

Noise reduction for functional time series

Cees Diks and Bram Wouters

University of Amsterdam

July 3, 2023

Abstract

A novel method for noise reduction in the setting of curve time series with error contamination is proposed, based on extending the framework of functional principal component analysis (FPCA). We employ the underlying, finite-dimensional dynamics of the functional time series to separate the serially dependent dynamical part of the observed curves from the noise. Upon identifying the subspaces of the signal and idiosyncratic components, we construct a projection of the observed curve time series along the noise subspace, resulting in an estimate of the underlying denoised curves. This projection is optimal in the sense that it minimizes the mean integrated squared error. By applying our method to simulated and real data, we show the denoising estimator is consistent and outperforms existing denoising techniques. Furthermore, we show it can be used as a pre-processing step to improve forecasting.

Contents

1	Introduction	3
1.1	Related literature	5
2	Methodology	6
2.1	A review of DFPCA	6
2.2	The structure of the noise space	7
2.3	MISE-optimal denoising	9
2.4	The noise level	11
3	Estimation and consistency	12
3.1	Estimation of \mathcal{M}	12
3.2	Estimation of \mathcal{M}_ε	12
3.3	Estimation of \mathcal{M}_\parallel and \mathcal{M}_\perp	13
3.4	Consistency of denoising procedure	14
4	Simulation	15
4.1	Setup	15
4.2	Selection of tuning parameters	17
4.3	Performance of the denoising procedure	17
4.4	Forecasting	21
5	Denoising empirical weather data	22
6	Conclusions and discussion	27
	References	29
A	Curve time series, FPCA and DFPCA	33
B	Vector autoregressive models	34
C	M(I)SE-optimal denoising	36
D	Estimation formulas for MISE-optimal denoising	42
E	Dependence of the lag order for finding $\Sigma_0^{(\eta)}$	44
F	Irreducible components of the noise curves	46

1 Introduction

Due to an abundance of data in our modern day and age, curve time series, also known as functional time series, are increasingly encountered across various disciplines of society (Bosq, 2000; Hörmann and Kokoszka, 2010; Hörmann and Kokoszka, 2012; Panaretos and Tavakoli, 2013; Aue *et al.*, 2015). Curve time series can either arise as functions observed at consecutive discrete moments in time, such as curves describing the term structure of interest rates (Hays *et al.*, 2012; Caldeira and Torrent, 2017; Andreasen *et al.*, 2019; Sen and Klüppelberg, 2019), return density curves (Bathia *et al.*, 2010) and near-infrared spectroscopy data (Yang *et al.*, 2022), or by splitting an underlying continuous-time process into consecutive equal-length time segments, such as periodic weather record charts (Shang and Hyndman, 2011), intraday energy consumption curves (Cho *et al.*, 2013) and hourly concentration patterns of pollutants (Hörmann *et al.*, 2015). The observed curve time series generally consist of a dynamical part, which is characterised by serial dependence across different curves in the time series, and a white noise part. These parts are latent, i.e. not separately observable, challenging the identification and modelling of the dynamical part of curve (or more generally, high-dimensional) time series, which currently is an active area of research (Qin *et al.*, 2020; Gao and Tsay, 2021; Chen *et al.*, 2022; Cubadda and Hecq, 2022; Dong *et al.*, 2022; Yang *et al.*, 2022; Qin, 2022; Chang *et al.*, 2023).

In this paper we propose a novel pre-processing methodology that filters out the noise from a functional times series, thereby giving access to the dynamics. Following Bathia *et al.* (2010) we consider a univariate curve time series

$$Y_t(u) = X_t(u) + \varepsilon_t(u), \quad (1.1)$$

where $t = 1, \dots, n$ labels the time steps and $u \in \mathcal{I}$, with \mathcal{I} being a bounded interval on which the curves are defined. The nature of this interval (e.g., temporal, spatial) depends on the context in which the time series arises. Only the curves $Y_t(\cdot) \in L^2(\mathcal{I})$ can be observed and are thought of as consisting of a sum of an unobservable signal curve $X_t(\cdot)$ and an unobservable noise curve $\varepsilon_t(\cdot)$ according to (1.1). By definition, the signal curve consists of the part of $Y_t(\cdot)$ that exhibits serial correlation and is in that sense dynamical. The remaining part of $Y_t(\cdot)$, which by assumption has zero autocorrelation, is associated with the noise curve. This so-called error contamination accounts for several potential sources of noise. For processes containing a component without serial dependence, this non-dynamical part will be attributed to the noise curves. If the signal is observed imperfectly, idiosyncratic measurement errors are also part of the noise curves. One can think of, for instance, experimental errors or the use of discrete grids giving rise to numerical round-off errors (Bathia *et al.*, 2010). Another example is when an observed signal requires estimation, thereby introducing estimation error.

A common approach to the analysis of curve time series is by functional principal component analysis (FPCA) (Dauxois *et al.*, 1982; Jones and Rice, 1992), where the signal curves are decomposed in terms of $\mathcal{M} = \text{span}\{\phi_1(u), \dots, \phi_d(u)\}$, a low-dimensional space of curves which we will call the dynamical space or signal space. This decomposition is

an estimate of \mathcal{M} , we can observe the perpendicular noise $\varepsilon_t^\perp(\cdot)$ component via orthogonal projection of the observed curves onto the dynamical space, and use the knowledge of the noise covariance $\Sigma_\varepsilon(\cdot, \cdot)$ to fit a linear regression model for the conditional expectation $\mathbb{E}[\varepsilon_t^\parallel(\cdot) | \varepsilon_t^\perp(\cdot)]$. This in turn is used to reconstruct the unobservable signal curves $X_t(\cdot)$ in a way that minimizes the mean integrated squared error (MISE) between the original signal curve and its reconstruction. We call this method, which is graphically illustrated in Figure 1, MISE-optimal denoising.

In a simulation study we show that MISE-optimal denoising is consistent, in the sense that the obtained MISE converges to its theoretical minimum as the time series length increases. Furthermore, we show that using a MISE-optimal denoised signal for forecasting outperforms other forecasting approaches. We also successfully apply MISE-optimal denoising to an empirical dataset of intraday temperature curves.

We emphasize that the proposed noise reduction method should be viewed as a pre-processing step, using only mild assumptions about the underlying dynamical model. After removing (part of) the noise from the observed curve time series, the remaining denoised signal curves can be used in, e.g., existing modelling and forecasting techniques. The idea is that the noise reduction step helps improve the estimation of the dynamics and as a result also improves forecasting performance.

This paper is organized as follows. After a discussion of the related literature, Section 2 builds on DFPCA Bathia *et al.* (2010) to identify the structure of the noise space \mathcal{M}_ε and to derive an estimator for MISE-optimal denoising. Section 3 discusses a number of practical issues related to estimation in the context of MISE-optimal denoising. In Section 4 MISE-optimal denoising is applied to simulated data and its performance is scrutinized in a variety of ways. An application to empirical data is given in Section 5.

1.1 Related literature

Dimension reduction in functional data analysis has a long history, documented in a vast body of literature; see, e.g., Ramsay and Silverman (2005) and references therein. In the presence of serial dependence we talk about curve or functional time series (Bosq, 2000; Hörmann and Kokoszka, 2010; Hörmann and Kokoszka, 2012; Panaretos and Tavakoli, 2013; Li *et al.*, 2020). For early literature about dimension reduction in this context, see (Bosq, 2000; Ferraty and Vieu, 2006). A major advancement in this field came with Bathia *et al.* (2010), which identifies the “dimensionality of a curve time series”, meaning the dimension and basis vectors of (what we call) the signal space. This is also the main inspiration of this paper. Ideas similar to Bathia *et al.* (2010) have also been applied to (latent) factor models for high-dimensional time series (Lam *et al.*, 2011; Lam and Yao, 2012), which are closely related to the curve time series setting we study. Cubadda and Hecq (2022) studied conditions for the existence of such a signal-noise decomposition in the context of high-dimensional time series. Other attempts at separating signal from noise have as disadvantage that they make strong additional assumptions about the structure of the noise (Yao *et al.*, 2005; Hall and Vial, 2006; Descary and Panaretos, 2019).

A related form of denoising was considered by Dong *et al.* (2022), who project a finite-dimensional time series onto a lower-dimensional time series, where the projection is determined by minimizing a forecast error. Chen *et al.* (2022) developed a method for functional linear regression, rather than noise reduction in a time series setting, using similar elements from Bathia *et al.* (2010) as we do. In order to fit a (scalar) response R_t with respect to a functional predictor $X_t(\cdot)$, corresponding to the signal curves in our setup, a slope function $\beta(\cdot)$ must be estimated such that

$$R_t = \int_{\mathcal{I}} X_t(u)\beta(u) \, du + \varepsilon'_t, \quad (1.3)$$

where ε'_t is some idiosyncrasy term. The main similarity lies in their usage of the fact that the autocovariance with nonzero lag of the signal curves $X_t(\cdot)$ is equal to the autocovariance of the observed curves $Y_t(\cdot)$. They use this to define a generalized method-of-moments estimator for the slope function of functional linear regression. Chang *et al.* (2023) put this approach in a broader framework and generalised it to multivariate functional time series.

2 Methodology

In this section we present our main theoretical results: a method to identify the structure of the noise space and a MISE-optimal denoising algorithm. We start with a brief review of DFPCA, as this is the starting point of our main contributions. To improve readability, technical details are often deferred to appendices.

2.1 A review of DFPCA

A naive approach to finding estimates of the dimension d and a basis of the dynamical space \mathcal{M} would consist of performing an eigensystem analysis of

$$\Sigma_X(u, v) := \text{Cov}[X_t(u), X_t(v)] = \sum_{i=1}^d \lambda_i \phi_i(u) \phi_i(v), \quad (2.1)$$

where (1.2) was used, as well as the following properties of the principal components ξ_{ti} as proven by the Karhunen-Loève theorem:

$$\mathbb{E}[\xi_{ti}] = 0, \quad \text{Var}[\xi_{ti}] = \lambda_i, \quad \text{and} \quad \mathbb{E}[\xi_{ti}\xi_{tj}] = 0 \quad \text{if } i \neq j, \quad (2.2)$$

which hold for all $t = 1, 2, \dots, n$. Note that by Mercer's theorem all eigenvalues in (2.1) are non-negative and for convenience we assume $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_d > 0$. The problem with this approach is that $\Sigma_X(\cdot, \cdot)$ cannot be estimated without bias, since the curves $X_t(\cdot)$ cannot be observed directly and $\Sigma_Y(u, v) := \text{Cov}[Y_t(u), Y_t(v)] = \Sigma_X(u, v) + \Sigma_\varepsilon(u, v)$. Using the observable curves $Y_t(\cdot)$ to estimate $\Sigma_X(\cdot, \cdot)$ leads to a bias, in particular for relatively

large noise. Hall and Vial (2006) circumvented this problem by assuming independence of the curves $Y_1(\cdot), \dots, Y_n(\cdot)$, as well as vanishing noise curves $\varepsilon_t(\cdot)$ in the limit of infinite sample size n . It should be noted that this method does not use the serial dependence of the curve time series.

Bathia *et al.* (2010) proposed an innovative method for estimating d and a basis for \mathcal{M} , based on the fact that the noise curves (by definition) do not exhibit any serial dependence. In other words, the autocovariance of the noise for nonzero lag is zero, $\text{Cov}[\varepsilon_t(u), \varepsilon_{t+k}(v)] = 0$ for any $k \neq 0$. This implies that the lag- k autocovariance of the signal curves, $M_k(u, v) := \text{Cov}[X_t(u), X_{t+k}(v)]$, equals the lag- k autocovariance of the observed curves:

$$M_k(u, v) = \text{Cov}[Y_t(u), Y_{t+k}(v)], \quad (2.3)$$

for any $k \neq 0$. This is crucial, as now the observed curves $Y_t(\cdot)$ can be used to estimate the autocovariance of the signal curves $X_t(\cdot)$ for nonzero lag. Bathia *et al.* (2010) then exploit the fact that, under certain assumptions (see Appendix A for details), the d -dimensional eigenspace of any operator

$$K(u, v) := \sum_{\ell=1}^q c_\ell N_\ell(u, v), \quad \text{with} \quad N_k(u, v) := \int_{\mathcal{I}} M_k(u, z) M_k(v, z) dz, \quad (2.4)$$

coincides with the dynamical space \mathcal{M} . The coefficients c_ℓ can be chosen (almost) arbitrarily, provided at least one of them is nonzero. The various operators $K(\cdot, \cdot)$ that are obtained by different choices of the coefficients c_ℓ , give different sets of eigenfunctions, but each set of eigenfunctions spans the same dynamical space \mathcal{M} . Bathia *et al.* (2010) choose $q = 5$ and $c_\ell = 1$ for $\ell = 1, \dots, 5$. We refer to their approach to finding d and a basis for \mathcal{M} as DFPCA, as it is using the dynamical properties encoded in the nonzero-lag autocovariances $M_{k \neq 0}(\cdot, \cdot)$. For more details about DFPCA, see Appendix A

2.2 The structure of the noise space

This section focuses primarily on the noise covariance $\Sigma_\varepsilon(\cdot, \cdot)$. Knowledge of this can provide insights in the structure of the noise via the FPCA framework. To be more concrete, it might enable us to identify a relatively small number of functions that account for most of the noise present in the curve time series. Because the noise curves $\varepsilon_t(\cdot)$ are not observable, it is not possible to directly estimate $\Sigma_\varepsilon(\cdot, \cdot)$. Since $\Sigma_\varepsilon(u, v) = \Sigma_Y(u, v) - \Sigma_X(u, v)$ and $\Sigma_Y(\cdot, \cdot)$ can be estimated directly, the question is whether we can get access to the covariance of the signal curves $\Sigma_X(\cdot, \cdot)$.

DFPCA provides an alternative basis $\psi_1(\cdot), \dots, \psi_d(\cdot)$ of the signal curve space \mathcal{M} that, unlike the basis of (1.2), can be estimated consistently. With respect to this alternative basis the demeaned signal curves can be expressed as,

$$X_t(u) - \mu(u) = \sum_{i=1}^d \xi_{ti} \phi_i(u) = \sum_{i=1}^d \eta_{ti} \psi_i(u), \quad \text{where} \quad \eta_{ti} = \int_{\mathcal{I}} [X_t(u) - \mu(u)] \psi_i(u) du, \quad (2.5)$$

implying that

$$M_k(u, v) = \boldsymbol{\phi}(u)^T \Sigma_k \boldsymbol{\phi}(v) = \boldsymbol{\psi}(u)^T \Sigma_k^{(\eta)} \boldsymbol{\psi}(v), \quad (2.6)$$

for any $k \in \mathbb{Z}$, where $\Sigma_k := \mathbb{E}[\boldsymbol{\xi}_t \boldsymbol{\xi}_{t+k}^T]$ and $\Sigma_k^{(\eta)} := \mathbb{E}[\boldsymbol{\eta}_t \boldsymbol{\eta}_{t+k}^T]$ are the finite-dimensional variance-covariance matrices of the principal components of the two respective bases of \mathcal{M} . Note that boldface greek letters denote vectors, e.g. $\boldsymbol{\phi}(u) := (\phi_1(u), \dots, \phi_d(u))$ and $\boldsymbol{\xi}_t := (\xi_{t1}, \dots, \xi_{td})$. This in particular means that $\Sigma_X(u, v) = M_0(u, v) = \boldsymbol{\psi}(u)^T \Sigma_0^{(\eta)} \boldsymbol{\psi}(v)$. Access to the matrix $\Sigma_0^{(\eta)}$ is what separates us from being able to estimate the covariance of the signal curves $\Sigma_X(\cdot, \cdot)$.

The sample versions of the principal components η_{ti} defined in (2.5) cannot be computed directly from the data, as the signal curves are unobservable. However, they have proxies

$$\chi_{ti} := \int_{\mathcal{I}} [Y_t(u) - \mu(u)] \psi_i(u) du = \eta_{ti} + \varepsilon'_{ti}, \quad \text{where } \varepsilon'_{ti} := \int_{\mathcal{I}} \varepsilon_t(u) \psi_i(u) du, \quad (2.7)$$

that can be computed given the observable curves $Y_t(\cdot)$. Conveniently, using the fact that the noise curves (by definition) do not have serial dependence, their covariance matrices are related via

$$\Sigma_k^{(\chi)} := \text{Cov}[\boldsymbol{\chi}_t, \boldsymbol{\chi}_{t+k}] = \begin{cases} \Sigma_k^{(\eta)} + \text{Cov}[\boldsymbol{\varepsilon}'_t, \boldsymbol{\varepsilon}'_t] & \text{if } k = 0, \\ \Sigma_k^{(\eta)} & \text{if } k \neq 0. \end{cases} \quad (2.8)$$

This means that the covariance matrices $\Sigma_k^{(\eta)}$ for any $k \neq 0$ can be estimated through the proxy principal components χ_{ti} .

To gain access to the lag-0 covariance matrix $\Sigma_0^{(\eta)}$, we wish to exploit the underlying dynamics. A simple assumption regarding the dynamics, is that the original principal components $\boldsymbol{\xi}_t$, defined in (1.2), follow the dynamics of a lag- p vector autoregressive process. We focus on the simplest case (lag-1) first, that is, the assumption that the time series of $\boldsymbol{\xi}_t$ is described by a VAR(1) model

$$\boldsymbol{\xi}_t = A \boldsymbol{\xi}_{t-1} + \mathbf{e}_t, \quad \text{where } \mathbf{e}_t \sim \text{IID}(\mathbf{0}, \Omega). \quad (2.9)$$

Note the absence of a constant term for the mean, since $\mathbb{E}[\boldsymbol{\xi}_{ti}] = 0$. As a consequence of this assumption the random coefficients $\boldsymbol{\eta}_t$ also follow a VAR(1) process. The lagged autocovariance matrices of a VAR(p)-process time series are related through the so-called Yule-Walker equations (see Appendix B for details). For the VAR(1)-process of the principal components $\boldsymbol{\eta}_t$ these Yule-Walker equations can be rearranged as

$$\Sigma_0^{(\eta)} = \Sigma_1^{(\eta)} \left(\Sigma_2^{(\eta)} \right)^{-1} \Sigma_1^{(\eta)}. \quad (2.10)$$

This is a crucial result, because this makes $\Sigma_0^{(\eta)}$ and thereby $\Sigma_X(\cdot, \cdot)$ accessible through the proxy principal components χ_{ti} defined in (2.7).

For VAR processes with lag order $p > 1$ it is possible to generalize (2.10) and express $\Sigma_0^{(\eta)}$ in terms of $\Sigma_p^{(\eta)}, \Sigma_{p+1}^{(\eta)}, \dots, \Sigma_{p+p}^{(\eta)}$. We have run the MISE-optimal denoising algorithm

on the simulated data of section 4 for $p = 2$ and $p = 3$. The results in terms of denoising performance and asymptotic behaviour were similar to the default case of $p = 1$ and (2.10). This indicates that our approach is insensitive to the assumption of an underlying VAR(1) process for the principal components $\boldsymbol{\xi}_t$, even when this model is misspecified (Dahlhaus and Wefelmeyer, 1996). Throughout this paper we assume a VAR(1) process by default, as this choice is expected to suffer the least from small-sample estimation noise. For more details, see Appendix E.

