

METHODS

Cyclic Nonlinear Correlation Analysis for Time Series

CHRISTOPHER M. A. BONENBERGER^{1,2}, FRIEDHELM SCHWENKER¹, (Member, IEEE),
WOLFGANG ERTEL², AND MARKUS SCHNEIDER²

¹Institute of Neural Information Processing, University of Ulm, 89081 Ulm, Germany

²Institute for Artificial Intelligence, Ravensburg-Weingarten University of Applied Sciences, 88250 Weingarten, Germany

Corresponding author: Christopher M. A. Bonenberger (christopher.bonenberger@uni-ulm.de)

The work of Friedhelm Schwenker was supported by the German Research Foundation (DFG) under Grant SCHW 623/7-1.


ABSTRACT Principal component analysis (PCA) and kernel PCA allow the decorrelation of data with respect to a basis that is found via variance maximization. However, these techniques are based on pointwise correlations. Especially in the context of time series analysis this is not optimal. We present a novel generalization of PCA that allows to imprint any desired correlation pattern. Thus the proposed method can be used to incorporate previously known statistical dependencies between input variables into the model which is increasing the overall performance. This is achieved by generalizing the projection onto the direction of maximum variance—as known from PCA—to a projection onto a multi-dimensional subspace. We focus on the use of cyclic correlation patterns, which is especially of interest in the domain of time series analysis. Beneath introducing the presented variation of PCA, we discuss the role of this method with respect to other well-known time series analysis techniques.

INDEX TERMS PCA, discrete Fourier transform, filter, correlation, time series, kernel PCA, circulant matrices.

I. INTRODUCTION

Principal component analysis (PCA) is a widely-used method that is well-established in machine learning, statistics and signal processing. Being at the core of machine learning and statistics, it is one of the best known data analysis techniques [1]. Hence, examining the basic concept of PCA enables a deeper understanding of the relations between all the different fields of application. From this perspective, the intention of this paper is twofold. First of all we introduce a generalization of PCA (respectively kernel PCA) for time series analysis. Secondly we aim to relate our theory to the classical methods of statistics and signal processing in order to deepen the understanding of the interrelations between the methods used.

In the context of machine learning PCA is typically used to achieve dimensionality reduction in order to overcome

The associate editor coordinating the review of this manuscript and approving it for publication was Mohamed M. A. Moustafa .

the “curse of dimensionality” [2]. As in many applications the data dimension is quite large, it is worthwhile seeking a low dimensional data representation. Such a representation is likely to be found, when the observed data stems from a process that is sparse with respect to an appropriately chosen basis. Principal component analysis can be understood as an attempt to find such a basis (cf. [1]). Practically, PCA formulates the optimization problem of finding a vector that points into direction of maximum variance (according to the data set at hand). As a result, we find a set of uncorrelated vectors that correspond to the eigenvectors of the empirical covariance matrix. Dimensionality reduction can then be achieved by projecting the data under consideration onto the subspace spanned by these vectors. Typically this subset of eigenvectors is chosen with respect to the corresponding eigenvalues. Notably, it is possible to formulate PCA as a kernel algorithm, which is widening the application spectrum considerably [3].

Beyond its applications in machine learning PCA is connected to many classical signal processing and time series

analysis techniques. As an example PCA is tightly linked to the Karhunen-Loève transform.¹ Due to the fact that the discrete Fourier transform decorrelates certain stochastic processes (see for example [5]) the connection between PCA and the discrete Fourier transform (DFT) becomes evident. In this regard, there are several interesting time series analysis methods that are closely related to PCA. Singular spectrum analysis (SSA, cf. [6]) is a technique that decomposes time series into meaningful components, such as trend, seasonality and noise. Dynamic PCA (DPCA, cf. [7]) is a variation of PCA that is based on data set augmentation. Hence, in the context of temporal data PCA is related to SSA, DFT and DPCA. These connections are research subject in different contexts (see for example [6], [8], [9], [10]). While classical PCA is based on pointwise correlations, especially in the context of time series analysis it is desirable to incorporate prior knowledge about the data or the generating process. As an example, DPCA picks up the idea of shift-invariance by augmenting the data set with cyclic permutations of itself. As shown in [11] this is equivalent to hypothesizing a shift-invariant model. Singular spectrum analysis embeds the data at hand into a (trajectory) space and only afterwards performs PCA on the *embedded* data set.² The embedding in SSA is related to the so-called delay embedding (Takens' embedding) known from dynamical systems theory [12]. Here previous knowledge about the dynamics of the underlying process can be incorporated via such an embedding.

Reference [11] generalizes PCA towards shift-invariance, by matching κ -circulant matrices to the data set under consideration and shows that this framework resembles PCA as a special case. The fact that these κ -circulant matrices actually implement FIR filtering indicates the similarity of this method to SSA (see [10], [13]). From an algebraic perspective matching a κ -circulant matrix, means that the data at hand is projected onto a multi-dimensional subspace instead of a single vector. This subspace is spanned by a single vector and its cyclic permutations. Practically, this can be understood as an analog to the step from classical neurons to those in convolutional neural networks, because regarding the underlying correlation structure image processing is very similar to time series analysis.

In this work, we pick up the idea of PCA via projections onto cyclic subspaces and generalize it towards freely definable correlation patterns. This way, prior knowledge about the data under consideration or the generating process can be incorporated into the model, thus increasing the overall performance. In fact, the proposed method allows to model any desired correlation pattern, subsequently enhancing the recognition of such patterns. Moreover, we introduce a way to perform the proposed method in a high-dimensional feature

¹The definitions of KLT and PCA are often used interchangeably. A distinction between the Karhunen-Loève transform and PCA is for example given by [4].

²Embedding and eigendecomposition of the data are the first two steps of SSA. The complete SSA algorithm also intends the reconstruction of the decomposed data based on a subset of eigentriples [10].

space without explicitly computing the feature map, i.e., we kernelize the proposed method.

To sum it up, the first contribution of this work is a generalization of PCA with respect to arbitrarily structured subspaces. As aforementioned, classical PCA projects data onto an optimal 1-dimensional subspace (see Section I-A). In Section II we show a generalization for cyclically structured subspaces. In this regard, we build on the work of [11], which is reviewed in Section II-A and extended in Section II-B. However, while [11] limits these subspaces to have a κ -circulant structure, we introduce an extension of this theory that allows the projection onto any P -dimensional subspace in Section III. Examples on this theory that allow to find guidelines on how to use the proposed method are given in Section V.

The second important contribution is the formulation of the presented theory as a kernel method in Section IV. Here we build on the idea of embedding the data set at hand such that the basic theory of kernel PCA (KPCA), which is reviewed in Section I-B is sufficient. In Section VI we briefly go through some examples followed by a discussion of the presented theory and a conclusion.

A. RELATED METHODS

In the following, we review classical PCA and kernel PCA in order to pave the way to a more general method.

