Structured Nonlinear Discriminant Analysis

Christopher Bonenberger[1,2(✉)][0000−0003−3186−3450], Wolfgang Ertel[1], Markus Schneider[1][0000−0003−0543−2593], and Friedhelm Schwenker[2][0000−0001−5118−0812]

[1] Ravensburg-Weingarten University of Applied Sciences (Institute for Artificial Intelligence), Weingarten, Germany
bonenbch@rwu.de

[2] University of Ulm (Institute of Neural Information Processing), James-Franck-Ring, 89081 Ulm, Germany

Abstract. Many traditional machine learning and pattern recognition algorithms—as for example linear discriminant analysis (LDA) or principal component analysis (PCA)—optimize data representation with respect to an information theoretic criterion. For time series analysis these traditional techniques are typically insufficient. In this work we propose an extension to linear discriminant analysis that allows to learn a data representation based on an algebraic structure that is tailored for time series. Specifically we propose a generalization of LDA towards shift-invariance that is based on cyclic structures. We expand this framework towards more general structures, that allow to incorporate previous knowledge about the data at hand within the representation learning step. The effectiveness of this proposed approach is demonstrated on synthetic and real-world data sets. Finally, we show the interrelation of our approach to common machine learning and signal processing techniques.

Keywords: Linear Discriminant Analysis · Time Series Analysis · Circulant Matrices · Representation Learning · Algebraic Structure.

1 Introduction

Often, when being confronted with temporal data, machine learning practitioners use feature transformations. Yet, mostly these feature transformations are not adaptive but rely on decomposition with respect to a fixed basis (Fourier, Wavelet, etc.). This is because simple data-adaptive methods as principal component analysis (PCA, [12]) or linear discriminant analysis (LDA, [22]) often lead to undesirable results for stationary time series [20]. Both methods, PCA and LDA, are based on successive projections onto optimal one-dimensional subspaces. For time series analysis via LDA this leads to problems, especially when the data at hand is not locally coherent in the corresponding vector space [10]—which is likely to be the case for high-dimensional (long) time series. In this paper we propose an adaption of LDA that relies on learning with algebraic structures. More precisely, we propose to learn a projection onto a structured multi-dimensional subspace instead of a single vector. The scope of this work is mainly to introduce the theoretical basics of *structured discriminant analysis* for time series, i.e., we focus on cyclic structures that incorporate shift-invariance and thus regularize supervised representation learning.

From an algebraic point of view the problem at hand is mainly an issue of data representation in terms of bases and frames [23]. In this setting we seek a basis (or a frame [5, 15]) for the input space, which is optimal with respect to some information-theoretic criterion. When it comes to time series the vital point is, that these algorithms can be tailored to meet the conditions of temporal data. *Basis pursuit* methods like dictionary learning (DL, optimize over-complete representations with respect to sparsity and reconstruction error) are often altered in order to yield shift-invariance and to model temporal dependencies [21, 17, 8]. Also convolutional neural networks are implicitly equipped with a mechanism to involve algebraic structure in the learning process, because this way "the architecture itself realizes a form of regularization" [3]. In fact both, shift-invariant DL [21, 8] and CNN, use cyclic structures, i.e., convolutions.

However, so far the idea of learning with cyclic structures has hardly been transferred to basic machine learning methods. The motive of this work is to transfer the idea of implicit shift-invariance to LDA by learning with algebraic structure. We strive for interpretable algorithms that go along with low computational complexity. This way we seek to bridge complex methods like convolutional neural networks and simple, well-understood techniques like LDA.

Recently [1] proposed a generalization of PCA that allows unsupervised representation learning with algebraic structure, which is tightly linked to methods like dynamic PCA [13], singular spectrum analysis [9] and spectral density estimation [1]. However, similarly to PCA this method does not allow to incorporate labeling information. Yet, representation learning can benefit from class-information. In this respect our main contribution is a formulation of linear discriminant analysis that involves cyclic structures, thus being optimized for stationary temporal data. We provide a generalization of this framework towards non-stationary time series and even arbitrary correlation structures. Moreover, the proposed technique is linked to classical signal processing methods.

2 Prerequisites

In the following we will briefly discuss the underlying theory of linear discriminant analysis, circulant matrices and linear filtering. In Section 2.2 we revisit principal component analysis and its generalization towards shift-invariance.

2.1 Circulant Matrices

We define a circulant matrix as a matrix of the form

$$\mathbf{G} = \begin{bmatrix} g_1 & g_D & g_{D-1} & g_{D-2} & \cdots & g_2 \\ g_2 & g_1 & g_D & g_{D-1} & \cdots & g_3 \\ g_3 & g_2 & g_1 & g_D & \cdots & g_4 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ g_D & g_{D-1} & g_{D-2} & g_{D-3} & \cdots & g_1 \end{bmatrix} \in \mathbb{R}^{D \times D}, \tag{1}$$

i.e., the circulant $\mathbf{G}$ is fully defined by its first column vector $\mathbf{g}$. In short, the i-th row of a circulant matrix contains the first row right-shifted by $i-1$. In the following, we write circulant matrices as a matrix polynomial of the form

$$\mathbf{G} = \sum_{l=0}^{L} g_l \mathbf{P}^{l-1} \tag{2}$$

with

$$\mathbf{P} = \begin{bmatrix} 0 & 0 & 0 & \cdots & 1 \\ 1 & 0 & 0 & & 0 \\ 0 & 1 & 0 & & 0 \\ \vdots & & & \ddots & \vdots \\ 0 & 0 & \cdots & 1 & 0 \end{bmatrix} \in \mathbb{R}^{D \times D}. \tag{3}$$

Note that $\mathbf{P}$ itself is also a circulant matrix.

Left-multiplying a signal $\mathbf{x}$ with a circulant matrix is equivalent to

$$\mathbf{G}\mathbf{x} = \mathbf{F}^{-1}\boldsymbol{\Lambda}\mathbf{F}\mathbf{x} \tag{4}$$

where $\mathbf{F} \in \mathbb{R}^{D \times D}$ is the Fourier matrix with coefficients

$$[\mathbf{F}]_{j,k} = \frac{1}{\sqrt{D}} \exp\left(-2\pi\mathfrak{i}\frac{(j-1)(k-1)}{D}\right) \tag{5}$$

where $\boldsymbol{\Lambda}$ is a diagonal matrix with the Fourier transform $\hat{\mathbf{g}} = \mathbf{F}\mathbf{g}$ of $\mathbf{g} \in \mathbb{R}^D$ on its diagonal and $\mathfrak{i}$ is the complex number, i.e., $\mathfrak{i}^2 = -1$. Hence, Eq. (4) describes a circular convolution

$$\mathbf{G}\mathbf{x} = \mathbf{F}^{-1}\boldsymbol{\Lambda}\hat{\mathbf{x}} = \mathbf{F}^{-1}(\hat{\mathbf{g}} \odot \hat{\mathbf{x}}) = \mathbf{g} \circledast \mathbf{x}.$$

Here, $\odot$ is the Hadamard product (pointwise multiplication) and $\circledast$ denotes the discrete circular convolution. Moreover, Eq. (4) describes the diagonalization of circulant matrices by means of the Fourier matrix.