The above analysis is a key step for our approach, as this enables us to express the noise covariance operator $\Sigma_\varepsilon(u, v) = \Sigma_Y(u, v) - \Sigma_X(u, v)$, where $\Sigma_X(u, v) = \boldsymbol{\psi}(u)^T \Sigma_0^{(\eta)} \boldsymbol{\psi}(v)$, in terms of quantities that can be estimated. Since $\Sigma_Y(\cdot, \cdot)$, $\boldsymbol{\psi}(\cdot)$ and $\Sigma_0^{(\eta)}$ can be estimated consistently, without bias from the noise curves $\varepsilon_t(\cdot)$, this provides a consistent estimator of the noise covariance operator $\Sigma_\varepsilon(\cdot, \cdot)$. This allows us to extend the FPCA framework to the noise space \mathcal{M}_ε . Assuming that the eigenspace of $\Sigma_\varepsilon(\cdot, \cdot)$ is of finite dimension, analogous to $X_t(\cdot)$ also the noise curves $\varepsilon_t(\cdot)$ can (approximately) be expanded in terms of a finite number of eigenfunctions, in this case of the noise covariance $\Sigma_\varepsilon(\cdot, \cdot)$. The orthonormal eigenbasis $\phi_1^{(\varepsilon)}(\cdot), \dots, \phi_{d_\varepsilon}^{(\varepsilon)}(\cdot)$ of $\Sigma_\varepsilon(\cdot, \cdot)$ forms a basis of the noise space \mathcal{M}_ε such that

$$\varepsilon_t(u) = \sum_{i=1}^{d_\varepsilon} \xi_{ti}^{(\varepsilon)} \phi_i^{(\varepsilon)}(u), \quad (2.11)$$

where the random variables $\xi_{ti}^{(\varepsilon)}$ have the same properties as in (2.2) due to the Karhunen-Loève theorem.

2.3 MISE-optimal denoising

Given an observed curve $Y_t(u) = X_t(u) + \varepsilon_t(u)$, the aim of denoising is to find a reconstruction $\widehat{X}_t^{\text{den}}(\cdot)$ of the unobservable signal curve $X_t(\cdot)$. An example of a denoising procedure is orthogonal denoising, where the observed curve $Y_t(\cdot)$ is projected orthogonally onto the dynamical space \mathcal{M} . In other words, for orthogonal denoising the reconstructed signal curve is given by $\widehat{X}_t^{\text{ortho}}(u) = Y_t^\parallel(u) := (P_{\mathcal{M}} Y_t)(u)$, where $P_{\mathcal{M}}(u, v) = \boldsymbol{\psi}(u) \cdot \boldsymbol{\psi}(v)$ is the operator of the orthogonal projection onto \mathcal{M} . Denoising performance can be measured in terms of the “mean integrated square error” (MISE)

$$\mathbf{MISE}_{\text{den}} := \overline{\mathbb{E}}[(X_t(\cdot) - \widehat{X}_t^{\text{den}}(\cdot))^2], \quad \text{where } \overline{\mathbb{E}}[g(\cdot)] := \frac{1}{\Delta_{\mathcal{I}}} \int_{\mathcal{I}} \mathbb{E}[g(u)] \, du, \quad (2.12)$$

in the case of a bounded interval of width $\Delta_{\mathcal{I}}$. As we will see, $\mathbf{MISE}_{\text{ortho}}$ associated with $\widehat{X}_t^{\text{ortho}}(\cdot)$ is generally not optimal in the sense that it does not minimize the MISE.

Let us now ask the question of finding a MISE-optimal denoising procedure. Concretely, the goal is to find an operator $P : L^2(\mathcal{I}) \rightarrow L^2(\mathcal{I})$ such that the denoised curve $X_t^{\text{opt}}(u) := (P Y_t)(u)$ minimizes the associated $\mathbf{MISE}_{\text{opt}}$ defined in (2.12). Note the slight abuse of notation here, as (2.12) already contains an estimator and here we are searching for the operator that minimizes the MISE in (2.12), which we will then call the MISE-optimal

denoising estimator. In this paper the space of operators P is restricted to (linear) projections onto the dynamical space, i.e. $(Pf)(u) = f(u)$ for any function $f(\cdot) \in \mathcal{M}$. Combined with the fact that a noise curve can be decomposed in a unique way in a part parallel to the dynamical space and a part orthogonal to the dynamical space, $\varepsilon_t(u) = \varepsilon_t^\parallel(u) + \varepsilon_t^\perp(u)$ where $\varepsilon_t^\parallel(\cdot) \in \mathcal{M}$ and $\varepsilon_t^\perp(\cdot)$ lies inside the orthogonal complement of \mathcal{M} , this implies that $X_t^{\text{opt}}(u) = Y_t^\parallel(u) + (P\varepsilon_t^\perp)(u)$, where $Y_t^\parallel(u) = X_t(u) + \varepsilon_t^\parallel(u)$.

The challenge is now to find a projection such that the parallel part of the noise curve in $Y_t^\parallel(\cdot)$ is cancelled by $(P\varepsilon_t^\perp)(\cdot)$ in a MISE-optimal fashion. For this purpose, the latter is expanded as

$$(P\varepsilon_t^\perp)(u) = \sum_{i=1}^{d_\parallel} \sum_{j=1}^{d_\perp} \alpha_{ij} \varepsilon_{t,j}^\perp \phi_i^\parallel(u), \quad (2.13)$$

in terms of an orthonormal eigenbasis $\phi_1^\parallel(\cdot), \phi_2^\parallel(\cdot), \dots, \phi_{d_\parallel}^\parallel(\cdot)$ of \mathcal{M}_\parallel , which is the space of noise curves $\varepsilon_t^\parallel(\cdot)$ parallel to \mathcal{M} . Here, the $\varepsilon_{t,j}^\perp$ are the coordinates of the perpendicular part of the noise curves with respect to an orthonormal basis $\phi_1^\perp(\cdot), \phi_2^\perp(\cdot), \dots, \phi_{d_\perp}^\perp(\cdot)$ of the space \mathcal{M}_\perp of these perpendicular noise curves,

$$\varepsilon_t^\perp(u) = \sum_{j=1}^{d_\perp} \varepsilon_{t,j}^\perp \phi_j^\perp(u). \quad (2.14)$$

The finite dimensionality of the subspaces \mathcal{M}_\parallel and \mathcal{M}_\perp is ensured by applying the FPCA assumption to the noise space \mathcal{M}_ε , as formulated in (2.11).

The idea behind (2.13) is that the perpendicular noise curves can be indirectly observed via $\varepsilon_t^\perp(u) = Y_t(u) - Y_t^\parallel(u)$. They can therefore be used as a predictor, with the parallel noise curves $\varepsilon_t^\parallel(\cdot)$ as response variable. From this perspective, the parameters α_{ij} are the regression coefficients and they are determined by minimizing the MISE in (2.12). We are essentially fitting a linear model for the conditional expectation $\mathbb{E}[\varepsilon_t^\parallel(\cdot) | \varepsilon_t^\perp(\cdot)]$ that minimizes the MISE by exploiting the covariance between the perpendicular and parallel part of the noise curves. See Figure 1 for a simplified illustration of this approach in the case of a bivariate time series. Note the absence of an intercept in (2.13), because the noise curves have zero mean.

The result is a convex optimization problem with solution $\hat{\alpha} = -\Omega_{\parallel\perp} (\Omega_\perp)^{-1}$, where

$$\Omega_\perp := \int_{\mathcal{I}} \int_{\mathcal{I}} \boldsymbol{\phi}^\perp(u) (\boldsymbol{\phi}^\perp(v))^T \Sigma_\varepsilon(u, v) \, du \, dv, \quad (2.15a)$$

$$\Omega_{\parallel\perp} := \int_{\mathcal{I}} \int_{\mathcal{I}} \boldsymbol{\phi}^\parallel(u) (\boldsymbol{\phi}^\perp(v))^T \Sigma_\varepsilon(u, v) \, du \, dv, \quad (2.15b)$$

and where $\boldsymbol{\phi}^\parallel(u) := (\phi_1^\parallel(u), \phi_2^\parallel(u), \dots, \phi_{d_\parallel}^\parallel(u))^T$ and $\boldsymbol{\phi}^\perp(u) := (\phi_1^\perp(u), \phi_2^\perp(u), \dots, \phi_{d_\perp}^\perp(u))^T$. The matrix Ω_\perp should be interpreted as the covariance of between the coordinates $\boldsymbol{\varepsilon}_t^\perp :=$

$(\varepsilon_{t,1}^\perp, \varepsilon_{t,2}^\perp, \dots, \varepsilon_{t,d_\perp}^\perp)^T$, whereas the matrix $\Omega_{\parallel\perp}$ is the covariance between the similarly defined coordinates $\boldsymbol{\varepsilon}_t^\parallel := (\varepsilon_{t,1}^\parallel, \varepsilon_{t,2}^\parallel, \dots, \varepsilon_{t,d_\parallel}^\parallel)^T$ and $\boldsymbol{\varepsilon}_t^\perp$. This leads to a MISE-optimal reconstruction of the signal curves $X_t(\cdot)$ given by

$$\widehat{X}_t^{\text{opt}}(u) = Y_t^\parallel(u) - \int_{\mathcal{I}} (\boldsymbol{\phi}^\parallel(u))^T \Omega_{\parallel\perp} (\Omega_\perp)^{-1} \boldsymbol{\phi}^\perp(v) (Y_t(v) - Y_t^\parallel(v)) \, dv. \quad (2.16)$$

The above formula for the MISE-optimal denoising estimator of the signal curves is the main result of this paper. More details about the derivation of (2.16) can be found in Appendix C, as well as an insightful illustration of the denoising formula in the context of finite-dimensional vector spaces.

Using this denoising approach, the minimum of the MISE is

$$\text{MISE}_{\text{opt}}^{\min} := \min_{\alpha \in \mathbb{R}^{(d_\parallel \times d_\perp)}} \text{MISE}_{\text{opt}} = \text{Tr} [\Omega_\parallel] - \text{Tr} [(\Omega_\perp)^{-1} (\Omega_{\parallel\perp})^T \Omega_{\parallel\perp}]. \quad (2.17)$$

It accounts for the irreducible components of the noise curves, which are present in the general case where the noise space and the dynamical space have overlap. In the special case of no overlap, i.e. $\mathcal{M}_\varepsilon \cap \mathcal{M} = \emptyset$, the minimum of the MISE is zero and perfect denoising is possible (at the population level). For more details, see Appendix F.

2.4 The noise level

An important property of the curve time series (1.1) is the noise level $\lambda \in [0, 1]$, which we define as the relative size of the noise with respect to the observed curve variance,

$$\lambda := \frac{\overline{\text{Var}}[\varepsilon_t(\cdot)]}{\overline{\text{Var}}[Y_t(\cdot)]}, \quad (2.18)$$

where $\overline{\text{Var}}[g(\cdot)]$ is defined analogous to $\overline{\mathbb{E}}[g(\cdot)]$ in (2.12). Note that λ is directly related to the (integrated) signal-to-noise ratio $\overline{\text{Var}}[X_t(\cdot)] / \overline{\text{Var}}[\varepsilon_t(\cdot)]$.

A consequence of the proposed MISE-optimal denoising procedure is the ability to estimate the noise level. Since the noise cannot be observed directly, estimating its relative size is a nontrivial problem. Using the MISE-optimal denoised signal $\widehat{X}_t^{\text{opt}}(\cdot)$, one can show that

$$\lambda = \frac{\text{MISE}_{\text{opt}}^{\min} + \overline{\mathbb{E}}[(Y_t(\cdot) - \widehat{X}_t^{\text{opt}}(\cdot))^2]}{\overline{\text{Var}}[Y_t(\cdot)]}, \quad (2.19)$$

consisting of quantities that all can be estimated. Note that, since the analysis of section 2.2 provides access to $\Sigma_\varepsilon(\cdot, \cdot)$, an alternative way of obtaining the noise level is through $\lambda = \text{Tr}[\Sigma_\varepsilon] / \overline{\text{Var}}[Y_t(\cdot)]$.

3 Estimation and consistency

In this section we describe how MISE-optimal denoising can be applied in the default realistic scenario, in which an observed curve time series $\{Y_t(u)\}_{t=1}^n$ of length n is available. In particular, we describe how orthonormal bases of the subspaces \mathcal{M} , \mathcal{M}_ε , \mathcal{M}_\parallel and \mathcal{M}_\perp can be estimated and how they can be used to define an estimated operator corresponding to MISE-optimal denoising. Furthermore, we define what we mean by consistency of a denoising method. This section focuses on the aspects of estimation that are specific for our denoising method, whereas more standard formulas for estimators can be found in Appendix D.

3.1 Estimation of \mathcal{M}

For the estimation of the dimension d of the dynamical space and the basis functions $\psi_1(\cdot), \dots, \psi_d(\cdot)$ we follow the approach of Bathia *et al.* (2010). The basis functions are the eigenfunctions of the estimate of the operator $K(\cdot, \cdot)$ defined in (2.4). No smoothing methods are used to improve estimation precision. This choice has the advantage that MISE-optimal denoising as presented here does not make any assumptions about the smoothness of the curves of the functional time series and can therefore also be applied to vector-valued high-dimensional time series, for which this smoothness is generally absent.

The dimension d is estimated through a series of bootstrap tests for the eigenvalues $\lambda_i^{(K)}$ of the operator $K(\cdot, \cdot)$. Alternative approaches, which will not be pursued here, include identifying a significant drop in the eigenvalues, minimizing forecast errors (Hyndman and Ullah, 2007) and information criteria approaches (Bai and Ng, 2002; Bai and Ng, 2007; Hallin and Liška, 2007). Each test works with a null hypothesis $H_0 : \lambda_{d_0+1}^{(K)} = 0$ for a different value of d_0 (see Appendix D for details). We start with a d_0 that is too large, for example corresponding to an estimated eigenvalue $\hat{\lambda}_{d_0+1}^{(K)}$ that is extremely small. Then we test and every time H_0 does not get rejected we lower d_0 by one, until H_0 gets rejected. Our multiple testing procedure makes it more likely to overestimate the dimension of \mathcal{M} than to underestimate it. This is preferable, because an overestimated dimension of \mathcal{M} generally leads to a smaller denoising error than an underestimated dimension. The intuition behind this is as follows. When you project onto an erroneous direction of \mathcal{M} the harm is relatively small, because the to-be-projected curves $Y_t(\cdot)$ do not have a component in that direction (apart potentially from a contribution of the noise curves). On the other hand, when \mathcal{M} is mistakenly missing a direction, then an actually existing component of $Y_t(\cdot)$ is lost during projection and this generically causes a larger denoising error.

3.2 Estimation of \mathcal{M}_ε

Estimation of the noise space \mathcal{M}_ε starts straightforward by estimating the noise covariance through $\Sigma_\varepsilon(u, v) = \Sigma_Y(u, v) - \Sigma_X(u, v)$. Details can be found in Appendix D.

A bootstrap test to select \hat{d}_ε is not available, unlike for selecting \hat{d} . Instead, we use the

FPCA-interpretation of the eigenvalues $\lambda_j^{(\varepsilon)}$ of $\Sigma_\varepsilon(\cdot, \cdot)$ as the variance of the j -th principal component. If the explained variance of the $(\hat{d}_\varepsilon + 1)$ -th eigenvalue falls below a certain threshold $\tau_{d_\varepsilon} \in (0, 1)$, we take \hat{d}_ε as the estimate for d_ε . To be more precise, the condition that determines the estimator \hat{d}_ε is

$$\frac{\hat{\lambda}_{\hat{d}_\varepsilon+1}^{(\varepsilon)}}{\sum_{j=1}^{\hat{d}_\varepsilon+1} \hat{\lambda}_j^{(\varepsilon)}} < \tau_{d_\varepsilon} \leq \frac{\hat{\lambda}_{\hat{d}_\varepsilon}^{(\varepsilon)}}{\sum_{j=1}^{\hat{d}_\varepsilon} \hat{\lambda}_j^{(\varepsilon)}}, \quad (3.1)$$

where the sum in the denominators is taken over all (positive) eigenvalues. Note that the estimated eigenvalues are assumed to be sorted in descending order. The threshold τ_{d_ε} is set by the practitioner.

A second subtlety is that in what follows $\widehat{\Sigma}_\varepsilon(u, v) = \widehat{\Sigma}_Y(u, v) - \widehat{M}_0(u, v)$ is replaced by

$$\widehat{\Sigma}_\varepsilon^+(u, v) := \sum_{j=1}^{\hat{d}_\varepsilon} \hat{\lambda}_j^{(\varepsilon)} \hat{\phi}_j^{(\varepsilon)}(u) \hat{\phi}_j^{(\varepsilon)}(v), \quad (3.2)$$

where $\hat{\phi}_j^{(\varepsilon)}(\cdot)$ is the eigenfunction of $\widehat{\Sigma}_\varepsilon(\cdot, \cdot)$ associated with eigenvalue $\hat{\lambda}_j^{(\varepsilon)}$. The reason is that due to estimation noise $\widehat{\Sigma}_\varepsilon(\cdot, \cdot)$ is not semi-positive definite. It is a known phenomenon (see Chen *et al.* (2021) and references therein) that even for small values of the estimation noise, the negative eigenvalues can remain relatively large. By removing these negative eigenvalues through (3.2) by hand, we reduce the estimation error and improve the denoising performance.

3.3 Estimation of \mathcal{M}_\parallel and \mathcal{M}_\perp

In order to estimate an orthonormal basis of \mathcal{M}_\parallel , we use the fact that we have access to the covariance of the noise curves $\varepsilon_t^\parallel(\cdot)$ parallel to \mathcal{M} via

$$\widehat{\Omega}_\parallel^{(\psi)} = \int_{\mathcal{I}} \int_{\mathcal{I}} \widehat{\psi}(u) \left(\widehat{\psi}(v) \right)^T \widehat{\Sigma}_\varepsilon^+(u, v) du dv. \quad (3.3)$$

This $(\hat{d} \times \hat{d})$ matrix is an estimate of the covariance matrix of the principal components of the curves $\varepsilon_t^\parallel(\cdot)$ with respect to the basis functions $\psi(\cdot)$ of \mathcal{M} . Diagonalization leads to \hat{d} orthonormal eigenvectors $\widehat{\Omega}_\parallel^{(\psi)} \hat{\mathbf{e}}_j^\parallel = \hat{\lambda}_j^\parallel \hat{\mathbf{e}}_j^\parallel$, with $\hat{\lambda}_1^\parallel \geq \hat{\lambda}_2^\parallel \geq \dots \geq \hat{\lambda}_{\hat{d}}^\parallel \geq 0$, from which an estimate for an orthonormal basis of \mathcal{M}_\parallel can be constructed as

$$\widehat{\mathcal{M}}_\parallel = \text{span}(\hat{\phi}_1^\parallel(\cdot), \dots, \hat{\phi}_{\hat{d}_\parallel}^\parallel(\cdot)), \quad \text{where } \hat{\phi}_j^\parallel(u) = \hat{\mathbf{e}}_j^\parallel \cdot \widehat{\psi}(u), \quad (3.4)$$

for $j = 1, 2, \dots, \hat{d}_\parallel$. Here, analogous to d_ε , the dimension of \mathcal{M}_\parallel is selected via the conditions

$$\frac{\hat{\lambda}_{\hat{d}_\parallel+1}^\parallel}{\text{Tr} \left[\widehat{\Sigma}_\varepsilon^+(\cdot, \cdot) \right]} < \tau_{d_\parallel} \leq \frac{\hat{\lambda}_{\hat{d}_\parallel}^\parallel}{\text{Tr} \left[\widehat{\Sigma}_\varepsilon^+(\cdot, \cdot) \right]}, \quad (3.5)$$

where $\tau_{d_{\parallel}} \in (0, 1)$ is a threshold set by the practitioner. Note that the denominator in these conditions represents the total variance of the noise curves. The j -th estimated eigenvalue represents the variance of the noise mode $\hat{\phi}_j^{\parallel}(\cdot)$. It is therefore prudent to compare these eigenvalues to the total variance of the noise. If, for example, (nearly) all noise is in the part perpendicular to \mathcal{M} , you do not wish to take many parallel modes into account for the regression analysis of MISE-optimal denoising. By mutually comparing the eigenvalues of only the parallel part of the noise, there is the risk of overestimating the number of relevant parallel modes and thereby introducing too much estimation uncertainty, leading to a poor performance in terms of denoising.