1) PRINCIPAL COMPONENT ANALYSIS

Suppose we have a data set consisting of N observations $\mathbf{x}_v \in \mathbb{R}^D$. We may also write this dataset as

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

Seeking a vector $\mathbf{u} \in \mathbb{R}^D$ that is most similar to the observations in \mathbf{X} leads to the optimization problem

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

Equating the derivative of the Lagrange function to zero leads to

$$\mathbf{X}\mathbf{X}^T \mathbf{u} = \lambda \mathbf{u} \iff \frac{\partial \mathbf{L}(\mathbf{u}, \lambda)}{\partial \mathbf{u}} = 2\mathbf{X}\mathbf{X}^T \mathbf{u} - 2\lambda \mathbf{u} = 0.$$

The sought vector \mathbf{u} is found as the eigenvector of $\mathbf{X}\mathbf{X}^T$ corresponding to the largest eigenvalue λ_{\max} as by definition the maximum of $\left\| \mathbf{u}^T \mathbf{X} \right\|_2^2$ is found as $\mathbf{u}^T \mathbf{X}\mathbf{X}^T \mathbf{u} = \lambda_{\max} \mathbf{u}^T \mathbf{u} = \lambda_{\max}$.

The complete set of eigenvectors of the symmetric matrix $\mathbf{X}\mathbf{X}^T$ forms an orthonormal basis for \mathbb{R}^D . The projection onto this basis, which can be considered as the *analysis* of some signal $\mathbf{x} \in \mathbb{R}^D$ with respect to the set of (eigen)vectors can be written as $\mathbf{U}^T \mathbf{x}$, where $\mathbf{X}\mathbf{X}^T \mathbf{U} = \mathbf{U} \Sigma \mathbf{U}^T$ with $\Sigma \in \mathbb{R}^{D \times D}$ being a diagonal matrix that holds the square roots of the eigenvalues of $\mathbf{X}\mathbf{X}^T$. When PCA is applied in order to reduce

dimensionality, data is projected onto Q eigenvectors. Typically only Q eigenvectors belonging to the Q largest eigenvalues are kept.

2) STATISTICAL INTERPRETATION OF PCA

A closer look at the Lagrange function $\mathbf{L}(\mathbf{u}, \lambda) = \mathbf{u}^T \mathbf{X} \mathbf{X}^T \mathbf{u} + \lambda \mathbf{u}^T \mathbf{u}$ resulting from (1) substantiates a statistical point of view, as $\mathbf{X} \mathbf{X}^T$ is proportional to the empirical covariance matrix $\mathbf{S} \in \mathbb{R}^{D \times D}$ if the data has zero mean,³ i.e.

$$[\mathbf{X} \mathbf{X}^T]_{j,k} = \sum_{v=1}^N x_{j,v} x_{k,v} \propto [\mathbf{S}]_{j,k}$$

where

$$[\mathbf{S}]_{j,k} = s_{jk} = \frac{1}{N-1} \sum_{v=1}^N (x_{j,v} - \bar{x}_j)(x_{k,v} - \bar{x}_k) \quad (2)$$

when the sample mean $\bar{x}_i = 0$ for all $i = 1, \dots, D$.

3) PCA AND SINGULAR VALUE DECOMPOSITION

Any matrix $\mathbf{X} \in \mathbb{R}^{D \times N}$ can be decomposed into $\mathbf{X} = \mathbf{U} \Sigma \mathbf{V}^T$, where $\Sigma \in \mathbb{R}^{D \times N}$ is a diagonal matrix holding the *singular values* (the roots of the eigenvalues of $\mathbf{X} \mathbf{X}^T$ and $\mathbf{X}^T \mathbf{X}$). The orthogonal matrices \mathbf{U} and \mathbf{V} hold the eigenvectors of $\mathbf{X} \mathbf{X}^T$ and $\mathbf{X}^T \mathbf{X}$ respectively, i.e. $\mathbf{X} \mathbf{X}^T \mathbf{U} = \mathbf{U} \Sigma \Sigma^T$ and $\mathbf{X}^T \mathbf{X} \mathbf{V} = \mathbf{V} \Sigma^T \Sigma$.

For zero-mean data the left eigenvectors \mathbf{U} of \mathbf{X} stem from the empirical covariance matrix, i.e., $\mathbf{S} \mathbf{U} \propto \mathbf{X} \mathbf{X}^T \mathbf{U} = \mathbf{U} \Sigma \Sigma^T$. Hence, the left eigenvectors \mathbf{U} correspond to the orthonormal basis that is found by PCA.

B. KERNEL PCA

The *kernelized* version of principal component analysis is—as any kernel method—based on a formulation of the algorithm by means of inner products between data set elements. Assuming zero-mean data, we may reformulate the principal components by means of the SVD $\mathbf{X} = \mathbf{U} \Sigma \mathbf{V}^T$, i.e.,

$$\mathbf{U}^T \mathbf{x} = (\Sigma^\dagger)^T \mathbf{V}^T \mathbf{X}^T \mathbf{x}. \quad (3)$$

Exploiting the fact that an inner product in a *reproducing kernel Hilbert space* \mathcal{H} (RKHS) can be written as $k(\mathbf{x}_i, \mathbf{x}_j) = \langle \phi(\mathbf{x}_i), \phi(\mathbf{x}_j) \rangle$ we may now compute the principal components in \mathcal{H} . Here $\phi: \mathbb{R}^D \rightarrow \mathcal{H}$ is the feature map associated to the kernel function k . In the following we use the index \mathcal{H} to indicate that a matrix or vector is associated to the RKHS \mathcal{H} defined by a kernel $k(\cdot, \cdot)$. Now (3) can be formulated in an RKHS as

$$\mathbf{U}_{\mathcal{H}}^T \phi(\mathbf{x}) = (\Sigma_{\mathcal{H}}^\dagger)^T \mathbf{V}_{\mathcal{H}}^T \begin{pmatrix} k(\mathbf{x}_1, \mathbf{x}) \\ \vdots \\ k(\mathbf{x}_N, \mathbf{x}) \end{pmatrix}, \quad (4)$$

where $\mathbf{U}_{\mathcal{H}}$, $\mathbf{V}_{\mathcal{H}}$ and $\Sigma_{\mathcal{H}}$ refer to the singular value decomposition of the mapped data $\mathbf{X}_{\mathcal{H}} = [\phi(\mathbf{x}_1), \dots, \phi(\mathbf{x}_N)]^T$. Note

³This is not a strong assumption. In machine learning data is often standardized to zero-mean and unit-variance by default.

that $\mathbf{U}_{\mathcal{H}}$ is computed via outer products, which means it is required to explicitly compute the map ϕ in order to find $\mathbf{U}_{\mathcal{H}}$. Only $\mathbf{V}_{\mathcal{H}}$ can be computed using the *kernel trick*, because

$$\mathbf{X}_{\mathcal{H}}^T \mathbf{X}_{\mathcal{H}} \mathbf{V}_{\mathcal{H}} = \mathbf{V}_{\mathcal{H}} \Sigma_{\mathcal{H}}^T \Sigma_{\mathcal{H}}$$

solely involves the data in terms of inner products. The matrix $\mathbf{X}_{\mathcal{H}}^T \mathbf{X}_{\mathcal{H}}$ is known as *kernel matrix*

$$\tilde{\mathbf{K}} = \begin{bmatrix} k(\mathbf{x}_1, \mathbf{x}_1) & \cdots & k(\mathbf{x}_1, \mathbf{x}_N) \\ \vdots & \ddots & \vdots \\ k(\mathbf{x}_N, \mathbf{x}_1) & \cdots & k(\mathbf{x}_N, \mathbf{x}_N) \end{bmatrix} \in \mathbb{R}^{N \times N}.$$

As $\mathbf{V}_{\mathcal{H}}$ and $\Sigma_{\mathcal{H}}$ can be found from the eigendecomposition of $\tilde{\mathbf{K}}$, we can compute the nonlinear projection in (4) without explicitly using ϕ .

