κ-Circulant Maximum Variance Bases

Christopher Bonenberger[1,2(✉)], Wolfgang Ertel[1], and Markus Schneider[1]

[1] Ravensburg-Weingarten University of Applied Sciences (Institut für Künstliche Intelligenz), Weingarten, Germany
bonenbch@rwu.de

[2] Universität Ulm (Institut für Neuroinformatik), James-Franck-Ring, 89081 Ulm, Germany

Abstract. Principal component analysis (PCA), a well-known technique in machine learning and statistics, is typically applied to time-independent data, as it is based on point-wise correlations. Dynamic PCA (DPCA) handles this issue by augmenting the data set with lagged versions of itself. In this paper, we show that both, PCA and DPCA, are a special case of κ-circulant maximum variance bases. We formulate the constrained linear optimization problem of finding such κ-circulant bases and present a closed-form solution that allows further interpretation and significant speed-up for DPCA. Furthermore, the relation of the proposed bases to the discrete Fourier transform, finite impulse response filters as well as spectral density estimation is pointed out.

Keywords: PCA · Circulant matrices · Toeplitz matrices · Linear filtering · Discrete fourier transform · Dynamic PCA · Representation learning

1 Introduction

The quest for an effective data representation is a driving force in machine learning. Often the data at hand has intrinsic regularities that are concealed in the original space. An appropriately chosen transform from the input space $\mathcal{I}$ to some feature space $\mathcal{F}$ can significantly improve the performance of machine learning algorithms. This is why feature engineering and feature learning are important tasks in machine learning [3].

Yet, a proper feature map is even more important, when there are regularities in the given data that are independent of time (or position) as it is the case for time series or image data [2]. In this context, feature engineering and feature learning are closely connected to signal processing and many of the well-known integral transforms (and their discrete counterparts [19]) are still used for feature engineering [8]. These integral transforms, as for example Fourier, Gabor

This work is supported by a grant from the German Ministry of Education and Research (BMBF; KMU-innovativ: Medizintechnik, 13GW0173E). We would like to express our gratitude to the reviewers for their efforts towards improving this work.

S. Edelkamp et al. (Eds.): KI 2021, LNAI 12873, pp. 17–29, 2021.
https://doi.org/10.1007/978-3-030-87626-5_2

and Wavelet transforms, rely on a fixed basis (or frame), whereas another well-established technique – principal component analysis [12] – can be used to learn an orthogonal basis based on available data by finding the direction of maximum variance. As it is desirable to have data-adaptive representations, principal component analysis and similar techniques for dimensionality reduction are frequently used in machine learning and pattern recognition [22,24].

State-of-the-art machine learning techniques, as for example convolutional neural networks (CNNs) [1] and sparse dictionary learning (SDL) [20] build on these ideas, i.e., signal representations are learned based on available data, while enforcing certain properties (e.g. discriminative or sparse representations). However, the convolutional layers in CNNs are intrinsically offering shift-invariance [7]. In contrast, in dictionary learning algorithms specific structures are imposed to the dictionary, in order to find a shift-invariant basis (or frame, e.g. [17]). Typically, these structures are introduced by means of Toeplitz or circulant matrices [9]. These matrices establish the link to CNNs, as both can be interpreted as a finite impulse response (FIR) filter, just like the filters in CNNs [15]. In both techniques (CNNs and SDL) the filter coefficients are learned adaptively.

The contribution of this work is a mathematical framework, that generalizes classical principal component analysis as a problem of matched[1] FIR filtering, by requiring the principal component to be shift-invariant. In particular, we present a constrained linear optimization problem, that formulates the classical optimization problem of PCA as a shift-invariant problem. This means, instead of finding the direction (a vector) of maximal variance in the input space $\mathcal{I}$, we seek a κ-circulant basis $\mathbf{G}_\kappa$ that maximizes variance (total dispersion [18]) in the feature space $\mathcal{F}$, where $\mathbf{G}_\kappa : \mathcal{I} \to \mathcal{F}$. The mathematical formulation of this optimization problem, which allows a closed-form solution and hence a better understanding of the results, is based on the decomposition of circulant matrices into a matrix polynomial of a simple circular permutation matrix.

As a result, we obtain a class of data-adaptive bases, that constitute classical PCA as well as discrete Fourier analysis (depending on the choice of parameters) and allows a data-adaptive time-frequency decomposition as known from wavelet analysis.

2 Preliminaries

Before we turn to the optimization problem of variance maximizing circulant bases, we recapture related methods. In this paper we focus on two specific orthogonal transforms, namely the discrete Fourier transform and principal component analysis.[2] Beneath classical PCA we also introduce dynamic PCA,

[1] Matched filters are learned in a supervised setting, while here we restrict ourselves to the unsupervised case. Hence, the "matching" of the filter coefficients is according to a variance criterion (similar to PCA).

[2] Principal component analysis is almost equivalent to the Karhunen-Loève transform (KLT) [14]. Further information regarding the relationship between PCA and KLT is given in [10].

because it is closely related to κ-circulant bases. Finally, in Sect. 2.3 κ-circulant matrices are introduced along with FIR filters.