2.2 (Circulant) Principal Component Analysis

Heading towards linear discriminant analysis, it is interesting to start with the Rayleigh quotient and its role in PCA (cf. [19]). The Rayleigh quotient of some vector $\mathbf{g} \in \mathbb{R}^D$ with respect to a symmetric matrix $\mathbf{S} \in \mathbb{R}^{D \times D}$ is defined as

$$\mathcal{R}(\mathbf{g}, \mathbf{S}) = \frac{\mathbf{g}^\mathsf{T}\mathbf{S}\mathbf{g}}{\mathbf{g}^\mathsf{T}\mathbf{g}}.$$

Maximizing $\mathcal{R}(\mathbf{g}, \mathbf{S})$ with respect to $\mathbf{g}$ leads to the eigenvalue problem $\mathbf{S}\mathbf{g} = \lambda\mathbf{g}$. The optimal vector $\mathbf{g}$ is the eigenvector of $\mathbf{S}$ with the largest corresponding eigenvalue.

Having a labeled data set $\{(\mathbf{x}_\nu, y_\nu)\}_{\nu=1,\ldots,N}$ with observations $\mathbf{x} \in \mathbb{R}^D$ and corresponding labels $y \in \{1, \ldots, C\}$, we define the overall data matrix as

$$\mathbf{X} = \begin{bmatrix} | & & | \\ \mathbf{x}_1 & \cdots & \mathbf{x}_N \\ | & & | \end{bmatrix} \in \mathbb{R}^{D \times N}.$$

Moreover we define class-specific data matrices $\mathbf{X}_c \in \mathbb{R}^{D \times N_c}$, where N_c is the number of observations $\mathbf{x}_\nu$ with corresponding label $y_\nu = c$.

The relation to principal component analysis becomes obvious when $\mathbf{S}$ is the empirical covariance matrix estimated from $\mathbf{X}$. Assuming zero-mean data, i.e., the expected value $\mathbb{E}\{\mathbf{x}\} = 0$, the matrix $\mathbf{S} = \mathbf{X}\mathbf{X}^\mathsf{T}$ is the empirical covariance matrix of $\mathbf{X}$. Thus

$$\max_{\mathbf{g} \in \mathbb{R}^D} \{\mathcal{R}(\mathbf{g}, \mathbf{S})\}$$

is equivalent to the linear constrained optimization problem

$$\max_{\mathbf{g} \in \mathbb{R}^D} \left\{ \left\| \mathbf{g}^\mathsf{T} \mathbf{X} \right\|_2^2 \right\} \text{ s.t. } \|\mathbf{g}\|_2^2 = 1, \tag{6}$$

which in turn formulates principal component analysis, where maximizing $\left\| \mathbf{g}^\mathsf{T} \mathbf{X} \right\|_2^2$ means to maximize variance (respectively power). As known the optimal *principal component vector(s)* are found from the eigenvalue problem (cf. [12])

$$\mathbf{S}\mathbf{g} = \lambda \mathbf{g}.$$

While classical PCA is based on a projection onto an optimal one-dimensional subspace [1] proposed a generalization of PCA which projects on a multi-dimensional subspace that is formed from cyclic permutations. This results in optimizing

$$\max_{\mathbf{g} \in \mathbb{R}^D} \left\{ \|\mathbf{G}\mathbf{X}\|_F^2 \right\} \text{ s.t. } \|\mathbf{g}\|_2^2 = 1, \tag{7}$$

with $\mathbf{G}$ being a κ-circulant matrix defined by the elements of $\mathbf{g}$ (see Section 2.1). The Frobenius norm $\|\mathbf{A}\|_F^2 = \mathrm{tr}\{\mathbf{A}^\mathsf{T}\mathbf{A}\}$. Solving Eq. (7) amounts to set the partial derivatives of the Lagrangian function

$$\mathbf{L}(\mathbf{g}, \lambda) = \mathrm{tr}\{\mathbf{X}^\mathsf{T}\mathbf{G}^\mathsf{T}\mathbf{G}\mathbf{X}\} + \lambda \left(\mathbf{g}^\mathsf{T}\mathbf{g} - 1\right)$$

to zero. Analogously to PCA this finally leads to the eigenvalue problem (see [1])

$$\mathbf{Z}\mathbf{g} = \lambda \mathbf{g}$$

with $[\mathbf{Z}]_{k,l} = \sum_\nu \mathbf{x}_\nu^\mathsf{T} \mathbf{P}^{l-k} \mathbf{x}_\nu$ using $\mathbf{P}$ as defined in Eq. (3), i.e., we write $\mathbf{G}$ as in Eq. (2).

2.3 Linear Discriminant Analysis

Relying on a decomposition of the overall empirical covariance matrix without using the class labels can result in a disadvantageous data representation. Yet, linear discriminant analysis exploits labeling information by maximizing the generalized Rayleigh quotient (cf. [22])

$$\mathcal{R}(\mathbf{g}, \mathbf{B}, \mathbf{W}) = \frac{\mathbf{g}^{\mathsf{T}}\mathbf{B}\mathbf{g}}{\mathbf{g}^{\mathsf{T}}\mathbf{W}\mathbf{g}}. \tag{8}$$

with the beneath-class scatter matrix

$$\mathbf{B} = \sum_{c=1}^{C} P_c \left(\overline{\mathbf{x}}_c - \overline{\mathbf{x}}_0\right) \left(\overline{\mathbf{x}}_c - \overline{\mathbf{x}}_0\right)^{\mathsf{T}} \tag{9}$$

and the within-class scatter matrix

$$\mathbf{W} = \sum_{c=1}^{C} P_c \left(\sum_{\nu \in \mathcal{I}_c} \left(\mathbf{x}_\nu - \overline{\mathbf{x}}_c\right) \left(\mathbf{x}_\nu - \overline{\mathbf{x}}_c\right)^{\mathsf{T}} \right), \tag{10}$$

where $\mathcal{I}_c$ is the index set for class c, i.e., $\mathcal{I}_c = \{\nu \in [1, \ldots, N] \,|\, y_\nu = c\}$. Moreover $\overline{\mathbf{x}}_c$ is the sample mean of observations from the class c and $\overline{\mathbf{x}}_0$ is the overall empirical mean value. The a priori class probabilities P_c have to be estimated as $P_c \approx N_c/N$. Note that the beneath-class scatter matrix is the empirical covariance matrix of class-specific sample mean values, while the within-class scatter matrix is a sum of the class-specific covariances. While typically $\operatorname{rank}\{\mathbf{W}\} = D$ the rank of the beneath-class scatter matrix is $\operatorname{rank}\{\mathbf{B}\} \leq C - 1$.