Again we may assume $\mathbf{X}_{\mathcal{H}} \mathbf{X}_{\mathcal{H}}^T \propto \mathbf{S}_{\mathcal{H}}$. However, in order to be sure that the sample covariance matrix $\mathbf{S}_{\mathcal{H}} \in \mathbb{R}^{N \times N}$ of the mapped data is found via

$$\mathbf{S}_{\mathcal{H}} \propto \sum_{v=1}^N \phi(\mathbf{x}_v) \phi(\mathbf{x}_v)^T$$

we have to assume zero-mean data. Yet, since the mapped data is not available, the kernel matrix has to be centered according to (see [14])

$$\mathbf{K} = \tilde{\mathbf{K}} - \tilde{\mathbf{K}} \mathbf{1}_N - \mathbf{1}_N \tilde{\mathbf{K}} + \mathbf{1}_N \tilde{\mathbf{K}} \mathbf{1}_N.$$

The matrix $\mathbf{1}_N \in \mathbb{R}^{N \times N}$ has elements $[\mathbf{1}_N]_{i,j} = 1/N$. Hence, the nonlinear mapping of a data set $\mathbf{X} \in \mathbb{R}^{D \times N}$ is found from

$$\mathbf{Y}_{\mathcal{H}} = (\Sigma_{\mathcal{H}}^\dagger)^T \mathbf{V}_{\mathcal{H}}^T \mathbf{K},$$

where

$$\mathbf{K} \mathbf{V}_{\mathcal{H}}^T = \mathbf{V}_{\mathcal{H}}^T \Lambda_{\mathcal{H}} \quad (5)$$

with $\Lambda_{\mathcal{H}} = \Sigma_{\mathcal{H}}^T \Sigma_{\mathcal{H}}$ holding the eigenvalues of \mathbf{K} . New data $\mathbf{X}' \in \mathbb{R}^{D \times N'}$ is mapped via

$$\mathbf{Y}'_{\mathcal{H}} = (\Sigma_{\mathcal{H}}^\dagger)^T \mathbf{V}_{\mathcal{H}}^T \mathbf{K}',$$

where $[\tilde{\mathbf{K}}']_{i,j} = k(\mathbf{x}'_i, \mathbf{x}_j)$ is centered as

$$\mathbf{K}' = \tilde{\mathbf{K}}' - \mathbf{1}'_N \tilde{\mathbf{K}}' - \tilde{\mathbf{K}}' \mathbf{1}_N + \mathbf{1}'_N \tilde{\mathbf{K}}' \mathbf{1}_N$$

and $\mathbf{1}'_N \in \mathbb{R}^{N' \times N}$ with $[\mathbf{1}'_N]_{i,j} = (1/N)$.

II. CYCLIC CORRELATION PATTERNS

We begin with the formulation of PCA for cyclic correlation patterns as proposed by [11], i.e., we formulate PCA for a circulant structure. However, we directly extend the definition of basic κ -circulant matrices by involving down-sampling in the input space via a parameter ρ and truncation of \mathbf{C} . While the parameters L , κ and ρ are known from Wavelet theory⁴

⁴The parameters κ and ρ realize down-sampling of the filtered or the input signal respectively. The latter can be used for up-sampling. All together, these structures are essential to an algebraic (polyphase) formulation of the discrete multi-level wavelet transform (cf. [15] or [16]).

and CNNs, the parameter M is used for DPCA, SSA and CNNs.

Possibly, the most common access to these parameters is via CNNs, as L corresponds to the filter width, ρ realizes a dilated convolution, κ is a down-sampling factor (stride) and M avoids undesired boundary effects⁵ (see [17]).

A. TRUNCATED κ -CIRCULANT STRUCTURES

As aforementioned instead of solving (1) for a vector we solve it for a circulant matrix, i.e., we project the data set onto a subspace that is defined via a truncated κ -circulant.

A basic circulant matrix \mathbf{C} can be written as

$$\mathbf{C} = \sum_{l=1}^L c_l \mathbf{P}^{l-1} \text{ with } \mathbf{P} = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & & 0 \\ \vdots & & & \ddots & \vdots \\ 0 & & & & 1 \\ 1 & 0 & 0 & \cdots & 0 \end{bmatrix} \in \mathbb{R}^{D \times D}. \tag{6}$$

This may be generalized to a truncated κ -circulant with parameters L, κ, M and ρ as

$$\mathbf{G} = \mathbf{M} \sum_{l=1}^L g_l \mathbf{P}^{\rho(l-1)}, \tag{7}$$

where the masking matrix \mathbf{M} is down-sampling by a factor κ and truncating from the M -th row onward, i.e.,

$$[\mathbf{M}]_{j,k} = \begin{cases} 1 & \text{if } M \geq j = k \in \mathbf{M} \\ 0 & \text{else} \end{cases} \tag{8}$$

with $\mathbf{M} = [1, \kappa + 1, 2\kappa + 1, \dots, \lfloor D/\kappa - 1 \rfloor \kappa + 1]$. Some examples are given in Fig. 1 (panels (a)-(f)).

Analogously to PCA, the solution of

$$\max_{\mathbf{g} \in \mathbb{R}^D} \left\{ \|\mathbf{G}\mathbf{x}\|_F^2 \right\} \text{ s.t. } \|\mathbf{g}\|_2^2 = 1. \tag{9}$$

is found from an eigenvalue problem. Yet, the covariance matrix is replaced by a scattering matrix $\mathbf{Z} \in \mathbb{R}^{L \times L}$ with

$$[\mathbf{Z}]_{j,k} = \sum_{v=1}^N \langle \mathbf{x}_v, \mathbf{P}^{-\rho(k-1)} \mathbf{M} \mathbf{P}^{\rho(j-1)} \mathbf{x}_v \rangle. \tag{10}$$

Solving

$$\mathbf{Z}\mathbf{g} = \lambda\mathbf{g} \tag{11}$$

leads to a set of L decorrelated⁶ vectors that span \mathbb{R}^L .

B. PROJECTIONS ONTO STRUCTURED SUBSPACES

For a reasonable choice of parameters L, M, κ and ρ the eigenvectors $\{\mathbf{g}_1, \dots, \mathbf{g}_L\}$ from (11) constitute a finite frame (cf. [18]). Each eigenvector \mathbf{g}_i defines a matrix \mathbf{G}_i according

⁵With $M = D - L + 1$ zero-padding is not necessary.

⁶Since \mathbf{Z} is necessarily symmetric its eigenvectors are orthogonal.

to (7), hence, the set of matrices $\{\mathbf{G}_i\}_{i=1,\dots,L}$ typically constitutes a frame⁷ for \mathbb{R}^D .