Estimation of the perpendicular noise space \mathcal{M}_{\perp} occurs as follows. First, the time series of the perpendicular noise curves is found through $\hat{\varepsilon}_t^{\perp}(u) = ((I - \hat{P}_{\mathcal{M}})\hat{\phi}_i^{(\varepsilon)})(u)$, where $\hat{P}_{\mathcal{M}}(u, v) := \hat{\psi}(u) \cdot \hat{\psi}(v)$ and $\hat{\phi}_i^{(\varepsilon)}(\cdot)$ was defined in (3.2). Based on this estimated time series the covariance $\hat{\Sigma}_{\varepsilon^{\perp}}(\cdot, \cdot)$ is computed in the usual way. Let this estimated covariance have orthonormal eigenfunctions $\hat{\phi}_i^{\perp}(\cdot)$ with corresponding eigenvalues $\hat{\lambda}_i^{\perp}$, sorted in descending order. The dimension of the perpendicular subspace \mathcal{M}_{\perp} is then selected via the conditions

$$\frac{\hat{\lambda}_{\hat{d}_{\perp}+1}^{\perp}}{\text{Tr} \left[\hat{\Sigma}_{\varepsilon^{\perp}}^+(\cdot, \cdot) \right]} < \tau_{d_{\perp}} \leq \frac{\hat{\lambda}_{\hat{d}_{\perp}}^{\perp}}{\text{Tr} \left[\hat{\Sigma}_{\varepsilon^{\perp}}^+(\cdot, \cdot) \right]}, \quad (3.6)$$

where $\tau_{d_{\perp}} \in (0, 1)$ is a threshold set by the practitioner, and the functions $\hat{\phi}_1^{\perp}(\cdot), \dots, \hat{\phi}_{\hat{d}_{\perp}}^{\perp}(\cdot)$ form an estimate of the basis of \mathcal{M}_{\perp} .

After having estimated the spaces $\mathcal{M}, \mathcal{M}_{\varepsilon}, \mathcal{M}_{\parallel}$ and \mathcal{M}_{\perp} , estimating the denoised signal curves through MISE-optimal denoising is a straightforward application of (2.15) and (2.16). For details, see Appendix D.

3.4 Consistency of denoising procedure

In the context of curve time series as considered in this paper, the aim of denoising is to reconstruct the signal curves $X_t(\cdot)$ from the observed curves $Y_t(\cdot)$, given the data of a curve time series $\{Y_t(u)\}_{t=1}^n$ of length n . A denoising procedure is consistent if the reconstruction $\hat{X}_t^{\text{den}}(\cdot)$ as estimated from the observed data converges to the true signal curve $X_t(\cdot)$, possibly up to an irreducible observational noise component, as the length of the available time series n increases. To be more precise, we call a denoising procedure consistent if

$$\lim_{n \rightarrow \infty} \mathbf{MISE}_{\text{den}} = \mathbf{MISE}_{\text{den}}^{\min}, \quad (3.7)$$

where $\mathbf{MISE}_{\text{den}}^{\min}$ is defined as the theoretically achievable minimum of the denoising procedure. This definition applies to both MISE-optimal denoising, with a minimum defined in (2.17), and to orthogonal denoising, with a minimum given by $\mathbf{MISE}_{\text{ortho}}^{\min} = \text{Tr} [\Omega_{\parallel}]$.

4 Simulation

This section illustrates MISE-optimal denoising with an application to simulated data. Section 4.3 analyzes its performance and compares it with orthogonal denoising. Section 4.4 uses both MISE-optimal and orthogonal denoising as a pre-processing step in a forecasting problem. We start by discussing the setup of the simulations.

4.1 Setup

The setup of our simulation is inspired by Bathia *et al.* (2010) and Chen *et al.* (2022). The data consists of a curve time series $\{Y_t(u)\}_{t=1}^n$ of n observations. Each curve is defined on the interval $\mathcal{I} = [0, 1]$ and represented on a grid of 200 equidistant points on that interval. They are constructed as the sum of a signal and a noise curve, $Y_t(u) = X_t(u) + \varepsilon_t(u)$, where the signal curves are defined as

$$X_t(u) = g_X(\lambda)\tilde{X}_t(u), \quad \tilde{X}_t(u) = \sum_{j=1}^d \xi_{tj}\phi_j(u). \quad (4.1)$$

Throughout our simulations, we work with three dynamical spaces \mathcal{M} with respective dimensions $d = 2, 4$ and 6 . The principal components ξ_{tj} are simulated according to the VAR(1)-process in (2.9). For $d = 2$ this is specified by¹

$$A = \begin{pmatrix} 0.14275022 & -0.61629756 \\ -0.4615736 & -0.49825869 \end{pmatrix}, \quad \Omega = \begin{pmatrix} 0.60977113 & -0.01529231 \\ -0.01529231 & 0.00121252 \end{pmatrix}. \quad (4.2)$$

The procedure by which we obtained these specific matrices A and Ω for the VAR(1)-process is as follows. We start from the requirement that the lag-0 autocovariance of the principal components is diagonal, as stipulated by the Karhunen-Loève theorem. We choose

$$\Sigma_0 = \mathbb{E}[\boldsymbol{\xi}_t \boldsymbol{\xi}_t^T] = \text{diag}(\lambda_1, \dots, \lambda_d), \quad \text{with} \quad \lambda_i = 0.2 \frac{i-1}{d-1} + 0.7 \frac{d-i}{d-1}, \quad (4.3)$$

for $i = 1, 2, \dots, d$. In the case $d = 2$ this means $\lambda_1 = 0.7$ and $\lambda_2 = 0.2$. The values of A are drawn i.i.d. from the standard normal distribution, after which the whole matrix is rescaled to make the absolute value of the largest eigenvalue equal to 0.8. This ensures stationarity of the VAR(1)-process, which requires that all eigenvalues of A are within the unit circle. Finally, we use the first Yule-Walker equation to determine the covariance of the noise term, $\Omega = \Sigma_0 - A\Sigma_0A^T$. Since Ω constructed in this way is not necessarily semi-positive definite, this is checked explicitly. If it fails to be so, the process of randomly generating entries of the matrix A is repeated until a semi-positive definite covariance matrix is obtained.

An alternative choice for our simulation setup would be to use the above procedure to generate a new VAR(1)-process every time we run our denoising algorithm. This has as an

¹See Appendix B for the specifications of the VAR(1)-processes for $d = 4$ and $d = 6$.

advantage that the results become independent of the particulars of the specific VAR(1)-processes for $d = 2, 4, 6$ that we are using. However, it will also introduce an additional source of variance in our results. Since in most practical situations one is dealing with a single (unknown) underlying VAR-process, this source of variance is rather unrealistic and therefore we do not randomize over different VAR models in our data generating process.

We use $\phi_j(u) = \cos(2\pi ju) + \sin(2\pi ju)$ with $j = 1, 2, \dots, d$ as orthonormal functions that form a basis of the dynamical space \mathcal{M} . To conclude our discussion of the signal curves, we define a normalization pre-factor

$$g_X(\lambda) = \sqrt{\frac{1 - \lambda}{\overline{\text{Var}}[\tilde{X}_t(\cdot)]}}, \quad \text{with} \quad \overline{\text{Var}}[\tilde{X}_t(\cdot)] = \sum_{j=1}^d \text{Var}[\xi_{tj}] = \sum_{j=1}^d \lambda_j. \quad (4.4)$$

Here the parameter $\lambda = 1 - \overline{\text{Var}}[X_t(\cdot)] \in [0, 1]$ plays the role of noise level, as defined in section 2.4.

The other part from which the observed curves $\{Y_t(u)\}_{t=1}^n$ are constructed are the noise curves,

$$\varepsilon_t(u) = g_\varepsilon(\lambda) \tilde{\varepsilon}_t(u), \quad \tilde{\varepsilon}_t(u) = \sum_{j=1}^{d_\varepsilon} \frac{Z_{tj}}{a^{j-1}} \phi_j^{(\varepsilon)}(u), \quad (4.5)$$

where $Z_{tj} \sim \text{N}(0, 1)$ are mutually independent. Throughout this simulation we choose $a = 1.5$ and $d_\varepsilon = 8$. This means that the noise mode with the smallest contribution, $\phi_{d_\varepsilon}^{(\varepsilon)}(\cdot)$, explains about 2% of the total (integrated) variance of the noise curves. As an (orthonormal) basis for the noise space \mathcal{M}_ε we use

$$\phi_j^{(\varepsilon)}(u) = [\cos \theta_j + \sin \theta_j] \cos(2\pi ju) + [\cos \theta_j - \sin \theta_j] \sin(2\pi ju), \quad (4.6)$$

where $j = 1, 2, \dots, d_\varepsilon$. This choice of parametrization allows us to control the orientation of the noise space with respect to the dynamical space. The angles between the basis vectors of \mathcal{M} and \mathcal{M}_ε with respect to the L^2 -norm (A.4) are

$$\angle(\phi_i(u), \phi_j^{(\varepsilon)}(u)) = \arccos\langle \phi_i(u), \phi_j^{(\varepsilon)}(u) \rangle = \delta_{i,j} \theta_j + (1 - \delta_{i,j}) \pi/2 \quad (4.7)$$

where $\delta_{i,j}$ is the Kronecker delta. By default we will take

$$\theta_j = \begin{cases} \pi/4 & j = 1, 2, \dots, \min(d, d_\varepsilon), \\ 0 & j = \min(d, d_\varepsilon) + 1, \dots, d_\varepsilon. \end{cases} \quad (4.8)$$

This means that each basis vector of \mathcal{M} is at an angle of $\pi/4$ with one basis vector of \mathcal{M}_ε , in a pairwise manner. All other pairs of basis vectors are perpendicular.

The normalization pre-factor for the noise curves is

$$g_\varepsilon(\lambda) = \sqrt{\frac{\lambda}{\overline{\text{Var}}[\tilde{\varepsilon}_t(\cdot)]}}, \quad \text{with} \quad \overline{\text{Var}}[\tilde{\varepsilon}_t(\cdot)] = \sum_{j=1}^{d_\varepsilon} \frac{1}{a^{2(j-1)}} = \frac{1 - \left(\frac{1}{a^2}\right)^{d_\varepsilon}}{1 - \frac{1}{a^2}}. \quad (4.9)$$

This parametrization implies $\overline{\text{Var}}[\varepsilon_t(\cdot)] = \lambda$. Since $\overline{\text{Var}}[Y_t(\cdot)] = \overline{\text{Var}}[X_t(\cdot)] + \overline{\text{Var}}[\varepsilon_t(\cdot)] = 1$, which confirms that λ indeed plays the role of noise level $\overline{\text{Var}}[\varepsilon_t(\cdot)] / \overline{\text{Var}}[Y_t(\cdot)]$ as defined in section 2.4. By controlling λ , we control the noise level in the simulated data. Another advantage of this parametrization is that $\overline{\text{Var}}[Y_t(\cdot)]$ is kept at a constant, enabling a more fair comparison of denoising performance for different values of the signal-to-noise ratio.

4.2 Selection of tuning parameters

The definition of the kernel $K(\cdot, \cdot)$ in (2.4) depends on a parameter $q \in \mathbb{N}^+$, which is the largest lag of the autocovariances that determine $K(\cdot, \cdot)$, as well as the coefficients c_1, \dots, c_q . It has been reported in different contexts (Bathia *et al.*, 2010; Lam *et al.*, 2011; Chen *et al.*, 2022) that estimation of the dynamical space \mathcal{M} is rather insensitive to the particular choice of $q > 1$, with the understanding that choosing q too large leads to a larger finite-sample noise in the estimator $\widehat{K}(\cdot, \cdot)$. Throughout this paper, we take $q = 2$ and $c_1 = c_2 = 1$, i.e. $K(u, v) = N_1(u, v) + N_2(u, v)$.

The thresholds introduced in Sections 3.2 and 3.3 for estimating orthonormal bases of \mathcal{M}_ε , \mathcal{M}_\parallel and \mathcal{M}_\perp are set at $\tau_{d_\varepsilon} = \tau_{d_\parallel} = \tau_{d_\perp} = 0.01$. We have experimented with different values of the thresholds and observed that the results of our simulations are always essentially the same. When the noise level λ is increased, the denoising performance seems to improve for larger values of the threshold. This makes sense, because the noise is relatively large and therefore it is easier to estimate $\Sigma_\varepsilon(\cdot, \cdot)$ more reliably. Furthermore, noise modes with relatively small eigenvalues do contribute notably when the noise level is relatively large. The above reasons make larger threshold values for large λ both robust for estimation noise and noticeable in the observed denoising performance. Finally, notice that when one of the thresholds is zero MISE-optimal denoising reduces to orthogonal denoising ($\widehat{X}_t^{\text{opt}}(u) = \widehat{X}_t^{\text{ortho}}(u) = Y_t^\parallel(u)$). We have verified this experimentally.

4.3 Performance of the denoising procedure

We study the denoising performance of MISE-optimal and orthogonal denoising in Figure 2, where the proportion of the (integrated) variance of the remaining noise after denoising, i.e. $\mathbf{MISE}_{\text{den}}/\lambda$, is plotted as a function of sample size n . The noise level is fixed at $\lambda = 0.2$ and $\mathbf{MISE}_{\text{den}}$ is estimated in-sample. For computational convenience the bootstrap estimate for d is replaced by the true value. This is reasonable, as increasing the bootstrap sample size will bring the proportion of erroneous estimates to zero. Every boxplot is based on 100 independent samples. Unless stated otherwise, we will use the same setup as described here.

Figure 2 also shows the performance of an ‘‘oracle’’ version of the MISE-optimal denoising approach, for which both \mathcal{M} and \mathcal{M}_ε are known exactly and do not need to be estimated. This enables us to disentangle how the estimation of these subspaces on the one hand, and the regression step of MISE-optimal denoising on the other hand, contribute to the overall performance of the algorithm.

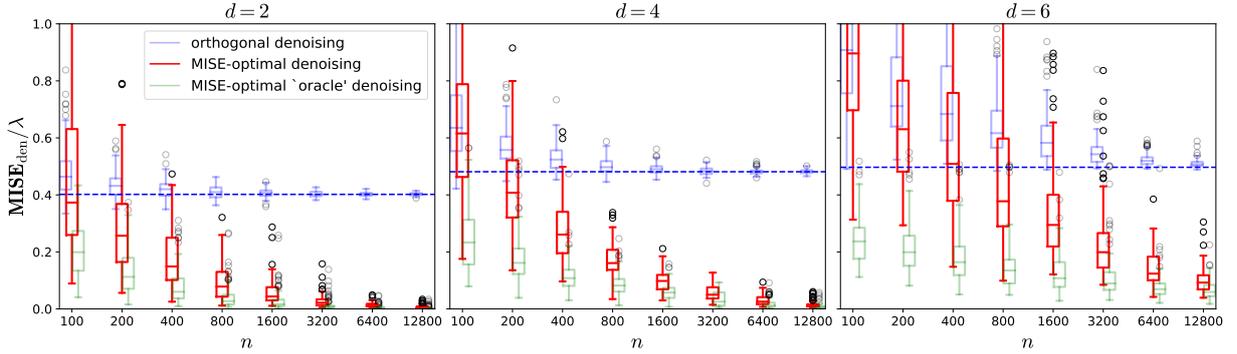


Figure 2: Proportion of the (integrated) variance of the remaining noise after MISE-optimal denoising (*red*) as a function of the sample size n . Same for orthogonal denoising (*blue*) and the “oracle” version (*green*). Also shown is the theoretical lower bound for orthogonal denoising (*blue, dashed line*).

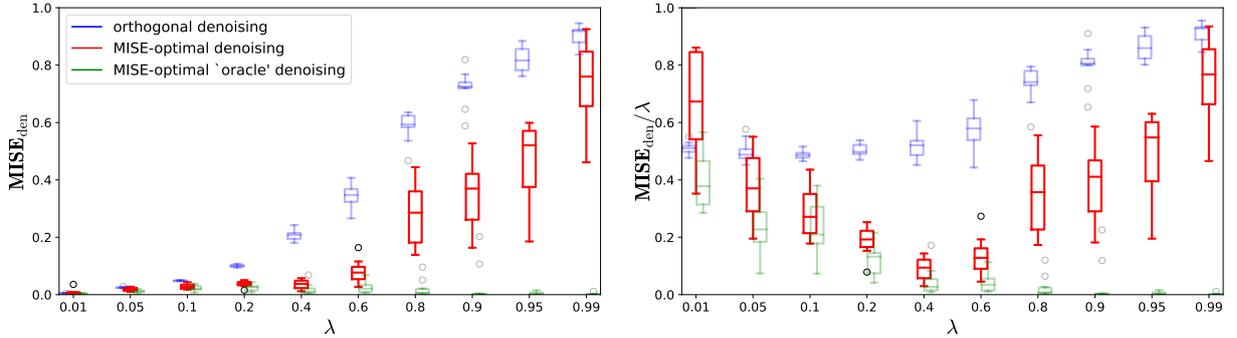


Figure 3: MISE (*left panel*) and proportion of the (integrated) variance of the remaining noise (*right panel*) after MISE-optimal denoising (*red*) as a function of the noise level λ . Same for orthogonal denoising (*blue*) and the “oracle” version (*green*).

The figure shows that MISE-optimal denoising is consistent, as all noise gets removed when n tends to infinity. For $d = 6$ convergence has not yet been achieved in the plot, but we verified it converges as well. Comparing MISE-optimal denoising with the “oracle” version, it is clear that for small n most of the improvement comes from estimating \mathcal{M} and \mathcal{M}_ε , while for larger n it comes from the regression step. For small n the performance of the “oracle” version does not increase much as a function of n (see $d = 4$ and $d = 6$), while for the same regime the performance of MISE-optimal denoising increases significantly as a function of n . For larger n both approaches converge at more or less the same speed.

It is worth mentioning that the performance of MISE-optimal denoising is particularly sensitive to a correct estimation of d , the dimension of \mathcal{M} , for which (multiple) bootstrap tests were used. Almost all outliers in the boxplots for MISE-optimal denoising are due to erroneous estimates of d .

MISE-optimal denoising almost always (except for very small n and large values of d) outperforms orthogonal denoising. Furthermore, orthogonal denoising does not lead to perfect denoising. To highlight this fact Figure 2 also exhibits a theoretical lower bound $\mathbf{MISE}_{\text{ortho}}^{\min}$ (for details, see Appendix F). This bound is due to the part of the noise that is parallel to \mathcal{M} and therefore irreducible if you project orthogonally onto \mathcal{M} .

Figure 3 displays the denoising performances as a function of λ , for fixed $n = 800$ and $d = 4$. Only for very small noise levels ($\lambda = 0.01$) our method is outperformed by the orthogonal denoising. This is explained by the fact that estimating \mathcal{M}_ε is rather noisy for such small noise levels (note that the ‘‘oracle’’ version still outperforms orthogonal denoising). However, the same plot also shows that the impact of this underperformance is small. For such small noise levels the observed curves $Y_t(\cdot)$ are already very close to the signal curves and denoising does not change that.

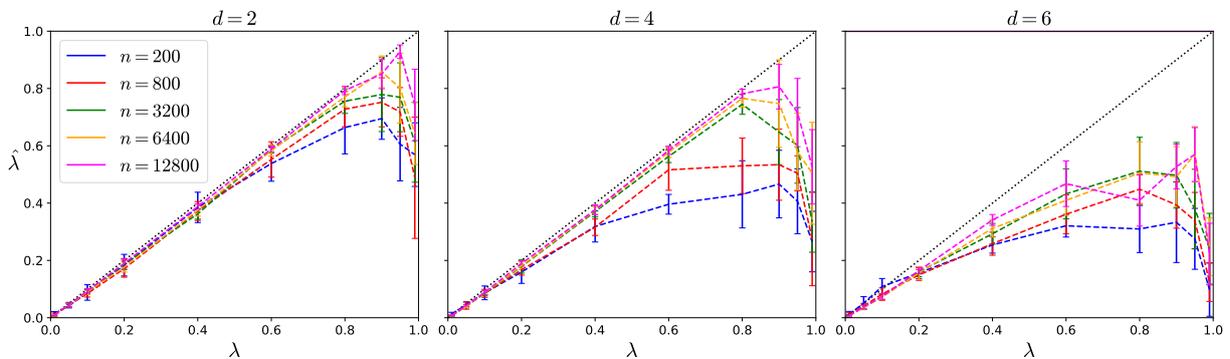


Figure 4: Estimation of the noise level λ , based on (2.19). The error bars signify 1 sample standard deviation.

