical distribution framework any scatter matrix can be used for the same purpose.

Another transformation is the so-called whitening transformation which, besides the scatter $\boldsymbol{S}(\boldsymbol{X}_n)$, needs a location $\boldsymbol{T}(\boldsymbol{X}_n)$. Whitened observations are obtained as

$$\boldsymbol{x}_i^{st} = \boldsymbol{S}(\boldsymbol{X}_n)^{-1/2}(\boldsymbol{x}_i - \boldsymbol{T}(\boldsymbol{X}_n)),$$

$i \in \{1, \dots, n\}$. These observations have the properties that $\boldsymbol{T}(\boldsymbol{X}_n^{st}) = \boldsymbol{0}$ and $\boldsymbol{S}(\boldsymbol{X}_n^{st}) = \boldsymbol{I}_p$ which means that compared to PCA, whitened transformations not only uncorrelate the components but also give them equal scales. Note however, that the whitened components are not necessarily just the scaled principle components but might have undergone an additional rotation. Actually, Ilmonen et al. [2012] mention five alternative ways to compute $\boldsymbol{S}(\boldsymbol{X}_n)^{-1/2}$ which might all differ by an orthogonal rotation. If not specified otherwise, we consider the symmetric variant $\boldsymbol{S}(\boldsymbol{X}_n)^{-1/2} = \boldsymbol{U}(\boldsymbol{X}_n)\boldsymbol{D}(\boldsymbol{X}_n)^{-1/2}\boldsymbol{U}(\boldsymbol{X}_n)^\top$ with $\boldsymbol{U}(\boldsymbol{X}_n)$ and $\boldsymbol{D}(\boldsymbol{X}_n)$ as above. Again, this transformation usually uses the regular mean vector and the covariance matrix but other locations and scatter functionals can also be used, in which case an elliptical model is tacitly assumed.

One of the first ideas regarding the use of two scatter matrices was then based on performing these two transformations one after the other, but using a different scatter matrix for each one. The idea is then, ignoring the location for a moment, that the data are first whitened with respect to a scatter $\boldsymbol{S}_1$ and then PCA is performed on the whitened data using another scatter $\boldsymbol{S}_2$. This can be formulated as the simultaneous diagonalization problem of finding the transformation matrix $\boldsymbol{W}(\boldsymbol{X}_n)$ such that

$$\boldsymbol{W}(\boldsymbol{X}_n)\boldsymbol{S}_1(\boldsymbol{X}_n)\boldsymbol{W}(\boldsymbol{X}_n)^\top = \boldsymbol{I}_p, \quad \boldsymbol{W}(\boldsymbol{X}_n)\boldsymbol{S}_2(\boldsymbol{X}_n)\boldsymbol{W}(\boldsymbol{X}_n)^\top = \boldsymbol{D}(\boldsymbol{X}_n),$$

where $\boldsymbol{D}(\boldsymbol{X}_n)$ is a diagonal matrix with decreasing elements. Note that in the following, when the context is clear, we drop the dependence on $\boldsymbol{X}_n$ for $\boldsymbol{W}, \boldsymbol{D}, \boldsymbol{S}_1$ and $\boldsymbol{S}_2$. Based on Section 3, it is clear that this is a generalized eigenvalue-eigenvector problem and $\boldsymbol{W}$ and $\boldsymbol{D}$ can be computed accordingly.

Thus, in a model-free context, this transformation can be considered as an investigation if, after removing the second order information as measured by $\boldsymbol{S}_1$, $\boldsymbol{S}_2$ can still find any structure in the data, which is, for example, not the case when the observations follow an elliptical distribution.

This transformation was first denoted generalized PCA in Caussinus and Ruiz [1990], Caussinus and Ruiz-Gazen [2007], Caussinus et al. [2003] but the more commonly acknowledged name at present is invariant coordinate selection (ICS) as established in Tyler et al. [2009]. Note that some special scatter combinations are considered under specific names. For example the combination $\boldsymbol{S}_1 = \text{Cov}$ and $\boldsymbol{S}_2 = \text{Cov}_{-1}$ is known as principal axis analysis [Critchley et al., 2006] and the combination $\boldsymbol{S}_1 = \text{Cov}$ and $\boldsymbol{S}_2 = \text{Cov}_4$ is known as fourth order blind identification (FOBI) [Cardoso, 1989] which may be one of the most popular combinations.

The name ICS is motivated based on the following equivariance property which holds when the elements of $\boldsymbol{D}$ are all distinct:

$$\boldsymbol{W}(\boldsymbol{X}_n)\boldsymbol{A}^{-1} = \boldsymbol{J}\boldsymbol{W}(\boldsymbol{X}_n\boldsymbol{A}^\top + \boldsymbol{1}_n\boldsymbol{b}^\top),$$

where $\boldsymbol{A}$ is a $p \times p$ matrix and $\boldsymbol{b}$ a p-vector. $\boldsymbol{J}$ denotes a sign change matrix, i.e., a diagonal matrix with ±1 on its diagonal. Thus, in connection with a location functional $\boldsymbol{T}$,

$$\boldsymbol{W}(\boldsymbol{X}_n)\left(\boldsymbol{x}_i - \boldsymbol{T}(\boldsymbol{X}_n)\right) = \boldsymbol{J}\boldsymbol{W}(\boldsymbol{X}_n\boldsymbol{A}^\top + \boldsymbol{1}_n\boldsymbol{b}^\top)\left((\boldsymbol{A}\boldsymbol{x}_i + \boldsymbol{b}) - \boldsymbol{T}(\boldsymbol{X}_n\boldsymbol{A}^\top + \boldsymbol{1}_n\boldsymbol{b}^\top)\right),$$

which means that the so called ICS-components $\boldsymbol{z}_i = \boldsymbol{W}(\boldsymbol{X}_n)\left(\boldsymbol{x}_i - \boldsymbol{T}(\boldsymbol{X}_n)\right)$ are affine invariant under linear transformations up to their signs. Therefore one can argue that the ICS components show the intrinsic structure of the data independent of the coordinate system in which the data were originally presented. This is quite different from the principal components and the whitened observations which do not have this type of invariance property. The differences between the transformations are illustrated in Fig. 1 where the observable 4-variate data are shown together with the corresponding principal components, whitened components and invariant coordinates, which are in this case based on FOBI. In this artificial example, the two latent clusters are best visible in the invariant coordinates (see the two modes of the density estimator on the 4th plot of the diagonal of panel D).

In their seminal paper Tyler et al. [2009] also provide an interpretation of the eigenvalues contained in the diagonal of $\boldsymbol{D}$. Let $\boldsymbol{w}$ be a p-vector. Then, one can consider $\boldsymbol{w}^\top\boldsymbol{S}_1(\boldsymbol{X})\boldsymbol{w}$ as a squared measure of the scale of $\boldsymbol{X}$ in the direction of $\boldsymbol{w}$. As the ratio of two squared scale measures can be seen as a kurtosis measure [Radojicic et al., 2020],

$$\kappa(\boldsymbol{w}^\top\boldsymbol{X}) = \frac{\boldsymbol{w}^\top\boldsymbol{S}_1(\boldsymbol{X})\boldsymbol{w}}{\boldsymbol{w}^\top\boldsymbol{S}_2(\boldsymbol{X})\boldsymbol{w}}$$

is therefore the kurtosis measured in the $\boldsymbol{S}_1 - \boldsymbol{S}_2$-sense of $\boldsymbol{w}^\top\boldsymbol{X}$. Hence, the diagonal elements in $\boldsymbol{D}$ are the (ordered) kurtosis values of the invariant coordinates. Tyler et al. [2009] show that the maximal/minimal kurtosis in the $\boldsymbol{S}_1 - \boldsymbol{S}_2$-sense that can be obtained for $\boldsymbol{X}$, corresponds to the first/last eigenvalue. Thus Tyler et al. [2009] state ICS can be "viewed as a projection pursuit without the pursuit effort". If $\boldsymbol{S}_1$ and $\boldsymbol{S}_2$ are normalized for the Gaussian distribution, then an eigenvalue $d_j = 1$, $j \in \{1, \dots, p\}$, can be taken as an indicator that the component z_j follows a Gaussian distribution.

Based on the invariance property of ICS and the properties of the eigenvalues ICS has been used for many purposes, mainly in an exploratory data analysis way.

4.1 ICS for descriptive statistics

As the ICS components show the intrinsic nature of the data, they are a natural start to describe the basic data features. Nordhausen et al. [2011a] actually suggest using an additional second location