The expression in Eq. (8), also known as Fisher's criterion, measures the separability of classes. Maximizing the Rayleigh quotient in Eq. (8) with respect to $\mathbf{g}$ defines LDA. Hence the optimal one-dimensional subspace of $\mathbb{R}^D$, where the optimality criterion is class separability due to the Rayleigh quotient, is found from

$$\max_{\mathbf{g} \in \mathbb{R}^D} \left\{\mathbf{g}^{\mathsf{T}}\mathbf{B}\mathbf{g}\right\} \text{ s.t. } \mathbf{g}^{\mathsf{T}}\mathbf{W}\mathbf{g} = 1. \tag{11}$$

Again this constrained linear optimization problem is solved by setting the partial derivatives of the corresponding Lagrangian $\mathbf{L}(\mathbf{g}, \lambda)$ to zero. Analogously to PCA we find a (generalized) eigenvalue problem

$$\frac{\partial \mathbf{L}(\mathbf{g}, \lambda)}{\partial \mathbf{g}} = 0 \iff \mathbf{B}\mathbf{g} = \lambda \mathbf{W}\mathbf{g}. \tag{12}$$

Assuming that $\mathbf{W}^{-1}$ exists, then

$$\mathbf{W}^{-1}\mathbf{B}\mathbf{g} = \lambda \mathbf{g}. \tag{13}$$

The projection $\mathbf{X}^{\perp} \in \mathbb{R}^{C-1 \times N}$ of $\mathbf{X} \in \mathbb{R}^{D \times N}$ onto the optimal subspace defined by the eigenvectors $\mathbf{g}_1, \ldots, \mathbf{g}_{C-1}$ of $\mathbf{W}^{-1}\mathbf{B}$ belonging to the $C-1$ largest eigenvalues is

$$\mathbf{X}^{\perp} = \begin{bmatrix} — & \mathbf{g}_1^{\mathsf{T}} & — \\ & \vdots & \\ — & \mathbf{g}_{C-1}^{\mathsf{T}} & — \end{bmatrix} \begin{bmatrix} | & & | \\ \mathbf{x}_1 & \cdots & \mathbf{x}_N \\ | & & | \end{bmatrix},$$

i.e., $\mathbf{X}^{\perp}$ is the mapping of $\mathbf{X}$ into a feature space that is optimal with respect to class discrimination.

3 Structured Discriminant Analysis

This section presents the main contribution of the paper, namely we introduce a generalization of linear discriminant analysis that allows for learning with algebraic structure. Instead of the projection onto a one-dimensional subspace we propose to learn the coefficients of a multi-dimensional structured subspace that is optimal with respect to class discrimination.

3.1 Circulant Discriminant Analysis

As introduced in Section 2.2 [1] proposes to modify Eq. (6) using κ-circulant matrices, which generalizes PCA towards shift invariance. In the following, we will adopt this approach and modify linear discriminant analysis as defined by Eq. (11) using circulant structures, i.e., we seek the coefficients of a circulant matrix $\mathbf{G} \in \mathbb{R}^{D\times D}$ of the form

$$\mathbf{G} = \sum_{l=1}^{L} g_l \mathbf{P}^{l-1} \tag{14}$$

instead of $\mathbf{g}$. Again $\mathbf{P}$ performs a cyclic permutation, as defined in Eq. (3). An example is depicted in Fig. 1, panel (1).

In this regard we use $\tilde{\mathbf{x}}_\nu = \mathbf{x}_\nu - \overline{\mathbf{x}}_c$ (class affiliation of $\mathbf{x}_\nu$ is unambiguous) and $\tilde{\overline{\mathbf{x}}}_c = \overline{\mathbf{x}}_c - \overline{\mathbf{x}}_0$ as an abbreviation, i.e.,

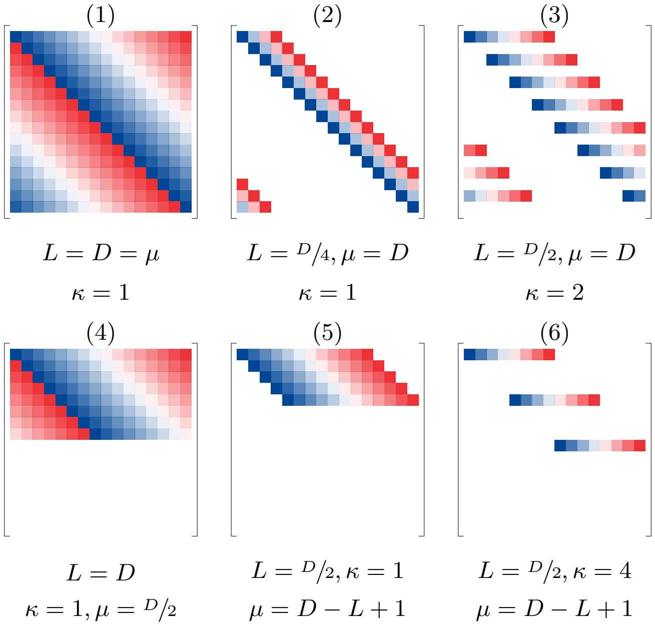
$$\mathbf{B} = \sum_{c=1}^{C} P_c \tilde{\overline{\mathbf{x}}}_c \tilde{\overline{\mathbf{x}}}_c^{\mathsf{T}}.$$

and

$$\mathbf{W} = \sum_{c=1}^{C} P_c \sum_{\nu\in\mathcal{I}_c} \tilde{\mathbf{x}}_\nu \tilde{\mathbf{x}}_\nu^{\mathsf{T}}.$$

The coefficients $\mathbf{g} \in \mathbb{R}^L$ of $\mathbf{G}$ that go along with optimal class separation are found from the linear constrained optimization problem

$$\max_{\mathbf{g}\in\mathbb{R}^D} \left\{ \sum_{c=1}^{C} P_c \left\| \mathbf{G}\tilde{\overline{\mathbf{x}}}_c \right\|_2^2 \right\} \text{ s.t. } \sum_{c=1}^{C} P_c \sum_{\nu\in\mathcal{I}_c} \left\| \mathbf{G}\tilde{\mathbf{x}}_\nu \right\|_2^2 = 1, \tag{15}$$



(1) $L = D = \mu$, $\kappa = 1$

(2) $L = {}^{D}\!/_{4}, \mu = D$, $\kappa = 1$

(3) $L = {}^{D}\!/_{2}, \mu = D$, $\kappa = 2$

(4) $L = D$, $\kappa = 1, \mu = {}^{D}\!/_{2}$

(5) $L = {}^{D}\!/_{2}, \kappa = 1$, $\mu = D - L + 1$

(6) $L = {}^{D}\!/_{2}, \kappa = 4$, $\mu = D - L + 1$

Fig. 1. Different examples on possible structures of $\mathbf{G}$ with $\mathbf{g}$ being a straight line from -1 to 1. These structures illustrate the dependencies associated to different parameter settings, e.g. in panel (1) each coordinate is related to each other while in (4) dependencies are restricted to $\pm{}^{D}\!/_{4}$.

which basically means to perform LDA on the data set $\mathbf{GX}$.