In Figure 4 we use (2.19) to estimate the noise level λ . Note that $\mathbf{MISE}_{\text{opt}}^{\min} = 0$ for this default simulation setup. For $d = 2$ and $d = 4$ we see that this approach leads to a consistent estimator of λ . The finite-sample bias leads to an underestimation of the noise level. An explanation might be the cutoffs we use to estimate the dimensions of \mathcal{M}_ε , \mathcal{M}_\parallel and \mathcal{M}_\perp . This is tantamount to neglecting a (small) portion of the noise. Similar to the results of Figure 2, convergence for $d = 6$ is much slower.

Note that with $\lambda = \text{Tr}[\Sigma_\varepsilon] / \sqrt{\text{Var}[Y_t(\cdot)]}$ we have an alternative approach to estimate λ . Applying this method to our simulations, with $\widehat{\Sigma}_\varepsilon^+(u, v)$ as an estimator for $\Sigma_\varepsilon(u, v)$, leads to a second consistent estimator of λ . However, the variance of this estimator is much larger (in particular for $d = 4$ and $d = 6$) and therefore we did not show it in the plots.

In Figure 5 we use exactly the same setup as in Figure 2, except that now $\theta_2 = 0$ (cf. (4.8)). This means that $\mathcal{M} \cap \mathcal{M}_\varepsilon \neq \emptyset$, i.e. the signal and noise curves have a principal component direction in common. As a consequence, also MISE-optimal denoising has an irreducible noise component (see Appendix F for details) given by $\mathbf{MISE}_{\text{opt}}^{\min} = g_\varepsilon^2(\lambda)/a^2$. This serves as a theoretical lower bound at the population level, and Figure 5 clearly exhibits convergence towards this lower bound. This shows that also in the more general case of $\mathcal{M} \cap \mathcal{M}_\varepsilon \neq \emptyset$ MISE-optimal denoising is consistent.

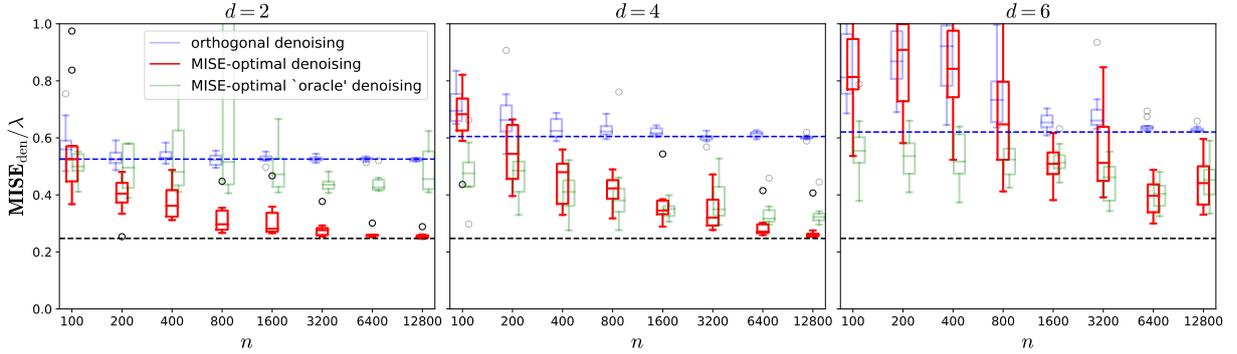


Figure 5: Proportion of the (integrated) variance of the remaining noise after denoising. Similar to Figure 2, except that now $\theta_2 = 0$ (cf. (4.8)). Also shown is the theoretical lower bound for orthogonal denoising (*blue, dashed line*) and MISE-optimal denoising (*black, dashed line*).

Note that the “oracle” version with exact knowledge of \mathcal{M} and \mathcal{M}_ε is performing worse than MISE-optimal denoising, in particular for $d = 2$. This might seem counter-intuitive, because the “oracle” version has more information about the underlying data. It can be understood however by observing that $\theta_2 = 0$ reduces the dimension of \mathcal{M}_\perp by one to $\dim \mathcal{M}_\perp = d_\varepsilon - 1$. Due to estimation noise the algorithm does not pick up this exact reduction and this makes $\widehat{\Omega}_\perp$ nearly singular, causing a poor denoising.

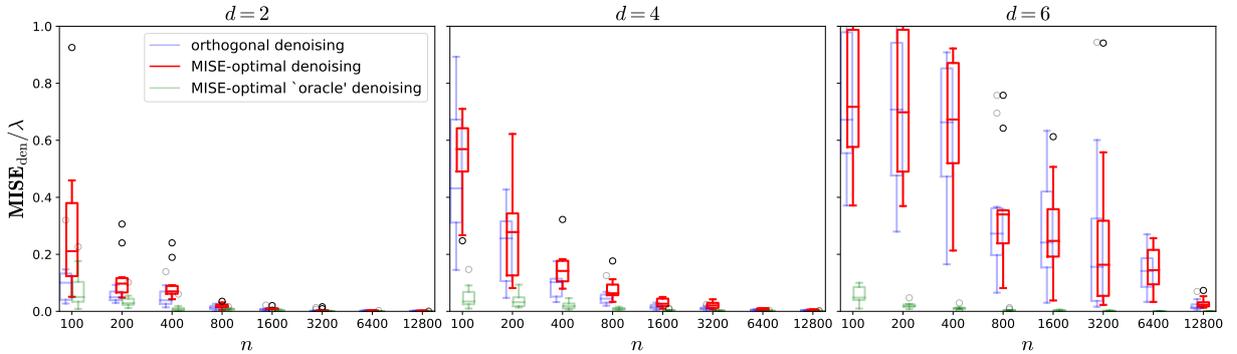


Figure 6: Proportion of the (integrated) variance of the remaining noise after denoising. Similar to Figure 2, except that now $\theta_j = \pi/2$ for $j = 1, \dots, \min(d, d_\varepsilon)$. This means that \mathcal{M}_ε is in the orthogonal complement of \mathcal{M} .

Finally, we investigate the case of all noise curves lying in the orthogonal complement of the signal curves: $\theta_j = \pi/2$ for $j = 1, \dots, \min(d, d_\varepsilon)$. In the absence of parallel noise components, there is no dependent variable in the regression step and MISE-optimal denoising is equivalent to orthogonal denoising. Indeed, Figure 6 shows that both methods have similar performances and convergence properties. MISE-optimal denoising even performs slightly worse than orthogonal projection, since due to estimation noise the result of

the regression step is not a perfectly orthonormal projection onto \mathcal{M} .

4.4 Forecasting

Forecasting of the simulated curve time series is possible by means of the FPCA framework and the extensive literature on forecasting of VAR models (Lütkepohl, 2005). We focus on one-step-ahead forecasting the signal curve $X_t(u)$. In the hypothetical situation where the signal curves are accessible by direct observation, one can use the covariance operator $\Sigma_X(\cdot, \cdot)$ to find the Karhunen-Loève expansion (1.2). The factor loadings $\boldsymbol{\xi}_t$ can be modelled with a VAR(p) process,

$$\boldsymbol{\xi}_t = \sum_{\ell=1}^p A_p \boldsymbol{\xi}_{t-\ell} + \mathbf{e}_t, \quad (4.10)$$

where $\mathbf{e}_t \sim \text{IID}(\mathbf{0}, \Omega)$, whose parameters can be least-square estimated. The one-step-ahead forecast of the signal curve at time t is then given by

$$\widehat{X}_t^{(p)}(u) = \sum_{i=1}^d \widehat{\xi}_{ti}^{(p)} \phi_i(u), \quad (4.11)$$

where $\widehat{\xi}_{ti}^{(p)} = \sum_{\ell=1}^p \widehat{A}_p \boldsymbol{\xi}_{t-\ell}$ is the optimal forecast of the factor loadings in the sense of minimizing the MSE of the forecast error (Lütkepohl, 2005). To evaluate forecasting performance, we define a normalized mean integrated squared error

$$\Delta_F := \frac{\overline{\mathbb{E}}[(X_t(\cdot) - \widehat{X}_t^{(p)}(\cdot))^2]}{\overline{\mathbb{E}}[(X_t(\cdot))^2]}. \quad (4.12)$$

In reality the signal curves $X_t(\cdot)$ are usually not observable. The default workaround would be to use the observed curves $Y_t(\cdot)$ to compute the eigenfunctions of the Karhunen-Loève expansion. We have seen earlier that this introduces a bias due to the noise covariance. Since MISE-optimal denoising reduces the noise term, it is expected that using the denoised curves $\widehat{X}_t^{\text{opt}}(\cdot)$ instead will lead to a better forecasting performance.

In Table 1 one-step ahead forecast errors Δ_F are computed in-sample, using five different approaches. On top of using the Karhunen-Loève expansion and the MISE-optimal denoised curves $\widehat{X}_t^{\text{opt}}(\cdot)$, we also use the orthogonally denoised curves $Y_t^{\parallel}(\cdot)$ and we introduce a “mean forecast” (using the mean of the observed curves as forecast) and a “naive forecast” (using the observed curve of the previous timestamp as forecast). Forecasting performance is also compared with the theoretical lower bound $g_X^2(\lambda) \text{Tr}[\Omega] / (1 - \lambda)$ of Δ_F due to the residual term in the VAR(p) process defined in (4.10). For the sake of simplicity, since we are mainly interested in differences in forecasting performance between different approaches, we take $p = 1$ and use for MISE-optimal and orthogonal denoising knowledge of the true value of d without estimating it. We find that MISE-optimal denoising outperforms the other forecasting strategies. The increase in performance compared to the Karhunen-Loève approach or the orthogonal-projection approach is not large in absolute

λ	0.05	0.2	0.4
mean forecast	$0.998 \pm 1.16e-4$	$0.999 \pm 1.15e-4$	$0.999 \pm 9.98e-5$
naive forecast	$2.335 \pm 5.83e-3$	$2.527 \pm 5.01e-3$	$2.935 \pm 6.00e-3$
Karhunen-Loève	$0.623 \pm 1.56e-3$	$0.669 \pm 1.62e-3$	$0.767 \pm 2.52e-3$
MISE-optimal denoising	$0.614 \pm 1.59e-3$	$0.618 \pm 1.56e-3$	$0.628 \pm 1.61e-3$
orthogonal denoising	$0.622 \pm 1.56e-3$	$0.654 \pm 1.53e-3$	$0.709 \pm 1.47e-3$
theoretical lower bound	0.617	0.617	0.617

Table 1: Estimated forecast errors Δ_F based on a VAR(1) model of the factor loadings (including standard error of the mean). Estimations are based on a sample of size 200 with parameters $n = 800$ and $d = 4$. The bottom row shows the theoretical lower bound of the forecast errors.

terms, but it can be large in relative terms compared to the theoretical lower bound of the forecast error (in particular for $\lambda = 0.4$).

Figure 7 gives more insights in the relevant factors affecting the forecasting performance Δ_F . In the top panel we compare forecasting by means for Karhunen-Loève, MISE-optimal denoising and orthogonal denoising, as a function of λ . The angles between the dynamical and noise space are the default choice: $\theta_1 = \theta_2 = \theta_3 = \theta_4 = \pi/4$. The figure shows results for the default VAR(1) model with $d = 4$, defined in (B.6), and an “alternative VAR(1) model” with $d = 4$ and eigenvalues of A that are much closer to the unit circle. The bottom panel only differs in the fact that $\theta_2 = 0$, meaning that there is an irreducible noise component.

There are several takeaways from Figure 7. First of all, forecasting based on MISE-optimal denoised signals outperforms the other approaches and the performance differences can be large (see for example $\lambda = 0.60$). When the underlying VAR(1) model of the factor loadings has eigenvalues closer to the unit circle, all forecasting strategies have an improved performance. This makes sense, as larger eigenvalues mean a relatively stronger serial dependence of the time series (or, equivalently, relatively smaller residuals). Finally, comparing the two panels of Figure 7, we see that the relative performance increase due to MISE-optimal denoising is less in the presence of irreducible noise components. This is also as expected, since in this case the amount of noise that MISE-optimal denoising can reduce is more limited in a relative sense.

5 Denoising empirical weather data

In this section MISE-optimal denoising is applied to empirical weather data. We consider hourly temperature measurements on 4015 consecutive days between 17 May 2012 and 14 May 2023. This means the curve time series has length 4015 and each curve consists of 24 measurements. The measurements were taken at a weather station in De Bilt, The Netherlands, and are made publicly available by the Royal Netherlands Meteorological Institute

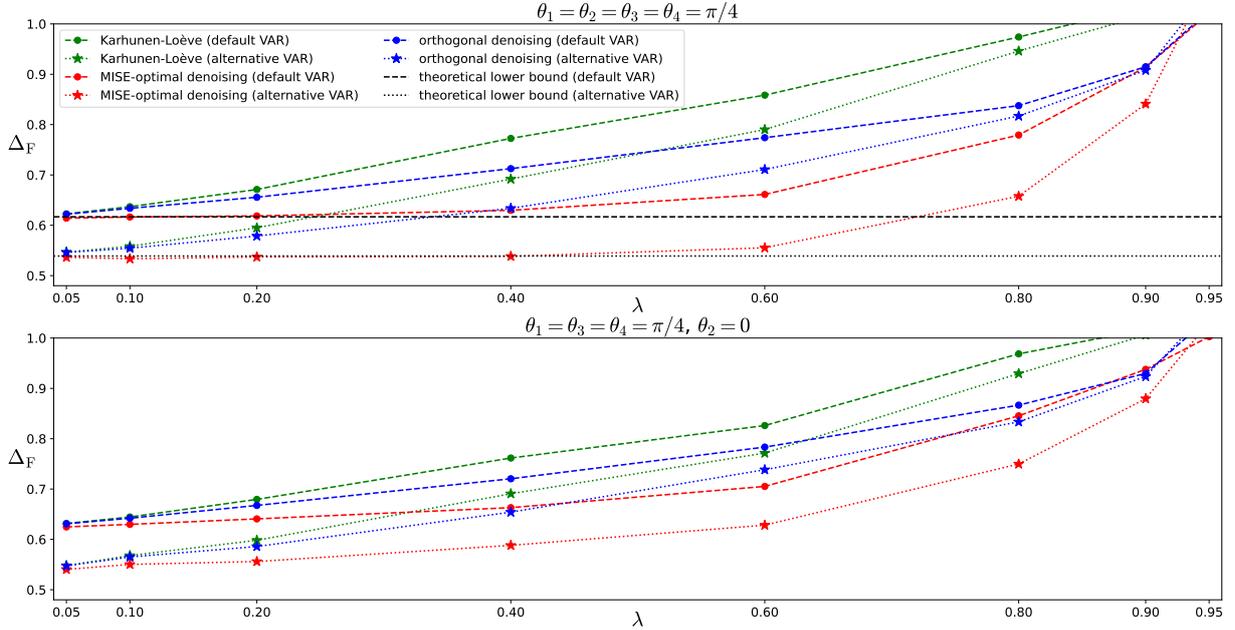


Figure 7: Estimated forecast errors Δ_F based on a VAR(1) model of the factor loadings, as a function of the noise level λ . Estimations are based on 50 simulations with parameters $n = 800$ and $d = 4$. Standard errors are negligible. The lower panel is the case with irreducible noise.

(KNMI).² The data coming from this particular weather station has been “homogenized” for eventual relocations of the weather station or changes in the measurement setup.

Before analyzing this data in the context of MISE-optimal denoising, three pre-processing steps are taken. Curve time series in the denoising methodology are assumed to be stationary. In order to make the empirical data more stationary, a correction is made for the seasonal trends in daily average temperature. At each day of the year an average daily temperature is estimated by means of a weighted Gaussian kernel with a bandwidth of 15 days and this average is subtracted from the data. Secondly, an hourly mean (the $\mu(u)$ in (1.2)) is estimated and subtracted from the data. Since this estimate turned out to be a smooth function, no additional smoothing was needed. Finally, the demeaned data is re-scaled such that it has a sample standard deviation of 1. This last step is strictly speaking not required by the proposed denoising method, but it makes interpreting the results more convenient. Note, for example, that the data now has the same variance as the observed curves of the simulated data in section 4.

Figures 8 and 9 show the estimated orthonormal bases of the dynamical space \mathcal{M} and the noise space \mathcal{M}_ε , respectively, with estimated dimensions $\hat{d} = 4$ and $\hat{d}_\varepsilon = 7$. Hyperparameters for estimation are the same as for the simulated data in section 4 (here we use a bootstrap sample size of 100). The first basis function of the dynamical space,

²<https://www.knmi.nl/nederland-nu/klimatologie/daggegevens>

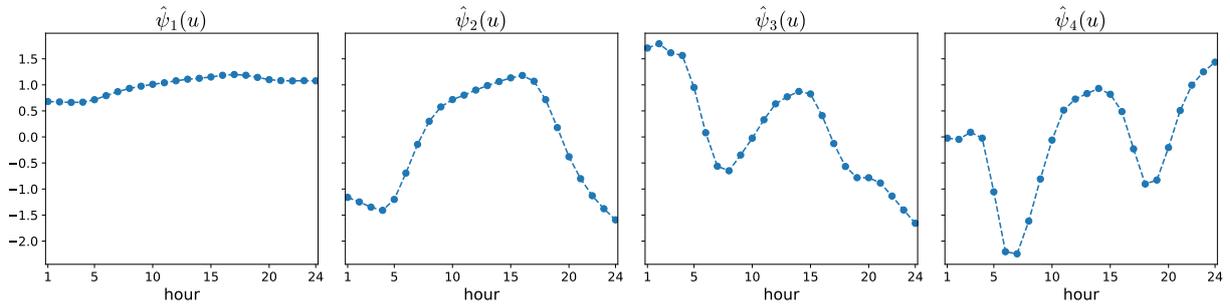


Figure 8: Estimated basis of the dynamical space \mathcal{M} .

$\hat{\psi}_1(\cdot)$, is more or less constant and can be interpreted as the persisting daily average away from the seasonal average that was subtracted during pre-processing. Days of relative hot or cold weather are often clustered. The second basis function, $\hat{\psi}_2(\cdot)$, can be interpreted as the persisting day-night difference in temperature, away from the hourly mean $\mu(u)$ that was subtracted from the data. It suggests that “diurnal air temperature variation” (the difference in minimum/maximum temperature during one day) is persistent during a series of consecutive days, which is in agreement with meteorological research (Cho *et al.*, 2020). Interpreting the other two basis functions seems less trivial, apart from the fact that they seem to be odd and even, and that most variation occurs during sunrise and sunset. It should be noted though that the eigenvalue of $\hat{\psi}_1(\cdot)$ (recall that the basis functions are eigenfunctions of the operator $K(\cdot, \cdot)$) covers 96.20% of the sum of the four eigenvalues, and the first two eigenvalues together cover 99.96%. Finally, there does not seem to be a non-expert interpretation of the basis functions of the noise space. Perhaps not surprisingly, their frequency seems to be increasing gradually.