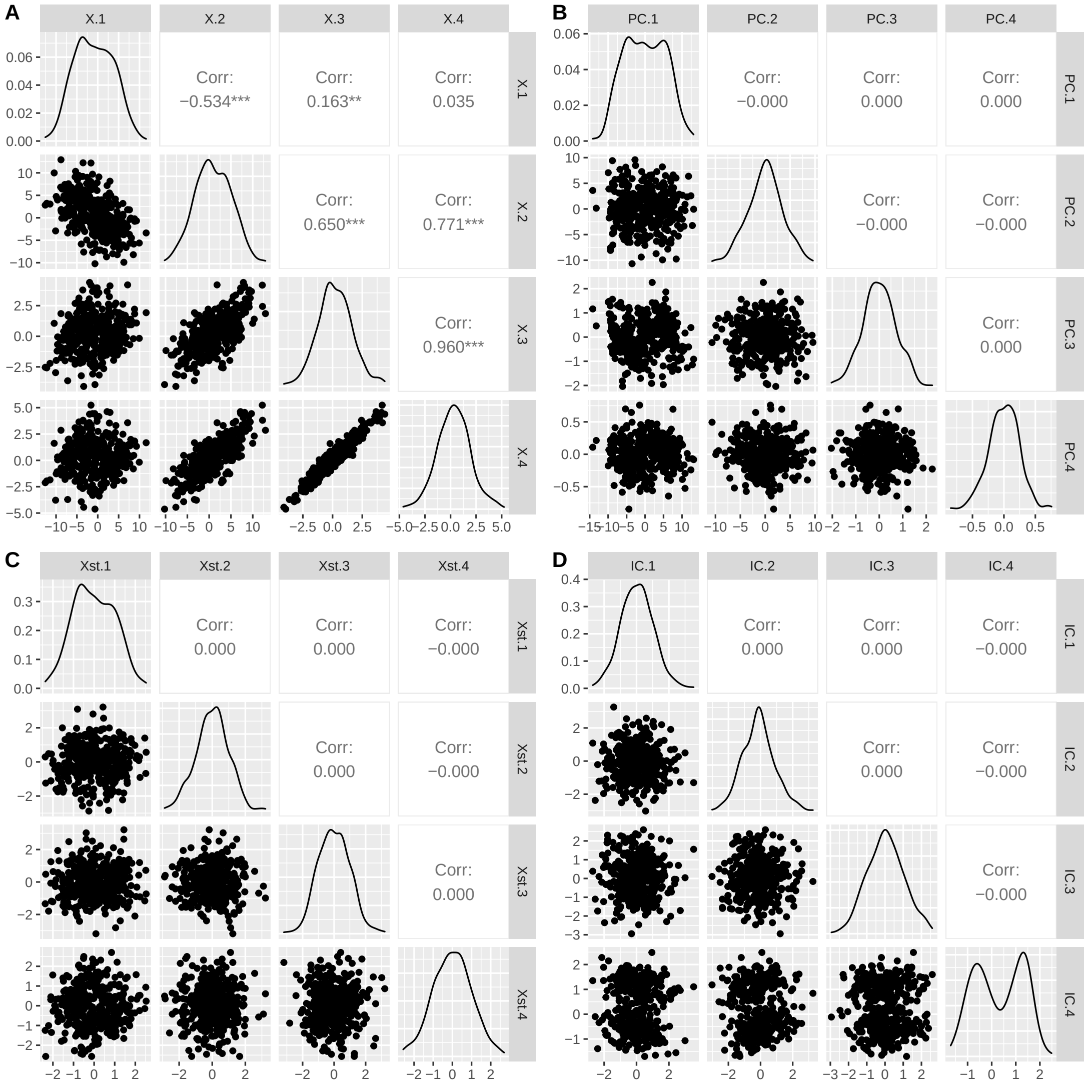

Fig. 1: Comparison of different data transformations. Panel A shows a matrix scatterplot with density estimators on the diagonal and correlations for the original data with $p = 4$, Panel B the corresponding principal components, Panel C the whitened components and Panel D the invariant coordinates where $\boldsymbol{S}_1 = \mathrm{Cov}$ and $\boldsymbol{S}_2 = \mathrm{Cov}_4$. Therefore all four panels give different views of the same dataset where in the ICS representation the two groups are best visible.

$\boldsymbol{T}_2$ and fixing the sign of the jth component of $\boldsymbol{z}_i$ so that $(\boldsymbol{T}_2(\boldsymbol{Z}_n))_j \geq 0$, which means that the difference in the locations is a measure of skewness of the components. Thus, if $\boldsymbol{T}_2(\boldsymbol{Z}_n) \approx \boldsymbol{0}$ the data are symmetric and if all eigenvalues of $\boldsymbol{D}$ are the same, it is an indication of ellipticity. Similarly the eigenvalues can give indications, together with the skewness measure, for other multivariate models such as skew-elliptical models as discussed, for example, in Nordhausen et al. [2011a], Loperfido [2021]. A more formal inference framework is given in Ilmonen et al. [2010] where the limiting distributions of $\boldsymbol{W}$, $\boldsymbol{D}$ and $\boldsymbol{T}_1 - \boldsymbol{T}_2$ are derived, based on the location and scatter functionals, especially when they are moment based. Kankainen et al. [2007] developed tests for multivariate normality based on these ideas when using the pair of scatter matrices Cov and Cov_4 (FOBI).

4.2 ICS for dimension reduction and outlier detection

Recently, data sets have been increasing in dimension and in sample size. Standard assumptions in modern multivariate statistics are such that data sets containing considerable noise and relevant features can be concentrated in a much smaller signal subspace. The goal of dimension reduction is then to estimate the signal subspace whose dimension is usually unknown. The question is how to define what makes a signal. For example, PCA says that the signal subspace is the one that contains most variation in the data, and there are many rules how to choose the subspace dimension [see for example Jolliffe, 2002]. PCA is likely the most commonly used dimension reduction method and seems to be quite successful in practice. It is from a theoretical point, however difficult, to argue why the directions in which, for example, groups are to be separated or outliers are to be identified, should be those with large variation. The construction of counterexamples is quite easy. Kurtosis, on the other hand, is a natural indicator of non-Gaussianity and, is one of the most popular projection pursuit indices [Huber, 1985]. In a mixture model framework, the classical kurtosis measure depends on the mixing proportion. If the group sizes are approximately equal, the kurtosis is small, while for unbalanced groups, with the extreme case of outliers, the kurtosis will be large. Thus, the eigenvalues contained in the matrix $\boldsymbol{D}$ give an indicator of interestingness of the components and allow, for example, the search for groups. In the above example, as shown in Fig. 1, the last invariant coordinate is most interesting as the groups in this example are of equal size. This makes component selection slightly more challenging, as first and last components might be of interest. This is different from PCA where usually only the first few components are of interest. Tyler et al. [2009] show that in a general framework with a mixture of elliptical distributions, ICS will find Fisher's linear discriminant without knowing the class labels, and ICS was considered a method for dimension reduction prior to group identification, for example, in Tyler et al. [2009], Peña et al. [2010], Alashwali and Kent [2016], Fekri and Ruiz-Gazen [2015], Fischer et al. [2017].

Similarly, the reduction of the dimension to make outlier detection easier via ICS was considered in Nordhausen et al. [2008b], Archimbaud et al. [2018], Archimbaud et al. [2018], especially in the context of reliability when it is known that the proportion of outliers is small. Archimbaud et al. [2018] show that it is easier to identify the outliers when they can be captured in a few invariant coordinates. If all invariant coordinates need to be selected for outlier detection, then the method corresponds to Mahalanobis-type outlier detection approach where the Malahanobis distances are computed with respect to $\boldsymbol{S}_1$.

The determination of which and how many components to retain is still often done visually or based on heuristics. However, when assuming a non-Gaussian component analysis framework where it is assumed that the data can be decomposed into a non-Gaussian (signal) subspace that is independent of the remaining (noise) Gaussian subspace, formal inference about the subspace dimensions was discussed in the context of FOBI in Luo and Li [2016], Nordhausen et al. [2017, 2021c], Luo and Li [2021], and for general scatter combinations, it was discussed in Radojicic and

Nordhausen [2020].

4.3 ICS as a transformation-retransformation method

As discussed above, in multivariate statistics, it is of key interest that the results of the analysis do not depend on the coordinate system used. Thus, estimates should have an appropriate equivariance property under affine transformations, and tests, for example, should be invariant. However, there are multivariate methods that are not affine equivariant/invariant, which is considered a major flaw. For example, multivariate methods based on marginal signs and ranks [see, for example, Puri and Sen, 1971] suffer from this disadvantage. ICS can help in this context as a transformation-retransformation approach. This means that multivariate methods are applied to the invariant coordinates and, if required, retransformed to the original scale. This was discussed, for example, in Nordhausen et al. [2006, 2008b].

As ICS is used with very different purposes in mind, Tyler et al. [2009] argued that there is no general best scatter combination for $\boldsymbol{S}_1$ and $\boldsymbol{S}_2$. Depending on the problem and data at hand, the scatter matrices might require different properties. In general, however, Tyler et al. [2009] argued that it would be advisable not to use two highly robust scatter matrices, as interesting features will not be detected when both scatter matrices focus too much on the same "inner" part of the data. Alashwali and Kent [2016] argued that it might be advisable that both scatter matrices are computed with respect to the same location functional, especially when subsequent clustering is the goal. Which scatter functional is used first and which second is also of minor consequence. The effect of changing the order is to invert the eigenvalues and reversing the order of the components, which then also have different scales. A common convention is, for example, to choose the order so that $\boldsymbol{S}_1$ is more robust than $\boldsymbol{S}_2$ and that the ICS components are centered with respect to the location functional which goes most naturally with $\boldsymbol{S}_1$. For further discussions regarding invariant transformations, we refer to Serfling [2010], Ilmonen et al. [2012], Serfling [2015].

To apply ICS and related methods in `R` R Core Team [2021], packages `ICS`Nordhausen et al. [2008b], `ICSOutlier` Archimbaud et al. [2018], `ICSShiny` Archimbaud et al. [2018] and `ICtest` Nordhausen et al. [2021d] are available.

5 Blind Source Separation

ICS is often seen as a mainly exploratory tool for multivariate analysis. A more model based approach where joint diagonalization plays a major role is blind source separation (BSS). The basic BSS model is

$$\boldsymbol{X} = \boldsymbol{A}\boldsymbol{Z} + \boldsymbol{\mu},$$

where $\boldsymbol{X}$ is a p-variate observable phenomenon that is seen as a linear mixture of a somewhat standardized latent p-variate source $\boldsymbol{Z}$, where the mixing is represented by the full rank $p \times p$ matrix $\boldsymbol{A}$ and the location of $\boldsymbol{X}$ is specified by the p-vector $\boldsymbol{\mu}$. Standard assumptions are that $E(\boldsymbol{Z}) = \boldsymbol{0}$ and $\mathrm{Cov}(\boldsymbol{Z}) = \boldsymbol{I}_p$ which indicates that the components of $\boldsymbol{Z}$ are at least uncorrelated. The goal in BSS is to estimate $\boldsymbol{Z}$ based on a realized sample $\boldsymbol{x}_1, \dots, \boldsymbol{x}_n$ of $\boldsymbol{X}$ alone. The location $\boldsymbol{\mu}$ in this case is mainly considered as a nuisance parameter and, in the following, we assume for simplicity $\boldsymbol{\mu} = \boldsymbol{0}$.

Clearly, without further assumptions, it is not possible to solve the BSS problem, and there must be at least one further structural component given for $\boldsymbol{Z}$ that can be exploited. Different BSS models have been suggested in the literature, with different additional assumptions. The ones we will consider here are: (i) The observations are independent and identically distributed (iid), and the components of $\boldsymbol{Z}$ are independent and non-Gaussian. This case is known as independent component analysis (ICA). (ii) The observed data are a p-variate time series, and the components of the latent time series are uncorrelated or independent. Then, additional information that can be exploited is serial dependence. (iii) The observed data come from a p-variate spatial random field where the p latent fields are again uncorrelated and independent, and the additional structure to be exploited is the spatial dependence.