2.1 Principal Component Analysis

Let $\mathbf{X} \in \mathbb{R}^{D \times N}$ be a data set consisting of N observations $\mathbf{x} \in \mathbb{R}^D$ with zero-mean. In order to find the principal component w.r.t. $\mathbf{X}$, the optimization problem

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

has to be solved. In a geometric meaning, we seek the vector $\mathbf{u}^*$ (the principal component) being most similar to the observations $\mathbf{x}$ in the data set.[3] The corresponding Lagrangian

$$\mathbf{L}(\mathbf{u}, \lambda) = \mathbf{u}^T \mathbf{X} \mathbf{X}^T \mathbf{u} - \lambda(\mathbf{u}^T \mathbf{u} - 1) \tag{2}$$

leads to the eigenvalue problem

$$\mathbf{X} \mathbf{X}^T \mathbf{u} = \lambda \mathbf{u}, \tag{3}$$

where $\mathbf{X}\mathbf{X}^T$ is proportional to the sample covariance matrix $\mathbf{S} \in \mathbb{R}^{D \times D}$ of the data at hand ($\mathbf{X}^T$ denotes the transpose of $\mathbf{X}$). The symmetric covariance matrix $\mathbf{S}$ can be diagonalized as $\mathbf{\Lambda} = \mathbf{U}^T \mathbf{S} \mathbf{U}$, i.e.,

$$\frac{1}{N-1} \mathbf{X} \mathbf{X}^T \mathbf{U} = \mathbf{S}\mathbf{U} = \mathbf{U}\mathbf{\Lambda}. \tag{4}$$

The column vectors of $\mathbf{U}$ are the eigenvectors (eigenbasis) of $\mathbf{S}$ and form an orthogonal basis for $\mathbb{R}^D$. The eigenvector corresponding to the largest eigenvalue is the principal component and points in the direction of maximum variance.

The representation $\mathbf{x}'$ of some signal $\mathbf{x} \in \mathcal{X}$ in a subspace $\mathcal{S}$ can be found as $\mathbf{U}_k^T \mathbf{x}$, where $\mathbf{U}_k^T : \mathcal{X} \to \mathcal{S}$ contains only the eigenvectors belonging to the k largest eigenvalues. The reconstruction $\mathbf{x}'$ of the signal $\mathbf{x}$ is then found as $\mathbf{U}_k \mathbf{U}_k^T \mathbf{x}$, i.e., $\mathbf{U}_k : \mathcal{S} \to \mathcal{X}$.

In the context of this work it is especially of interest, that depending on the underlying stochastic process, the covariance matrix may have a Toeplitz-like structure (cf. Sect. 4.1) and the corresponding eigenbasis approximates a Fourier basis [21].

2.2 Dynamic Principal Component Analysis

In [13] Ku et al. proposed dynamic PCA for statistical process monitoring, where the original data set is augmented by lagged versions of itself. Hereby the number of lags L is a free parameter. This method of data augmentation is used in order to achieve circular permutation invariance and to overcome the static behavior

[3] The dot product $\mathbf{u}^T \mathbf{x}$ serves as measure of similarity.

of classical PCA. In fact, a classical PCA is performed, yet instead of the original data set $\mathbf{X}$ the augmented data set $\mathbf{X}_A$ is used. This can be formalized as

$$\mathbf{X}_A = \begin{bmatrix} \mathbf{P}^0\mathbf{X} & \mathbf{P}^1\mathbf{X} & \cdots & \mathbf{P}^{L-1}\mathbf{X} \end{bmatrix}, \tag{5}$$

where $\mathbf{P}$ is the permutation matrix

$$\mathbf{P} = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & & 0 \\ \vdots & & & \ddots & \vdots \\ 0 & & & & 1 \\ 1 & 0 & 0 & \cdots & 0 \end{bmatrix} \in \mathbb{R}^{D \times D}. \tag{6}$$

Hence, for DPCA the principal component is the eigenvector $\mathbf{u}^*$ of $\mathbf{X}_A\mathbf{X}_A^T$ with the largest corresponding eigenvalue, which can also be found as a solution to the optimization problem

$$\max_{\mathbf{u} \in \mathbb{R}^D} \left\{ \left\| \mathbf{u}^T\mathbf{P}^0\mathbf{X} \right\|_2^2 + \cdots + \left\| \mathbf{u}^T\mathbf{P}^{L-1}\mathbf{X} \right\|_2^2 \right\} \text{ s.t. } \|\mathbf{u}\|_2^2 = 1. \tag{7}$$

2.3 κ-Circulant Matrices

A circulant matrix $\mathbf{C} \in \mathbb{R}^{D \times D}$ is formed by a vector $\mathbf{c} \in \mathbb{R}^W$ and its lagged versions, e.g. when $W = D$ it is a matrix of the form

$$\mathbf{C} = \begin{bmatrix} c_0 & c_1 & \cdots & c_{W-1} \\ c_{W-1} & c_0 & \cdots & c_{W-2} \\ \vdots & & \ddots & \vdots \\ c_1 & c_2 & \cdots & c_0 \end{bmatrix} = \begin{bmatrix} — & \mathbf{c}^T & — \\ — & \mathbf{c}^T\mathbf{P} & — \\ & \vdots & \\ — & \mathbf{c}^T\mathbf{P}^{D-1} & — \end{bmatrix} = \sum_{w=0}^{W-1} c_w \mathbf{P}^w, \tag{8}$$

where $\mathbf{P}$ corresponds to a circular permutation as defined in Eq. 6. An example of a simple circulant is shown in the top left graph of Fig. 1.