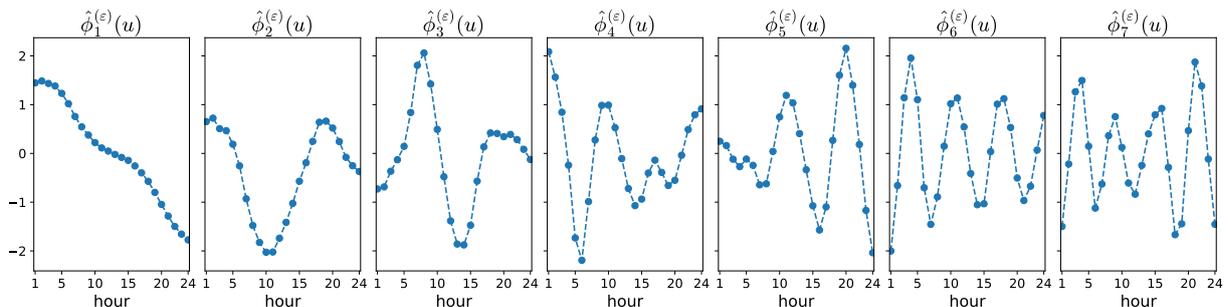


Figure 9: Estimated basis of the noise space \mathcal{M}_ε .

We have applied MISE-optimal denoising to this data and find that $\hat{d}_\parallel = 3$ and $\hat{d}_\perp = 5$. It is difficult to assess the validity of these estimates, as the data-generating process is unknown. However, it seems reassuring that the estimates are not at their theoretical minimum (0) and maximum ($\hat{d}_\varepsilon = 7$) values. In Figure 10 the MISE-optimal denoised curves are plotted for five consecutive days, alongside the original data, the orthogonally denoised curves and curves that are obtained by applying the Karhunen-Loève expansion directly

to the observed data without denoising. For the latter we worked with 5 eigenfunctions, which together explain 98% of the variance of the observed data. Note that this choice is rather arbitrary; for an increasing number of eigenfunctions the curves will converge to the original data. Also note that, unlike MISE-optimal and orthogonal denoising, this approach does not attempt to separate the dynamical, persistent part of the time series from the white-noise component.

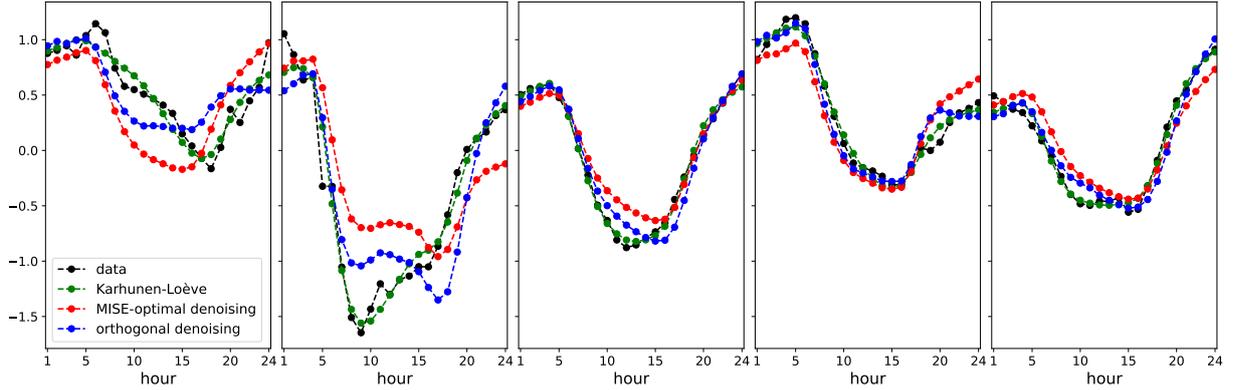


Figure 10: MISE-optimal denoised curves for five consecutive days, alongside the original data, the orthogonally denoised curves and curves that are obtained by applying the Karhunen-Loève expansion to the observed data.

Despite the absence of knowledge of the data-generating process, it is possible to estimate the proportion of the noise that has been reduced and the noise level in the original data (i.e. λ). Consider the most general case, similar to the setup of Figure 5 of the simulations, where a part of the noise is irreducible. Furthermore, assume that MISE-optimal denoising has converged, i.e. all reducible noise has been removed. Since here $\hat{d} = 4$ and $n = 4015$, looking at Figure 5 this assumption seems reasonable but not entirely accurate. We have noted however that the first two basis functions of \mathcal{M} cover 99.96% of the total variance in the dynamical space, making \mathcal{M} effectively two-dimensional and making the assumption more appropriate. As a consequence, we can estimate the MISE of the parts of the noise that remain after denoising and that have been removed by denoising:

$$\text{MISE}_{\text{opt}}^{(\text{remaining})} = \text{MISE}_{\text{opt}}^{\min} = \text{Tr} [\Omega_{\parallel}] - \text{Tr} \left[(\Omega_{\perp})^{-1} (\Omega_{\parallel\perp})^T \Omega_{\parallel\perp} \right] \hat{=} 0.0759, \quad (5.1a)$$

$$\text{MISE}_{\text{opt}}^{(\text{removed})} = \bar{\text{E}}[(Y_t(\cdot) - \hat{X}_t^{\text{opt}}(\cdot))^2] \hat{=} 0.0884. \quad (5.1b)$$

The proportion of the variance of the noise removed by MISE-optimal denoising is then

$$\frac{\text{MISE}_{\text{opt}}^{(\text{removed})}}{\text{MISE}_{\text{opt}}^{(\text{removed})} + \text{MISE}_{\text{opt}}^{(\text{remaining})}} \hat{=} 0.538, \quad (5.2)$$

and noise level is (cf. (2.19))

$$\lambda = \text{MISE}_{\text{opt}}^{(\text{removed})} + \text{MISE}_{\text{opt}}^{(\text{remaining})} \hat{=} 0.164, \quad (5.3)$$

where one should recall that the data has been re-scaled to have unit variance. In other words, 53.8% of the variance of the noise is removed by MISE-optimal denoising. In the same fashion orthogonal denoising removes 27.2% of the variance of the noise and gives the same estimate of λ . We thus see that MISE-optimal denoising has a considerable advantage over orthogonal denoising in the case of this empirical dataset.

Finally, the five one-step-ahead forecasting approaches of section 4.4 were *mutatis mutandis* applied to this dataset. Figure 11 shows the resulting forecasts for five consecutive days. In the absence of knowledge of the signal curves, assessing the forecasting performance is not straightforward. As a proxy, we replace the signal curves $X_t(\cdot)$ by the observed curves $Y_t(\cdot)$ in (4.12). This leads to an extra contribution to the performance measure, coming from the noise curves, which is expected to be the same for all forecasting methods and therefore comparing forecasting performances is still possible.

mean forecast	1.000
naive forecast	0.687
Karhunen-Loève	0.401
MISE-optimal denoising	0.418
orthogonal denoising	0.406

Table 2: Estimated forecast errors Δ_F , with $X_t(\cdot)$ replaced by $Y_t(\cdot)$, for the empirical weather data, based on five forecasting methods and using a VAR(1) model for the factor loadings.

Table 2 shows that the forecasting performances of the Karhunen-Loève approach and the two denoising approaches are comparable, while clearly outperforming the mean and naive forecasts. The fact that MISE-optimal denoising clearly outperformed orthogonal denoising and the Karhunen-Loève approach, which essentially does not remove any noise, in terms of noise reduction, is not translated in a better forecasting performance. An explanation could be that the amount of reduced noise is only one of the many factors that influence forecasting performance, as Figure 7 indicated for the simulated data. Considering that for the temperature data $\lambda \hat{=} 0.164$ and the proportion of irreducible noise is higher than in the lower panel of Figure 7, it seems to make sense that the forecasting performances of the three approaches are close. Furthermore, the relative serial dependence within this curve time series could also play a role; if the serial dependence is small relative to $1/\lambda$, it will be difficult for denoising approaches to improve on the forecasting performance.

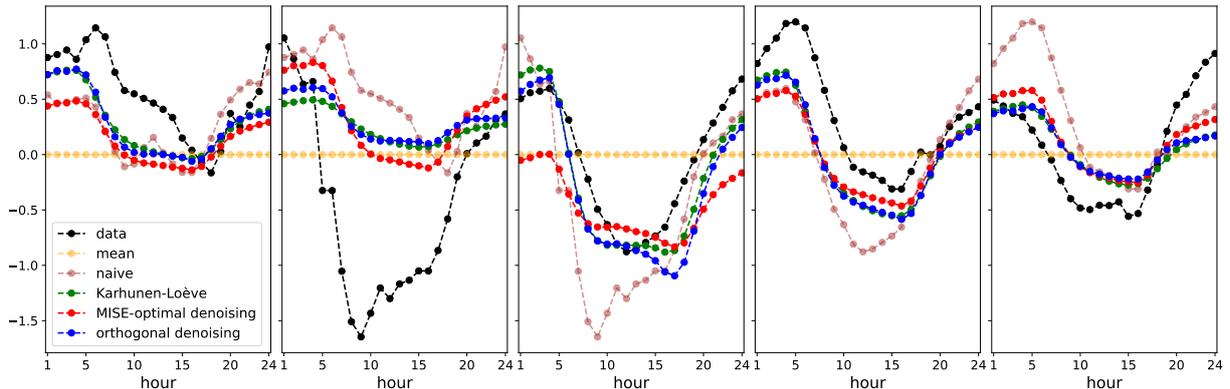


Figure 11: One-day-ahead forecasts for five consecutive days based on five different methods, alongside the original data.

6 Conclusions and discussion

By building on the DFPCA method of Bathia *et al.* (2010), this paper has introduced a method to disclose the structure of the noise space of functional time series and developed a MISE-optimal denoising procedure. By using simulated and real data, it was shown that this method outperforms DFPCA-based orthogonal denoising and can also be used as a pre-processing step to improve forecasting. We believe that this result can be seen as a “blessing of dimensionality” (Gorban and Tyukin, 2018), since we can exploit the high dimensionality of the observed data to disentangle the dynamics and the noise in a curve time series.

An interesting open question stems from the restrictions imposed on the denoising operators we consider for minimizing the MISE. Only $Y_t(\cdot)$ is used to predict the underlying signal $X_t(\cdot)$. Adding information about the dynamics, for example $Y_{t-1}(\cdot)$, could potentially improve the denoising further.

We should also point out a hidden assumption about the noise. We apply the FPCA framework to the noise curves, assuming they can be approximated by a low-dimensional decomposition. It is unclear to what extent this assumption is valid for real data, or which impact a violation of this assumption has on MISE-optimal denoising.

Finally, let us mention a number of potential extensions and applications of MISE-optimal denoising. Although this paper considers denoising of functional time series, nowhere have we assumed anything about the smoothness of the curves of the time series. In other words, the ordering of the curves along the coordinate u is irrelevant. Our method can therefore easily be extended to vector-valued, high-dimensional time series for which the factor model approach (Lam and Yao, 2012; Cubadda and Hecq, 2022; Dong *et al.*, 2022) is applicable.

To verify this explicitly, we have considered denoising of curve time series where the coordinate u is shuffled. In the simulations in this paper the spatial coordinate is represented by a grid of $N = 200$ equidistant points u_1, u_2, \dots, u_N on the interval $\mathcal{I} = [0, 1]$. Given a

curve time series $Y_t(\cdot)$, with a slight abuse of notation we can then introduce a “shuffled time series” $(\sigma Y)_t(u_i) = Y_t(u_{\sigma(i)})$, where σ is a permutation of the integers $\{1, 2, \dots, N\}$. In the same fashion one can define $(\sigma X)_t(u_i)$ and $(\sigma \varepsilon)_t(u_i)$. We have verified numerically that MISE-optimal denoising applied to the shuffled time series leads to exactly the same denoising as when applied to the original curve time series. To be more concrete, we found that $\widehat{(\sigma X)}_t^{\text{opt}}(u) = (\sigma \widehat{X}^{\text{opt}})_t(u)$, or equivalently $\widehat{X}_t^{\text{opt}}(u) = (\sigma^{-1} \widehat{(\sigma X)}_t^{\text{opt}})_t(u)$.

Surface time series, which are relevant for e.g. climate and environmental sciences (Guillas and Lai, 2010; Martínez-Hernández and Genton, 2020), are another potential direction for extending the MISE-optimal denoising approach.

Another potential application is to use MISE-optimal denoising as a pre-processing step for functional linear regression. The hypothesis would be that using a denoised signal outperforms existing methods, which are essentially based on DFPCA and orthogonal denoising (Chen *et al.*, 2022; Chang *et al.*, 2023).

References

- Andreasen, M. M., Christensen, J. H. E. and Rudebusch, G. D. (2019). Term Structure Analysis with Big Data: One-Step Estimation Using Bond Prices. *Journal of Econometrics*, **212**, number 1, 26–46.
- Aue, A., Norinho, D. D. and Hörmann, S. (2015). On the Prediction of Stationary Functional Time Series. *Journal of the American Statistical Association*, **110**, number 509, 378–392.
- Bai, J. and Ng, S. (2002). Determining the Number of Factors in Approximate Factor Models. *Econometrica*, **70**, number 1, 191–221.
- Bai, J. and Ng, S. (2007). Determining the Number of Primitive Shocks in Factor Models. *Journal of Business & Economic Statistics*, **25**, number 1, 52–60.
- Bathia, N., Yao, Q. and Ziegelmann, F. (2010). Identifying the finite dimensionality of curve time series. *The Annals of Statistics*, **38**, number 6, 3352–3386.
- Bosq, D. (2000). *Linear processes in function spaces: theory and applications*, volume 149. Berlin: Springer.
- Caldeira, J. and Torrent, H. (2017). Forecasting the US Term Structure of Interest Rates Using Nonparametric Functional Data Analysis. *Journal of Forecasting*, **36**, number 1, 56–73.
- Chang, J., Chen, C., Qiao, X. and Yao, Q. (2023). An autocovariance-based learning framework for high-dimensional functional time series. *Journal of Econometrics*.
- Chen, C., Guo, S. and Qiao, X. (2022). Functional Linear Regression: Dependence and Error Contamination. *Journal of Business & Economic Statistics*, **40**, number 1, 444–457.
- Chen, Y., Chi, Y., Fan, J. and Ma, C. (2021). Spectral Methods for Data Science: A Statistical Perspective. *Foundations and Trends® in Machine Learning*, **14**, number 5, 566–806.
- Cho, D., Yoo, C., Im, J. and Cha, D.-H. (2020). Comparative Assessment of Various Machine Learning-Based Bias Correction Methods for Numerical Weather Prediction Model Forecasts of Extreme Air Temperatures in Urban Areas. *Earth and Space Science*, **7**, number 4.
- Cho, H., Goude, Y., Brossat, X. and Yao, Q. (2013). Modeling and Forecasting Daily Electricity Load Curves: A Hybrid Approach. *Journal of the American Statistical Association*, **108**, number 501, 7–21.

- Cubadda, G. and Hecq, A. (2022). Dimension Reduction for High-Dimensional Vector Autoregressive Models. *Oxford Bulletin of Economics and Statistics*, **84**, number 5, 1123–1152.
- Dahlhaus, R. and Wefelmeyer, W. (1996). Asymptotically optimal estimation in misspecified time series models. *The Annals of Statistics*, **24**, number 3, 952–974.
- Dauxois, J., Pousse, A. and Romain, Y. (1982). Asymptotic theory for the principal component analysis of a vector random function: Some applications to statistical inference. *Journal of Multivariate Analysis*, **12**, number 1, 136–154.
- Descary, M.-H. and Panaretos, V. M. (2019). Functional data analysis by matrix completion. *The Annals of Statistics*, **47**, number 1, 1–38.
- Dong, Y., Qin, S. J. and Boyd, S. P. (2022). Extracting a low-dimensional predictable time series. *Optimization and Engineering*, **23**, number 2, 1189–1214.
- Ferraty, F. and Vieu, P. (2006). *Nonparametric functional data analysis: theory and practice*. Springer Series in Statistics. Berlin: Springer.
- Gao, Z. and Tsay, R. S. (2021). Modeling High-Dimensional Time Series: A Factor Model With Dynamically Dependent Factors and Diverging Eigenvalues. *Journal of the American Statistical Association*, 1–17.
- Gorban, A. N. and Tyukin, I. Y. (2018). Blessing of dimensionality: mathematical foundations of the statistical physics of data. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, **376**, number 2118, 20170237.
- Guillas, S. and Lai, M.-J. (2010). Bivariate splines for spatial functional regression models. *Journal of Nonparametric Statistics*, **22**, number 4, 477–497.
- Hall, P. and Vial, C. (2006). Assessing the finite dimensionality of functional data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **68**, number 4, 689–705.
- Hallin, M. and Liška, R. (2007). Determining the Number of Factors in the General Dynamic Factor Model. *Journal of the American Statistical Association*, **102**, number 478, 603–617.
- Hallin, M., Nisol, G. and Tavakoli, S. (2023). Factor models for high-dimensional functional time series I: Representation results. *Journal of Time Series Analysis*.
- Hays, S., Shen, H. and Huang, J. Z. (2012). Functional dynamic factor models with application to yield curve forecasting. *The Annals of Applied Statistics*, **6**, number 3, 870–894.

- Hyndman, R. J. and Ullah, M. S. (2007). Robust forecasting of mortality and fertility rates: A functional data approach. *Computational Statistics & Data Analysis*, **51**, number 10, 4942–4956.
- Hörmann, S., Kidziński, L. and Hallin, M. (2015). Dynamic functional principal components. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **77**, number 2, 319–348.
- Hörmann, S. and Kokoszka, P. (2010). Weakly dependent functional data. *The Annals of Statistics*, **38**, number 3, 1845–1884.
- Hörmann, S. and Kokoszka, P. P. (2012). Functional Time Series. In *Time Series Analysis* (ed. T. S. Rao), Handbook of Statistics, volume 30, pp. 157–186. Elsevier B.V.
- Jones, M. C. and Rice, J. A. (1992). Displaying the Important Features of Large Collections of Similar Curves. *The American Statistician*, **46**, number 2, 140–145.
- Lam, C. and Yao, Q. (2012). Factor modeling for high-dimensional time series: Inference for the number of factors. *The Annals of Statistics*, **40**, number 2, 694–726.
- Lam, C., Yao, Q. and Bathia, N. (2011). Estimation of latent factors for high-dimensional time series. *Biometrika*, **98**, number 4, 901–918.
- Li, D., Robinson, P. M. and Shang, H. L. (2020). Long-Range Dependent Curve Time Series. *Journal of the American Statistical Association*, **115**, number 530, 957–971.
- Lütkepohl, H. (2005). *New Introduction to Multiple Time Series Analysis*. Berlin: Springer.
- Martínez-Hernández, I. and Genton, M. G. (2020). Recent developments in complex and spatially correlated functional data. *Brazilian Journal of Probability and Statistics*, **34**, number 2, 204–229.
- Pan, J. and Yao, Q. (2008). Modelling multiple time series via common factors. *Biometrika*, **95**, number 2, 365–379.
- Panaretos, V. M. and Tavakoli, S. (2013). Fourier analysis of stationary time series in function space. *The Annals of Statistics*, **41**, number 2, 568–603.
- Pena, D. and Box, G. E. P. (1987). Identifying a Simplifying Structure in Time Series. *Journal of the American Statistical Association*, **82**, number 399, 836.
- Qin, S. J. (2022). Latent vector autoregressive modeling and feature analysis of high dimensional and noisy data from dynamic systems. *AIChE Journal*, **68**, number 6.
- Qin, S. J., Dong, Y., Zhu, Q., Wang, J. and Liu, Q. (2020). Bridging systems theory and data science: A unifying review of dynamic latent variable analytics and process monitoring. *Annual Reviews in Control*, **50**, 29–48.

- Ramsay, J. O. and Silverman, B. W. (2005). *Functional Data Analysis*. Springer Series in Statistics. New York: Springer.
- Sen, R. and Klüppelberg, C. (2019). Time series of functional data with application to yield curves. *Applied Stochastic Models in Business and Industry*, **35**, number 4, 1028–1043.
- Shang, H. L. and Hyndman, R. J. (2011). Nonparametric time series forecasting with dynamic updating. *Mathematics and Computers in Simulation*, **81**, number 7, 1310–1324.
- Tavakoli, S., Nisol, G. and Hallin, M. (2023). Factor models for high-dimensional functional time series II: Estimation and forecasting. *Journal of Time Series Analysis*.
- Tjøstheim, D. and Paulsen, J. (1983). Bias of some commonly-used time series estimates. *Biometrika*, **70**, number 2, 389–399.
- Yang, Y., Yang, Y. and Shang, H. L. (2022). Feature extraction for functional time series: Theory and application to NIR spectroscopy data. *Journal of Multivariate Analysis*, **189**, 104863.
- Yao, F., Müller, H.-G. and Wang, J. (2005). Functional Data Analysis for Sparse Longitudinal Data. *Journal of the American Statistical Association*, **100**, number 470, 577–590.