In a geometrical understanding we seek a multi-dimensional cyclically structured subspace of $\mathbb{R}^D$ that is optimal with respect to class separability, while in classical LDA the sought subspace is one-dimensional. This implies that instead of $\|\mathbf{g}^\mathsf{T}\mathbf{x}\|_2^2$ (variance) we measure the length $\|\mathbf{Gx}\|_2^2$ of the projection onto the subspace defined by $\mathbf{G}$ (*total variation*[1] with respect to the variable under consideration, $\tilde{\tilde{\mathbf{x}}}_c$ or $\tilde{\mathbf{x}}_\nu$). However, according to Eq. (4) $\|\mathbf{Gx}\|_2^2$ can also be understood as the power of the filtered signal $\mathbf{G}^\mathsf{T}\mathbf{x}$, while $\mathbf{G}$ is an optimally matched filter. In a two-class setting $\mathbf{G}$ can even be understood as a Wiener filter [24].

The Lagrangian for Eq. (15) is

$$\mathbf{L}(\mathbf{g}, \lambda) = \sum_{c=1}^{C} P_c \tilde{\tilde{\mathbf{x}}}_c^\mathsf{T} \mathbf{G}^\mathsf{T} \mathbf{G} \tilde{\tilde{\mathbf{x}}}_c - \lambda \sum_{c=1}^{C} P_c \sum_{\nu \in \mathcal{I}_c} \tilde{\mathbf{x}}_\nu^\mathsf{T} \mathbf{G}^\mathsf{T} \mathbf{G} \tilde{\mathbf{x}}_\nu - \lambda. \qquad (16)$$

Due to $(\mathbf{P}^i)^\mathsf{T}\mathbf{P}^j = \mathbf{P}^{j-i}\ \forall i, j \in \mathbb{N}$ with $\mathbf{P}$ according to Eq. (14) we find

$$\mathbf{x}^\mathsf{T}\mathbf{G}^\mathsf{T}\mathbf{G}\mathbf{x} = \mathbf{x}^\mathsf{T}(g_1^2\mathbf{P}^0 + \cdots + g_1 g_L \mathbf{P}^{L-1} + \cdots + g_L g_1 \mathbf{P}^{-L+1} + \cdots + g_L^2 \mathbf{P}^0)\mathbf{x}$$

[1] The total variation is the trace of the covariance matrix (cf. [18]).

for any vector $\mathbf{x} \in \mathbb{R}^D$. The derivative with respect to g_k is

$$\frac{\mathrm{d}\mathbf{x}^\mathsf{T}\mathbf{G}^\mathsf{T}\mathbf{G}\mathbf{x}}{\mathrm{d}g_k} = \mathbf{x}^\mathsf{T} \sum_{l=1}^{L} g_l \left(\mathbf{P}^{l-k} + \mathbf{P}^{k-l}\right) \mathbf{x} = 2\mathbf{x}^\mathsf{T} \sum_{l=1}^{L} g_l \mathbf{P}^{l-k} \mathbf{x}. \tag{17}$$

The second equality in Eq. (17) is using the symmetry of real inner products and $(\mathbf{P}^i)^\mathsf{T} = \mathbf{P}^{-i}$, which leads to $\mathbf{x}^\mathsf{T}\mathbf{P}^{-i}\mathbf{x} = (\mathbf{P}^i\mathbf{x})^\mathsf{T}\mathbf{x} = \mathbf{x}^\mathsf{T}\mathbf{P}^i\mathbf{x}$. Using Eq. (17) the partial derivative of Eq. (16) w.r.t. g_k can be written as

$$\frac{\partial \mathbf{L}(\mathbf{g}, \lambda)}{\partial g_k} = 2\sum_{c=1}^{C} P_c \overline{\mathbf{x}}_c^\mathsf{T} \textstyle\sum_l g_l \mathbf{P}^{l-k} \overline{\mathbf{x}}_c - 2\lambda \displaystyle\sum_{c=1}^{C} P_c \sum_{\nu \in \mathcal{I}_c} \tilde{\mathbf{x}}_\nu^\mathsf{T} \textstyle\sum_l g_l \mathbf{P}^{l-k} \tilde{\mathbf{x}}_\nu. \tag{18}$$

Setting Eq. (18) to zero leads to the generalized eigenvalue problem

$$\mathbf{Z}_B \mathbf{g} = \lambda \mathbf{Z}_W \mathbf{g}, \tag{19}$$

where

$$[\mathbf{Z}_B]_{k,l} = \sum_{c=1}^{C} P_c \tilde{\overline{\mathbf{x}}}_c^\mathsf{T} \mathbf{P}^{l-k} \tilde{\overline{\mathbf{x}}}_c \tag{20}$$

and

$$[\mathbf{Z}_W]_{k,l} = \sum_{c=1}^{C} P_c \sum_{\nu \in \mathcal{I}_c} \tilde{\mathbf{x}}_\nu^\mathsf{T} \mathbf{P}^{l-k} \tilde{\mathbf{x}}_\nu. \tag{21}$$

Analogously to classical LDA we find the eigenvalue problem

$$\mathbf{Z}_W^{-1} \mathbf{Z}_B \mathbf{g} = \lambda \mathbf{g}. \tag{22}$$

In contrast to LDA, for non-trivial $\tilde{\overline{\mathbf{x}}}_c$ rank $\{\mathbf{Z}_B\}$ = rank $\{\mathbf{Z}_W\} = L$, which is due to the permutations in Eqs. (20) and (21). Every eigenvector $\mathbf{g}_q$ defines a circulant matrix $\mathbf{G}_q$ and a corresponding subspace. In accordance with Eq. (15) the length of the projection onto this subspace is

$$\mathbf{x}^\perp = \|\mathbf{G}_q \mathbf{x}\|_2^2 . \tag{23}$$

Note that using the nonlinear projection in Eq. (23) yields a nonlinear algorithm. Of course this is not a necessity, since it would be viable to proceed with the linear representation $\mathbf{G}_q\mathbf{x}$. However, Eq. (23) fits in with linear discriminant analysis and is easily interpretable in terms of linear filtering.