Before going into detail, we point out that general overviews for BSS are, for example, Cichocki and Amari [2002], Comon and Jutten [2010], Adali et al. [2014], Nordhausen and Oja [2018] and that BSS approaches that are based on joint diagonalization are often called algebraic BSS methods.

All approaches make use of the following key result Miettinen et al. [2015]. Let $\boldsymbol{X}^{st} = \mathrm{Cov}(\boldsymbol{X})^{-1/2}(\boldsymbol{X} - \boldsymbol{E}(\boldsymbol{X}))$ be the standardized version of $\boldsymbol{X}$, then

$$\boldsymbol{X}^{st} = \boldsymbol{U}^\top \boldsymbol{Z},$$

where $\boldsymbol{U}$ is some orthogonal $p \times p$ matrix. Thus, after whitening, the BSS problem can be reduced to the problem of finding an orthogonal matrix.

The strategy of all algebraic BSS methods described below makes use of the generalized concept of a scatter functional which only requires affine equivariance but relaxes the positive definiteness requirement. Then, the approach is to select $K \geq 1$ scatter functionals $\boldsymbol{S}_1, \ldots, \boldsymbol{S}_K$ for which

$$\boldsymbol{S}_i(\boldsymbol{Z}) = \boldsymbol{D}_i, \quad i \in i, \ldots, K,$$

holds with $\boldsymbol{D}_i$ being diagonal matrices. Thus, all the scatter matrices used are diagonal when computed for the sources. Then, the approach is to find the orthogonal matrix $\boldsymbol{U}$ such that:

$$\boldsymbol{U}\boldsymbol{S}_i(\boldsymbol{X}^{st})\boldsymbol{U}^\top = \boldsymbol{D}_i, \quad i \in i, \ldots, K,$$

which yields the unmixing matrix $\boldsymbol{W} = \boldsymbol{U}\,\mathrm{Cov}(\boldsymbol{X})^{-1/2}$. Thus, algebraic BSS methods consists of a joint diagonalization problem, where the unmixing matrix $\boldsymbol{W}$ diagonalizes the $K+1$ scatter matrices $\mathrm{Cov}(\boldsymbol{X}), \boldsymbol{S}_1(\boldsymbol{X}), \ldots, \boldsymbol{S}_K(\boldsymbol{X})$ under the constraint that $\boldsymbol{W}\,\mathrm{Cov}(\boldsymbol{X})\boldsymbol{W}^\top = \boldsymbol{I}_p$.

In the following, we will discuss the different additional structural requirements made on the latent components and which scatter functionals are suitable.

Before this, let us discuss why we should perform BSS:

1. Often, it is assumed that the latent components have either physical meanings (BSS was suggested first in the signal processing literature) or that they are easier to interpret than the original components.

2. Another motivation is that often only a few components are considered interesting and the remainder noise, thus, it can be used for dimension reduction.

3. As the latent components are assumed uncorrelated or even independent, each component can be modelled in a univariate way, and instead of fitting a p-variate model, one could fit p univariate models, which is often considered much simpler. For spatial data, such a benefit is demonstrated in Muehlmann et al. [2020] in the context of prediction .

5.1 Independent component analysis

ICA is the best known BSS approach. ICA methods are designed for iid data but are also often applied for dependent observations in which case, however, not all available information is exploited. In the ICA, it is assumed that:

(IC1): $E(\boldsymbol{Z}) = \boldsymbol{0}$ and $\text{Cov}(\boldsymbol{Z}) = \boldsymbol{I}_p$

(IC2): The components of $\boldsymbol{Z}$=$(Z_1, \ldots, Z_p)$, are mutually independent.

(IC3): At most one component of $\boldsymbol{Z}$ is Gaussian.

The first algebraic ICA approach is FOBI Cardoso [1989], as described above, which means that $K = 1$ and $\boldsymbol{S}_1 = \text{Cov}_4$ in the BSS framework. FOBI yields an unmixing matrix if all independent components have distinct kurtosis values. The statistical properties of FOBI, in the ICA model, were derived in Miettinen et al. [2015]. Oja et al. [2006] generalized then FOBI by showing that Cov and $\boldsymbol{S}_1$ in FOBI can be replaced by any scatter functionals that have the independence property, which is also further investigated in Nordhausen et al. [2008a]. Therefore, one can say, given a suitable choice of scatter functionals, that ICS can also solve the ICA problem when kurtosis values of $\boldsymbol{Z}$, in the sense of the two scatter functionals involved, are distinct.

In ICA, however, assuming that the components have distinct kurtosis (in the sense of the involved scatter matrices) is considered as a strong constraint. Therefore, Nordhausen et al. [2012] suggested using $K + 1$ scatter functionals that have all the independence property, and then using joint diagonalization as described in Section 3 to obtain the unmixing matrix. This is more flexible in the sense that this solves the ICA problem, where for each component there is at least one scatter combination $\boldsymbol{S}_0, \boldsymbol{S}_j$, $j \in \{1, \ldots, K\}$ with a distinct kurtosis compared with other components. Thus, the K-Scatter ($K > 1$) approach is more flexible than the 2-scatter approach. However, this approach still cannot separate components that have the same distribution. An ICA approach that also works for identical distributed components and is based on joint diagonalization is JADE (joint diagonalization of eigenmatrices), which uses cumulant matrices. Therefore, the method does not fully fit in the framework presented here, as no scatter functionals are diagonalized but certain cumulant matrices. We refer the reader to Cardoso and Souloumiac [1996], Miettinen et al. [2015] for further details. The 2-scatter ICA approach is often not optimal and FOBI is, for example, always less efficient than JADE Miettinen et al. [2015]. The 2-scatter ICA approach, compared to the K-scatter approach or JADE, is usually much easier to compute, especially FOBI which is often the start of an ICA analysis.

An extension of ICA is independent subspace analysis (ISA), which is also known as multivariate ICA. In this framework $\boldsymbol{Z}$ does not have p "univariate" independent components but only c components which may be multivariate. Thus, $\boldsymbol{Z} = (\boldsymbol{Z}_1^\top, \ldots, \boldsymbol{Z}_c^\top)^\top$ where $\boldsymbol{Z}_i$ has dimension p_i, $i \in \{1, \ldots, c\}$ with $p_1 + \cdots + p_c = p$. In ISA, the individual components cannot be recovered but only their subspaces. An approach based on joint diagonalization is first to perform 2-scatter ICA, then compute a third scatter that has the block independence property for the obtained components, and finally blockdiagonalize this third scatter. For details, see for example Nordhausen and Oja [2011].

5.2 BSS for time series

In ICA, the extra feature exploited was non-Gaussianity for BSS. In time series, serial dependence can be exploited, which in turn allows multiple Gaussian components. Different BSS approaches imply different assumptions regarding the time series. For simplicity, we continue using $\boldsymbol{X}$ for the stochastic process $\boldsymbol{X}_t$.

1. Second order source separation (SOS) makes the following model assumptions:

   (SOS): $E(\boldsymbol{Z}) = \boldsymbol{0}$ and $\text{Cov}(\boldsymbol{Z}) = \boldsymbol{I}_p$.

   (SOS2): $\text{ACov}_\tau(\boldsymbol{Z}) = \boldsymbol{D}_\tau$ for all $\tau \geq 1$ where $\boldsymbol{D}_\tau$ are all diagonal matrices.

Therefore, in the SOS model, the latent components are uncorrelated throughout time and it is usually assumed that the latent components are linear processes.

The first SOS method is known as AMUSE (algorithm for multiple unknown signals extraction) Tong et al. [1990], which chooses $K = 1$ and $\boldsymbol{S}_1 = \mathrm{ACov}_\tau^S$ for some lag τ, which is very often 1. AMUSE can solve the SOS problem if all autocorrelations at the used lag are distinct. AMUSE is very sensitive to the chosen lag and was extended to SOBI (second order blind identification) in Belouchrani et al. [1997]. The method consists in choosing K symmetrized autocovariance matrices with different lags $\tau_1, \dots, \tau_K$, which are then jointly diagonalized. This is again more flexible, and different autocovariance matrices can contribute to the separation of different lags. The statistical properties of AMUSE are discussed in Miettinen et al. [2012] and those of SOBI in Illner et al. [2015], Miettinen et al. [2014, 2016]. Note, however, that SOBI is not always better than AMUSE but it is in most cases. The choice of lags is, however, an open question that has a large impact in practice, where the default is often to use simply the first 12 lags. For more sophisticated considerations for lag selection, see, for example, Tang et al. [2005], Taskinen et al. [2016]. Replacing Cov and ACov's in AMUSE or SOBI with robust alternatives is discussed, for example, in Ilmonen et al. [2015] but requires that the time series be symmetric.

2. Independent component time series (IC time series) model. In this model, one makes the assumptions:

(IC time series 1): $E(\boldsymbol{Z}) = \boldsymbol{0}$ and $\mathrm{Cov}(\boldsymbol{Z}) = \boldsymbol{I}_p$

(IC time series 2): The latent time series contained in $\boldsymbol{Z}$ are all independent.

Note that (IC time series 1) implies stationarity which is assumed in the methods presented in the following. It could, however, be relaxed and there is some overlap with the BSS approach considering nonstationary data described subsequently.

The main difference when assuming the IC time series model compared with the SOS model is that independence between components is required, and one usually has components with stochastic volatility, such as GARCH components. Statistics measuring second order information do not necessarily carry information, and higher order information is therefore commonly used. The main method, in our context, is the generalized FOBI (gFOBI) Matilainen et al. [2015] which extends FOBI by defining the scatter functional for the mean zero process

$$\mathrm{Cov}_{4,\tau}(\boldsymbol{X}) = E(\boldsymbol{X}_{t+\tau}\boldsymbol{X}_t^\top \mathrm{Cov}(\boldsymbol{X}_t)^{-1}\boldsymbol{X}_t\boldsymbol{X}_{t+\tau}^\top),$$

which therefore can be seen as a lagged fourth moment matrix. For gFOBI, one selects a set of lags $\tau_1, \dots, \tau_K$ used in the joint diagonalization approach for $\boldsymbol{S}_j = \mathrm{Cov}_{4,\tau_j}$, $j \in \{1, \dots, j\}$. If the set consists only of $\tau_1 = 0$ the method reduces to FOBI. gFOBI therefore can solve the BSS problem if all fourth moments are finite and the set of lags contains a lag τ for which the ith and jth diagonal elements of $\mathrm{Cov}_{4,\tau}(\boldsymbol{X})$ are distinct, for all pairs $i \neq j$. Note that JADE was similarly extended for this setting to gJADE in Matilainen et al. [2015].