A circulant matrix can be diagonalized as

$$\mathbf{C} = \mathbf{F}\mathbf{\Lambda}\mathbf{F}^{-1}, \tag{9}$$

where $\mathbf{F} \in \mathbb{R}^{D \times D}$ is the discrete Fourier matrix with components $f_{j,k}$ that are given as $f_{j,k} = \exp(-2\pi ijk/D)/\sqrt{D}$ for all $0 \leq j, k < D$ with $i^2 = -1$ [11]. The eigenvalue matrix $\mathbf{\Lambda} = \sqrt{D}\,\mathrm{diag}(\mathbf{Fc})$ is a diagonal matrix with the discrete Fourier transform of $\mathbf{c}$ on its diagonal. Hence, the eigenvectors of a circulant matrix are the Fourier modes and the eigenvalues can be computed from the DFT $\hat{\mathbf{c}} = \sqrt{D}\,\mathbf{Fc}$ of the vector $\mathbf{c}$, i.e., $\lambda_j = \hat{c}_j$.

This can also be understood by means of the convolution theorem: multiplying a circulant matrix $\mathbf{C} \in \mathbb{R}^{D \times D}$ (that is based on the vector $\mathbf{c} \in \mathbb{R}^D$) with some vector $\mathbf{x} \in \mathbb{R}^D$ realizes a discrete circular convolution[4], i.e., $\mathbf{Cx} = \mathbf{x} \circledast \mathbf{c} = \mathbf{F}^{-1}\mathbf{\Lambda}^H\mathbf{Fx}$ with $\mathbf{\Lambda}^H$ being the conjugate transpose of $\mathbf{\Lambda}$.

[4] The discrete circular convolution of two sequences $\mathbf{x}, \mathbf{y} \in \mathbb{R}^D$ is written as $\mathbf{x} \circledast \mathbf{y}$, while the linear convolution is written as $\mathbf{x} * \mathbf{y}$.

A linear convolution can be expressed as a circular convolution, when zero-padding is used. Hence, $\mathbf{C}$ can describe a simple FIR filter [23]. For a filter-kernel $\mathbf{c} \in \mathbb{R}^W$ and data $\mathbf{x} \in \mathbb{R}^D$ the linear convolution $\mathbf{c} * \mathbf{x}$ can be written as

$$\mathbf{C}\mathbf{x} = \mathbf{c} * \mathbf{x} = \left(\sum_{w=0}^{W-1} c_w \mathbf{P}^w \right) \mathbf{x} \quad \text{where} \quad \mathbf{x}, \mathbf{P} \in \mathbb{R}^{W+D-1}. \tag{10}$$

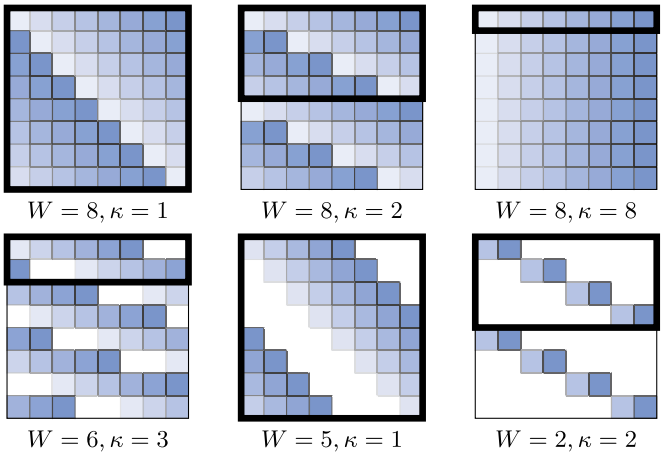


Fig. 1. Examples of κ-circulant matrices $\mathbf{C}_\kappa$ according to Eq. 11 for $D = 8$ and the equivalent (sub)matrices $\mathbf{G}_\kappa$ according to Eq. 20 (framed black). Equation 19 formulates the problem of finding the optimal kernel $\mathbf{g}$ given a fixed structure of $\mathbf{G}_\kappa$. For ease of visualization these exemplary matrices are based on a kernel $\mathbf{c} \in \mathbb{R}^W$ that is a simple ramp (e.g. $\mathbf{c}^T = [0.125\ 0.25\ 0.375\ \cdots\ 1]$ is the first row in the top left matrix). Note that the bottom right matrix is a wavelet-like structure.

A κ-circulant matrix $\mathbf{C}_\kappa \in \mathbb{R}^{D \times D}$ is the generalization of a standard circulant and has the form [4]

$$\mathbf{C}_\kappa = \begin{bmatrix} c_0 & c_1 & \cdots & c_{D-1} \\ c_{D-\kappa} & c_{D-\kappa+1} & \cdots & c_{D-\kappa-1} \\ \vdots & & \ddots & \vdots \\ c_\kappa & c_{\kappa+1} & \cdots & c_{\kappa-1} \end{bmatrix} = \begin{bmatrix} — & \mathbf{c}^T & — \\ — & \mathbf{c}^T\mathbf{P}^\kappa & — \\ & \vdots & \\ — & \mathbf{c}^T\mathbf{P}^{\kappa(D-1)} & — \end{bmatrix}, \tag{11}$$

where the subscripts are modulo D. Some examples are shown in Fig. 1.