3.2 Computational Aspects for Circulant Structures

Since both matrices $\mathbf{Z}_B, \mathbf{Z}_W \in \mathbb{R}^{L \times L}$ have a symmetric Toeplitz structure[2] the computational complexity of Eqs. (20) and (21) can be reduced to $\mathcal{O}(L)$ as both matrices are fully determined by their first row vectors $\mathbf{z}_B, \mathbf{z}_W \in \mathbb{R}^L$ respectively.

[2] $[\mathbf{Z}]_{i,j}$ is constant for constant $i - j$ and $\mathbf{x}^\mathsf{T}\mathbf{P}^{j-i}\mathbf{x} = \mathbf{x}^\mathsf{T}\mathbf{P}^{i-j}\mathbf{x}$ (cf. [1]).

Beyond that, the term $\mathbf{x}^\mathsf{T}\mathbf{P}^{l-k}\mathbf{y}$ realizes a circular convolution $\mathbf{x} \circledast \mathbf{y}$. A circular convolution in turn can be expressed by means of the (fast) Fourier transform (FFT, cf. [24]), i.e., $\mathbf{x} \circledast \mathbf{y} = \mathbf{F}^{-1}(\mathbf{F}(\mathbf{x}) \odot \mathbf{F}(\mathbf{y}))$, where $\mathbf{F}$ denotes the Fourier transform and $\mathbf{F}^{-1}$ its inverse. This allows to compute $\mathbf{z}_B$ and $\mathbf{z}_W$ in $\mathcal{O}(D \log D)$ using the fast Fourier transform, i.e.,

$$[\mathbf{z}_B]_l = \left[\sum_{c=1}^{C} P_c \mathbf{F}^{-1}\left(\mathbf{F}\tilde{\bar{\mathbf{x}}}_c \odot \mathbf{F}\tilde{\bar{\mathbf{x}}}_c\right)\right]_l, \quad l = 1\ldots, L \tag{24}$$

and

$$[\mathbf{z}_W]_l = \left[\sum_{c=1}^{C} P_c \mathbf{F}^{-1} \sum_{\nu \in \mathcal{I}_c} \mathbf{F}\tilde{\mathbf{x}}_\nu \odot \mathbf{F}\tilde{\mathbf{x}}_\nu\right]_l, \quad l = 1\ldots, L. \tag{25}$$

Using these insights the projection according to Eq. (23) can be accelerated via

$$\mathbf{x}^\perp = \left\|\mathbf{F}^{-1}\left(\mathbf{F}\mathbf{g}_q \odot \mathbf{F}\mathbf{x}\right)\right\|_2^2, \tag{26}$$

where $\mathbf{g}_q$ has to be zero-padded such that $\mathbf{g}_q \in \mathbb{R}^D$.

Beneath the low complexity of estimating $\mathbf{Z}_B$ and $\mathbf{Z}_W$ via the FFT, there is a considerable reduction of computational complexity in solving Eq. (22) because L can be chosen much smaller than D. In fact, $L \ll D$ is typically a reasonable choice, because for large L the localization in frequency domain is inappropriately precise (cf. Figs. 4 and 5 and Section 3.3).

3.3 Harmonic solutions

As can be seen from Figs. 4 and 5 for circulant structures with $L = D$ the optimal solution to Eq. (15) is Fourier mode. Investigating Eqs. (20) and (21) for $L = D$ we can see that both, $\mathbf{Z}_W$ and $\mathbf{Z}_B$ are circulant matrices for $L = D$. Generally both matrices have coefficients of the form

$$[\mathbf{Z}]_{k,l} = \mathbf{x}\mathbf{P}^{l-k}\mathbf{x},$$

with some $\mathbf{x} \in \mathbb{R}^D$. Hence, when $L = D$ the first row of $\mathbf{Z}$ is palindromic, i.e., $\mathbf{Z}_{k,l} = \mathbf{Z}_{k,D-l}$ because $\mathbf{P}^{-l} = \mathbf{P}^{D-l}$. Thus $\mathbf{Z}$, respectively $\mathbf{Z}_W$ and $\mathbf{Z}_B$ are symmetric circulant Toeplitz matrices and both admit an eigendecomposition according to Eq. (4), i.e.,

$$\mathbf{Z} = \begin{bmatrix} z_{1,1} & z_{1,2} & z_{1,3} & \cdots & z_{1,3} & z_{1,2} \\ z_{1,2} & z_{1,1} & z_{1,2} & \cdots & z_{1,4} & z_{1,3} \\ z_{1,3} & z_{1,2} & z_{1,1} & \cdots & z_{1,5} & z_{1,4} \\ \vdots & & & & & \vdots \\ z_{1,2} & z_{1,3} & z_{1,4} & \cdots & z_{1,2} & z_{1,1} \end{bmatrix} = \mathbf{F}^{-1}\mathbf{\Lambda}\mathbf{F} \in \mathbb{R}^{D \times D}.$$

Notably, the inverse of a circulant (Toeplitz) matrix is again a circulant matrix [14]. Thus we can conclude that for $L = D$ we have

$$\mathbf{Z}_W^{-1}\mathbf{Z}_B\mathbf{F} = \mathbf{F}\mathbf{\Lambda},$$

with $\mathbf{F}$ being the Fourier matrix (cf. Eq. (5)). We observe that independently of the data at hand the optimal solutions to Eq. (22) respectively Eq. (15) are Fourier modes, i.e., for stationary data Fourier modes maximize the Rayleigh quotient.

3.4 Truncated κ-Circulants

Although circulant structures are beneficial in terms of computational complexity their use is tied to the assumption of *stationarity*[3]. In the following we slightly change the definition of $\mathbf{G}$ to a more general "cyclic" matrix $\boldsymbol{\Gamma}$ in order to gain more flexibility when incorporating dependencies into the structure of $\boldsymbol{\Gamma}$. We refer to a cyclic matrix, when Eq. (4) is not full-filled, i.e., the matrix is based on cyclic permutations, but is not strictly circulant. [17] describes κ-circulants as down-sampled versions of simple circulant. This approach can be generalized to truncated κ-circulant matrices

$$\boldsymbol{\Gamma} = \mathbf{M} \sum_{l=1}^{L} g_l \mathbf{P}^{l-1} \tag{27}$$

with $\mathbf{M}$ performing the down-sampling (with a factor κ) and truncation of all rows following the μ-th row, i.e.,

$$[\mathbf{M}]_{i,j} = \begin{cases} 1 & \text{if } \mu \geq i = j \in [1, \kappa+1, 2\kappa+1, \cdots, \lfloor D/\kappa + 1 \rfloor \kappa] \\ 0 & \text{else.} \end{cases}$$