3. Nonstationary source separation (NSS). Thus far, stationary data are considered. In the NSS models this is relaxed slightly and the assumptions are:

(NSS 1): $E(\boldsymbol{Z}) = \boldsymbol{0}$ and $\mathrm{Cov}(\boldsymbol{Z}_t) = \boldsymbol{D}_t$, where $\boldsymbol{D}_t$ is a diagonal matrix depending on t.

(NSS 2): $\mathrm{ACov}_\tau(\boldsymbol{Z}_t) = \boldsymbol{D}_{\tau,t}$ for all $\tau \geq 1$ where $\boldsymbol{D}_{\tau,t}$ are all diagonal matrices.

In this model, the mean is stationary, but not the second moment. To make the model formulation easier, it is often assumed that, for the observed time series $\boldsymbol{X}_T$, the latent components are scaled such that $\mathrm{Cov}(\boldsymbol{Z}_T) = \boldsymbol{I}_p$. The main idea for NSS is to divide the observed time span T into K non-overlapping intervals $T_1, \dots, T_K$, and then compute the scatter matrices separately for each interval, and jointly diagonalize them.

The NSS method.SD Choi and Cichocki [2000b] uses $K = 2$ and the unmixing matrix W corresponds to the matrix that simultaneously diagonalizes $\mathrm{Cov}(\boldsymbol{X}_{T_1})$ and $\mathrm{Cov}(\boldsymbol{X}_{T_2})$. Similar to 2-scatter ICA and AMUSE, the performance of this approach is sensitive to the division, and requires that each component has a distinct variance in the intervals. A straightforward extension is NSS.JD Choi and Cichocki [2000b] that chooses $K > 2$ and jointly diagonalizes $\mathrm{Cov}(\boldsymbol{X}^{st}_{T_1}), \dots, \mathrm{Cov}(\boldsymbol{X}^{st}_{T_K})$, where $\boldsymbol{X}^{st} = \mathrm{Cov}(\boldsymbol{X}_T)^{-1/2}(\boldsymbol{X}_T - \boldsymbol{1}_T \bar{\boldsymbol{X}}_{\boldsymbol{T}}{}^\top)$. NSS.SD and NSS.JD only require a temporal ordering of the observations but do not otherwise exploit the serial dependence. The approach NSS.TD.JD Choi and Cichocki [2000a] therefore chooses K intervals and a set of L lags $\tau_1, \dots, \tau_L$ and jointly diagonalizes the $K \times L$ autocovariance matrices $\mathrm{ACov}^S_{\tau_i}(\boldsymbol{X}^{st}_{T_j})$, $i \in \{1, \dots, L\}$ $j \in \{1, \dots, K\}$. Thus, for $K = 1$, this approach reduces to SOBI, and the general idea is that the data follow a block stationary model. NSS was first considered in the context of audio signals, and K was chosen such that there are sufficient observations within an interval, so that the scatter matrices can be computed with sufficient precision. Another framework is that, on K subjects, the same experiment was performed and produces for each subject a p-variate time series. Then, assuming that for all subjects the same "mixing" occurred, one can concatenate the K time series and apply an NSS approach, where the intervals correspond to the concatenation points. Such an approach is often referred to as groupICA. NSS with robust scatter functionals, was, for example considered in Nordhausen [2014].

As it is not always clear which of the three time series BSS models is suitable, there exist generalizations combining different approaches. In general, approaches such as gSOBI Miettinen et al. [2020], cannot be expressed in a joint diagonalization framework. Nordhausen et al. [2021a] used almost all scatter matrices, as described above, including the subdivision into intervals, for joint diagonalization to cover all three models. This is, however, very challenging as the different scatter functionals are of different magnitudes, and it is not very clear how to weight them. This is still an area for further research. More details about general BSS approaches for time series are reviewed in Comon and Jutten [2010], Pan et al. [2021].

5.3 BSS for spatial data

Most areas where BSS was applied to date produced time series data. Therefore, the focus of BSS was mainly on time series methods. However, recently, BSS was also considered in the context of spatial data. In that case, $\boldsymbol{X} = \boldsymbol{X}(\boldsymbol{s})$ is a p-variate random field specified on the domain $\mathcal{S}$, where the domain can be 1, 2 or 3 dimensional. To estimate the latent components, one has a sample of n points $\boldsymbol{X}(\boldsymbol{s}_i)$, sampled at the distinct locations $\boldsymbol{s}_i \in \mathcal{S}$, $i \in \{1, \dots, n\}$. $\boldsymbol{X}_S$ denotes then the data matrix with the sampled observations.

Two spatial settings that have been considered so far in a BSS framework, can be seen as spatial counterpart to the SOS and NSS time series models where the role of the autocovariance matrices will be taken on by local covariance matrices (see the definition in Section 2).

1. Spatial blind source separation model (SBSS) makes the following model assumptions:

   (SBSS1): $E(\boldsymbol{Z}) = \boldsymbol{0}$ and $\mathrm{Cov}(\boldsymbol{Z}) = \boldsymbol{I}_p$.

(SBSS2): $E(\boldsymbol{Z}(\boldsymbol{s}), \boldsymbol{Z}(\boldsymbol{s}')^\top) = \boldsymbol{D}_h$, where $h = ||\boldsymbol{s} - \boldsymbol{s}'||$ for all $\boldsymbol{s}$ and $\boldsymbol{s}' \in \mathcal{S}$ with $\boldsymbol{s} \neq \boldsymbol{s}'$ and the diagonal matrix $\boldsymbol{D}_h$ contains the univariate covariance functions corresponding to the latent fields.

Thus, in SBSS, the latent fields are assumed to be uncorrelated / independent stationary random fields.

Nordhausen et al. [2015] suggested a 2-scatter approach that jointly diagonalizes Cov and one LCov_f for SBSS. The performance of this approach depends again heavily on the chosen kernel f. Bachoc et al. [2020] then suggested a joint diagonalization approach with Cov and $\boldsymbol{S}_1 = \text{LCov}_{f_1}, \ldots, \boldsymbol{S}_K = \text{LCov}_{f_K}$, for $K \geq 2$, where the so-called ring kernels, with different radii, are the most natural kernels considered so far. The statistical properties of the two approaches are given in Bachoc et al. [2020] in the case of latent Gaussian random fields and show again that the joint diagonalization approach seems preferable.

2. Spatial nonstationary source separation (SNSS). This model assumes

(SNSS 1): $E(\boldsymbol{Z}) = \boldsymbol{0}$ and $\text{Cov}(\boldsymbol{Z}(\boldsymbol{s})) = \boldsymbol{D}_{\boldsymbol{s}}$, where $\boldsymbol{D}_{\boldsymbol{s}}$ is a diagonal matrix depending on $\boldsymbol{s} \in \mathcal{S}$.

(SNSS 2): $E(\boldsymbol{Z}(\boldsymbol{s}), \boldsymbol{Z}(\boldsymbol{s}')^\top) = \boldsymbol{D}_{\boldsymbol{s},\boldsymbol{s}'}$ or all $\boldsymbol{s}$ and $\boldsymbol{s}' \in \mathcal{S}$ with $\boldsymbol{s} \neq \boldsymbol{s}'$ and the diagonal matrix $\boldsymbol{D}_{\boldsymbol{s},\boldsymbol{s}'}$ depends on the locations.

Thus, the location is stationary and all latent fields are uncorrelated or independent. However, for all latent fields, the spatial covariance is non-stationary.

The idea for algebraic BSS in this model is quite similar to NSS in the time series case. The domain is divided into K non-overlapping subdomains $\mathcal{S}_1, \ldots, \mathcal{S}_K$ such that $\mathcal{S}_1 \cup \cdots \cup \mathcal{S}_K = \mathcal{S}$. Then, analogously to the time series setting, Muehlmann et al. [2021a] suggested the methods SNSS.SD, SNSS.JD and SNSS.SJD.

For SNSS.SD, $K = 2$ and the covariance matrices of the two domains are simultaneously diagonalized, which is again very sensitive to division into subdomains. SNSS.JD, which whitens the data using $\text{Cov}(\boldsymbol{X}_S)$ and then jointly diagonalizes the $K > 2$ covariance matrices obtained for the subdomains, i.e., $\text{Cov}(\boldsymbol{X}^{st}_{S_1}), \ldots, \text{Cov}(\boldsymbol{X}^{st}_{S_K}))$, is less sensitive to division into subdomains. Both, SNSS.SD and SNSS.JD, are based only on the spatial ordering of the points. If it is assumed that there would be some kind of block-stationary model underlying, SNSS.SJD suggests to compute L local covariance matrices $\text{LCov}_{f_j}(\boldsymbol{X}^{st}_{S_i})$, $i \in \{1, \ldots, K\}$, $j \in \{1, \ldots, L\}$, for all subdomains, and then jointly diagonalizes these $K \times L$ matrices.

All the above BSS methods, either simultaneously diagonalize two scatter functionals or jointly diagonalize $K + 1$ scatter matrices with one scatter playing a special role. There exist many other BSS models or methods where joint diagonalization plays a role. Some are for example summarized in Theis and Inouye [2006], Chabriel et al. [2014]. BSS is still an active research area, and spatial BSS is currently actively developed. A schematic overview of the BSS models covered in the present review and of ICS is given in Fig. 2.