A Curve time series, FPCA and DFPCA

This appendix contains some technical details about the definition of the curve time series we consider, FPCA and the DFPCA approach of Bathia *et al.* (2010). It should be read as supplemental to sections 1 and 2.1.

Let's start by making the definition of the curve time series (also called “functional time series”) under consideration more precise. We have defined the noise term in (1.1) as the part of the observable curves $Y_t(\cdot)$ that does not have serial dependence. In particular, we will assume that $\varepsilon_t(\cdot)$ is a white noise sequence, as was done in, for example, Bathia *et al.* (2010); Chen *et al.* (2022). This means that $E[\varepsilon_t(u)] = 0$ for all t and any $u \in \mathcal{I}$, and that $\text{Cov}[\varepsilon_t(u), \varepsilon_s(v)] = 0$ for any $u, v \in \mathcal{I}$ and all $t \neq s$.

Furthermore, we will assume that both the signal $X_t(\cdot)$ and the noise $\varepsilon_t(\cdot)$ are square integrable on the bounded interval \mathcal{I} , and that

$$\int_{\mathcal{I}} E[X_t(u)^2 + \varepsilon_t(u)^2] du \quad (\text{A.1})$$

is finite and well-defined. Finally, we assume stationarity of the curve time series. This means that both

$$\mu(u) = E[X_t(u)] \quad \text{and} \quad M_k(u, v) = \text{Cov}[X_t(u), X_{t+k}(v)], \quad (\text{A.2})$$

where $k \in \mathbb{Z}$, are independent of t .

We apply the framework of FPCA to the signal curves $X_t(\cdot)$. As a starting point, we write the zero-lag autocovariance of the signal curves in terms of its spectral decomposition (2.1), where $\lambda_1 \geq \lambda_2 \geq \dots \geq 0$ are the eigenvalues and $\{\phi_i(\cdot)\}_{i=1}^{\infty}$ are the corresponding orthonormal eigenfunctions of the linear Hilbert-Schmidt operator

$$\Sigma_X : L^2(\mathcal{I}) \rightarrow L^2(\mathcal{I}), \quad \Sigma_X(f)(u) = \int_{\mathcal{I}} \Sigma_X(u, v) f(v) dv. \quad (\text{A.3})$$

The inner product on the Hilbert space is defined as

$$\langle f(\cdot), g(\cdot) \rangle = \int_{\mathcal{I}} f(v) g(v) dv. \quad (\text{A.4})$$

Note that we are using the same symbol Σ_X for the covariance function (2.1) and the linear operator (A.3), in order to keep notation simple. The existence of the expansion in (2.1) is guaranteed by Mercer's theorem, which is applicable because $\Sigma_X(\cdot, \cdot)$ is a continuous, symmetric, non-negative definite kernel.

The assumption of FPCA is that the (centered) curves $X_t(u) - \mu(u)$ can be approximately represented by a finite number of eigenfunctions,

$$X_t(u) \approx \mu(u) + \sum_{i=1}^m \xi_{ti} \phi_i(u) =: X_t^{(m)}(u), \quad (\text{A.5})$$

and that this approximation improves for increasing m , in the sense that

$$\lim_{m \rightarrow \infty} \mathbb{E} \left[\left(X_t(u) - X_t^{(m)}(u) \right)^2 \right] = 0, \quad \text{for all } u \in \mathcal{I}. \quad (\text{A.6})$$

Inspired by the framework of FPCA and similarly to Bathia *et al.* (2010), we assume that the signal $X_t(\cdot)$ of the curve time series (1.1) under consideration is “ d -dimensional”, by which we mean that $\lambda_i = 0$ for $i > d$. In other words, we assume that the exact identity (1.2) holds for a certain integer d .

In section 2.1 it was stated that almost any choice of the operator $K(\cdot, \cdot)$, defined in (2.4), gives an eigenspace that coincides with the dynamical space \mathcal{M} . In Bathia *et al.* (2010) it was namely proven that for any $k \geq 1$ for which the matrix $\Sigma_k = \mathbb{E}[\boldsymbol{\xi}_t \boldsymbol{\xi}_{t+k}^T]$ is full-rank, the operator

$$N_k(u, v) := \int_{\mathcal{I}} M_k(u, z) M_k(v, z) dz \quad (\text{A.7})$$

has exactly d nonzero eigenvalues and their corresponding eigenfunctions span the dynamical space \mathcal{M} . Since it can be argued (Pena and Box, 1987; Pan and Yao, 2008) that always for some $k \geq 1$, $\text{rank}(\mathbb{E}[\boldsymbol{\xi}_t \boldsymbol{\xi}_{t+k}^T]) = d$, this always provides an approach to estimating d and \mathcal{M} . As a consequence, the same properties hold for any operator of the form (2.4), as long as there is a $k \in \{1, 2, \dots, q\}$ for which $c_k \neq 0$ and Σ_k is full-rank. For the sake of simplicity, we will assume throughout this paper that Σ_k is of full rank for all $k \geq 1$.

The advantage of using $K(\cdot, \cdot)$ instead of $N_k(\cdot, \cdot)$ in the estimation procedure is that the former combines information about the dynamics of the underlying process from different lags. It is an interesting question what the optimal values for q and c_ℓ in (2.4) are, in order to estimate d and \mathcal{M} . Intuitively, a small value for q discards useful information stored in larger lags, while a large value for q introduces more noise in the estimation coming from the larger lags. It has been reported however in different contexts (Bathia *et al.*, 2010; Lam *et al.*, 2011; Chen *et al.*, 2022) that the precise choice of $K(\cdot, \cdot)$ does not have much impact on the quality of the estimates of d and \mathcal{M} . We will therefore not investigate this question any further in the remainder of this paper.

B Vector autoregressive models

This appendix contains some technical details about vector autoregressive models relevant for our analysis, as well as a specification of the VAR(1) models used in the simulations of section 4. See, for example, Lütkepohl (2005) for more details about VAR models.

A general VAR(1) model is defined by

$$\vec{\xi}_t = \vec{\alpha} + A \vec{\xi}_{t-1} + \vec{e}_t, \quad (\text{B.1})$$

where $\vec{\xi}_t \in \mathbb{R}^d$, $\vec{\alpha} \in \mathbb{R}^d$ and the d -dimensional $\vec{e}_t \sim \text{IID}(\vec{0}, \Omega)$. The matrix A is not necessarily symmetric, while the matrix Ω is. A VAR process is called stationary if

$$\mathbb{E}[\vec{\xi}_t], \quad \text{Var}[\vec{\xi}_t] \quad \text{and} \quad \Sigma_k := \text{Cov}[\vec{\xi}_t, \vec{\xi}_{t+k}] \quad (\text{B.2})$$

are independent of t . If A has all its eigenvalues within the unit circle, then it is a stationary process.

Let us now take the perspective of DFPCA as summarized in section 2.1. Assume that we have chosen an operator $K(\cdot, \cdot)$ and that $\psi_1(\cdot), \psi_2(\cdot), \dots, \psi_d(\cdot)$ is the corresponding orthonormal set of eigenfunctions. This implies that

$$\mathcal{M} = \text{span}\{\phi_1(u), \dots, \phi_d(u)\} = \text{span}\{\psi_1(u), \dots, \psi_d(u)\}. \quad (\text{B.3})$$

and, furthermore, that there is an orthogonal matrix C such that $\phi(u) = C\psi(u)$, where $\phi(u) = (\phi_1(u), \dots, \phi_d(u))^T$ and $\psi(u) = (\psi_1(u), \dots, \psi_d(u))^T$. Considering the two expansions of the demeaned signal curves in (2.5), this means that $\boldsymbol{\eta}_t = C^T \boldsymbol{\xi}_t$, where $\boldsymbol{\xi}_t$ was defined earlier as $\boldsymbol{\xi}_t = (\xi_{t1}, \dots, \xi_{td})^T$ and $\boldsymbol{\eta}_t$ is defined similarly. We also have that $\Sigma_k^{(\eta)} = C^T \Sigma_k C$, where the covariance matrices were defined below (2.6). Recall that by the Karhunen-Loève theorem $\Sigma_0 = \text{diag}(\lambda_1, \dots, \lambda_d)$ (see (2.2)).

In section 2.2 we devised a method for estimating the $\Sigma_\varepsilon(\cdot, \cdot)$, using the assumption (2.9) that the principal components $\boldsymbol{\xi}_t$ follow a VAR(1)-process. As a consequence of the assumption (2.9), the random coefficients $\boldsymbol{\eta}_t$ also follow a VAR(1) process,

$$\boldsymbol{\eta}_t = A^{(\eta)} \boldsymbol{\eta}_{t-1} + \mathbf{e}_t^{(\eta)}, \quad \text{where } \mathbf{e}_t^{(\eta)} \sim \text{IID}(\mathbf{0}, \Omega^{(\eta)}), \quad (\text{B.4})$$

and where $A^{(\eta)} = C^T A C$ and $\Omega^{(\eta)} = C^T \Omega C$. The constant term is also absent here, because $E[\boldsymbol{\eta}_t] = \mathbf{0}$ as well.

As a final step in section 2.2 we employ the Yule-Walker equations for the VAR(1) process of the $\boldsymbol{\eta}_t$:

$$\Sigma_0^{(\eta)} = A^{(\eta)} \Sigma_0^{(\eta)} (A^{(\eta)})^T + \Omega^{(\eta)}, \quad (\text{B.5a})$$

$$\Sigma_k^{(\eta)} = \Sigma_{k-1}^{(\eta)} (A^{(\eta)})^T \quad \text{for } k > 0. \quad (\text{B.5b})$$

The above equations are conventionally used to estimate the model parameters of the VAR(1) process. This so-called Yule-Walker estimator has the same asymptotic properties as the least-squares estimator, although for small samples it sometimes performs worse (Lütkepohl, 2005; Tjøstheim and Paulsen, 1983). We use the Yule-Walker equations differently, namely by “reasoning backwards” and computing $\Sigma_0^{(\eta)}$ through (2.10) from the lag-1 and lag-2 autocovariance matrices, which both can be estimated consistently.

VAR(1) models used in simulations

As explained in section 4.1, throughout our simulations we work with three dynamical spaces \mathcal{M} with dimensions $d = 2, 4$ and 6 . The principal components ξ_{tj} are simulated according to the VAR(1)-process in (2.9). The model for $d = 2$ is specified by in the main

text. For $d = 4$ we use

$$A = \begin{pmatrix} -0.40475218 & 0.56881667 & -0.01251201 & -0.33319225 \\ 0.36328118 & 0.23656237 & 0.17826015 & 0.47609812 \\ 0.04062105 & -0.13439131 & -0.3596354 & -0.24931481 \\ -0.31412948 & 0.08911365 & -0.36549673 & 0.20076313 \end{pmatrix}, \quad (\text{B.6a})$$

$$\Omega = \begin{pmatrix} 0.39050087 & 0.06370578 & 0.0340153 & -0.10433378 \\ 0.06370578 & 0.35412048 & 0.05387206 & 0.07341199 \\ 0.0340153 & 0.05387206 & 0.2960237 & -0.02286662 \\ -0.10433378 & 0.07341199 & -0.02286662 & 0.06964716 \end{pmatrix}. \quad (\text{B.6b})$$

For $d = 6$ we use

$$A = \begin{pmatrix} 0.37504966 & 0.08142893 & -0.07435684 & -0.03887785 & 0.25655029 & 0.25170869 \\ -0.14126954 & -0.19192149 & -0.0982056 & -0.37670302 & 0.16884435 & -0.38686508 \\ 0.00451676 & -0.32514261 & -0.22975774 & 0.12353677 & 0.27258333 & 0.26566839 \\ 0.44140703 & -0.08094657 & 0.05391765 & -0.09386828 & 0.03307928 & -0.14231888 \\ 0.3419833 & -0.20556356 & 0.19934397 & 0.08967538 & 0.0027988 & 0.22842928 \\ 0.01997925 & 0.10989784 & 0.29140585 & -0.007507 & 0.38542961 & 0.19185898 \end{pmatrix}, \quad (\text{B.7a})$$

$$\Omega = \begin{pmatrix} 0.56177209 & 0.04343599 & -0.0262747 & -0.10676641 & -0.0826483 & -0.03922044 \\ 0.04343599 & 0.46386005 & -0.02291322 & 0.01014503 & 0.05098028 & 0.02312988 \\ -0.0262747 & -0.02291322 & 0.36764981 & -0.00149778 & -0.03508077 & 0.01351097 \\ -0.10676641 & 0.01014503 & -0.00149778 & 0.25032321 & -0.11118431 & -0.00733749 \\ -0.0826483 & 0.05098028 & -0.03508077 & -0.11118431 & 0.15925535 & -0.02909279 \\ -0.03922044 & 0.02312988 & 0.01351097 & -0.00733749 & -0.02909279 & 0.09806406 \end{pmatrix}. \quad (\text{B.7b})$$

C M(I)SE-optimal denoising

This appendix provides a detailed derivation and discussion of the MISE-optimal denoising formula (2.16). It starts with a discussion of a much simpler case, namely MSE-optimal denoising for a two-dimensional time series. Then the n -dimensional generalization is considered, after which we discuss the case of functional time series. Because all cases are similar in essence, starting with more simple situations provides further insights in MISE-optimal denoising.

MSE-optimal denoising in \mathbb{R}^2

Consider a stationary time series of a bivariate random variable $\vec{Y}_t = \vec{X}_t + \vec{\varepsilon}_t$, with $E[\vec{\varepsilon}_t] = 0$ and $\Omega_\varepsilon = \text{Var}[\vec{\varepsilon}_t]$ a (2x2)-matrix. As in the rest of this paper, assume that all persistence of the time series is contained in the ‘‘signal’’ part \vec{X}_t and that $\vec{\varepsilon}_t$ is white noise. Furthermore, assume that the \vec{X}_t lie in a one-dimensional ‘‘dynamical space’’ $\mathcal{M} = \text{span}(\vec{u})$, where $\vec{u} \in \mathbb{R}^2$ is a unit vector. In other words, this is the case where $d = \dim \mathcal{M} = 1$. Note that this

setup corresponds to the illustration of MISE-optimal denoising in Figure 1. We do not consider the case $d = 2$, because then necessarily $\mathcal{M}_\varepsilon \subset \mathcal{M}$ and (as we will see below) MSE-optimal denoising is not applicable.

Given an (observed) \vec{Y}_t , the goal is to reconstruct the corresponding \vec{X}_t in an MSE-optimal fashion. In other words, we would like to find a (linear) projection P onto \mathcal{M} that performs this reconstruction, $\vec{X}_t^{\text{opt}} = P\vec{Y}_t$, and such that $\text{E}[|\vec{X}_t^{\text{opt}} - \vec{X}_t|^2]$ is minimized. Observe that the noise can be uniquely decomposed into a part parallel to \mathcal{M} and an orthogonal part, $\vec{\varepsilon}_t = \vec{\varepsilon}_t^\parallel + \vec{\varepsilon}_t^\perp$. The idea behind MSE-optimal denoising is as follows. With knowledge of \mathcal{M} and given a \vec{Y} , you can compute the orthogonal part of the noise and use this to reconstruct \vec{X} .

Because P is a projection onto the dynamical space, $P\vec{X}_t = \vec{X}_t$ and thus $\vec{X}_t^{\text{opt}} = \vec{X}_t + P\vec{\varepsilon}_t$. Let's write $\vec{\varepsilon}_t^\perp = \varepsilon_t^\perp \vec{v}$, where $\vec{v} \in \mathbb{R}^2$ is a unit vector perpendicular to \vec{u} . Note that \vec{v} is uniquely defined up to a sign. Since $P\vec{\varepsilon}_t^\parallel = \vec{\varepsilon}_t^\parallel$ and considering a specific form of the projection of the perpendicular noise, $P\vec{\varepsilon}_t^\perp = \alpha \varepsilon_t^\perp \vec{u}$, the objective of minimizing the MSE then translates into minimizing

$$\text{E}[|\vec{X}_t^{\text{opt}} - \vec{X}_t|^2] = \text{E}[|P\vec{\varepsilon}_t|^2] = \text{Var}[\varepsilon_t^\parallel] + 2\alpha \text{Cov}[\varepsilon_t^\parallel, \varepsilon_t^\perp] + \alpha^2 \text{Var}[\varepsilon_t^\perp], \quad (\text{C.1})$$

where it was used that $\text{E}[\vec{\varepsilon}_t] = 0$ and ε_t^\parallel was defined such that $\vec{\varepsilon}_t^\parallel = \varepsilon_t^\parallel \vec{u}$. The (co)variances in (C.1) are related to the original covariance matrix via an orthonormal basis transformation of the noise subspace \mathcal{M}_ε ,

$$\begin{pmatrix} \text{Var}[\varepsilon_t^\parallel] & \text{Cov}[\varepsilon_t^\parallel, \varepsilon_t^\perp] \\ \text{Cov}[\varepsilon_t^\parallel, \varepsilon_t^\perp] & \text{Var}[\varepsilon_t^\perp] \end{pmatrix} = S^T \Omega_\varepsilon S, \quad \text{where } S = (\vec{u} \ \vec{v}). \quad (\text{C.2})$$

The minimum of the convex cost function (C.1) is given by

$$\hat{\alpha} = -\frac{\text{Cov}[\varepsilon_t^\parallel, \varepsilon_t^\perp]}{\text{Var}[\varepsilon_t^\perp]}, \quad (\text{C.3})$$

leading to the MSE-optimal denoised signal

$$\hat{X}_t^{\text{opt}} = \vec{X}_t + \hat{P}\vec{\varepsilon}_t \quad (\text{C.4a})$$

$$= \vec{Y}_t^\parallel + \hat{P}\vec{\varepsilon}_t^\perp \quad (\text{C.4b})$$

$$= P_{\mathcal{M}}\vec{Y}_t + \hat{\alpha} \varepsilon_t^\perp \vec{u} \quad (\text{C.4c})$$

$$= \left[P_{\mathcal{M}} - \vec{u} \frac{\text{Cov}[\varepsilon_t^\parallel, \varepsilon_t^\perp]}{\text{Var}[\varepsilon_t^\perp]} \vec{v}^T (I - P_{\mathcal{M}}) \right] \vec{Y}_t, \quad (\text{C.4d})$$

where \vec{Y}_t^\parallel is the part of \vec{Y}_t parallel to the signal subspace \mathcal{M} and $P_{\mathcal{M}} = \vec{u}\vec{u}^T$ is the matrix of the orthogonal projection onto \mathcal{M} . In the final line we used that $\varepsilon_t^\perp = \vec{v} \cdot \vec{\varepsilon}_t = \vec{v} \cdot \vec{\varepsilon}_t^\perp =$

$\vec{v} \cdot (\vec{Y}_t - \vec{Y}_t^\parallel)$. The particular order of factors in (C.4d) was chosen with the prospect of generalizing this expression to higher dimensional time series. Finally, it should be noted that in essence the MSE-optimal denoising projection amounts to finding the MSE-optimal linear regression function of the conditional mean $E[\varepsilon_t^\parallel | \varepsilon_t^\perp] = -\alpha \varepsilon_t^\perp$, without an intercept because $E[\vec{\varepsilon}_t] = 0$.