This especially allows to model dependencies for non-stationary data (see Section 4.2). The idea of truncation is important, as it allows a simple handling of the boundaries by setting $\mu = D - L + 1$ (as known from singular spectrum analysis [2, 9]). On the other hand using some $\mu > 1$ along with $L = D$ is the LDA-equivalent to dynamic PCA (cf. [1, 2, 13]). Setting $\kappa > 1$ implements down-sampling and is equivalent to *stride* in CNNs. Some examples are given in Fig. 1.[4] Using $\boldsymbol{\Gamma}$ instead of $\mathbf{G}$ leads to

$$[\mathbf{Z}_B]_{k,l} = \sum_{c=1}^{C} P_c \tilde{\bar{\mathbf{x}}}_c^{\mathsf{T}} \mathbf{P}^{1-l} \mathbf{M} \mathbf{P}^{k-1} \tilde{\bar{\mathbf{x}}}_c \tag{28}$$

and

$$[\mathbf{Z}_W]_{k,l} = \sum_{c=1}^{C} P_c \sum_{\nu \in \mathcal{I}_c} \tilde{\mathbf{x}}_\nu^{\mathsf{T}} \mathbf{P}^{1-l} \mathbf{M} \mathbf{P}^{k-1} \tilde{\mathbf{x}}_\nu . \tag{29}$$

All derivations are analogous to Section 3.1, except for the binary diagonal matrix $\mathbf{M}$ with $\mathbf{M}^{\mathsf{T}}\mathbf{M} = \mathbf{M}$.[5] Note that with $\mu = D$ and $\kappa = 1$ Eqs. (28) and (29)

[3] The distribution of stationary signals is invariant with respect to time ($\mathbb{E}\{x_t\}$ is constant for all t and the covariance $C(x_t, x_s)$ solely depends on the index/time difference $|t - s|$) [16].

[4] κ-circulant structures can also be used to model Wavelet-like structures [24].

[5] Hence we can fully simplify analogously to the step from Eq. (17) to Eq. (18).

are equal to Eqs. (20) and (21) (circulant discriminant analysis). Moreover, for $L = D$ and $\mu = 1$ (or $\kappa = D$) Eqs. (28) and (29) coincide with $\mathbf{B}$ and $\mathbf{W}$ (Eqs. (9) and (10)). In that sense, circulant discriminant analysis and classical LDA are a special case of truncated κ-circulant structures.

3.5 Non-cyclic Structures

Using circulant structures as proposed in Section 2.1 is an adequate approach for (weakly) stationary data sets. The generalization to truncated κ-circulant matrices allows to embed more complex dependencies into the structure of $\mathbf{G}$ (respectively $\mathbf{\Gamma}$) and hence allows to work with non-stationary data. Yet, when choosing the correlation structure, of course, one is not limited to cyclic structures. As the truncated κ-circulant structure in Eq. (27) can be given as $\mathbf{\Gamma} = \sum_l g_l \mathbf{M}\mathbf{P}^{l-1}$, clearly we can formulate Eq. (15) using an arbitrary structure $\mathbf{\Gamma}_A \in \mathbb{R}^{D \times D}$ which is modeled as

$$\mathbf{\Gamma}_A = \sum_{l=1}^{L} g_l \mathbf{\Pi}_l.$$

Here, the coefficients of $\mathbf{\Pi}_l$ model the dependencies of the i-th variable.

The corresponding solution is equivalent to the above derivations, i.e., we find the generalized eigenvalue problem of Eq. (19). However, the matrices $\mathbf{Z}_B$ and $\mathbf{Z}_W$ are defined as

$$[\mathbf{Z}_B]_{k,l} = \sum_{c=1}^{C} P_c \tilde{\bar{\mathbf{x}}}_c^{\mathsf{T}} \mathbf{\Pi}_l^{\mathsf{T}} \mathbf{\Pi}_k \tilde{\bar{\mathbf{x}}}_c$$

and

$$[\mathbf{Z}_W]_{k,l} = \sum_{c=1}^{C} P_c \sum_{\nu \in \mathcal{I}_c} \tilde{\mathbf{x}}_\nu^{\mathsf{T}} \mathbf{\Pi}_l^{\mathsf{T}} \mathbf{\Pi}_k \tilde{\mathbf{x}}_\nu.$$

This very general formulation also allows for more complex structures that can explicitly model statistical dependencies for non-temporal data. In the field of time series analysis this general approach can be used to build over-complete multi-scale models.

4 Examples and Interpretation

In this section we illustrate the proposed method at the example of different real-world and synthetic data sets.

4.1 (Quasi-)Stationary Data

As a start we use synthetic data generated from different auto-regressive moving average models (ARMA model, cf. [16]) corresponding to the different classes. In the left panel of Fig. 2 realizations from these four different models are depicted.

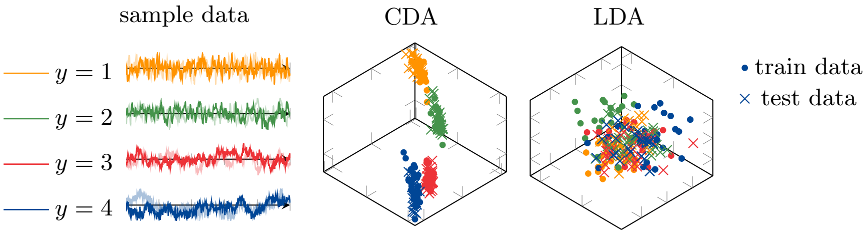


Fig. 2. A simple example demonstrating the performance of circulant discriminant analysis compared to classical linear discriminant analysis according to Section 4.1.

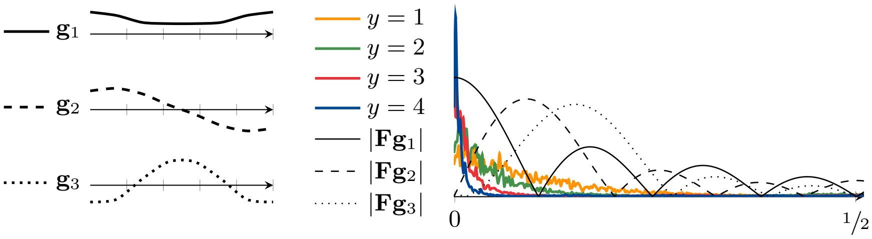


Fig. 3. This figure shows the first three eigenvectors of $\mathbf{Z}_W^{-1}\mathbf{Z}_B$ for the ARMA-process data set (cf. Section 4.1 and Fig. 2) and their spectrum. The right panel shows the spectral density of the four largest classes (colored) along with Fourier transformed (filter) coefficients. The x axis in the right panel is the frequency axis in half cycles per sample.