Note that, as mentioned earlier, the goal of BSS is to estimate the latent components, and for that purpose, we used simultaneous and joint diagonalization (see Section 3) to obtain an unmixing matrix $\boldsymbol{W}$ such that

$$\boldsymbol{z}_i = \boldsymbol{W}(\boldsymbol{x}_i - \boldsymbol{T}(\boldsymbol{X}_n)), \quad i \in \{i, \ldots, n\}.$$

This is however rather imprecise, as none of the models described above are identifiable in a strict sense. In all the BSS models described above, one can write

$$\boldsymbol{X} = \boldsymbol{A}\boldsymbol{Z} = (\boldsymbol{A}\boldsymbol{J}\boldsymbol{P})(\boldsymbol{P}^\top \boldsymbol{J}\boldsymbol{Z}) = \boldsymbol{A}^*\boldsymbol{Z}^*,$$

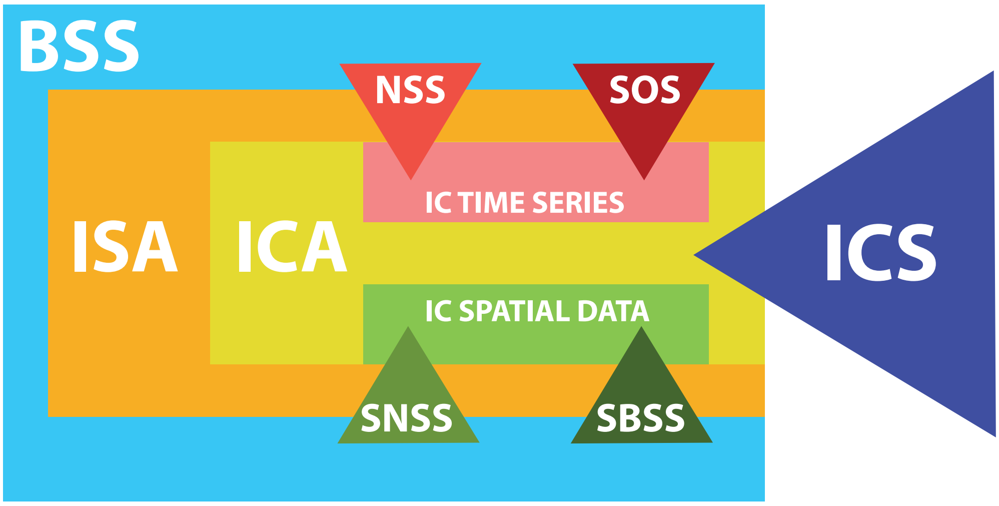

Fig. 2: Schematic overview of the different BSS models and ICS. For the definition of the models see Section 5.

where $\boldsymbol{J}$ is $p \times p$ sign-change matrix and $\boldsymbol{P}$ a $p \times p$ permutation matrix. Thus, the signs and order of the components cannot be fixed. Consequently, for any unmixing matrix $\boldsymbol{W}$, the matrix $\boldsymbol{JPW}$ is also an unmixing matrix for all permutation matrices $\boldsymbol{P}$ and all sign-change matrices $\boldsymbol{J}$. However, these identifiability issues are usually not considered to be a problem and the order of the components is, for example, usually fixed based on the diagonal elements of $\boldsymbol{D}$ in the case of simultaneous diagonalization, and based on $\operatorname{diag}(\sum_{i=1}^{K} \boldsymbol{W}\boldsymbol{S}_i(\boldsymbol{X}_n)\boldsymbol{W}^\top)$ in the case of joint diagonalization.

In performance studies, these indeterminancies have naturally to be taken into account and an overview of BSS performance measures is given for example in Nordhausen et al. [2011b].

Most algebraic BSS methods described above are implemented in R via the R packages `ICS`, `BSSasymp`, `JADE` Miettinen et al. [2017] , `tsBSS` Nordhausen et al. [2021b] and `SpatialBSS` Muehlmann et al. [2021d] where `JADE` contains also some performance measures.

6 Joint diagonalization in the context of supervised multivariate methods

PCA, ICS and BSS are often used as dimension reduction methods, which means that some components are selected and used in further modelling. However, if there is a response $\boldsymbol{Y}$ to be modelled, no direct information is used when computing the new directions in an unsupervised manner. Such dimension reduction methods are therefore called unsupervised dimension reduction methods. When information about the target is used in the dimension reduction process, one refers to it as supervised dimension reduction (SDR). Surprisingly, many SDR methods can be seen within a joint diagonalization framework.

For example, linear discriminant analysis (LDA) Fisher [1936] can be seen as a supervised method in a classification context. Let us consider a data set $\boldsymbol{X} = (\boldsymbol{x}_1, \dots, \boldsymbol{x}_n)$, with n observations and p variables, which is partitioned into K subpopulations or groups. Fisher's idea was to look for the best linear function of the p variables which maximized the ratio of the between-groups covariance to the within-groups covariance. The intuition is that groups are more easily visible when the between-groups variability is large in comparison with the within-groups variability. We use the index i for the group and j for the observation in each group, so that $\boldsymbol{x}_{ij}$ denotes the jth observation in group i, for $i \in \{1, \ldots, K\}$, $j \in \{1, \ldots, n_i\}$, where n_i denotes the number of observations in group i. Using the analysis of variance equation, we can decompose the total covariance matrix Cov, which does not take into account the groups, into the between scatter matrix $\mathrm{Cov}_{\boldsymbol{B}}$ and the within scatter matrix $\mathrm{Cov}_{\boldsymbol{W}}$ defined by:

$$\mathrm{Cov}_{\boldsymbol{B}} = \frac{1}{n}\sum_{i=1}^{K} n_i(\bar{\boldsymbol{x}}_i - \bar{\boldsymbol{x}})(\bar{\boldsymbol{x}}_i - \bar{\boldsymbol{x}})^\top \tag{6}$$

$$\mathrm{Cov}_{\boldsymbol{W}} = \frac{1}{n}\sum_{i=1}^{K}\sum_{j=1}^{n_i}(\boldsymbol{x}_{ij} - \bar{\boldsymbol{x}}_i)(\boldsymbol{x}_{ij} - \bar{\boldsymbol{x}}_i)^\top \tag{7}$$

where $\bar{\boldsymbol{x}}_i$ denotes the mean of the ith group and $\bar{\boldsymbol{x}}$ the overall mean. The Fisher's linear discriminant vectors are the eigenvectors of ${\mathrm{Cov}_{\boldsymbol{W}}}^{-1}\,\mathrm{Cov}_{\boldsymbol{B}}$ [see Mardia et al., 1979, for more details]. Both $\mathrm{Cov}_{\boldsymbol{B}}$ and $\mathrm{Cov}_{\boldsymbol{W}}$ are scatter matrices in the sense that they are affine equivariant and semipositive definite. Note however that the between-matrix is of rank $K-1$ and thus is generally singular. As a consequence, in general, there are $K-1$ nontrivial linear discriminant vectors. As detailed in Section 3, the Fisher's discriminant vectors are obtained as the solution of a generalized eigendecomposition.

Also canonical correlation analysis (CCA) Hotelling [1936] can be formulated as the simultaneous diagonalization of two scatter functionals when using

$$\boldsymbol{S}_1(\boldsymbol{X}) = \mathrm{Cov}(\boldsymbol{X}) \quad \text{and} \quad \boldsymbol{S}_2(\boldsymbol{X}) = \mathrm{Cov}_{CCA} = \mathrm{Cov}(\boldsymbol{X}, \boldsymbol{Y})\,\mathrm{Cov}(\boldsymbol{Y})^{-1/2}\,\mathrm{Cov}(\boldsymbol{Y}, \boldsymbol{X}).$$

Most SDR methods are developed in a regression context where, for simplicity, we assume from now on that the response Y is univariate. In the spirit of a BSS model, we assume that:

$$\boldsymbol{X} = \boldsymbol{A}\boldsymbol{Z} + \boldsymbol{\mu},$$

where $\boldsymbol{A}$ is the $p \times p$ full rank mixing matrix and $\boldsymbol{\mu}$ the p-variate location vector. For the latent p-vector, we assume there exists a partition $\boldsymbol{Z} = \left({\boldsymbol{Z}^{(1)}}^\top, {\boldsymbol{Z}^{(2)}}^\top\right)^\top$ with respective dimensions k and $p-k$. The assumptions are:

(SDR 1): $E(\boldsymbol{Z}) = \boldsymbol{0}$ and $\mathrm{Cov}(\boldsymbol{Z}) = \boldsymbol{I}_p$.

(SDR 2): $(Y, {\boldsymbol{Z}^{(1)}}^\top)^\top$ and $\boldsymbol{Z}^{(2)}$ are independent.

Thus all information on Y is contained in $\boldsymbol{Z}^{(1)}$. Note that there are naturally many partitions of $\boldsymbol{Z}$ fulfilling these assumptions. But the partition of interest is the one with the smallest value k that needs to be estimated together with the unmixing matrix $\boldsymbol{W}$. Note also that $\boldsymbol{Z}^{(1)}$ is only identifiable up to an orthogonal transformation, which means that the "unmixing" matrix can only recover the subspace of interest, which is however sufficient.

Liski et al. [2014] defined supervised invariant coordinate selection (SICS) as the joint diagonalization of one unsupervised scatter functional ($\boldsymbol{S}_1$) and one supervised scatter functional ($\boldsymbol{S}_2$).

Many well established supervised dimension reduction methods, like sliced inverse regression (SIR) Li [1991], sliced average variance estimation (SAVE) Cook [2000], principal Hessian directions (pHd) Li [1992], or directional regression (DR) Li and Wang [2007], can be seen as special cases of SICS. All these methods use $\boldsymbol{S}_1 = \mathrm{Cov}$ and differ regarding $\boldsymbol{S}_2$. SIR, for example, uses $\boldsymbol{S}_{SIR}$ as defined in Section 2. For the exact forms of $\boldsymbol{S}_{SAVE}$, $\boldsymbol{S}_{pHd}$ and $\boldsymbol{S}_{DR}$ we refer to Liski et al. [2014], where many other possibilities for supervised scatter functionals are listed. The advantage of an SDR approach over unsupervised methods is demonstrated in Fig. 3, where there is a response Y which is to be explained by four possible predictors. Panel A gives the original data where no clear relationship between any of the predictors and Y is visible. The PCs based on $\boldsymbol{X}$ given in panel B are not more informative regrding their relationship with Y. The invariant coordinates in panel C give some idea about the relationship when looking at IC.3. But the relationship is very clearly visible in panel D where SICS is displayed.

The performance of the SDR methods depends a lot on the true relationship between response and predictors, and different methods are more suitable to recognize certain types of dependencies than others. For detailed discussions about SDR methods, we refer, for example, to Li [2018], Ma and Zhu [2013]. Note that many of these methods also can make weaker assumptions than (SDR2).

In the example in Fig. 3, the true k is 1. In practice this needs to estimated. In SDR, depending on the scatter used, the theoretical value of eigenvalues which correspond to $\boldsymbol{Z}^{(2)}$, i.e., the values in $\boldsymbol{D}$, are known, and therefore tests and estimators can be based on these eigenvalues, as for example discussed in Bura and Cook [2001], Bura and Yang [2011], Luo and Li [2016, 2021], Nordhausen et al. [2021c].