3 Maximum Variance Bases

In this section, we present a novel optimization problem, that generalizes PCA and DPCA. This more general optimization problem is based on the idea of searching a basis, that is inherently translation-invariant. This is enforced, by choosing circulant structures, that are based on circular permutations. We refer to those as "matched" κ-circulant matrices.

3.1 Simple Matched Circulants

For the sake of simplicity, before treating the general case of κ-circulant bases, in this section we restrict ourselves to the case of standard circulant matrices ($\kappa = 1$). In order to find a basis that maximizes variance we formulate a constrained linear optimization problem analogously to PCA (cf. Eq. 1). However, instead of a single vector we seek a circulant matrix $\mathbf{G}$ (a filter, cf. Sect. 2.3) that solves

$$\max_{\mathbf{g}\in\mathbb{R}^W} \left\{ \|\mathbf{GX}\|_F^2 \right\} \text{ s.t. } \|\mathbf{g}\|_2^2 = 1, \tag{12}$$

where $\|\cdot\|_F$ denotes the Frobenius norm and

$$\mathbf{G} = \sum_{w=0}^{W-1} g_w \mathbf{P}^w. \tag{13}$$

According to Eq. 10 $\mathbf{G}$ is a FIR filter with coefficients $\mathbf{g}$. Hence Eq. 12 may also be understood as the problem of designing a FIR filter such that the transmitted energy is maximized.[5] Hence, Eq. 12 poses power spectral density estimation as a constrained linear optimization problem. The corresponding Lagrangian is

$$\begin{aligned}\mathbf{L}(\mathbf{g},\lambda) &= \sum_{\nu=1}^{N} \langle \textstyle\sum_w g_w \mathbf{P}^w \mathbf{x}_\nu, \sum_w g_w \mathbf{P}^w \mathbf{x}_\nu \rangle - \lambda \left(\mathbf{g}^T\mathbf{g} - 1\right) \\ &= \sum_{\nu=1}^{N} \mathbf{x}_\nu^T \left(\textstyle\sum_w g_w \mathbf{P}^{-w}\right) \left(\sum_w g_w \mathbf{P}^w\right) \mathbf{x}_\nu - \lambda \left(\mathbf{g}^T\mathbf{g} - 1\right),\end{aligned}$$

where the product $\mathbf{G}^T\mathbf{G} = \left(\sum_w g_w \mathbf{P}^{-w}\right)\left(\sum_w g_w \mathbf{P}^w\right)$ is

$$\mathbf{G}^T\mathbf{G} = g_0^2\mathbf{P}^0 + \cdots + g_0 g_{W-1}\mathbf{P}^{W-1} + \cdots + g_{W-1}g_0\mathbf{P}^{1-W} + \cdots + g_{W-1}^2\mathbf{P}^0.$$

Thus the derivative of $\mathbf{L}(\mathbf{g},\lambda)$ w.r.t. g_k can be given as

$$\frac{\partial \mathbf{L}}{\partial g_k} = -\sum_{\nu=1}^{N} \left\langle \mathbf{x}_\nu, \left(\textstyle\sum_w g_w \left(\mathbf{P}^{w-k} + \mathbf{P}^{k-w}\right)\right)\mathbf{x}_\nu \right\rangle + 2\lambda g_k. \tag{14}$$

Due to $\langle \mathbf{Px}, \mathbf{x}\rangle = \langle \mathbf{x}, \mathbf{P}^T\mathbf{x}\rangle$ and because the transpose $\mathbf{P}^T$ of a circulant matrix equals its inverse (i.e., $\mathbf{P}^T = \mathbf{P}^{-1}$) we can write $\langle \mathbf{x}, \mathbf{P}^{k-w}\mathbf{x}\rangle + \langle \mathbf{x}, \mathbf{P}^{w-k}\mathbf{x}\rangle = \langle \mathbf{x}, \mathbf{P}^{k-w}\mathbf{x}\rangle + \langle \mathbf{P}^{k-w}\mathbf{x}, \mathbf{x}\rangle$ and hence using the symmetry of the scalar product we find

$$\frac{\partial \mathbf{L}}{\partial g_k} = -2\sum_\nu \left\langle \mathbf{x}_\nu, \textstyle\sum_w g_w \mathbf{P}^{w-k}\mathbf{x}_\nu \right\rangle + 2\lambda g_k. \tag{15}$$

With the abbreviation $z_{k,w} = \sum_\nu \left\langle \mathbf{x}_\nu, \mathbf{P}^{k-w}\mathbf{x}_\nu \right\rangle$ for the components of the matrix $\mathbf{Z} \in \mathbb{R}^{W\times W}$ this leads to

$$\begin{bmatrix} z_{0,0} & \cdots & z_{0,W-1} \\ \vdots & \ddots & \vdots \\ z_{W-1,0} & \cdots & z_{W-1,W-1} \end{bmatrix} \mathbf{g} = \mathbf{Zg} = \lambda\mathbf{g}. \tag{16}$$

[5] Due to the constraint $\|\mathbf{g}\|_2^2 = 1$ this is not trivial.