The fact, that mapping to $\{\mathbf{G}_i\}_{i=1,\dots,L}$ potentially leads to an over-complete representation results in a major difference to classical PCA. Namely, a single data point $\mathbf{x} \in \mathbb{R}^D$ may have a set of counterparts, i.e., $\mathbf{y}_1, \dots, \mathbf{y}_P$ where

$$P = \lceil M/\kappa \rceil.$$

First of all, this contradicts the idea of dimensionality reduction. However, as we optimize with respect to variance, it is evident that the new axis hold $\|\mathbf{G}_i\mathbf{x}\|_2^2 \in \mathbb{R}$ instead of $\mathbf{G}_i\mathbf{x} \in \mathbb{R}^P$ when $P > 1$. Noting that $\mathbf{G}_i\mathbf{x}$ is related to linear filtering of \mathbf{x} with respect to the filter kernel⁸ \mathbf{g} allows a simple interpretation: each eigenvector corresponds to a band-filter \mathbf{G}_i and $\|\mathbf{G}_i\mathbf{x}\|_2^2$ is the power of \mathbf{x} within this frequency band. Geometrically $\mathbf{G}_i\mathbf{x}$ is the projection of \mathbf{x} onto the subspace defined by \mathbf{G}_i and $\|\mathbf{G}_i\mathbf{x}\|_2^2$ is the corresponding variance. Here, actually our measure for variance is the total dispersion (cf. [19]). Note that this is non-linear and not invertible.

From the statistical point of view, the structure of \mathbf{G} encodes dependencies between different (time) coordinates. As an example, let $L = D$ and $M = 3$ and $\rho = \kappa = 1$, i.e.,

$$\mathbf{G} = \begin{bmatrix} g_1 & g_2 & g_3 & g_4 & \cdots & g_L \\ g_L & g_1 & g_2 & g_3 & \cdots & g_{L-1} \\ g_{L-1} & g_L & g_1 & g_2 & \cdots & g_{L-2} \\ 0 & 0 & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & & \vdots \end{bmatrix}.$$

In fact, the above matrix hypothesizes, that each coordinate is coupled to its direct left and right neighbor. In other words, with such a structure of \mathbf{G} we encode the assumption that (temporally) neighboring observations are correlated.

More examples on possible patterns are given in Fig. 1.⁹ A mentionable special case arises when using a structure that prospects for global coupling (Fig. 1, panel (a)) between all variables. In this case the matrix \mathbf{Z} becomes a symmetric circulant (the auto-covariance matrix) whose eigenvectors resemble the discrete Fourier basis. Thus under the assumption that all variables are coupled we return to Fourier analysis—independently of the data under consideration (cf. [11]).

III. ARBITRARY CORRELATION PATTERNS

Using a cyclically structured matrix \mathbf{G} is a constraint when modeling correlation patterns. Although circulant structures are reasonable in the domain of time series analysis, for other types of data different patterns could be desirable. Especially

⁷For special choices of the parameters κ, L, M and ρ such a linear map might be singular, i.e., not a frame. However, such considerations are out of the scope of this work.

⁸The filter kernel should not be confused with a kernel function associated to a RKHS (cf. Section I-B).

⁹Regarding Fig. 1 the question arises of whether a matrix is cyclic or non-cyclic. We refer to a “cyclic matrix” in the context of (7) using the parameters L, M, κ and ρ , i.e., any truncated κ -circulant matrix is referred to as cyclic.

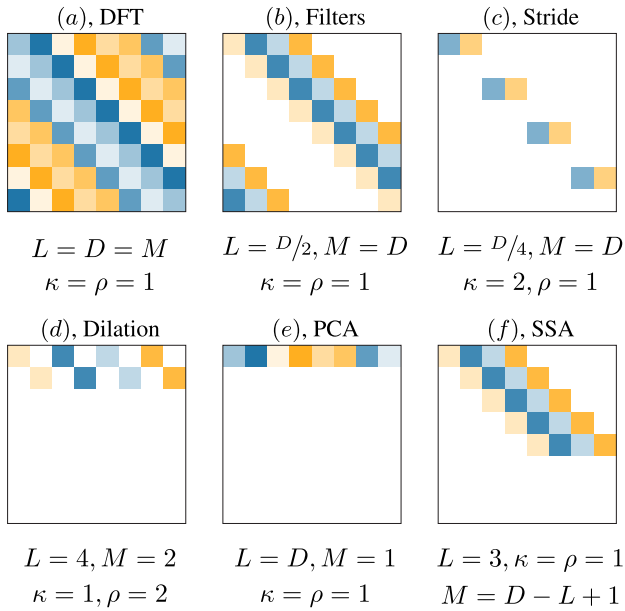


FIGURE 1. Some examples of possible matrices \mathbf{G} encoding different correlation patterns. For each matrix the defining vector $\mathbf{g} \in \mathbb{R}^L$ has randomly chosen components g_j in order to visualize the hypothesized coupling between (time) coordinates $[\mathbf{x}]_j$.

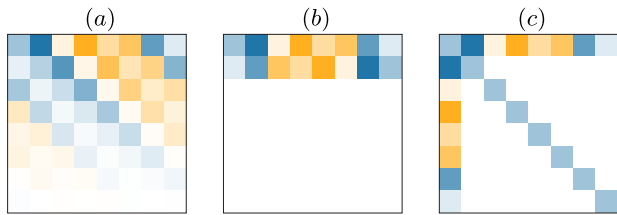


FIGURE 2. Examples on arbitrary correlation structures (visualized via randomly chosen components g_j).

the condition of global shift-invariance on a certain scale—typically useful in image and time series analysis—is not always met. In the following, we propose a generalization that drops these constraints.

Instead of using a cyclic permutation matrix $\mathbf{P} \in \mathbb{R}^{D \times D}$ (cf. (6)) we use a set of arbitrary real-valued matrices $\Pi_l \in \mathbb{R}^{P \times D}$. Now we solve

$$\max_{\mathbf{g} \in \mathbb{R}^D} \left\{ \left\| \sum_{l=1}^L g_l \Pi_l \mathbf{X} \right\|_F^2 \right\} \quad \text{s.t.} \quad \|\mathbf{g}\|_2^2 = 1. \quad (12)$$

Some examples on non-circulant¹⁰ correlation patterns are shown in Fig. 2, where all examples except for panel (a) visualizes binary coupling, i.e., the corresponding matrices Π_l are (0, 1)-matrices. The matrix in panel (a) is the same as in panel Fig. 1, (a), yet, correlation is fading out over time/distance, i.e., the closer the closer the variables, the stronger the coupling. Panel (b) shows a symmetric correlation and in

¹⁰Actually all correlation patterns depicted in Fig. 1 can be applied to (12), because (12) generalizes (9).

panel (c) the underlying hypothesis is that all variables are coupled to the first. The example in panel (c) shows that variables that occur often (here the first variable) will pronounced strongly compared to the others (more details in Section V).

Solving (12) follows the solution to (9). The partial derivative of the corresponding Lagrangian $\mathbf{L}(\mathbf{g}, \lambda)$ is

$$\begin{aligned} \frac{\partial \mathbf{L}}{\partial g_k} &= \sum_{v=1}^N \mathbf{x}_v^T \left(g_1 \left(\Pi_1^T \Pi_k + \Pi_k^T \Pi_1 \right) + \dots \right. \\ &\quad \left. \dots + g_L \left(\Pi_L^T \Pi_k + \Pi_k^T \Pi_L \right) \right) \mathbf{x}_v + 2\lambda g_k. \end{aligned}$$

Again the optimal vector $\mathbf{g} \in \mathbb{R}^L$ is found from an eigenvalue problem $\mathbf{Z}\mathbf{g} = \lambda\mathbf{g}$. Using distributivity and symmetry of the dot product we find $\mathbf{x}^T (\Pi_i^T \Pi_j + \Pi_j^T \Pi_i) \mathbf{x} = 2\mathbf{x}^T \Pi_i^T \Pi_j \mathbf{x}$ such that the components of $\mathbf{Z} \in \mathbb{R}^{L \times L}$ can be written as

$$[\mathbf{Z}]_{i,j} = \sum_{v=1}^N \langle \mathbf{x}_v, \Pi_i^T \Pi_j \mathbf{x}_v \rangle.$$

Note that (12) generalizes (9) and hence also may resemble (linear) PCA, i.e., when $[\Pi_i]_{1,i} = 1 \forall i \in [1, D]$ and all other entries are zero (12) is analogue to PCA.