Supervised dimension reduction methods are of course also of interest in the context of time series and spatial data. The effect of the predictors on the response might, for example, be delayed in the time series case, or depend on neighbouring values in the spatial setting. However, it is straightforward to adjust the SDR assumptions from above, and to formulate an appropriate BSS-SDR framework for dependent data. To take the temporal delay and spatial proximity into account, supervised temporal and spatial scatter functionals should be used, and more than two scatter matrices might be used. Matilainen et al. [2017] define time series SIR (TSIR), which is based on $\boldsymbol{S}_{TSIR,\tau}(\boldsymbol{X}) = \mathrm{Cov}(E(\boldsymbol{X}_t|Y_{t+\tau}))$, where τ is some lag. Then, TSIR whitens the data using Cov and jointly diagonalizes $\boldsymbol{S}_{TSIR,\tau_i}(\boldsymbol{X}^{st})$ with $\tau_i \in \{\tau_1, \dots, \tau_K\}$. Time series SAVE (TSAVE) is suggested in Matilainen et al. [2019], and jointly diagonalizes Cov, and K so-called time series SAVE matrices $\boldsymbol{S}_{TSAVE,\tau_i}$, using K different lags. Spatial SIR (SSIR) was so far only considered for lattice data in Muehlmann et al. [2021c], and jointly diagonalizes Cov and $\boldsymbol{S}_{SSIR,\boldsymbol{\tau}}(\boldsymbol{X}) = \mathrm{Cov}(E(\boldsymbol{X}_{\boldsymbol{s}}|Y_{\boldsymbol{s}+\boldsymbol{\tau}}))$, where $\boldsymbol{\tau}$ is now a d-dimensional lag. These approaches are all fairly new and inference tools are still missing.

Various SDR approaches discussed here are, for example, implemented in R in the packages dr Weisberg [2002], ICS and tsBSS.

7 Conclusions

Many multivariate statistical methods make use of the joint diagonalization of several scatter matrices as illustrated in the previous sections. Table 1 summarizes the different models, methods and scatter functionals that are jointly diagonalized. However, the overview we propose in this paper is far from exhaustive. For example, Theis and Inouye [2006], Chabriel et al. [2014] give an overview of algebraic BSS methods that contains models and approaches not mentioned here. Then, there are also completely different multivariate statistical methods that use joint diagonalization, but that

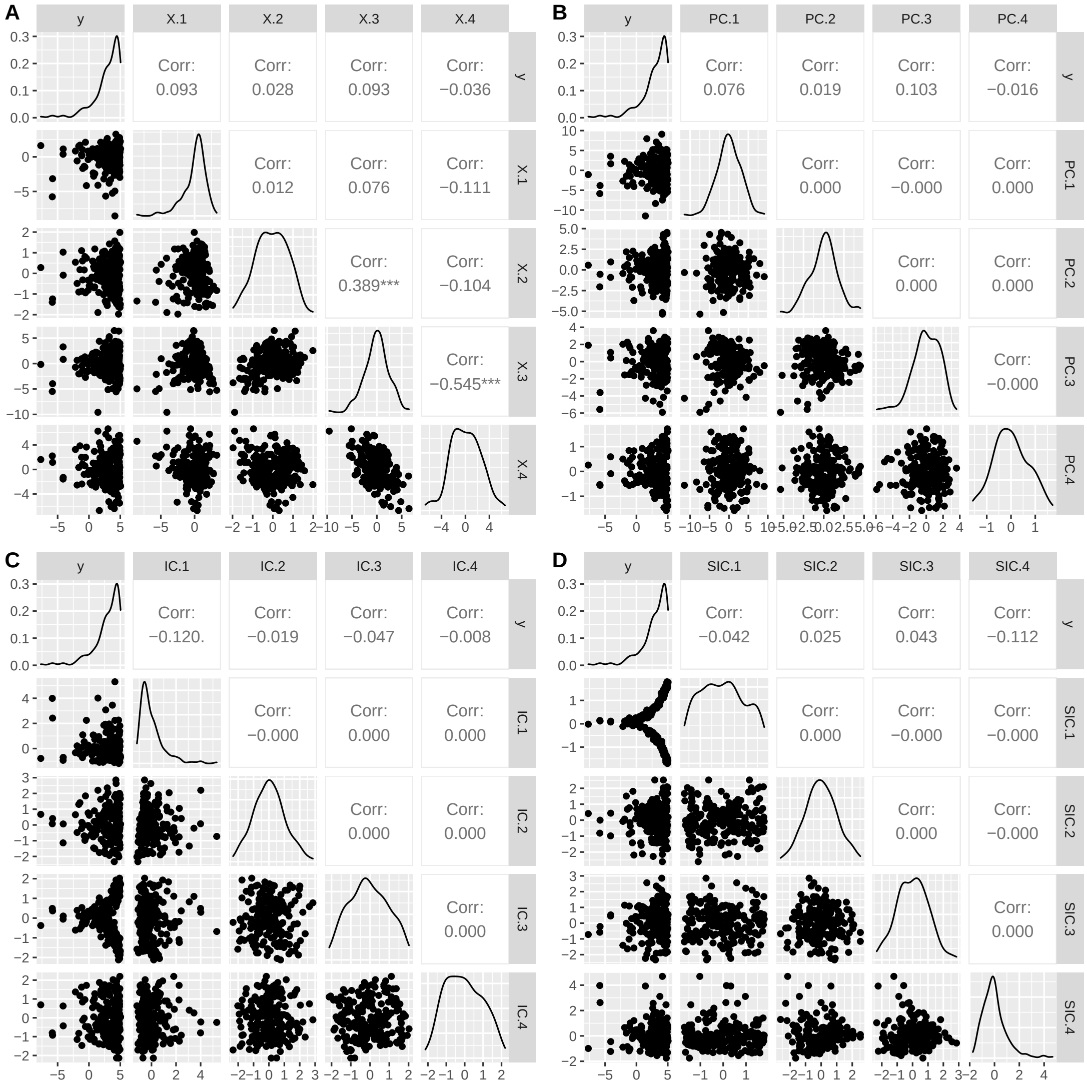

Fig. 3: Comparison of different data transformations. Panel A shows a matrix scatterplot with density estimators on the diagonal and correlations for the original data and where Y is the response to be modeled by the 4 predictors, Panel B the principal components based on $\boldsymbol{X}_n$, Panel C the invariant coordinates (FOBI) based on $\boldsymbol{X}_n$ and Panel D the supervised invariant coordinates based on $\boldsymbol{X}_n$ and Y. Clearly the supervised components make it easiest to see a relationship between reponse and predictors.

Table 1: Multivariate methods which are based on the joint diagonalization of two or more scatter matrices.

Name	Family	Primary data type	Scatters
ICS Tyler et al. [2009]	ICS	non-elliptical iid data	two diffe
PAA Critchley et al. [2006]	ICS	non-elliptical iid data	$\boldsymbol{S}_1 =$ Co
LDA Fisher [1936]	SDR	multigroup iid data	$\boldsymbol{S}_1 =$ Co
CCA Hotelling [1936]	SDR	two group data	$\boldsymbol{S}_1 =$ Co
FOBI Cardoso [1989], Nordhausen and Virta [2019]	ICS, ICA	non-elliptical / ICA iid data	$\boldsymbol{S}_1 =$ Co
2-Scatter-ICA Oja et al. [2006]	ICA	ICA iid data	Two sca erty
k-Scatter-ICA Nordhausen et al. [2012]	ICA	ICA iid data	k scatter
3-scatter-ISA Nordhausen and Oja [2011]	ISA	ISA iid data	Three d (block) i
SICS Liski et al. [2014]	SDR	regression data	a unsupe functiona
SIR Li [1991]	SDR	regression data	$\boldsymbol{S}_1 =$ Co
SAVE Cook [2000]	SDR	regression data	$\boldsymbol{S}_1 =$ Co
pHd Li [1992]	SDR	regression data	$\boldsymbol{S}_1 =$ Co
DR Li and Wang [2007]	SDR	regression data	$\boldsymbol{S}_1 =$ Co
TSIR Matilainen et al. [2017]	SDR	time series regression data	$\boldsymbol{S}_1 =$ Co
TSAVE Matilainen et al. [2019]	SDR	time series regression data	$\boldsymbol{S}_1 =$ Co
SSIR Muehlmann et al. [2021c]	SDR	spatial regression data	$\boldsymbol{S}_1 =$ Co
AMUSE Miettinen et al. [2012], Tong et al. [1990]	SOS	stationary time series	$\boldsymbol{S}_1 =$ Co
SOBI Belouchrani et al. [1997], Miettinen et al. [2015, 2016]	SOS	stationary time series	$\boldsymbol{S}_1 =$ Co
gFOBI Matilainen et al. [2015]	IC-time series	time series with for example stochastic volatilty	$\boldsymbol{S}_1 =$ Co ces
NSS.SD Choi and Cichocki [2000b]	NSS	non-stationary time series	2 Covs
NSS.JD Choi and Cichocki [2000b]	NSS	non-stationary time series	K+1 Co
NSS.TD.JD Choi and Cichocki [2000a]	NSS	block stationary time series	$K \times L$ A
SBSS Bachoc et al. [2020], Nordhausen et al. [2015]	SBSS	stationary spatial data	Cov and
SNSS.SD Muehlmann et al. [2021a]	SNSS	non-stationary spatial data	2 Covs
SNSS.JD Muehlmann et al. [2021a]	SNSS	non-stationary spatial data	K+1 Co
SNSS.TD.JD Muehlmann et al. [2021a]	SNSS	block stationary spatial data	$K \times L$ L

were not considered here, such as for example common principal component analysis Flury [1988]. Additionally, the methods have been extended in several directions to tackle more complex data, such as tensors, functional data or composition data [see for example Virta et al., 2017, 2020, Muehlmann et al., 2021b], and similarities and different approaches are discussed in Cook [2021], Nordhausen et al. [2021c], Fischer et al. [2020]. Finally, let us mention the problem of high-dimensional data and the sparsity question that needs further development.

Acknowledgements

We are grateful for the comments which helped to improve the paper and we are honoured to have the opportunity to contribute to this special issue and celebrate 50 years of JMVA!