Note that $\mathbf{Z}$ is a symmetric Toeplitz matrix that is fully determined by its first column vector $\mathbf{z}$, because the value of $z_{k,w}$ only depends on the difference of the indeces $k - w$ but not on the absolute value of k and w. The components $z_{k,w}$ of the matrix $\mathbf{Z}$ are given as (cf. Eq. 15)

$$z_{k,w} = \sum_\nu \left\langle \mathbf{x}_\nu, \mathbf{P}^{k-w}\mathbf{x}_\nu \right\rangle \tag{17}$$

This means the value of $z_{k,w}$ depends on the similarity of $\mathbf{x}_\nu$ and the lagged versions $\mathbf{P}^{k-w}\mathbf{x}_\nu$. In case of zero-mean data, this corresponds to the circular sample autocorrelation[6] $\mathbf{r} \in \mathbb{R}^D$, which can be estimated by

$$\mathbf{r} \propto \sum_{\nu=1}^{N} \left[\mathbf{F}^{-1}\left(\mathbf{FX} \odot \overline{\mathbf{FX}}\right)\right]_{w,\nu}, \tag{18}$$

where $\odot$ denotes the Hadamard product (pointwise matrix multiplication). Due to $\mathbf{z} \propto \mathbf{r}$ the sample autocorrelation matrix $\mathbf{Z}$ can be computed in $\mathcal{O}(D \log D)$ by means of the fast Fourier transform (FFT).

3.2 Matched κ-Circulant Matrices

The generalization of Eq. 12 for κ-circulant matrices is

$$\max_{\mathbf{g}\in\mathbb{R}^W} \left\{ \|\mathbf{G}_\kappa \mathbf{X}\|_F^2 \right\} \text{ s.t. } \|\mathbf{g}\|_2^2 = 1. \tag{19}$$

The matrix $\mathbf{G}_\kappa$ is defined as

$$\mathbf{G}_\kappa = \mathbf{M}_\kappa \sum_{w=0}^{W-1} g_w \mathbf{P}^w, \tag{20}$$

where the "masking" matrix $\mathbf{M}_\kappa$ has components

$$m_{j,k} = \begin{cases} 1 & \text{if } j = k \in [0, \kappa, 2\kappa, \cdots, \lfloor D/\kappa - 1 \rfloor \kappa] \\ 0 & \text{else.} \end{cases} \tag{21}$$

Left multiplication of this masking matrix to a circulant matrix $\mathbf{G}$ essentially preserves only every κ-th rows of $\mathbf{G}$ (cf. [17]). The effect of $\mathbf{M}_\kappa$ and the resulting structure of the matrix $\mathbf{G}_\kappa$ is shown in Fig. 1.

The derivative of the Lagrangian resulting from Eq. 19 is

$$\frac{\partial \mathbf{L}}{\partial g_k} = \sum_{\nu=1}^{N} \langle \mathbf{x}_\nu, \left(\sum_w g_w \left(\mathbf{P}^{-w}\mathbf{M}_\kappa^T\mathbf{M}_\kappa\mathbf{P}^k + \mathbf{P}^{-k}\mathbf{M}_\kappa^T\mathbf{M}_\kappa\mathbf{P}^w\right)\right)\mathbf{x}_\nu \rangle - 2\lambda g_k. \tag{22}$$

[6] This interpretation is only valid under the assumptions mentioned in Sect. 3.3. Furthermore, the normalization of the autocorrelation (autocovariance) is to be performed as $\mathbf{r}' = \frac{\mathbf{r}}{r_0}$, with the first component r_0 of $\mathbf{r}$ being the variance [16].

Analogously to Eq. 16 we find an eigenvalue problem

$$\mathbf{Z}_\kappa \mathbf{g} = \lambda \mathbf{g}, \tag{23}$$

where the symmetric matrix $\mathbf{Z}_\kappa \in \mathbb{R}^{W \times W}$ has components

$$z_{k,w} = \sum_{\nu=1}^{N} \langle \mathbf{x}_\nu, \mathbf{P}^{-w} \mathbf{M}_\kappa \mathbf{P}^k \mathbf{x}_\nu \rangle. \tag{24}$$

Note that the resulting matrix $\mathbf{Z}_\kappa$ is a consequence of the structure of $\mathbf{G}_\kappa$, in other words, fixing the parameters κ and W means to hypothesize a certain model. Assume – as an example – a circulant data matrix: in this case the sample covariance matrix $\mathbf{S}$ becomes a symmetric circulant matrix and is equal to the sample autocorrelation matrix. Hence, the parameters κ and W imply a certain model.

3.3 Relation to PCA, DPCA and DFT

The closed-form solution for a κ-circulant variance maximizing bases presented in the previous section has many implications. In the following the relation of the presented solution to PCA, DPCA and DFT is described.

PCA. When $\kappa = W = D$ the optimization problem formulated in Eq. 19 is equivalent to the problem of finding principal components as described in Eq. 1. In this case the only non-zero component of $\mathbf{M}_\kappa$ is $m_{0,0} = 1$, such that $\mathbf{Z}_\kappa = \mathbf{Z}_D \propto \mathbf{S}$, i.e., the matrix $\mathbf{Z}$ is proportional to the sample covariance matrix (as for usual PCA this only holds when the underlying data set has zero-mean data). This can also be understood from Fig. 1, where this case is depicted by example ($W = D = \kappa = 8$).