As there are no other restrictions put on the matrices Π_i than being real-valued, any statistical dependency can be incorporated to the model. More specifically, it is possible to hypothesize specific statistical dependencies between different variables (coordinates) by coupling the corresponding entries in \mathbf{G} . Notably, one is not restricted to binary coupling, i.e., an arbitrary real coupling factor can be chosen in order to encode a certain strength of correlation.

IV. KERNELIZED CORRELATION ANALYSIS

Our next step is the kernelization of (9), i.e., we formulate the above linear optimization problem respectively its solution by means of inner products between data points. This allows to find vectors in some RKHS \mathcal{H} , that are optimal with respect to a certain correlation pattern. We begin with cyclically structured correlation patterns that are defined by the parameters κ, L, M and ρ .

A straightforward approach can be deduced from Section I-B after realizing that the product $\mathbf{G}\mathbf{X}$ is equivalent to the matrix-vector multiplication $\mathbf{g}^T \mathbf{W}$, where \mathbf{W} is a restructured data set. The restructuring according to Section II-A simply has to follow the correlation-pattern defined by \mathbf{G} . In fact,

$$\begin{aligned} \|\mathbf{G}\mathbf{X}\|_F^2 &= \left\| \mathbf{M} \sum_{l=1}^L g_l \mathbf{P}^{\rho(l-1)} \begin{bmatrix} x_{1,1} & \dots & x_{1,N} \\ \vdots & \ddots & \vdots \\ x_{D,1} & \dots & x_{D,N} \end{bmatrix} \right\|_F^2 \\ &= \left\| \mathbf{g}^T \mathbf{W} \right\|_2^2 \end{aligned}$$

with

$$\mathbf{W} = \begin{bmatrix} x_{1,1} & \cdots & x_{D,1} & \cdots & x_{D,N} \\ x_{1+\rho,1} & \cdots & x_{D+\rho,1} & \cdots & x_{D+\rho,N} \\ \vdots & & \vdots & & \vdots \\ x_{1+\rho(L-1),1} & \cdots & x_{D+\rho(L-1),N} \end{bmatrix} \mathbf{M}_R,$$

where $\mathbf{M}_R \in \mathbb{R}^{DN \times DN}$ performs a reduction of the augmented matrix according to the choice of κ and M . It is a block-diagonal matrix with sub-matrices being equal to \mathbf{M} on its diagonal.

We refer to the above restructuring of \mathbf{X} as “embedding” (cf. Section I). We formalize this as $\mathcal{E} : \mathbb{R}^D \rightarrow \mathbb{R}^{L \times P}$. More specifically, a single observation $\mathbf{x} \in \mathbb{R}^D$ is embedded as

$$\begin{pmatrix} x_1 \\ \vdots \\ x_D \end{pmatrix} \mapsto \begin{bmatrix} x_1 & x_\kappa & \cdots & x_P \\ x_{1+\rho} & x_{\kappa+\rho} & \cdots & x_{P+\rho} \\ \vdots & \vdots & \ddots & \vdots \\ x_{1+(L-1)\rho} & x_{\kappa+(L-1)\rho} & \cdots & x_{P+(L-1)\rho} \end{bmatrix} \quad (13)$$

Generalizing this embedding process to arbitrary correlation patterns based on a set of matrices $\{\Pi_l\}_{l=1,\dots,L}$ —as proposed in Section III—leads to the map

$$\mathbf{x} \mapsto \mathcal{E}(\mathbf{x}) = \begin{bmatrix} | & & | \\ \Pi_1 \mathbf{x} & \cdots & \Pi_L \mathbf{x} \\ | & & | \end{bmatrix} \quad (14)$$

Rewriting $\|\mathbf{G}\mathbf{x}\|_F^2$ according to (14), the optimization problem in (9) takes the form of classical PCA (cf. (1)). Hence, we may proceed as described in Section I-B, i.e., we rewrite the projection onto the loadings¹¹ in \mathcal{H} using the eigendecomposition of a centered kernel matrix \mathbf{K} associated to \mathbf{W} (respectively $\mathbf{W}_{\mathcal{H}}$). Clearly, mapping new data points again requires the embedding of those.

In Section II-A the result of the projection of an observation \mathbf{x} onto a subspace defined by an eigenvector \mathbf{g}_i of \mathbf{Z} was given as $\|\mathbf{G}_i \mathbf{x}\|_2^2$. It is of particular importance, that this can not be adopted in \mathcal{H} , because the embedding $\mathcal{E}(\mathbf{x})$ is not a single vector. Instead, the result is computed from the set of embedded vectors. More precisely, assume we have a single observation \mathbf{x}' that is embedded as

$$\mathcal{E}(\mathbf{x}') = \mathbf{W}' = \begin{bmatrix} | & & | \\ \mathbf{w}'_1 & \cdots & \mathbf{w}'_P \\ | & & | \end{bmatrix}$$

and a data set consisting of observations $\{\mathbf{x}_v\}_{v=1,\dots,N}$ with embeddings $\{\mathbf{w}_v\}_{v=1,\dots,NP}$. The total variance within a subspace defined by $\mathbf{v}_{\mathcal{H}}$ is then given as

$$\|\mathbf{y}_{\mathcal{H}}\|_2^2 = \|\sigma_{\mathcal{H}} \mathbf{v}_{\mathcal{H}}^T \mathbf{K}'\|_2^2$$

where \mathbf{K}' is the centered version of the matrix $\tilde{\mathbf{K}}' \in \mathbb{R}^{P \times NP}$ with components

$$[\tilde{\mathbf{K}}']_{i,v} = k(\mathbf{w}'_i, \mathbf{x}_v).$$

¹¹According to [1] we refer to the eigenvectors of \mathbf{S} as loadings.

V. THEORETICAL EXAMPLES

Before continuing with numerical examples, we recapitulate the presented theory by means of theoretical examples. For this purpose let \mathbf{x} be a D -dimensional real-valued random vector with zero-mean, i.e., $\mathbf{x}^T = [x_1 \ x_2 \ \cdots \ x_D]$ with $\bar{x}_i = 0 \ \forall i \in [1, D]$.