The work of KN was supported by the Austrian Science Fund (FWF) under grant P31881-N32. ARG acknowledges funding from ANR under grant ANR-17-EURE-0010 (Investissements d'Avenir program).

References

T. Adali, M. Anderson, and G.-S. Fu. Diversity in independent component and vector analyses: Identifiability, algorithms, and applications in medical imaging. *IEEE Signal Processing Magazine*, 31:18–33, 2014. doi: 10.1109/MSP.2014.2300511.

F. Alashwali and J. T. Kent. The use of a common location measure in the invariant coordinate selection and projection pursuit. *Journal of Multivariate Analysis*, 152:145–161, 2016. doi: 10.1016/j.jmva.2016.08.007.

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

A. Archimbaud, J. May, K. Nordhausen, and A. Ruiz-Gazen. *ICSShiny: ICS via a Shiny Application*, 2018. URL `https://CRAN.R-project.org/package=ICSShiny`. R package version 0.5.

A. Archimbaud, K. Nordhausen, and A. Ruiz-Gazen. ICS for multivariate outlier detection with application to quality control. *Computational Statistics & Data Analysis*, 128:184–199, 2018.

A. Archimbaud, K. Nordhausen, and A. Ruiz-Gazen. Unsupervized outlier detection with `ICSOutlier`. *The R Journal*, 10(1):234–250, 2018.

F. Bachoc, M. G. Genton, K. Nordhausen, A. Ruiz-Gazen, and J. Virta. Spatial blind source separation. *Biometrika*, 107:627–646, 2020.

A. Belouchrani, K. Abed Meraim, J.-F. Cardoso, and E. Moulines. A blind source separation technique based on second order statistics. *IEEE Transactions on Signal Processing*, 45:434–444, 1997.

M. Bilodeau and D. Brenner. *Theory of Multivariate Statistics.* Springer, New York, 2008.

E. Bura and R. Cook. Extending sliced inverse regression: The weighted chi-squared test. *Journal of the American Statistical Association*, 96:996–1003, 2001.

E. Bura and J. Yang. Dimension estimation in sufficient dimension reduction: A unifying approach. *Journal of Multivariate Analysis*, 102(1):130–142, 2011.

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

J.-F. Cardoso and A. Souloumiac. Jacobi angles for simultaneous diagonalization. *SIAM Journal on Matrix Analysis and Applications*, 17:161–164, 1996.

H. Caussinus and A. Ruiz. Interesting projections of multidimensional data by means of generalized principal component analyses. In K. Momirović and V. Mildner, editors, *Compstat*, pages 121–126, Heidelberg, 1990. Physica-Verlag HD.

H. Caussinus and A. Ruiz-Gazen. Classification and generalized principal component analysis. In P. Brito, G. Cucumel, P. Bertrand, and F. de Carvalho, editors, *Selected Contributions in Data Analysis and Classification*, pages 539–548. Springer, Berlin, 2007.

H. Caussinus, M. Fekri, S. Hakam, and A. Ruiz-Gazen. A monitoring display of multivariate outliers. *Computational Statistics & Data Analysis*, 44(1):237–252, 2003.

G. Chabriel, M. Kleinsteuber, E. Moreau, H. Shen, P. Tichavsky, and A. Yeredor. Joint matrices decompositions and blind source separation: A survey of methods, identification, and applications. *IEEE Signal Processing Magazine*, 31(3):34–43, 2014.

S. Choi and A. Cichocki. Blind separation of nonstationary and temporally correlated sources from noisy mixtures. In *Neural Networks for Signal Processing X. Proceedings of the 2000 IEEE Signal Processing Society Workshop*, volume 1, pages 405–414. IEEE, 2000a.

S. Choi and A. Cichocki. Blind separation of nonstationary sources in noisy mixtures. *Electronics Letters*, 36:848–849, 2000b.

A. Cichocki and S.-I. Amari. *Adaptive Blind Signal and Image Processing: Learning Algorithms and Applications*. John Wiley & Sons, New York, 2002. doi: 10.1002/0470845899.

D. Clarkson. A least squares version of algorithm AS 211: The F-G diagonalization algorithm. *Applied Statistics*, 37:317–321, 1988.

P. Comon and C. Jutten. *Handbook of Blind Source Separation: Independent Component Analysis and Applications*. Academic Press, Oxford, 2010. doi: 10.1016/C2009-0-19334-0.

R. D. Cook. SAVE: A method for dimension reduction and graphics in regression. *Communications in Statistics - Theory and Methods*, 29:2109–2121, 2000.

R. D. Cook. A slice of multivariate dimension reduction. *Journal of Multivariate Analysis*, (online first):104812, 2021. doi: https://doi.org/10.1016/j.jmva.2021.104812.

F. Critchley, A. Pires, and C. Amado. Principal axis analysis. Technical Report 06/14, The Open University Milton Keynes, 2006. URL `http://stats-www.open.ac.uk/technicalreports/PAA.pdf`.

C. Croux and G. Haesbroeck. Principal component analysis based on robust estimators of the covariance or correlation matrix: influence functions and efficiencies. *Biometrika*, 87(3):603–618, 2000.

M. Fekri and A. Ruiz-Gazen. A b-robust non-iterative scatter matrix estimator: Asymptotics and application to cluster detection using invariant coordinate selection. In K. Nordhausen and S. Taskinen, editors, *Modern Nonparametric, Robust and Multivariate Methods: Festschrift in Honour of Hannu Oja*, pages 395–423. Springer International Publishing, Cham, 2015.

D. Fischer, M. Honkatukia, M. Tuiskula-Haavisto, K. Nordhausen, D. Cavero, R. Preisinger, and J. Vilkki. Subgroup detection in genotype data using invariant coordinate selection. *BMC Bioinformatics*, 18:173–181, 2017.

D. Fischer, K. Nordhausen, and H. Oja. On linear dimension reduction based on diagonalization of scatter matrices for bioinformatics downstream analyses. *Heliyon*, 6:e05732, 2020. doi: https://doi.org/10.1016/j.heliyon.2020.e05732.

R. A. Fisher. The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7(2): 179–188, 1936. doi: https://doi.org/10.1111/j.1469-1809.1936.tb02137.x.

B. Flury. *Common Principal Components & Related Multivariate Models.* John Wiley & Sons, Chichester, 1988.

B. N. Flury and W. Gautschi. An algorithm for simultaneous orthogonal transformation of several positive definite symmetric matrices to nearly diagonal form. *SIAM Journal on Scientific and Statistical Computing*, 7(1):169–184, 1986.

H. Hotelling. Relations between two sets of variates. *Biometrika*, 28:321–377, 1936. doi: 10.1093/biomet/28.3-4.321.

P. Huber and E. Ronchetti. *Robust Statistics.* Wiley, Hoboken, 2011. ISBN 9781118210338.

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

K. Illner, J. Miettinen, C. Fuchs, S. Taskinen, K. Nordhausen, H. Oja, and F. J. Theis. Model selection using limiting distributions of second-order blind source separation algorithms. *Signal Processing*, 113:95–103, 2015.

P. Ilmonen, J. Nevalainen, and H. Oja. Characteristics of multivariate distributions and the invariant coordinate system. *Statistics & Probability Letters*, 80(23):1844–1853, 2010.

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

P. Ilmonen, K. Nordhausen, H. Oja, and F. Theis. An affine equivariant robust second-order BSS method. In E. Vincent, A. Yeredor, Z. Koldovský, and P. Tichavský, editors, *Latent Variable Analysis and Signal Separation. LVA/ICA 2015. Lecture Notes in Computer Science*, volume 9237, pages 328–335, Cham, 2015. Springer. doi: 10.1007/978-3-319-22482-4_38.

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

A. Kankainen, S. Taskinen, and H. Oja. Tests of multinormality based on location vectors and scatter matrices. *Statistical Methods & Applications*, 16:357–379, 2007.

B. Li. *Sufficient Dimension Reduction Methods and Applications with R.* Chapman and Hall/CRC, Boca Raton, 2018.

B. Li and S. Wang. On directional regression for dimension reduction. *Journal of the American Statistical Association*, 102(479):997–1008, 2007.

K.-C. Li. Sliced inverse regression for dimension reduction. *Journal of the American Statistical Association*, 86(414):316–327, 1991.

K.-C. Li. On principal Hessian directions for data visualization and dimension reduction: Another application of Stein's lemma. *Journal of the American Statistical Association*, 87(420):1025–1039, 1992.

E. Liski, K. Nordhausen, and H. Oja. Supervised invariant coordinate selection. *Statistics: A Journal of Theoretical and Applied Statistics*, 4:711–731, 2014.

N. Loperfido. Some theoretical properties of two kurtosis matrices, with application to invariant coordinate selection. *Journal of Multivariate Analysis*, page 104809, 2021.

W. Luo and B. Li. Combining eigenvalues and variation of eigenvectors for order determination. *Biometrika*, 103(4):875–887, 2016.

W. Luo and B. Li. On order determination by predictor augmentation. *Biometrika*, 108:557–574, 2021. doi: 10.1093/biomet/asaa077.

Y. Ma and L. Zhu. A review on dimension reduction. *International Statistical Review*, 81(1):134–150, 2013. doi: 10.1111/j.1751-5823.2012.00182.x.

K. Mardia, J. Kent, and J. Bibby. *Multivariate Analysis*. Academic Press, London, 1979. ISBN 9780124712508.

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

R. A. Maronna and V. J. Yohai. Robust estimation of multivariate location and scatter. In N. Balakrishnan, T. Colton, B. Everitt, W. Piegorsch, F. Ruggeri, and J. Teugels, editors, *Wiley StatsRef: Statistics Reference Online*, pages 1–12. Wiley, 2016. doi: https://doi.org/10.1002/9781118445112.stat01520.pub2. URL `https://onlinelibrary.wiley.com/doi/abs/10.1002/9781118445112.stat01520.pub2`.

R. A. Maronna, R. D. Martin, V. J. Yohai, and M. Salibián-Barrera. *Robust Statistics: Theory and methods (with R)*. John Wiley & Sons, New York, 2019.

M. Matilainen, K. Nordhausen, and H. Oja. New independent component analysis tools for time series. *Statistics & Probability Letters*, 105:80–87, 2015.

M. Matilainen, C. Croux, K. Nordhausen, and H. Oja. Supervised dimension reduction for multivariate time series. *Econometrics and Statistics*, 4:57–69, 2017.