DPCA. When choosing $\kappa = 1$ and $W = D$ we can identify the number of lags L for DPCA with the kernel width W. This can be seen, when comparing Eq. 7 with the equivalent problem

$$\max_{\mathbf{g} \in \mathbb{R}^D} \left\{ \left\| (\mathbf{P}^0 \mathbf{g})^T \mathbf{X} \right\|_2^2 + \cdots + \left\| (\mathbf{P}^{W-1} \mathbf{g})^T \mathbf{X} \right\|_2^2 \right\} \text{ s.t. } \|\mathbf{g}\|_2^2 = 1, \tag{25}$$

which in turn is equivalent to Eq. 12. However, computing $\mathbf{Z}_1$ as $\mathbf{Z}_1 = \mathbf{X}_A \mathbf{X}_A^T$ in $\mathcal{O}(D^3)$ is less efficient than using Eq. 18 with $\mathcal{O}(D \log D)$. This case is shown by example in Fig. 1 in the top left matrix ($W = D = 8, \kappa = 1$).

Introducing another masking matrix $\mathbf{M}_\gamma$ similar to $\mathbf{M}_\kappa$ it is possible to augment Eq. 12, such that DPCA with a smaller number of lags ($L < D$) can be modeled. This matrix $\mathbf{M}_\gamma$ has to be a diagonal matrix with the first $L-1$ diagonal entries being one and the rest being zero. The solution of this problem is analogously to Eq. 23 with $\mathbf{M}_\kappa$ in Eq. 24 being replaced by $\mathbf{M}_\gamma$. Yet, when narrow range dependencies are expected, it is more efficient to choose $\kappa = 1$ and $W < D$.

DFT. Assuming equidistantly sampled zero-mean data that stems from a wide-sense stationary stochastic process, in the case $D = W$ and $\kappa = 1$, the matrix $\mathbf{Z}_1$ is proportional to the circular sample autocorrelation matrix [6]. The circular autocorrelation is symmetric and hence the resulting matrix $\mathbf{Z}_1$ is a symmetric circulant, i.e., according to Eq. 9 we have $\mathbf{Z}_1 = \mathbf{F}\mathbf{\Lambda}\mathbf{F}^{-1}$ with $\mathbf{\Lambda} = \sqrt{D}\,\mathrm{diag}\,(\mathbf{Fr})$. This means that the optimal D-dimensional shift-invariant basis (maximizing the total dispersion[7]) is the discrete Fourier basis $\mathbf{F}$, because this is the eigenbasis of the matrix $\mathbf{Z}_1$. Furthermore, under the assumptions stated above, the vector $\mathbf{p} = \sqrt{D}\,\mathbf{Fr}$ is an estimate of the power spectral density, which is by definition the discrete time Fourier transform of the sample autocorrelation [16].

4 Numerical Results

In the following the presented theory is demonstrated using the example of a stationary random process, i.e., data points $\mathbf{x} \in \mathbb{R}^{128}$ are sampled from a moving-average (MA) process (cf. [16]). This moving average process is based on a modulated and truncated Gaussian $f(t) = \sin(2\pi t)\exp(-t^2)$ (cf. Fig. 2) that is sampled on the interval $[-2, 2]$ (with sampling frequency 16 Hz such that $\mathbf{f} \in \mathbb{R}^{64}$). Additionally we visualize the results based on a stationary process that realizes random shifts of a fixed signal. Although these settings are rather specific, they serve the purpose of demonstrating the proposed framework and visualize the main results.

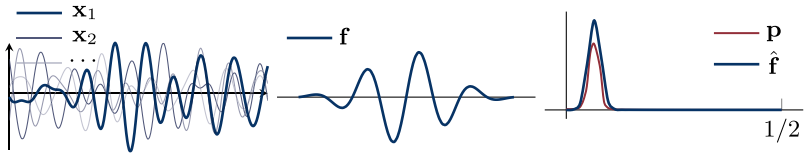


Fig. 2. Observations of the data (left graph) taken from a moving average process with the kernel $\mathbf{f}$ shown in the middle graph. The right graph shows the absolute value of the discrete Fourier transform $\hat{\mathbf{f}}$ of $\mathbf{f}$ and its estimate $\mathbf{p}$, i.e., the (single-sided) spectral density of the corresponding random process (the x-axis is the frequency axis showing halfcycles/sample).

In the examples presented here, the parameters W and κ are always chosen appropriately although such a clear setup does not always naturally arise. Yet, in most cases this is a minor problem. If W, D and κ are inconsistent, zero-padding can be used to overcome this issue (as it is done for discrete Fourier transform). However, as shown in Fig. 1 in the bottom left graph, the effect of $\mathbf{M}_\kappa$ is not always as desired. In fact κ and W should be chosen such that $\mathrm{mod}(D, \kappa) =$

[7] Let $\mathbf{Y} = \mathbf{G}_\kappa\mathbf{X}$. Maximizing $\|\mathbf{Y}\|_F^2$ (cf. Eq. 19) means maximizing the trace of the covariance matrix $\mathbf{S} \propto \mathbf{Y}\mathbf{Y}^T$, which in turn is a measure for the total dispersion [18].

$\mathrm{mod}(D, W) = 0$. This is not a strong restriction, when zero-padding (or another kind of padding, e.g. symmetric) is used.