A. CYCLIC PATTERNS

We begin with the simple circulant correlation structure

$$\mathbf{G} = \begin{bmatrix} g_1 & g_2 & g_3 \\ g_3 & g_1 & g_2 \\ g_2 & g_3 & g_1 \end{bmatrix} = \begin{bmatrix} -\mathbf{g}_1^T & - \\ -\mathbf{g}_2^T & - \\ -\mathbf{g}_3^T & - \end{bmatrix}. \quad (15)$$

As stated in Section II, this structure hypothesizes that there are no positional dependencies. This becomes more obvious when noting that:

$$\|\mathbf{G}\mathbf{x}\|_2^2 = \left\| (g_1 \ g_2 \ g_3) \begin{bmatrix} x_1 & x_3 & x_2 \\ x_2 & x_1 & x_3 \\ x_3 & x_2 & x_1 \end{bmatrix} \right\|_2^2.$$

Hence the corresponding covariance matrix is

$$\mathbf{S} = \begin{bmatrix} s_{11}+s_{22}+s_{33} & s_{12}+s_{23}+s_{31} & s_{13}+s_{21}+s_{32} \\ s_{21}+s_{32}+s_{13} & s_{22}+s_{33}+s_{11} & s_{23}+s_{31}+s_{12} \\ s_{31}+s_{12}+s_{23} & s_{32}+s_{13}+s_{21} & s_{33}+s_{11}+s_{22} \end{bmatrix},$$

where s_{ij} is the covariance between the variables x_i and x_j (cf. (2), note that $s_{ij} = s_{ji}$). From this example it can be seen that each row \mathbf{g}_i^T of \mathbf{G} corresponds to a covariance matrix with a certain structure, i.e.,

$$\mathbf{S} = \begin{bmatrix} s_{11} & s_{12} & s_{13} \\ s_{21} & s_{11} & s_{23} \\ s_{31} & s_{32} & s_{33} \end{bmatrix} + \begin{bmatrix} s_{22} & s_{23} & s_{21} \\ s_{32} & s_{33} & s_{31} \\ s_{13} & s_{13} & s_{11} \end{bmatrix} + \begin{bmatrix} s_{33} & s_{31} & s_{32} \\ s_{13} & s_{11} & s_{12} \\ s_{23} & s_{21} & s_{22} \end{bmatrix}.$$

This shows by example how the structure of \mathbf{G} is reflected in \mathbf{S} . For larger D it becomes obvious, that for a structure as in (15) (which stems from $L = D = M, \kappa = \rho = 1$, cf. Fig. 1, panel (a)) the covariance matrix becomes the cyclic auto-covariance matrix, i.e., a symmetric circulant matrix (cf. Section II-B).

B. ARBITRARY PATTERNS

All the structures in Fig. 1 are well-known from time series analysis and signal processing. Yet, it suggests itself to generalize these approaches. We start with an example, that is still motivated by time series analysis, namely the structure shown in Fig. 2, panel (a). This is equivalent to a structure as in Section V-A only that the time shifts are weighted. We define the new structure \mathbf{G} with row vectors $\mathbf{g}_1 := \mathbf{g}_1$, $\mathbf{g}_2 := 0.5\mathbf{g}_2$ and $\mathbf{g}_3 := \mathbf{0}$, i.e.,

$$\mathbf{G} = \begin{bmatrix} g_1 & g_2 & g_3 \\ \frac{g_3}{2} & \frac{g_1}{2} & \frac{g_2}{2} \\ 0 & 0 & 0 \end{bmatrix} = \begin{bmatrix} -\mathbf{g}_1^T & - \\ -\mathbf{g}_2^T & - \\ -\mathbf{g}_3^T & - \end{bmatrix}.$$

such that

$$\|\mathbf{G}\mathbf{x}\|_2^2 = \left\| (g_1 \ g_2 \ g_3) \begin{bmatrix} x_1 & \frac{x_3}{2} & 0 \\ x_2 & \frac{x_1}{2} & 0 \\ x_3 & \frac{x_2}{2} & 0 \end{bmatrix} \right\|_2^2.$$

Hence the corresponding covariance matrix is

$$\mathbf{S} = \begin{bmatrix} s_{11} & s_{12} & s_{13} \\ s_{21} & s_{22} & s_{23} \\ s_{31} & s_{32} & s_{33} \end{bmatrix} + \frac{1}{4} \begin{bmatrix} s_{22} & s_{23} & s_{21} \\ s_{32} & s_{33} & s_{31} \\ s_{13} & s_{13} & s_{11} \end{bmatrix}.$$

Of course, this is more useful for larger D , yet, we chose $D = 3$ for the sake of layout.

In the previous examples, each variable appeared only once per row. However, this is not necessary. Although a further discussion is out of the scope of this work, we give a short example: let

$$\mathbf{G} = \begin{bmatrix} g_1 & g_2 & g_3 \\ g_2 & 0 & g_2 \end{bmatrix},$$

then

$$\mathcal{E}(\mathbf{x}) = \begin{bmatrix} x_1 & 0 \\ x_2 & x_1 + x_3 \\ x_3 & 0 \end{bmatrix}$$

such that

$$\mathbf{S} = \begin{bmatrix} s_{11} & s_{12} & s_{13} \\ s_{21} & s_{22} + s_{11} + 2s_{31} + s_{33} & s_{23} \\ s_{31} & s_{32} & s_{33} \end{bmatrix}.$$

One possible application for non-cyclic structures is in image processing, where more complex correlation patterns can be used in order to define templates for pattern matching.

VI. NUMERICAL EXAMPLES

First of all, it is important to note that the implementation of the proposed method relies on classical (kernel) PCA using a data set that is restructured according to (13) (or more generally (14)). Hence, the algorithmic steps are analogous to classical (kernel) PCA, except for preliminary embedding of the data set at hand via $\mathbf{X} \mapsto \mathcal{E}(\mathbf{X})$ and subsequent “de-embedding” by means of the 2-norm ($\|\mathbf{G}\mathbf{x}\|_2^2$ or $\|\sigma_{\mathcal{H}}\mathbf{v}_{\mathcal{H}}^T \mathbf{K}'\|_2^2$ for kernel PCA). This his makes all results easily reproducible.

In the following, we demonstrate the theory presented above based on a few different data sets. We use two rather different data sets from the UCR Time Series Archive [20], synthetic data drawn from a stochastic process and a toy example in \mathbb{R}^3 for visualization of the proposed method. In all examples, if a kernel is used, we use the radial basis function kernel, i.e., $k(\mathbf{x}_1, \mathbf{x}_2) = \exp(-\|\mathbf{x}_1 - \mathbf{x}_2\|_2^2 / \sigma)$ with $\sigma = D$ where $\mathbf{x}_1, \mathbf{x}_2 \in \mathbb{R}^D$.

For all time series classification data sets, we give the accuracy of a simple 1-nearest neighbor classifier and as a baseline we compare our method to the classical PCA respectively kernel PCA. The classifier is trained on the transformed data

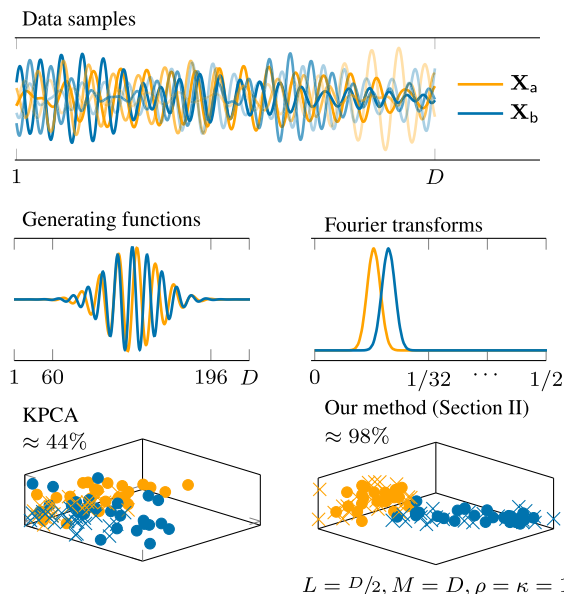


FIGURE 3. Comparison of KPCA and cyclic correlation patterns according to Section II. The data set consists of observations $\mathbf{X}_a \in \mathbb{R}^{256 \times 50}$ and $\mathbf{X}_b \in \mathbb{R}^{256 \times 50}$ defined by two differently parameterized MA processes (examples in the left panels). The generating functions and the corresponding spectral densities of these processes are shown in the middle two panels. The projections of the original data and test data—a 50/50 split—are shown in the right panels (test data points are marked with a cross). The given values are test-data classification accuracies for a 1-nearest-neighbor classifier (50/50-split).