The MSE-optimal denoising formula (C.4d) exists provided that $\text{Var}[\varepsilon_t^\perp] \neq 0$. If this condition is not satisfied, the noise space \mathcal{M}_ε is (effectively) one-dimensional and parallel to \mathcal{M} , i.e. $\mathcal{M}_\varepsilon = \mathcal{M}$. Intuitively it makes sense that a denoising projection is not possible in this situation. Given \vec{Y}_t , since $\varepsilon_t^\perp = 0$ we cannot use the part of the noise perpendicular to \mathcal{M} to make a prediction for the noise parallel to \mathcal{M} . The only remaining, sensible denoising option is subtracting a noise bias term from \vec{Y}_t , but since we assumed $E[\vec{\varepsilon}_t] = 0$ this bias is zero.

Observe that when the noise parallel and perpendicular to the dynamical space \mathcal{M} are uncorrelated the MSE-optimal projection reduces to the orthogonal projection onto \mathcal{M} . This is as expected, since in this case ε_t^\perp does not contain any information about ε_t^\parallel and therefore cannot be used to project out (part of) ε_t^\parallel .

Also note that the minimized MSE is given by $E[|\hat{P}\vec{\varepsilon}_t|^2] = \text{Var}[\varepsilon_t^\parallel] (1 - \rho^2)$, where ρ is the correlation coefficient between ε_t^\parallel and ε_t^\perp . We get a perfect denoising (i.e. zero error) in the special case of $\rho = \pm 1$, which generally corresponds to the situation $\varepsilon_t^\parallel = \gamma \varepsilon_t^\perp$ for some finite $\gamma \neq 0$. In other words, perfect denoising occurs when there is perfect correlation between the parallel and perpendicular parts of the noise. The noise space is then (effectively) one-dimensional, $\mathcal{M}_\varepsilon = \text{span}(\vec{w})$ with \vec{w} a unit vector, and makes an angle with \mathcal{M} that is given by $\tan^{-1}(1/\gamma)$. In this special case the MSE-optimal projection can take an alternative form $\hat{X}_t = \tilde{P}\vec{Y}_t$, where

$$\tilde{P} = \begin{pmatrix} \vec{u} & \vec{0} \end{pmatrix} (A^T A)^{-1} A^T \quad \text{and} \quad A = \begin{pmatrix} \vec{u} & \vec{w} \end{pmatrix}. \quad (\text{C.5})$$

Assuming $\rho = \pm 1$ it is not difficult to show that the two projection formulas correspond, i.e. $\hat{P} = \tilde{P}$. From a geometric viewpoint, \tilde{P} corresponds to a projection of \vec{Y}_t parallel to \mathcal{M}_ε onto \mathcal{M} . This makes sense, as the noise is located on the line spanned by the vector \vec{w} .

Note that within this special situation there exist two particularly special cases. When $\gamma = 0$ there is no noise parallel to \mathcal{M} , implying that $\vec{w} \perp \vec{u}$. The projection corresponds to the orthogonal projection onto \mathcal{M} , meaning that $\tilde{P} = P_{\mathcal{M}}$, because $A^T A = I$. The other special case is when $\gamma \rightarrow \infty$. The noise then becomes parallel to \mathcal{M} , implying that $A^T A$ becomes a singular matrix and the projection formula for \tilde{P} breaks down. This corresponds to the situation $\mathcal{M}_\varepsilon = \mathcal{M}$ which was discussed earlier and for which no MSE-optimal denoising exists.

MSE-optimal denoising in \mathbb{R}^n

This section generalizes the situation of the previous section to a time series of an n -dimensional random variable $\vec{Y}_t = \vec{X}_t + \vec{\varepsilon}_t$. Unless stated otherwise, the assumptions of the

previous section are still valid. It is now assumed that the signal \vec{X}_t lies in a d -dimensional dynamical space \mathcal{M} with $0 < d < n$. The case where $d = n$ is not of interest, because in that case necessarily $\mathcal{M}_\varepsilon \subseteq \mathcal{M}$ and this renders MSE-optimal denoising impossible.

As in the previous section, the noise vector is (uniquely) decomposed as $\vec{\varepsilon}_t = \vec{\varepsilon}_t^\parallel + \vec{\varepsilon}_t^\perp$, but now

$$\vec{\varepsilon}_t^\parallel = \sum_{i=1}^{d_\parallel} \tilde{\varepsilon}_{t,i}^\parallel \vec{u}_i, \quad \vec{\varepsilon}_t^\perp = \sum_{j=1}^{d_\perp} \tilde{\varepsilon}_{t,j}^\perp \vec{v}_j, \quad (\text{C.6})$$

where the vectors $\vec{u}_1, \dots, \vec{u}_{d_\parallel}$ and $\vec{v}_1, \dots, \vec{v}_{d_\perp}$ form orthonormal basis of parallel and perpendicular noise spaces, \mathcal{M}_\parallel and \mathcal{M}_\perp , respectively. It is important to realize that the two bases combined do not necessarily form a basis of the total noise space \mathcal{M}_ε . For example, in the case of perfect correlations between some parallel noise components $\tilde{\varepsilon}_{t,i}^\parallel$ and some perpendicular noise components $\tilde{\varepsilon}_{t,j}^\perp$, the total noise space will have a dimension $d_\varepsilon < d_\parallel + d_\perp$. We will investigate this case further below. The current analysis is valid both in the absence of perfect correlations ($d_\varepsilon = d_\parallel + d_\perp$) and in the presence of perfect correlations ($d_\varepsilon < d_\parallel + d_\perp$).

The covariances of the parallel and perpendicular components of the noise can be computed from Ω_ε by means of coordinate transformations,

$$\Omega_\parallel := \text{Cov}[\tilde{\varepsilon}_t^\parallel, \tilde{\varepsilon}_t^\parallel] = U^T \Omega_\varepsilon U, \quad (\text{C.7a})$$

$$\Omega_\perp := \text{Cov}[\tilde{\varepsilon}_t^\perp, \tilde{\varepsilon}_t^\perp] = V^T \Omega_\varepsilon V, \quad (\text{C.7b})$$

$$\Omega_{\parallel\perp} := \text{Cov}[\tilde{\varepsilon}_t^\parallel, \tilde{\varepsilon}_t^\perp] = U^T \Omega_\varepsilon V, \quad (\text{C.7c})$$

where

$$U = (\vec{u}_1 \ \vec{u}_2 \ \dots \ \vec{u}_{d_\parallel}), \quad V = (\vec{v}_1 \ \vec{v}_2 \ \dots \ \vec{v}_{d_\perp}). \quad (\text{C.8})$$

Note that $\Omega_{\parallel\perp}$ is a matrix of dimensions $d_\parallel \times d_\perp$.

We again look for a projection P that minimizes the MSE $\text{E}[|\vec{X}_t^{\text{opt}} - \vec{X}_t|^2]$ for for this purpose we model the projection of the perpendicular noise component as

$$P\vec{\varepsilon}_t^\perp = \sum_{i=1}^{d_\parallel} \sum_{j=1}^{d_\perp} \alpha_{ij} \tilde{\varepsilon}_{t,j}^\perp \vec{u}_i = U \alpha \tilde{\varepsilon}_t^\perp. \quad (\text{C.9})$$

The MSE then takes the form

$$\mathbb{E}[\|\vec{X}_t^{\text{opt}} - \vec{X}_t\|^2] = \mathbb{E}[\|P\vec{\varepsilon}_t\|^2] \quad (\text{C.10a})$$

$$\begin{aligned} &= \sum_{i=1}^{d_{\parallel}} \text{Var}[\tilde{\varepsilon}_{t,i}^{\parallel}] + 2 \sum_{i=1}^{d_{\parallel}} \sum_{j=1}^{d_{\perp}} \alpha_{ij} \text{Cov}[\tilde{\varepsilon}_{t,i}^{\parallel}, \tilde{\varepsilon}_{t,j}^{\perp}] \\ &\quad + \sum_{i=1}^{d_{\parallel}} \sum_{j=1}^{d_{\perp}} \sum_{k=1}^{d_{\perp}} \alpha_{ij} \alpha_{ik} \text{Cov}[\tilde{\varepsilon}_{t,j}^{\perp}, \tilde{\varepsilon}_{t,k}^{\perp}] \end{aligned} \quad (\text{C.10b})$$

$$= \text{Tr}[\text{Var}[\tilde{\varepsilon}_t^{\parallel}]] + 2 \text{Tr}[\alpha^T \text{Cov}[\tilde{\varepsilon}_t^{\parallel}, \tilde{\varepsilon}_t^{\perp}]] + \text{Tr}[\alpha \text{Cov}[\tilde{\varepsilon}_t^{\perp}, \tilde{\varepsilon}_t^{\perp}] \alpha^T] \quad (\text{C.10c})$$

$$= \text{Tr}[\Omega_{\parallel}] + 2 \text{Tr}[\alpha^T \Omega_{\parallel\perp}] + \text{Tr}[\alpha \Omega_{\perp} \alpha^T], \quad (\text{C.10d})$$

This optimization problem is convex, since Ω_{\perp} is positive semi-definite and therefore $\text{Tr}[\alpha \Omega_{\perp} \alpha^T] \geq 0$ for any nonzero matrix α . First order conditions are solved by $\hat{\alpha} = -\Omega_{\parallel\perp} (\Omega_{\perp})^{-1}$, provided Ω_{\perp} is invertible. This results in an MSE-optimal denoised signal

$$\hat{X}_t^{\text{opt}} = \vec{X}_t + \hat{P} \vec{\varepsilon}_t \quad (\text{C.11a})$$

$$= \vec{Y}_t^{\parallel} + \hat{P} \vec{\varepsilon}_t^{\perp} \quad (\text{C.11b})$$

$$= P_{\mathcal{M}} \vec{Y}_t + U \hat{\alpha} \tilde{\varepsilon}_t^{\perp} \quad (\text{C.11c})$$

$$= [P_{\mathcal{M}} - U \Omega_{\parallel\perp} (\Omega_{\perp})^{-1} V^T (I - P_{\mathcal{M}})] \vec{Y}_t, \quad (\text{C.11d})$$

where in the last line we used that $\tilde{\varepsilon}_{t,j}^{\perp} = \vec{v}_j \cdot \vec{\varepsilon}_t^{\perp} = \vec{v}_j \cdot (\vec{Y}_t - \vec{Y}_t^{\parallel})$. This is the general MSE-optimal denoising result for n -dimensional time series.

Three special cases are worth to be mentioned here. First, there is the possibility that $\mathcal{M}_{\varepsilon} \subseteq \mathcal{M}$. This means the absence of a noise component perpendicular to the dynamical space, $\vec{\varepsilon}_t = \vec{\varepsilon}_t^{\parallel}$, and therefore the impossibility of MSE-optimal denoising. In the formulation of (C.11d), the matrix Ω_{\perp} is not defined in this situation. Secondly, there is the possibility that $\mathcal{M}_{\varepsilon} \subseteq \mathcal{M}^{\perp}$. The noise is fully perpendicular to the dynamical space, $\vec{\varepsilon}_t = \vec{\varepsilon}_t^{\perp}$, and MSE-optimal denoising reduces to orthogonal denoising: $\hat{P} = P_{\mathcal{M}}$.

At last, there is the already mentioned case of perfect correlations between parallel and perpendicular components of the noise. This implies $d_{\varepsilon} < d_{\perp} + d_{\parallel}$. The general result (C.11d) is still valid here, but there is an alternative formulation of MSE-optimal denoising. For those components of the parallel noise that are perfectly correlated with $\vec{\varepsilon}_t^{\perp}$, it is now possible to project them out (completely) by projection in those directions parallel to the noise space $\mathcal{M}_{\varepsilon}$.

Let's illustrate this in the particular case where all parallel noise components are perfectly correlated with $\vec{\varepsilon}_t^{\perp}$. In other words, consider the case where $\vec{\varepsilon}_t^{\parallel} = \Gamma \vec{\varepsilon}_t^{\perp}$ for some $(d_{\parallel} \times d_{\perp})$ -dimensional matrix Γ . This is equivalent to saying that the noise space and the dynamical space do not have any directions in common, i.e. $\mathcal{M}_{\varepsilon} \cap \mathcal{M} = \emptyset$. Note that in

this case $d_\varepsilon = \max(d_\parallel, d_\perp)$. If $\vec{w}_1, \vec{w}_2, \dots, \vec{w}_{d_\varepsilon}$ is an orthonormal basis of \mathcal{M}_ε , an alternative MSE-optimal projection equivalent to (C.11d) is then given by

$$\tilde{P} = \left(\vec{u}_1 \dots \vec{u}_{d_\parallel} \vec{0} \dots \vec{0} \right) (A^T A)^{-1} A^T \quad \text{where} \quad A = \left(\vec{u}_1 \dots \vec{u}_{d_\parallel} \vec{w}_1 \dots \vec{w}_{d_\varepsilon} \right). \quad (\text{C.12})$$

In this case the denoising is perfect in the sense that the minimized MSE is equal to zero.

MISE-optimal denoising in $L^2(\mathcal{I})$

In this section details about the derivation of MISE-optimal denoising for curve time series is being discussed. It should be read as a supplement to section 2.3, where the main steps of the derivation were presented. It can also be seen as a generalization of the previous sections of this appendix, from finite-dimensional time series to functional time series.

As soon as the parallel and perpendicular noise spaces are properly defined, as was done in section 2.3, the derivation of MISE-optimal denoising is rather similar to the finite n -dimensional case of the previous section. The analogue of (C.9) was defined in (2.13). Using that $X_t^{\text{opt}}(u) = X_t(u) + \varepsilon_t^\parallel(u) + (P\varepsilon_t^\perp)(u)$, the orthonormality of the bases of \mathcal{M}_\parallel and \mathcal{M}_\perp , as well as the fact that the noise curves have zero mean, one can easily find that

$$\mathbf{MISE}_{\text{opt}} = \text{Tr} [\Omega_\parallel] + 2 \text{Tr} [\alpha^T \Omega_{\parallel\perp}] + \text{Tr} [\alpha \Omega_\perp \alpha^T], \quad (\text{C.13})$$

where Ω_\perp and $\Omega_{\parallel\perp}$ were defined in (2.15) and

$$\Omega_\parallel := \int_{\mathcal{I}} \int_{\mathcal{I}} \phi^\parallel(u) (\phi^\parallel(v))^T \Sigma_\varepsilon(u, v) du dv. \quad (\text{C.14})$$

As in the finite-dimensional case this is a convex optimization problem and the first-order conditions are solved by $\hat{\alpha} = -\Omega_{\parallel\perp} (\Omega_\perp)^{-1}$, provided Ω_\perp is invertible. The solution of the first-order conditions lead to the MISE-optimal denoising formula (2.16), the main result of this paper. Note that the curves $Y_t^\parallel(\cdot)$ in this expression can be conveniently expressed in terms of the basis vectors of \mathcal{M} :

$$Y_t^\parallel(u) := (P_{\mathcal{M}} Y_t)(u) = \int_{\mathcal{I}} [\boldsymbol{\psi}(u) \cdot \boldsymbol{\psi}(v)] Y_t(v) dv \quad \text{where} \quad P_{\mathcal{M}}(u, v) := \boldsymbol{\psi}(u) \cdot \boldsymbol{\psi}(v) \quad (\text{C.15})$$

is the operator for the orthogonal projection onto \mathcal{M} .

Similar to the finite-dimensional case, there are three special situations that need to be mentioned. In the absence of noise perpendicular to the dynamical space ($\mathcal{M}_\varepsilon \subseteq \mathcal{M}$) there is no MISE-optimal denoising possible. In the absence of noise parallel to the dynamical space, MISE-optimal denoising is equivalent to orthogonal denoising. And finally, in the case of perfect correlations ($d_\varepsilon < d_\parallel + d_\perp$) some parallel noise components can be removed completely.

In the special case that all parallel noise components are perfectly correlated with the perpendicular noise ($\varepsilon_t^\parallel = \Gamma \varepsilon_t^\perp$), one can project an observed curve $Y_t(\dots)$ parallel to the

noise space \mathcal{M}_ε onto the dynamical space \mathcal{M} and thereby remove all noise. The projection operator that achieves this is given by

$$\mathcal{P}_i(u, v) := \phi_i(u) \sum_{j=1}^{d+d_\varepsilon} (\mathcal{A}^T \mathcal{A})_{ij}^{-1} \tilde{\phi}_j(v) \quad (\text{C.16})$$

where

$$\tilde{\phi}(u) := \left(\phi_1(u), \dots, \phi_d(u), \phi_1^{(\varepsilon)}(u), \dots, \phi_{d_\varepsilon}^{(\varepsilon)}(u) \right)^T \quad \text{and} \quad (\mathcal{A}^T \mathcal{A})_{ij} := \int_{\mathcal{I}} \tilde{\phi}_i(u) \tilde{\phi}_j(u) du. \quad (\text{C.17})$$

The denoised signal is then

$$\tilde{X}_t^{\text{opt}}(u) = \sum_{i=1}^d \int_{\mathcal{I}} \mathcal{P}_i(u, v) Y_t(v) dv. \quad (\text{C.18})$$

At the population level this alternative denoising procedure is equivalent to our MISE-optimal denoising result (2.16). However, it turns out that at the level of finite-size samples this method performs less well, in the sense that it produces a larger MISE. The explanation is that $\mathcal{A}^T \mathcal{A}$ can quickly become nearly-singular due to estimation noise, causing large denoising errors.

D Estimation formulas for MISE-optimal denoising

This appendix contains the (rather standard) expressions for estimators needed for MISE-optimal denoising. It should be read as supplemental to section 3.

Estimation of \mathcal{M}

Defining an estimator for the operator $K(\cdot, \cdot)$ is straightforward,

$$\hat{K}(u, v) := \sum_{\ell=1}^p c_\ell \hat{N}_\ell(u, v), \quad (\text{D.1})$$

with

$$\hat{N}_k(u, v) = \int_{\mathcal{I}} \hat{M}_k(u, z) \hat{M}_k(v, z) dz \quad (\text{D.2})$$

and where

$$\hat{M}_k(u, v) = \frac{1}{n-k-1} \sum_{t=1}^{n-k} (Y_t(u) - \bar{Y}(u)) (Y_{t+k}(v) - \bar{Y}(v)), \quad (\text{D.3})$$

for $k = 1, 2, \dots$ and where $\bar{Y}(u) = \frac{1}{n} \sum_{t=1}^n Y_t(u)$. Note that for $k = -1, -2, \dots$ we have $\hat{M}_k(u, v) = \hat{M}_{-k}(v, u)$. The operator $\hat{K}(\cdot, \cdot)$ has orthonormal eigenfunctions given by

$$(\hat{K} \hat{\psi}_j)(u) = \int_{\mathcal{I}} \hat{K}(u, z) \hat{\psi}_j(z) dz = \hat{\lambda}_j^{(K)} \hat{\psi}_j(u) \quad (\text{D.4})$$

where $j = 1, 2, \dots$ and $\hat{\lambda}_1^{(K)} \geq \hat{\lambda}_2^{(K)} \geq \dots \geq 0$. The estimated dynamical space is then given by $\widehat{\mathcal{M}} = \text{span}(\widehat{\psi}_1(u), \widehat{\psi}_2(u), \dots, \widehat{\psi}_d(u))$, where the dimension of the dynamical space is estimated via a bootstrap test (as was also done in Bathia *et al.* (2010); Chen *et al.* (2022)).

For this, we do multiple tests of the form $H_0 : \lambda_{d_0+1}^{(K)} = 0$ with significance level α . The hypothesis is rejected when $\hat{\lambda}_{d_0+1}^{(K)} > c_\alpha$, where c_α is the boundary of the rejection region. In order to apply the bootstrap procedure we define $\widehat{Y}_t(u) = \sum_{j=1}^{d_0} \widehat{\chi}_{tj} \widehat{\psi}_j(u)$, where

$$\widehat{\chi}_{tj} := \int_{\mathcal{I}} [Y_t(u) - \bar{Y}(u)] \widehat{\psi}_j(u) du, \quad (\text{D.5})$$

and then define $\widehat{\varepsilon}_t(u) = Y_t(u) - \widehat{Y}_t(u)$. We then generate a bootstrap sample $Y_t^*(u) = \widehat{Y}_t(u) + \widehat{\varepsilon}_t^*(u)$, where $\widehat{\varepsilon}_t^*(u)$ is drawn with replacement from the set $\{\widehat{\varepsilon}_t(u)\}_{t=1}^n$. Based on this bootstrap sample, we compute the (d_0+1) -th largest eigenvalue $\widehat{\lambda}_{d_0+1}^{(K,*)}$ of the associated operator $\widehat{K}^*(u, v)$, similar to (D.1) and (D.4). We repeat this for B bootstrap samples and count how often $\widehat{\lambda}_{d_0+1}^{(K)} > \widehat{\lambda}_{d_0+1}^{(K,*)}$ occurs. If this is more than $(1 - \alpha)B$ times, we reject H_0 .