4.1 MA Process

In Fig. 2 an overview about the data and the generating process is given. In the right graph of Fig. 2 the power spectral density and its estimate from the sample autocorrelation is shown. Figure 3 shows four different configurations of the matrix $\mathbf{Z}$. As the underlying process is a MA process, the covariance matrix $\mathbf{S}$ (which stems from Eq. 24 with $\kappa = D$) is almost Toeplitz-structured. Yet, estimating the process characteristic incorporating shift-invariance ($\kappa = 1$) to the model leads to a better approximation. This can also be seen from Fig. 4, where the "principal component" is depicted as a function of W. In Fig. 5 the first 5 eigenvectors are depicted for $W = D/2$ and $W = D/4$. Finally it is interesting to note that the basic optimization problem stated in Eq. 12 is strongly related to a MA model.

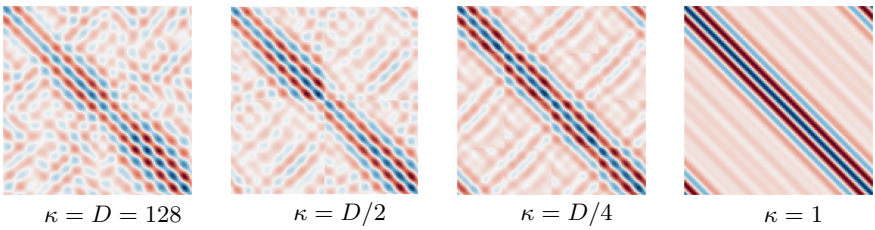


Fig. 3. The matrix $\mathbf{Z}_\kappa$ for different values of κ (with $W = D$). The left matrix corresponds to the covariance matrix (PCA) and the right matrix is equivalent to the autocovariance matrix. The matrices in between show two "intermediate" steps.

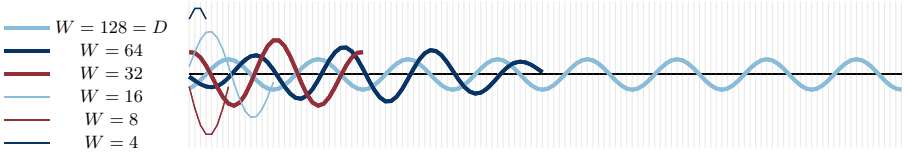


Fig. 4. The eigenvector $\mathbf{g}^* \in \mathbb{R}^W$ belonging to the largest eigenvalue for varying W.

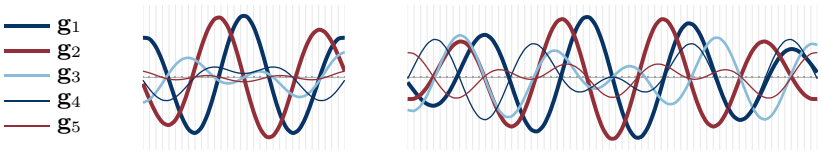


Fig. 5. The first 5 eigenvectors $\mathbf{g}_1, \ldots, \mathbf{g}_5$ corresponding to the 5 largest eigenvalues for $W = 32 = D/4$ (left graph) and $W = 64 = D/2$.

4.2 Circular Process

The characteristic of the process and example data is shown in Fig. 6, whereas the generating function is chosen as in the previous section, i.e., we observe random shifts of a noisy modulated gaussian. The corresponding matrices $\mathbf{Z}_\kappa$ are shown in Fig. 8 for a variety of parameters. In Fig. 7 the reconstruction of a sample signal is shown for different parameter settings. The example shown in Fig. 7 is restricted to the case $W = \kappa$, which means PCA is performed on various scales. The result is less trivial when $\kappa < W$, as then a frame (cf. [5]) is resulting instead of an orthogonal basis.

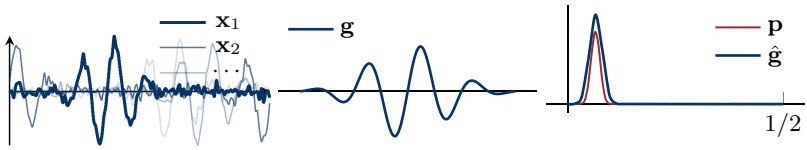


Fig. 6. Examples of the data (left graph) generated by random shifts of the kernel $\mathbf{g}$ shown in the middle graph with additive noise. The right graph shows the power spectral density $\hat{\mathbf{g}}$ of $\mathbf{g}$ and its estimate $\mathbf{p}$.

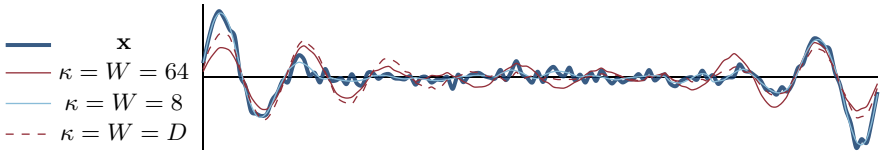
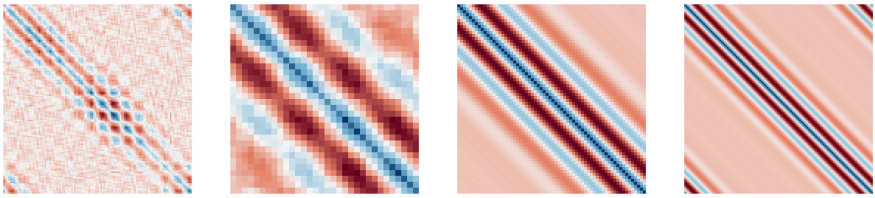


Fig. 7. Example of the reconstruction of a signal (from the two first principal components) drawn from the process described with four different settings. As expected the reconstruction improves with decreasing W (increasing resolution).