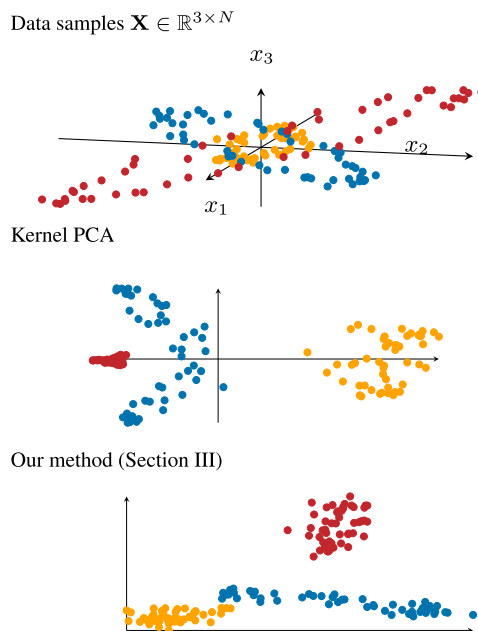


FIGURE 4. In this toy example the data set (upper panel) consists of three clusters, namely the observations lie around three concentric circles with different radius and inclination. In the middle panel the result using kernel PCA is depicted. The bottom panel shows the decorrelation according to the structure defined in (16). As can be seen, the coupling between the first two variables is decisive.

set (in \mathbb{R}^3 , i.e., $Q = 3$) using a 50/50 train-test-split, i.e., 50% of the data set are used to determine the adaptive feature map. The accuracy is evaluated from the test-data.

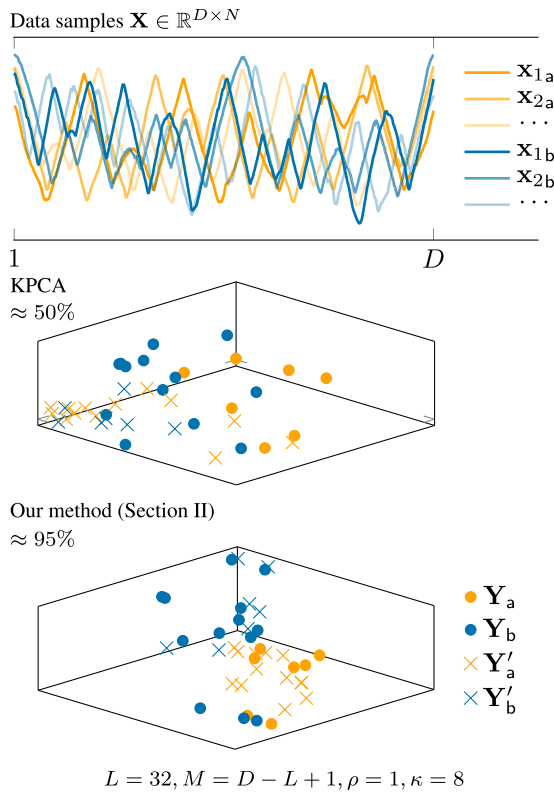


FIGURE 5. Comparison of kernel PCA and circulant component analysis at the example of the Beetle Fly data set from the UCR Archive. $L = 32$ as the data points seem to be correlated within this range ($D = 512$). Sample data is plotted in the left graph.

The data in Fig. 3 is obtained from moving-average (MA) processes (cf. [21]). The two different MA processes have generating functions of the form

$$\mathbf{p} = \sin(2\pi f \mathbf{t}) \odot \exp(-\mathbf{t}^2)$$

where $\mathbf{t} = [-4, \dots, 4]^T \in \mathbb{R}^D$. The first cluster in Fig. 3 has a generating function \mathbf{p}_a with $f = 2$ while the second cluster is based on \mathbf{p}_b with $f = 2.5$.

Notably, the two classes are hard to distinguish in time domain¹² (see KPCA in Fig. 3) and the distributions are overlapping in frequency domain, yet, we are able to separate the classes. This is because our method allows a compromise between time and frequency resolution.

Fig. 4 shows a toy example that is well-known in the context of kernel PCA. The data includes three clusters that correspond to samples drawn in the surrounding of three concentric circles around the x_3 -axis. Hence, the first two variables x_1 and x_2 are coupled. Knowing, that the coupling between x_1 and x_2 is of importance we set up the following matrix

$$\mathbf{G} = \sum_{l=1}^3 g_l \Pi_l = g_1 \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} + \dots \\ \dots + g_2 \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \end{bmatrix} + g_3 \begin{bmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix}$$

¹²Note that KPCA fails on the test data.

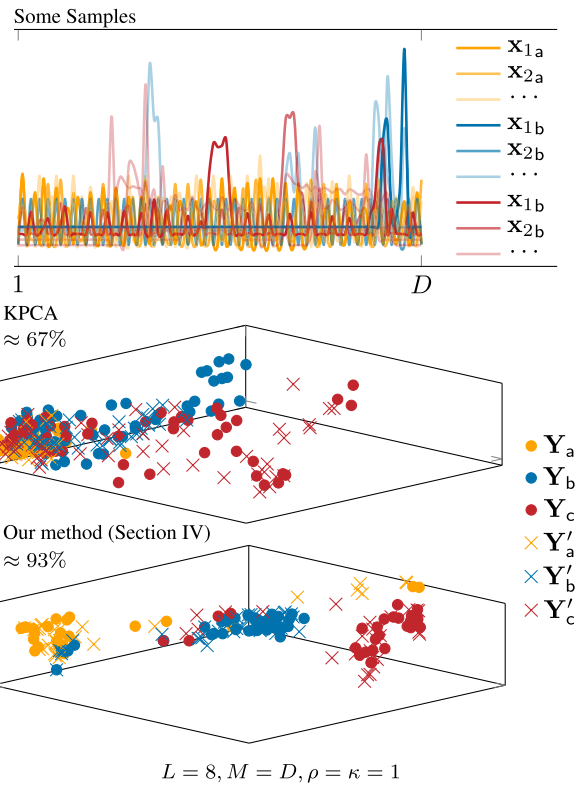


FIGURE 6. Example on kernel PCA and nonlinear circulant component analysis used on a subset of the “Electric Devices” data set of the UCR Archive ($Q = 3$). For better visualization here only three of the originally 7 classes are used.

$$= \begin{bmatrix} g_1 & g_2 & g_3 \\ g_2 & g_1 & 0 \end{bmatrix}. \tag{16}$$

Via Π_1 and Π_2 we hypothesize that there is a relevant coupling between the first two variables while the third variable is assumed to be independent. Note that we ignore the fact that the third variable is also coupled to the second (observe the inclination). Of course, this knowledge could be included by setting $[\Pi_3]_{2,2} = 1$.