M. Matilainen, C. Croux, K. Nordhausen, and H. Oja. Sliced average variance estimation for multivariate time series. *Statistics*, 53:630–655, 2019. doi: 10.1080/02331888.2019.1605515.

J. Miettinen. Alternative diagonality criteria for SOBI. In K. Nordhausen and S. Taskinen, editors, *Modern Nonparametric, Robust and Multivariate Methods: Festschrift in Honour of Hannu Oja*, pages 455–469. Springer, Cham, 2015.

J. Miettinen, K. Nordhausen, H. Oja, and S. Taskinen. Statistical properties of a blind source separation estimator for stationary time series. *Statistics & Probability Letters*, 82(11):1865–1873, 2012.

J. Miettinen, K. Nordhausen, H. Oja, and S. Taskinen. Deflation-based separation of uncorrelated stationary time series. *Journal of Multivariate Analysis*, 123:214–227, 2014.

J. Miettinen, S. Taskinen, K. Nordhausen, and H. Oja. Fourth moments and independent component analysis. *Statistical Science*, 30(3):372–390, 08 2015. doi: 10.1214/15-STS520. URL `https://doi.org/10.1214/15-STS520`.

J. Miettinen, K. Illner, K. Nordhausen, H. Oja, S. Taskinen, and F. Theis. Separation of uncorrelated stationary time series using autocovariance matrices. *Journal of Time Series Analysis*, 37:337–354, 2016.

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. doi: 10.18637/jss.v076.i02.

J. Miettinen, M. Matilainen, K. Nordhausen, and S. Taskinen. Extracting conditionally heteroskedastic components using independent component analysis. *Journal of Time Series Analysis*, 41: 293–311, 2020. doi: https://doi.org/10.1111/jtsa.12505.

C. Muehlmann, K. Nordhausen, and M. Yi. On cokriging, neural networks, and spatial blind source separation for multivariate spatial prediction. *IEEE Geoscience and Remote Sensing Letters*, pages 1–5, 2020. doi: 10.1109/LGRS.2020.3011549.

C. Muehlmann, F. Bachoc, and K. Nordhausen. Spatial nonstationary source separation. *Arxiv*, page https://arxiv.org/abs/2107.01916, 2021a.

C. Muehlmann, K. Fačevicová, A. Gardlo, H. Janečková, and K. Nordhausen. Independent component analysis for compositional data. In A. Daouia and A. Ruiz-Gazen, editors, *Advances in Contemporary Statistics and Econometrics: Festschrift in Honor of Christine Thomas-Agnan*, pages 525–545. Springer, Cham, 2021b. doi: 10.1007/978-3-030-73249-3_27.

C. Muehlmann, K. Nordhausen, and H. Oja. Sliced inverse regression for spatial data. In E. Bura and B. Li, editors, *Festschrift in Honor of R. Dennis Cook: Fifty Years of Contribution to Statistical Science*, pages 87–107. Springer, Cham, 2021c.

C. Muehlmann, K. Nordhausen, and J. Virta. *`SpatialBSS`: Blind Source Separation for Multivariate Spatial Data*, 2021d. URL `https://CRAN.R-project.org/package=SpatialBSS`. R package version 0.11-0.

K. Nordhausen. On robustifying some second order blind source separation methods for nonstationary time series. *Statistical Papers*, 55(1):141–156, 2014.

K. Nordhausen and H. Oja. Scatter matrices with independent block property and ISA. In *2011 19th European Signal Processing Conference*, pages 1738–1742. IEEE, 2011.

K. Nordhausen and H. Oja. Independent component analysis: A statistical perspective. *WIREs: Computational Statistics*, 10:e1440, 2018. doi: 10.1002/wics.1440.

K. Nordhausen and D. E. Tyler. A cautionary note on robust covariance plug-in methods. *Biometrika*, 102(3):573–588, 2015. ISSN 0006-3444. doi: 10.1093/biomet/asv022.

K. Nordhausen and J. Virta. An overview of properties and extensions of FOBI. *Knowledge-Based Systems*, 173:113–116, 2019.

K. Nordhausen, H. Oja, and D. E. Tyler. On the efficiency of invariant multivariate sign and rank test. In E. P. Liski, J. Isotalo, J. Niemelä, S. Puntanen, and G. P. H. Styan, editors, *Festschrift for Tarmo Pukkila on his 60th Birthday*, pages 217–231. University of Tampere, Tampere, 2006.

K. Nordhausen, H. Oja, and E. Ollila. Robust independent component analysis based on two scatter matrices. *Austrian Journal of Statistics*, 37:91–100, 2008a.

K. Nordhausen, H. Oja, and D. E. Tyler. Tools for exploring multivariate data: The package `ICS`. *Journal of Statistical Software*, 28:1–31, 2008b.

K. Nordhausen, H. Oja, and E. Ollila. Multivariate models and the first four moments. In *Nonparametric Statistics and Mixture Models*, pages 267–287. World Scientific, Hackensack, 2011a. doi: 10.1142/9789814340564_0016.

K. Nordhausen, E. Ollila, and H. Oja. On the performance indices of ICA and blind source separation. In *2011 IEEE 12th International Workshop on Signal Processing Advances in Wireless Communications*, pages 486–490, 2011b. doi: 10.1109/SPAWC.2011.5990458.

K. Nordhausen, H. W. Gutch, H. Oja, and F. J. Theis. Joint diagonalization of several scatter matrices for ICA. In F. Theis, A. Cichocki, A. Yeredor, and M. Zibulevsky, editors, *Latent Variable Analysis and Signal Separation: 10th International Conference*, pages 172–179, Berlin, 2012. Springer.

K. Nordhausen, H. Oja, P. Filzmoser, and C. Reimann. Blind source separation for spatial compositional data. *Mathematical Geosciences*, 47(7):753–770, 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, G. Fischer, and P. Filzmoser. Blind source separation for compositional time series. *Mathematical Geosciences*, 53:905–924, 2021a. doi: 10.1007/s11004-020-09869-y.

K. Nordhausen, M. Matilainen, J. Miettinen, J. Virta, and S. Taskinen. Dimension reduction for time series in a blind source separation context using R. *Journal of Statistical Software*, 98:1–30, 2021b. doi: 10.18637/jss.v098.i15.

K. Nordhausen, H. Oja, and D. E. Tyler. Asymptotic and bootstrap tests for subspace dimension. *Journal of Multivariate Analysis*, (online first):104830, 2021c. doi: https://doi.org/10.1016/j.jmva.2021.104830.

K. Nordhausen, H. Oja, D. E. Tyler, and J. Virta. *`ICtest`: Estimating and Testing the Number of Interesting Components in Linear Dimension Reduction*, 2021d. URL `https://CRAN.R-project.org/package=ICtest`. R package version 0.3-4.

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. doi: 10.17713/ajs.v35i2\&3.364.

Y. Pan, M. Matilainen, S. Taskinen, and K. Nordhausen. A review of second-order blind identification methods. *WIREs Computational Statistics*, n/a:e1550, 2021. doi: https://doi.org/10.1002/wics.1550.

D. Peña, F. J. Prieto, and J. Viladomat. Eigenvectors of a kurtosis matrix as interesting directions to reveal cluster structure. *Journal of Multivariate Analysis*, 101(9):1995–2007, 2010.

M. L. Puri and P. K. Sen. *Nonparametric Methods in Multivariate Analysis.* John Wiley & Sons, New York, USA, 1971.

U. Radojicic and K. Nordhausen. Non-gaussian component analysis: Testing the dimension of the signal subspace. In M. Maciak, M. Pesta, and M. Schindler, editors, *Analytical Methods in Statistics. AMISTAT 2019*, pages 101–123. Springer, Cham, 2020.

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

R Core Team. *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria, 2021. URL `https://www.R-project.org/`.

J. R. Schott. *Matrix Analysis for Statistics.* John Wiley & Sons, Hoboken, 2005.

R. Serfling. Equivariance and invariance properties of multivariate quantile and related functions, and the role of standardisation. *Journal of Nonparametric Statistics*, 22:915–936, 2010.

R. Serfling. On invariant within equivalence coordinate system (IWECS) transformations. In K. Nordhausen and S. Taskinen, editors, *Modern Nonparametric, Robust and Multivariate Methods: Festschrift in Honour of Hannu Oja*, pages 445–457. Springer International Publishing, Cham, 2015.

A. C. Tang, J.-Y. Liu, and M. T. Sutherland. Recovery of correlated neuronal sources from EEG: The good and bad ways of using SOBI. *NeuroImage*, 28:507–519, 2005. doi: https://doi.org/10.1016/j.neuroimage.2005.06.062.

S. Taskinen, J. Miettinen, and K. Nordhausen. A more efficient second order blind identification method for separation of uncorrelated stationary time series. *Statistics & Probability Letters*, 116: 21–26, 2016. doi: https://doi.org/10.1016/j.spl.2016.04.007.

F. J. Theis and Y. Inouye. On the use of joint diagonalization in blind signal processing. In *IEEE International Symposium on Circuits and Systems*, pages 3589–3593. IEEE, 2006.

L. Tong, V. Soon, Y. Huang, and R. Liu. AMUSE: A new blind identification algorithm. In *Proceedings of IEEE International Symposium on Circuits and Systems*, pages 1784–1787. IEEE, 1990.

D. E. Tyler, F. Critchley, L. Dümbgen, and H. Oja. Invariant coordinate selection. *Journal of the Royal Statistical Society: Series B*, 71(3):549–592, 2009.

J. Virta. One-step M-estimates of scatter and the independence property. *Statistics & Probability Letters*, 110:133–136, 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. Independent component analysis for multivariate functional data. *Journal of Multivariate Analysis*, 176:104568, 2020.

S. Weisberg. Dimension reduction regression in R. *Journal of Statistical Software*, 7(1):1–22, 2002.

A. Yeredor. Non-orthogonal joint diagonalization in the least-squares sense with application in blind source separation. *IEEE Transactions on Signal Processing*, 50(7):1545–1553, 2002. doi: 10.1109/TSP.2002.1011195.

A. Ziehe, P. Laskov, G. Nolte, and K.-R. Müller. A fast algorithm for joint diagonalization with non-orthogonal transformations and its application to blind source separation. *Journal of Machine Learning Research*, 5(Jul):777–800, 2004.