$\kappa = D = W = 128$ $\kappa = W = 32 = D/4$ $\kappa = 1, W = 64$ $\kappa = 1, W = D$

Fig. 8. The matrix $\mathbf{Z}_\kappa$ for different values of κ estimated from $N = 32$ samples of a circular random process generated from random shifts of the vector $\mathbf{g}$ depicted in Fig. 6 (middle graph). The left matrix corresponds to PCA ($\mathbf{Z}_1 = \mathbf{S}$ when $W = D = \kappa$), the right matrix is an estimate of the autocovariance matrix and the two graphs in the middle show matrices $\mathbf{Z}_\kappa \in \mathbb{R}^{W \times W}$ of reduced dimension $W < D$.

5 Conclusion

PCA and linear filtering (FIR filters) are both relevant topics with many applications in machine learning. Our work presents a mathematical framework that establishes a link between those techniques and hence allows a better understanding. Beyond that, our formulation allows to estimate models that may incorporate time-frequency trade-offs in data-adaptive representations. Inherent to the presented theory is a mathematical formulation, that generalizes PCA in terms of shift-invariance. This way the relation between PCA and FIR filters, DFT and DPCA is made explicit and an FFT-based implementation of DPCA can be proposed. A signal processing point-of-view is provided along with examples on stationary stochastic processes.

References

1. Albawi, S., Mohammed, T.A., Al-Zawi, S.: Understanding of a convolutional neural network. In: 2017 International Conference on Engineering and Technology (ICET), pp. 1–6. IEEE (2017)
2. Bagnall, A., Lines, J., Bostrom, A., Large, J., Keogh, E.: The great time series classification bake off: a review and experimental evaluation of recent algorithmic advances. Data Min. Knowl. Disc. **31**(3), 606–660 (2016). https://doi.org/10.1007/s10618-016-0483-9
3. Bengio, Y., Courville, A., Vincent, P.: Representation learning: a review and new perspectives. IEEE Trans. Pattern Anal. Mach. Intell. **35**(8), 1798–1828 (2013)
4. Bose, A., Saha, K.: Random Circulant Matrices. CRC Press (2018)
5. Casazza, P.G., Kutyniok, G., Philipp, F.: Introduction to finite frame theory. Finite Frames, pp. 1–53 (2013)
6. Chatfield, C.: The Analysis of Time Series: An Introduction. Chapman and Hall/CRC (2003)
7. Ismail Fawaz, H., Forestier, G., Weber, J., Idoumghar, L., Muller, P.-A.: Deep learning for time series classification: a review. Data Min. Knowl. Disc. **33**(4), 917–963 (2019). https://doi.org/10.1007/s10618-019-00619-1

8. Fulcher, B.D.: Feature-based time-series analysis. In: Feature Engineering for Machine Learning and Data Analytics, pp. 87–116. CRC Press (2018)
9. Garcia-Cardona, C., Wohlberg, B.: Convolutional dictionary learning: a comparative review and new algorithms. IEEE Trans. Comput. Imaging **4**(3), 366–381 (2018)
10. Gerbrands, J.J.: On the relationships between SVD, KLT and PCA. Pattern Recogn. **14**(1–6), 375–381 (1981)
11. Gray, R.M.: Toeplitz and Circulant Matrices: A Review (2006)
12. Jolliffe, I.T.: Principal components in regression analysis. In: Jolliffe, I.T. (ed.) Principal Component Analysis. SSS, pp. 129–155. Springer, New York (1986). https://doi.org/10.1007/978-1-4757-1904-8_8
13. Ku, W., Storer, R.H., Georgakis, C.: Disturbance detection and isolation by dynamic principal component analysis. Chemom. Intell. Lab. Syst. **30**(1), 179–196 (1995)
14. Orfanidis, S.: SVD, PCA, KLT, CCA, and all that. Optimum Signal Processing, pp. 332–525 (2007)
15. Papyan, V., Romano, Y., Elad, M.: Convolutional neural networks analyzed via convolutional sparse coding. J. Mach. Learn. Res. **18**(1), 2887–2938 (2017)
16. Pollock, D.S.G., Green, R.C., Nguyen, T.: Handbook of Time Series Analysis, Signal Processing, and Dynamics. Elsevier (1999)
17. Rusu, C.: On learning with shift-invariant structures. Digit. Signal Process. **99**, 102654 (2020)
18. Seber, G.A.: Multivariate Observations, vol. 252. Wiley, Hoboken (2009)
19. Strang, G., Nguyen, T.: Wavelets and Filter Banks. SIAM (1996)
20. Tošić, I., Frossard, P.: Dictionary learning. IEEE Signal Process. Mag. **28**(2), 27–38 (2011)
21. Unser, M.: On the approximation of the discrete Karhunen-Loeve transform for stationary processes. Signal Process. **7**(3), 231–249 (1984)
22. Vaswani, N., Narayanamurthy, P.: Static and dynamic robust PCA and matrix completion: a review. Proc. IEEE **106**(8), 1359–1379 (2018)
23. Vetterli, M., Kovačević, J., Goyal, V.K.: Foundations of Signal Processing. Cambridge University Press (2014)
24. Zhao, D., Lin, Z., Tang, X.: Laplacian PCA and its applications. In: 2007 IEEE 11th International Conference on Computer Vision, pp. 1–8. IEEE (2007)