For the sake of simplicity all model parameters are chosen depending on the class index, i.e., observations belonging to class c stem from a ARMA(p, q) model with $p = q = c$ and coefficients $\theta_i = \phi_i = 1/(c + 1)$ for all $i = 1, \dots, c$, where θ_i and ϕ_i are AR and MA coefficients respectively. Each class comprises $N_c = 50$ samples, with a 50/50 train-test split. The data dimension is $D = 256$. In the middle and right panel of Fig. 2 a comparison of circulant discriminant analysis (CDA, according to Section 3.1) and linear discriminant analysis based on this data is shown. For CDA we used $L = 8$. For both methods the projection onto the first three subspaces is used. More precisely, for CDA the projection is according to Eq. (26). Note that CDA is considerably faster than LDA, due to the computational simplifications proposed in Section 3.2.

In Fig. 6 the "user identification from walking activity" data set (cf. [4]) from the UCR machine learning repository ([7]) is used. The data set contains accelerometer data from 22 different individuals, each walking the same predefined path. For each class, there are x, y and z measurements of the accelerometer forming three time series. For further use, we use sub-series of equal length D from a single variable (acceleration in x-direction).

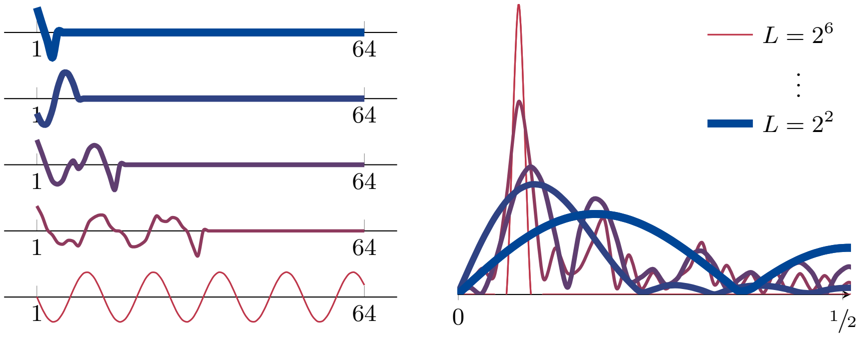


Fig. 4. The left panel shows the optimal solution to Eq. (15) within a parameter over the filter kernel width L based on data that stems from the ARMA process described in Section 4.1. Here $D = 64$. For the special case $L = D = 64$ the solution is a pure harmonic oscillation, as the discrete Fourier basis is the optimal basis in this configuration—independently of the data set under consideration (see Section 3.3).

While the synthetic data used for Fig. 2 is strictly stationary, real-world data—as the gait pattern data used in the examples of Figs. 5 to 7—can be assumed to be stationary on (small) intervals [11]. For the visualization in Fig. 6 we used a 50/50 train-test-split based on observations of length $D = 64$ which corresponds to approximately 2 seconds window width. For the sake of simplicity we used only one variable (the x-coordinate). The accuracies in Fig. 6 are based on a feature vector with 3 elements, i.e., classification is performed on the depicted data. The overall 1-nearest neighbor accuracy on the complete data set with 22 classes on a single variable (x-coordinate) is 46% (CDA) and 20% (LDA) respectively.

4.2 Non-stationary data

In Section 3.4 we introduced κ-circulant structures, that account for non-stationary data. Here we demonstrate the use of such structures using the "Plane" data set (cf. [6]).

Often for time series the assumption of stationarity does not hold. In one example, the data at hand is triggered, i.e., all observations start at a certain point in time (space, ...). This results in non-stationarity, because distinct patterns are likely to be found at a certain index. The "Plane" data set from the UCR Time Series Archive (see [6]) is such a triggered data set. It contains seven different classes that encode the outline of different planes as a function of angle. The triggering stems from the fact, that the outline is captured using the identical starting angle. Hence, the "Plane" data set is an example for non-stationary data, that nevertheless shows temporal (spatial) correlations.

Fig. 8 shows a comparison of stationary and non-stationary parameter settings. The difference between these settings is shown in Fig. 1. A structure for stationary data is visualized in panel (2), while the non-stationary setting is shown in panel

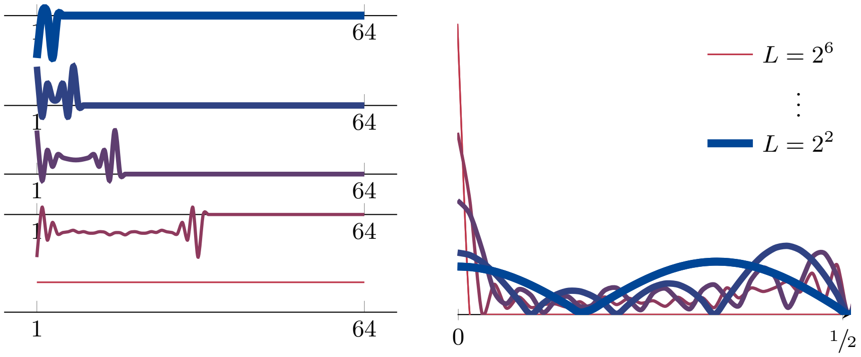


Fig. 5. Illustration of a parameter sweep over L according to Fig. 4. However, here the underlying dataset is the "user identification from walking activity" data set (cf. Figs. 6 and 7). Again for $L = D = 64$ we find a Fourier mode as optimal solution (cf. Section 3.3).

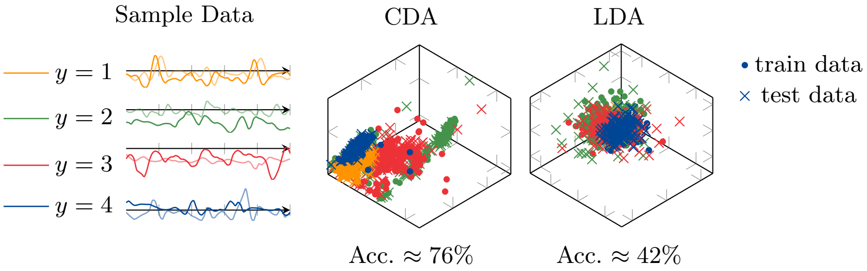


Fig. 6. Comparison of LDA and CDA at the example of the according to Section 4.1. For this figure the four largest classes of the data set are depicted.

(5) of Fig. 1. The former is equivalent to a FIR-filter with data-adaptive coefficients, while the latter is similar to singular spectrum analysis. A detailed analysis of interrelations between theses techniques is provided in [2].