Finally, the examples in Fig. 5 and Fig. 6 are based on the “Beetle Fly” data set and the “Electric Devices” data set out of the UCR Time Series Archive. These examples demonstrate that the proposed approach also allows meaningful decorrelation of data without knowing the generating process, i.e., visual patterns in the data can be utilized. However, the choice of the two data sets is not arbitrary. Here we have chosen data sets with a (weakly) shift invariant characteristic as the projection onto cyclic subspaces is a tool that is made for this kind of data.

VII. CONCLUSION

Provided that certain correlations or structures of the data under consideration are known, it should be possible to use this knowledge in order to improve results. We have proposed a generalization of PCA that can be used to incorporate prior knowledge into adaptive data analysis. With regard to time

series analysis we presented a framework that allows to use cyclic structures that are tightly linked to common signal processing techniques, which enables a simple interpretation and application. Beyond that, we generalized this theory to arbitrary structures making it possible to involve any kind previous knowledge about dependencies in the data. Finally, we formulated this method as a kernel algorithm thus enlarging the field of applications.

ACKNOWLEDGMENT

The authors would like to thank Ulm University for continuous support.

REFERENCES

- [1] I. T. Jolliffe, *Principal Component Analysis*, vol. 2. Springer, 2002.
- [2] F. Kherif and A. Latypova, "Principal component analysis," in *Machine Learning*, A. Mechelli and S. Vieira, Eds. New York, NY, USA: Academic, 2020, pp. 209–225.
- [3] B. Schölkopf, A. Smola, and K.-R. Müller, "Kernel principal component analysis," in *Proc. Int. Conf. Artif. Neural Netw.* Springer, 1997, pp. 583–588.
- [4] J. J. Gerbrands, "On the relationships between SVD, KLT and PCA," *Pattern Recognit.*, vol. 14, nos. 1–6, pp. 375–381, 1981.
- [5] A. S. Willsky, G. W. Wornell, and J. H. Shapiro, "Stochastic processes, detection and estimation," *Course Notes MIT*, vol. 6, p. 109, 2003.
- [6] N. Golyandina and A. Zhigljavsky, *Singular Spectrum Analysis for Time Series*, vol. 2. Springer, 2020.
- [7] W. Ku, R. H. Storer, and C. Georgakakis, "Disturbance detection and isolation by dynamic principal component analysis," *Chemometrics Intell. Lab. Syst.*, vol. 30, no. 1, pp. 179–196, 1995.
- [8] E. Bozzo, R. Carniel, and D. Fasino, "Relationship between singular spectrum analysis and Fourier analysis: Theory and application to the monitoring of volcanic activity," *Comput. Math. Appl.*, vol. 60, no. 3, pp. 812–820, Aug. 2010.
- [9] J. B. Elsner and A. A. Tsonis, *Singular Spectrum Analysis: A New Tool in Time Series Analysis*. Springer, 1996.
- [10] C. Bonenberger, W. Ertel, F. Schwenker, and M. Schneider, "Singular spectrum analysis and circulant maximum variance frames," *Adv. Data Sci. Adapt. Anal.*, Jul. 2022, Art. no. 2250008.
- [11] C. Bonenberger, W. Ertel, and M. Schneider, " κ -circulant maximum variance bases," in *Proc. German Conf. Artif. Intell.*, S. Edelkamp, R. Möller, and E. Rueckert, Eds. Springer, 2021, pp. 17–29.
- [12] H. Kantz and T. Schreiber, *Nonlinear Time Series Analysis*, vol. 7. Cambridge, U.K.: Cambridge Univ. Press, 2004.
- [13] T. J. Harris and H. Yuan, "Filtering and frequency interpretations of singular spectrum analysis," *Phys. D, Nonlinear Phenomena*, vol. 239, nos. 20–22, pp. 1958–1967, Oct. 2010.
- [14] B. Schölkopf, A. Smola, and K.-R. Müller, "Nonlinear component analysis as a kernel eigenvalue problem," *Neural Comput.*, vol. 10, no. 5, pp. 1299–1319, Jul. 1998.
- [15] G. Strang and T. Nguyen, *Wavelets and Filter Banks*, vol. 24. Cambridge, U.K.: Cambridge Univ. Press, 1996, p. 27.
- [16] M. Vetterli, J. Kovačević, and V. K. Goyal, *Foundations of Signal Processing*. Cambridge, U.K.: Cambridge Univ. Press, 2014.
- [17] V. Dumoulin and F. Visin, "A guide to convolution arithmetic for deep learning," 2016, *arXiv:1603.07285*.
- [18] O. Christensen, *An Introduction to Frames and Riesz Bases*. Springer, 2016.
- [19] G. A. Seber, *Multivariate Observations*. Hoboken, NJ, USA: Wiley, 2009.
- [20] H. A. Dau, A. Bagnall, K. Kamgar, C.-C. M. Yeh, Y. Zhu, S. Gharghabi, C. A. Ratanamahatana, and E. Keogh, "The UCR time series archive," *IEEE/CAA J. Autom. Sinica*, vol. 6, no. 6, pp. 1293–1305, Nov. 2019.
- [21] D. S. G. Pollock, R. C. Green, and T. Nguyen, *A Handbook of Time-Series Analysis, Signal Processing and Dynamics*. Amsterdam, The Netherlands: Elsevier, 1999.



CHRISTOPHER M. A. BONENBERGER received the M.Eng. degree in electrical engineering from the Ravensburg-Weingarten University of Applied Sciences, in 2016. He is currently pursuing the Ph.D. degree in computer science with Ulm University.

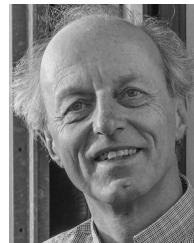
His research interests include machine learning, signal processing, and statistical learning theory with a focus on time series analysis.



FRIEDHELM SCHWENKER (Member, IEEE) received the Diploma and Ph.D. degrees from the University of Osnabrück.

He is currently a Professor of computer science with the Institute of Neural Information Processing, Ulm University. He has (co-)edited over 20 special issues and workshop proceedings published in international journals and publishing companies, and published more than 250 papers at international conferences and journals. His research interests include artificial neural networks, machine learning, statistical learning theory, data mining, pattern recognition, information fusion, and affective computing.

Dr. Schwenker served as the Co-Chair for the IAPR TC3 on Neural Networks and Computational Intelligence. He was the chair from 2016 to 2020. He has founded the IAPR TC9 on Pattern Recognition in human–computer interaction.



WOLFGANG ERTEL received the degree in physics and mathematics from University Konstanz and the Ph.D. degree from Technical University Munich. He is currently a Professor of AI at Ravensburg-Weingarten University, where he was a founder and a leader at the Institute for Artificial Intelligence.

His research interests include machine learning, AI, robot learning, service robotics, and sustainability. He is a member of the Scientists for Future who take efforts to mitigate climate change.



MARKUS SCHNEIDER received the Ph.D. degree from the University of Ulm. He is currently a Professor of computer science and the Head of the Institute for Artificial Intelligence, Ravensburg-Weingarten University of Applied Sciences.

His research interests include machine learning, artificial intelligence, statistical learning theory, and intelligent robotics.

...