Estimation of \mathcal{M}_ε

Principal components $\widehat{\chi}_{tj}$ associated with the basis functions $\widehat{\psi}(\cdot)$ were defined in (D.5). The lagged covariance matrices of the principal components η_{tj} can now be estimated via

$$\widehat{\Sigma}_k^{(\eta)} = \frac{1}{n - k - 1} \sum_{t=1}^{n-k} \widehat{\chi}_t \widehat{\chi}_{t+k}^T, \quad \text{for } k=1, 2, \dots \quad (\text{D.6})$$

Note that for negative lags $k = -1, -2, \dots$ we have $\widehat{\Sigma}_k^{(\eta)} = \left(\widehat{\Sigma}_{-k}^{(\eta)}\right)^T$. By means of the Yule-Walker equations the estimator of the lag-0 covariance of the principal components η_{tj} is then

$$\widehat{\Sigma}_0^{(\eta)} = \frac{1}{2} \left[\widehat{\Sigma}_1^{(\eta)} \left(\widehat{\Sigma}_2^{(\eta)}\right)^{-1} \widehat{\Sigma}_1^{(\eta)} + \left(\dots\right)^T \right], \quad (\text{D.7})$$

where on the \dots there is a copy of the first term, such that $\widehat{\Sigma}_0^{(\eta)}$ is symmetric. Note that $\Sigma_0^{(\eta)}$ is necessarily symmetric, but due to finite-sample noise this is not the case for the estimator unless we explicitly symmetrize.

We are now in a position to estimate the covariance of the noise curves $\varepsilon_t(\cdot)$ through $\widehat{\Sigma}_\varepsilon(u, v) = \widehat{\Sigma}_Y(u, v) - \widehat{M}_0(u, v)$, where $\widehat{M}_0(u, v) = \widehat{\psi}(u)^T \widehat{\Sigma}_0^{(\eta)} \widehat{\psi}(v)$ and

$$\widehat{\Sigma}_Y(u, v) = \frac{1}{n-1} \sum_{t=1}^n (Y_t(u) - \bar{Y}(u)) (Y_t(v) - \bar{Y}(v)). \quad (\text{D.8})$$

In parallel to how the dynamical space \mathcal{M} was estimated, we can then estimate the noise space in terms of orthonormal eigenfunctions,

$$\widehat{\mathcal{M}}_\varepsilon = \text{span}(\widehat{\phi}_1^{(\varepsilon)}(\cdot), \dots, \widehat{\phi}_{d_\varepsilon}^{(\varepsilon)}(\cdot)), \quad \text{where} \quad \left(\widehat{\Sigma}_\varepsilon^+ \widehat{\phi}_j^{(\varepsilon)}\right)(u) = \widehat{\lambda}_j^{(\varepsilon)} \widehat{\phi}_j^{(\varepsilon)}(u) \quad (\text{D.9})$$

and $\hat{\lambda}_1^{(\varepsilon)} \geq \hat{\lambda}_2^{(\varepsilon)} \geq \dots \geq 0$. First of all, note that in the above expression we have replaced $\widehat{\Sigma}_\varepsilon(\cdot, \cdot)$ by $\widehat{\Sigma}_\varepsilon^+(\cdot, \cdot)$, which was defined in (3.2). See section 3.2 for a discussion about this replacement.

Estimating the denoising operator

It is now straightforward to find the estimated MISE-optimal denoised signal curves,

$$\widehat{X}_t^{\text{opt}}(u) = \widehat{Y}_t^{\parallel}(u) - \int_{\mathcal{I}} \left(\widehat{\phi}^{\parallel}(u) \right)^T \widehat{\Omega}_{\parallel\perp} \left(\widehat{\Omega}_{\perp} \right)^{-1} \widehat{\phi}^{\perp}(v) \left(Y_t(v) - \widehat{Y}_t^{\parallel}(v) \right) dv, \quad (\text{D.10})$$

where $\widehat{Y}_t^{\parallel}(u) = (\widehat{P}_{\mathcal{M}} Y_t)(u) = \int_{\mathcal{I}} \left[\widehat{\psi}(u) \cdot \widehat{\psi}(v) \right] Y_t(v) dv$,

$$\widehat{\Omega}_{\perp} := \int_{\mathcal{I}} \int_{\mathcal{I}} \widehat{\phi}^{\perp}(u) \left(\widehat{\phi}^{\perp}(v) \right)^T \widehat{\Sigma}_\varepsilon^+(u, v) du dv \quad (\text{D.11})$$

and

$$\widehat{\Omega}_{\parallel\perp} := \int_{\mathcal{I}} \int_{\mathcal{I}} \widehat{\phi}^{\parallel}(u) \left(\widehat{\phi}^{\perp}(v) \right)^T \widehat{\Sigma}_\varepsilon^+(u, v) du dv. \quad (\text{D.12})$$

E Dependence of the lag order for finding $\Sigma_0^{(\eta)}$

In Section 2.2 an estimation procedure for the autocovariance $\Sigma_\varepsilon(\cdot, \cdot)$ of the noise curves was derived by assuming that the principal components $\boldsymbol{\xi}_t$ of the signal curves $X_t(\cdot)$ follow a VAR(1) process. Here we investigate the impact of the assumption that the lag order is equal to 1. We first show that using a more general assumption of a VAR(p) process with $p > 1$ is possible and then analyze by numerical simulation the impact of choosing a larger lag order on MISE-optimal denoising.

If we assume that the principal components $\boldsymbol{\xi}_t$ follow a VAR(p) process (instead of the VAR(1) process of (2.9)), the loadings $\boldsymbol{\eta}_t$ also follow a VAR(p) process:

$$\boldsymbol{\eta}_t = \sum_{\ell=1}^p A_\ell^{(\eta)} \boldsymbol{\eta}_{t-\ell} + \mathbf{e}_t^{(\eta)}, \quad \text{where } \mathbf{e}_t^{(\eta)} \sim \text{IID}(\mathbf{0}, \Omega^{(\eta)}). \quad (\text{E.1})$$

The Yule-Walker equations are then given by

$$\Sigma_0^{(\eta)} = \sum_{\ell=1}^p A_\ell^{(\eta)} \Sigma_\ell^{(\eta)} + \Omega^{(\eta)}, \quad (\text{E.2a})$$

$$\Sigma_k^{(\eta)} = \sum_{\ell=1}^p \Sigma_{k-\ell}^{(\eta)} \left(A_\ell^{(\eta)} \right)^T, \quad \text{for } k > 0. \quad (\text{E.2b})$$

We would like to express $\Sigma_0^{(\eta)}$ in terms of autocovariance matrices $\Sigma_k^{(\eta)}$ of non-zero lag, as the latter can be estimated using the proxy principal components χ_{tj} . For this, we

need to eliminate the model parameter matrices $A_1^{(\eta)}, A_2^{(\eta)}, \dots, A_p^{(\eta)}$ from the Yule-Walker equations. This is alike the standard Yule-Walker estimate, except that here we cannot include $\Sigma_0^{(\eta)}$ in our estimator.

We consider the Yule-Walker equations (E.2b) for $k = p + 1, p + 2, \dots, p + p$, and write them as a system of linear equations for the model parameters:

$$\begin{pmatrix} \Sigma_{p+1}^{(\eta)} \\ \Sigma_{p+2}^{(\eta)} \\ \vdots \\ \Sigma_{p+p}^{(\eta)} \end{pmatrix} = \begin{pmatrix} \Sigma_p^{(\eta)} & \Sigma_{p-1}^{(\eta)} & \cdots & \Sigma_1^{(\eta)} \\ \Sigma_{p+1}^{(\eta)} & \Sigma_p^{(\eta)} & \cdots & \Sigma_2^{(\eta)} \\ \vdots & \vdots & & \vdots \\ \Sigma_{p+p-1}^{(\eta)} & \Sigma_{p+p-2}^{(\eta)} & \cdots & \Sigma_p^{(\eta)} \end{pmatrix} \begin{pmatrix} \left(A_1^{(\eta)} \right)^T \\ \left(A_2^{(\eta)} \right)^T \\ \vdots \\ \left(A_p^{(\eta)} \right)^T \end{pmatrix}. \quad (\text{E.3})$$

This system can be solved for the model parameters $A_1^{(\eta)}, A_2^{(\eta)}, \dots, A_p^{(\eta)}$ and subsequently we can use the Yule-Walker equation (E.2b) with $k = p$ to find

$$\Sigma_0^{(\eta)} = \left[\Sigma_p^{(\eta)} - \sum_{\ell=1}^{p-1} \Sigma_{p-\ell}^{(\eta)} \left(A_\ell^{(\eta)} \right)^T \right] \left(\left(A_p^{(\eta)} \right)^T \right)^{-1}. \quad (\text{E.4})$$

Hereby we have expressed $\Sigma_0^{(\eta)}$ in terms of quantities that can be estimated using the accessible principal components $\boldsymbol{\chi}_t$ defined in (2.7). This means that it is indeed possible to generalize the VAR(1) assumption to higher lag orders.

As an alternative to (E.4) we could have used the Yule-Walker equations (E.2b) for $k = 2, 3, \dots, p + 1$ to arrive at the (almost) standard Yule-Walker estimates for the model parameters, which are then expressed in terms of the unknown $\Sigma_0^{(\eta)}$. The Yule-Walker equation (E.2b) for $k = 1$ then leads to a highly non-linear equation for $\Sigma_0^{(\eta)}$, which could potentially be solved numerically. This has as an advantage that you need covariance matrices up to lag $p + 1$ instead of lag $2p$, but as disadvantage that there is no closed-form expression for $\Sigma_0^{(\eta)}$. We do not pursue this option further.

Using the right-hand side of (E.4) to estimate $\Sigma_0^{(\eta)}$, one may question its estimation properties. A full analysis is beyond the scope of this paper, but it is worthwhile to point out that the regular Yule-Walker estimator for the model parameters of a VAR(p) process has the same asymptotic properties as the least-squares estimator, although for small samples it sometimes performs worse (Lütkepohl, 2005; Tjøstheim and Paulsen, 1983). Furthermore, also for misspecified models the estimator is asymptotically optimal (Dahlhaus and Wefelmeyer, 1996). This gives reason to believe that (E.4) is a consistent estimator of $\Sigma_0^{(\eta)}$.

We investigate further the dependence of our denoising algorithm on the choice of the lag order p by re-doing the simulations of Section 4, but now assuming that the principal components $\boldsymbol{\xi}_t$ obey a VAR(p) process with $p = 1, 2, 3$. The case $p = 1$ is the default and analyzed in Section 4. The cases $p = 2$ and $p = 3$ are new and have as main novelty that (E.4) is used to estimate $\Sigma_0^{(\eta)}$. Apart from this, we use the same simulation setup as was used in Figure 2.

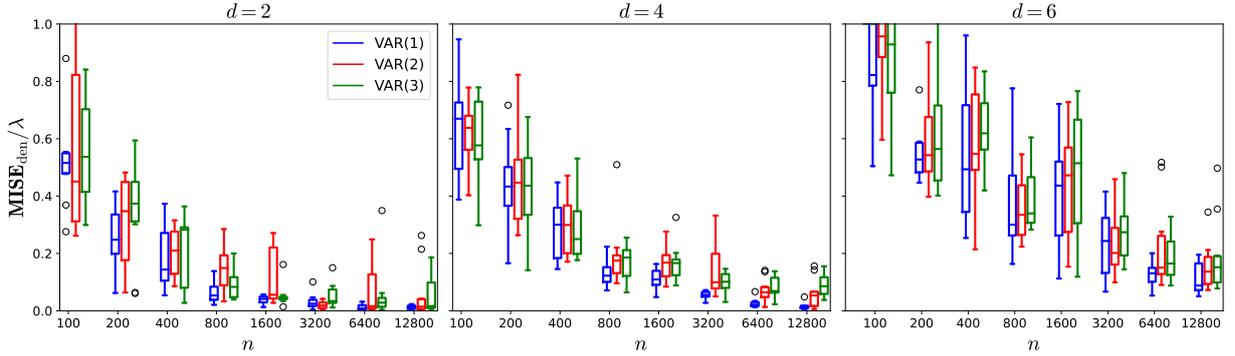


Figure 12: Proportion of the (integrated) variance of the remaining noise after MISE-optimal denoising as a function of the sample size n , for different lag orders ($p = 1, 2, 3$) of the assumed underlying $\text{VAR}(p)$ process of the principal components ξ_t .

The results are plotted in Figure 12. Although the denoising performance for $p = 1$ seems marginally better (in particular for small sample size n), the order of magnitude of the proportion of the (integrated) variance of the remaining noise and its asymptotic trend are similar for $p = 1, 2, 3$. This suggests that MISE-optimal denoising is insensitive to the choice of lag order p of the assumed $\text{VAR}(p)$ model for the principal components.

Of course, making p even larger will introduce more estimation noise, as the estimation of $\Sigma_0^{(\eta)}$ requires estimates of $\Sigma_k^{(\eta)}$ for $k = p + 1, \dots, p + p$, and autocovariance matrices with larger lags will suffer more from finite-sample estimation noise. This seems to be the reason that the denoising performance for $p = 1$ is marginally better. Furthermore, this is why we choose the $\text{VAR}(1)$ assumption by default.

F Irreducible components of the noise curves

In this section we compute a theoretical lower bound for MISE_{opt} , defined in (2.12), in the context of the simulation of section 4 in the case of $\theta_2 = 0$ (cf. (4.8)). This means that $\mathcal{M} \cap \mathcal{M}_\varepsilon \neq \emptyset$, i.e. the signal and noise curves have a principal component direction in common. Since MISE-optimal denoising cannot filter out this component, the noise has an irreducible component and there will be a non-zero lower bound for the MISE. Let's focus on the simulation setup of section 4, and in particular on the basis functions of \mathcal{M} and \mathcal{M}_ε . They were defined as follows:

$$\phi_j(u) = \cos(2\pi ju) + \sin(2\pi ju), \quad j = 1, \dots, d, \quad (\text{F.1})$$

$$\phi_k^{(\varepsilon)}(u) = [\cos \theta_k + \sin \theta_k] \cos(2\pi ku) + [\cos \theta_k - \sin \theta_k] \sin(2\pi ku), \quad j = 1, \dots, d_\varepsilon, \quad (\text{F.2})$$

For simplicity we assume $d < d_\varepsilon$, as is the case in all our simulations. If we decompose the noise modes as $\phi_k^{(\varepsilon)}(u) = \phi_k^\parallel(u) + \phi_k^\perp(u)$ with respect to \mathcal{M} , we find

$$\phi_k^\parallel(u) = \begin{cases} \cos(\theta_k)\phi_k(u), & k = 1, \dots, d, \\ 0 & k = d+1, \dots, d_\varepsilon, \end{cases} \quad (\text{F.3a})$$

$$\phi_k^\perp(u) = \begin{cases} \sin(\theta_k) [\cos(2\pi ku) - \sin(2\pi ku)], & k = 1, \dots, d, \\ \phi_k^{(\varepsilon)}(u) & k = d+1, \dots, d_\varepsilon. \end{cases} \quad (\text{F.3b})$$

Note first that if $\theta_j \neq \pi/2$ for $j = 1, \dots, d$, we have that $\mathcal{M}_\parallel = \mathcal{M}$ and we can use the functions $\phi_j(\cdot)$ as a basis for \mathcal{M}_\parallel . Let's assume this is the case here. Furthermore, let's for now assume that $\theta_j \neq 0$ for $j = 1, \dots, d$, which implies that $\dim(\mathcal{M}_\perp) = d_\varepsilon$ and as an orthonormal basis for \mathcal{M}_\perp we can use

$$\begin{cases} \cos(2\pi ku) - \sin(2\pi ku), & k = 1, \dots, d, \\ \phi_k^{(\varepsilon)}(u) & k = d+1, \dots, d_\varepsilon. \end{cases} \quad (\text{F.4})$$

Eventually we are interested in the case $\theta_2 = 0$, for which $\dim(\mathcal{M}_\perp) = d_\varepsilon - 1$ and the second basis function in the above basis is removed. But for now we assume $\theta_2 \neq 0$.

Using that

$$\Sigma_\varepsilon(u, v) = \mathbb{E}[\varepsilon_t(u)\varepsilon_t(v)] = g_\varepsilon^2(\lambda) \sum_{j=1}^{d_\varepsilon} \frac{1}{a^{2(j-1)}} \phi_j^{(\varepsilon)}(u)\phi_j^{(\varepsilon)}(v) \quad (\text{F.5})$$

we can find

$$(\Omega_\parallel)_{i,j} = \int_{\mathcal{I}} \int_{\mathcal{I}} \phi_i(u)\phi_j(v)\Sigma_\varepsilon(u, v) du dv = \delta_{i,j} \frac{g_\varepsilon^2(\lambda)}{a^{2(j-1)}} \cos^2(\theta_j) \quad (\text{F.6})$$

for $i, j = 1, \dots, d$, and similarly

$$(\Omega_\perp)_{i,j} = \delta_{i,j} \frac{g_\varepsilon^2(\lambda)}{a^{2(j-1)}} \times \begin{cases} \sin^2(\theta_j) & j = 1, \dots, d, \\ 1 & j = d+1, \dots, d_\varepsilon, \end{cases} \quad (\text{F.7})$$

for $i = 1, \dots, d_\varepsilon$, and

$$(\Omega_{\parallel\perp})_{i,j} = \delta_{i,j} \frac{g_\varepsilon^2(\lambda)}{a^{2(j-1)}} \cos(\theta_j) \sin(\theta_j), \quad (\text{F.8})$$

where in the last expression $i = 1, \dots, d$ and $j = 1, \dots, d_\varepsilon$.

The minimal value for the MISE is given in (2.17). Plugging in the above expressions we find that all terms in $\text{Tr}[\Omega_\parallel]$ are canceled by the second trace and $\mathbf{MISE}_{\text{opt}}^{\min} = 0$. In this case perfect denoising is possible (at the population level).

Let's now focus on the case where $\theta_j = 0$ for all $j \in \mathcal{J}_0 \subseteq \{1, \dots, d\}$. The term $\text{Tr}[\Omega_\parallel]$ in $\mathbf{MISE}_{\text{opt}}^{\min}$ is unaltered, but in the second trace the terms corresponding to the $j \in \mathcal{J}_0$ are missing. This leads to

$$\mathbf{MISE}_{\text{opt}}^{\min} = g_\varepsilon^2(\lambda) \sum_{j \in \mathcal{J}_0} \frac{1}{a^{2(j-1)}}. \quad (\text{F.9})$$

The case discussed in section 4.3 is when $\mathcal{J}_0 = \{2\}$.

Denoising by means of orthogonal projection onto the dynamical space \mathcal{M} also leads to an irreducible noise component, provided the noise space is not entirely inside the orthogonal complement of the dynamical space. Assuming the general simulation setup of section 4 this lower bound can be computed analogously to how (F.9) was computed and is given by

$$\mathbf{MISE}_{\text{ortho}}^{\min} = \text{Tr} [\Omega_{\parallel}] = g_{\varepsilon}^2(\lambda) \sum_{j \in \mathcal{J}_0} \frac{\cos^2(\theta_j)}{a^{2(j-1)}}. \quad (\text{F.10})$$