5 Conclusion

Linear discriminant analysis is a core technique in machine learning and statistics. In this work we introduced an adaption of linear discriminant analysis that is optimized for stationary time series. This approach is based on the idea of projecting data onto cyclically structured subspaces, which is related to adaptive linear filtering. We generalize this approach towards non-stationary data and show how arbitrary correlation structures can be modeled. This reconnects to classical LDA, which is a special case of circulant discriminant analysis with truncated κ-circulants. The effectiveness of this approach is demonstrated on

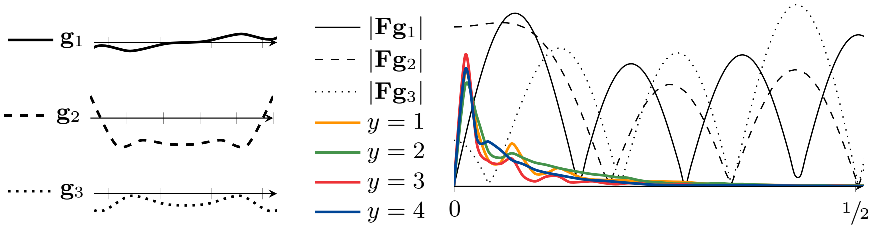


Fig. 7. The left panel shows the first three solutions to Eq. (22) for the "user identification" data set according to Fig. 6 with $L = 8$ (cf. Section 4.1). The right panel shows the spectral density of the four largest classes (colored) along with Fourier transformed (filter) coefficients. The x axis in the right panel is the frequency axis in half cycles per sample.

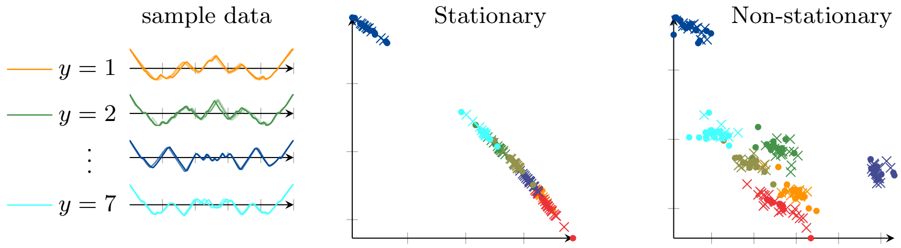


Fig. 8. Non-stationary analysis via truncated κ-circulant structures (using $L = 5, \kappa = 1, \mu = D - L + 1$) compared to (stationary) circulant discriminant analysis ($L = 5$) based on the "Plane" data set ($D = 144$). Here the projection onto the first two subspace is shown.

synthetic stationary data and temporal data from benchmark data sets. Finally, we discussed the connection between circulant discriminant analysis and linear filtering as well as Fourier analysis.

Acknowledgements We are grateful for the careful review of the manuscript. We thank the reviewers for corrections and helpful comments. Additionally we would like to thank the maintainers of the UCI Machine Learning Repository and the UCR Time Series Archive for providing benchmark data sets.

References

1. Bonenberger, C., Ertel, W., Schneider, M.: κ-circulant maximum variance bases. In: German Conference on Artificial Intelligence (Künstliche Intelligenz). pp. 17–29. Springer (2021)

2. Bonenberger, C., Ertel, W., Schwenker, F., Schneider, M.: Singular spectrum analysis and circulant maximum variance frames. Advances in Data Science and Adaptive Analysis (2022)
3. Bouvrie, J.: Notes on convolutional neural networks (2006)
4. Casale, P., Pujol, O., Radeva, P.: Personalization and user verification in wearable systems using biometric walking patterns. Personal and Ubiquitous Computing **16**(5), 563–580 (2012)
5. Christensen, O., et al.: An introduction to frames and Riesz bases. Springer (2016)
6. Dau, H.A., Bagnall, A., Kamgar, K., Yeh, C.C.M., Zhu, Y., Gharghabi, S., Ratanamahatana, C.A., Keogh, E.: The ucr time series archive. IEEE/CAA Journal of Automatica Sinica **6**(6), 1293–1305 (2019)
7. Dua, D., Graff, C.: UCI machine learning repository (2017), http://archive.ics.uci.edu/ml
8. Garcia-Cardona, C., Wohlberg, B.: Convolutional dictionary learning: A comparative review and new algorithms. IEEE Transactions on Computational Imaging **4**(3), 366–381 (2018)
9. Golyandina, N., Zhigljavsky, A.: Singular Spectrum Analysis for time series. Springer Science & Business Media (2013)
10. Hastie, T., Tibshirani, R., Friedman, J.H., Friedman, J.H.: The elements of statistical learning: data mining, inference, and prediction, vol. 2. Springer (2009)
11. Hoffmann, R., Wolff, M.: Intelligente Signalverarbeitung 1: Signalanalyse. Springer-Verlag (2014)
12. Jolliffe, I.T.: Principal component analysis, vol. 2. Springer (2002)
13. Ku, W., Storer, R.H., Georgakis, C.: Disturbance detection and isolation by dynamic principal component analysis. Chemometrics and intelligent laboratory systems **30**(1), 179–196 (1995)
14. Lv, X.G., Huang, T.Z.: A note on inversion of toeplitz matrices. Applied Mathematics Letters **20**(12), 1189–1193 (2007)
15. Morgenshtern, V.I., Bölcskei, H.: A short course on frame theory. arXiv preprint arXiv:1104.4300 (2011)
16. Pollock, D.S.G., Green, R.C., Nguyen, T.: Handbook of time series analysis, signal processing, and dynamics. Elsevier (1999)
17. Rusu, C., Dumitrescu, B., Tsaftaris, S.A.: Explicit shift-invariant dictionary learning. IEEE Signal Processing Letters **21**(1), 6–9 (2013)
18. Seber, G.A.: Multivariate observations. John Wiley & Sons (2009)
19. Serpedin, E., Chen, T., Rajan, D.: Mathematical foundations for signal processing, communications, and networking. CRC Press (2011)
20. Shumway, R.: Discriminant analysis for time series. Handbook of statistics **2**, 1–46 (1982)
21. Sulam, J., Papyan, V., Romano, Y., Elad, M.: Multilayer convolutional sparse modeling: Pursuit and dictionary learning. IEEE Transactions on Signal Processing **66**(15), 4090–4104 (2018)
22. Theodoridis, S., Koutroumbas, K.: Pattern recognition. Elsevier (2006)
23. Tosic, I., Frossard, P.: Dictionary learning: What is the right representation for my signal? IEEE Signal Processing Magazine **28**(ARTICLE), 27–38 (2011)
24. Vetterli, M., Kovačević, J., Goyal, V.K.: Foundations of signal processing. Cambridge University Press (2014)