NONLINEAR WAVELET THRESHOLD ESTIMATION OF TIME-VARYING COVARIANCE MATRICES IN A LOG-EUCLIDEAN MANIFOLD

Gabriel Bailly, Rainer von Sachs

LIDAM Discussion Paper ISBA
2024 / 04



UCLouvain

ISBA | LIDAM
Louvain Institute of Data Analysis and Modeling in economics and statistics

ISBA

Voie du Roman Pays 20 - L1.04.01

B-1348 Louvain-la-Neuve

Email : lidam-library@uclouvain.be

https://uclouvain.be/en/research-institutes/lidam/isba/publication.html

Nonlinear wavelet threshold estimation of time-varying covariance matrices in a log-Euclidean manifold

Gabriel Bailly Rainer von Sachs

Institute of Statistics, Biostatistics and Actuarial Sciences (ISBA), LIDAM
UCLouvain

December 13, 2024

Abstract

We tackle the problem of estimating time-varying covariance matrices (TVCM; i.e. covariance matrices with entries being time-dependent curves) whose elements show inhomogeneous smoothness over time (e.g. pronounced local peaks). To address this challenge, wavelet denoising estimators are particularly appropriate. Specifically, we model TVCM using a signal-noise model within the Riemannian manifold of symmetric positive definite matrices (endowed with the log-Euclidean metric) and use the intrinsic wavelet transform, designed for curves in Riemannian manifolds. Within this non-Euclidean framework, the proposed estimators preserve positive definiteness. Although linear wavelet estimators for smooth TVCM achieve good results in various scenarios, they are less suitable if the underlying curve features singularities. Consequently, our estimator is designed around a nonlinear thresholding scheme, tailored to the characteristics of the noise in covariance matrix regression models. The effectiveness of this novel nonlinear scheme is assessed by deriving mean-squared error consistency and by numerical simulations, and its practical application is demonstrated on TVCM of electroencephalography (EEG) data showing abrupt transients over time.

Keywords: Nonlinear wavelet thresholding, non-Euclidean geometry, sample covariance matrices, time-varying second-order structure, log-Wishart distribution.

1 Introduction

In this paper, we tackle the problem of estimating the time-varying covariance matrix (TVCM) of a multivariate time series, which features sharp local peaks or general inhomogeneous smoothness. We consider that the theoretical variance-covariance matrix of the multivariate time series is evolving as a function of time and can be modelled by a continuous curve with values in the set $\mathrm{Sym}^+(d)$ of symmetric positive definite (SPD) $d \times d$ matrices. Then, we can estimate the covariance matrix locally in time by calculating the sample covariance matrix over short disjoint windows. Obviously, in order to control the variability of this estimator, we still need to care for an appropriate smoothing method to reduce the noise of this preliminary sequence of covariance estimators over time - without smoothing out important time-localised features in one (or various) element(s) of the TVCM.

In a univariate situation of smoothing variance curves, methodology accompanied by asymptotic results on mean-squared error convergence has been derived in a nonparametric regression framework (such as in [2], using linear wavelet smoothing). However, in a multivariate context, to maintain positive-definiteness of our estimators at each and every point in time, and to prevent other undesirable effects such as "swelling" [5], one should not use classical regression methods because the considered curves do not live in a Euclidean vector space – we rather need to consider $\mathrm{Sym}^+(d)$ as a Riemannian manifold by equipping it with an appropriate metric. This non-Euclidean property of the space of interest motivates an intrinsic regression methodology, exploiting tools from Riemannian geometry.

Extensions of nonparametric regression methods to non-Euclidean data have been investigated up to now by only a few works: [7] with intrinsic cubic splines, [28] with intrinsic local polynomials, [24, 11] with intrinsic wavelets. In particular, the latter (which primarily focused on spectral density estimation of multivariate data) had been developed in order to estimate manifold-valued curves with inhomogeneous smoothness.

Most of the mentioned literature on smoothing or denoising trajectories of covariance matrices (over time or space) deals with *linear* smoothing approaches. In our work, however, we concentrate on *nonlinear* wavelet thresholding estimators of the underlying SPD matrix-valued curve. This allows us to benefit from all advantageous properties of denoising schemes known from univariate first-generation wavelet estimation, i.e. adapting to possibly inhomogeneous spatial smoothness of the underlying curve and allowing for fast implementations. As our object of interest is SPD, the uttermost paradigm of our approach is that at every point in time the estimators remain SPD. Moreover, we wish to keep some equivariance properties being suited to the multivariate situation, e.g. a permutation in the ordering of the time series dimensions should entail the same permutation in the estimated covariance matrix (compare also the discussion in [11]). To do so, we choose to work in the *log-Euclidean* manifold, by application of the matrix logarithm to empirical and population TVCMs and subsequent development of finite-sample and asymptotic properties of this approach.

Compared to the aforementioned literature, the originality of our paper is two-fold.

- On the one hand, we deliver a complete modelling and estimation approach for denoising *empirical* TVCMs of weakly dependent time series. As a motivation, we recall that in the univariate situation of Gaussian independent observations it suffices to take the logarithm of an empirical variance estimator to turn a multiplicative χ^2 noise model into a homoscedastic additive signal-plus-noise "log χ^2" model. However, developing an appropriate similar set-up for the general d-variate situation of serially dependent data requires to solve problems inherent to, among others, the non-commutative nature of products of general symmetric matrices (as the matrix logarithm of the product of two symmetric matrices is not equal to the sum of their logarithms). Additionally, proving the asymptotic properties of our approach required a lot of non-trivial matrix calculus (especially Taylor expansion of matrix-valued functions), which is more difficult than for scalar and Euclidean cases.

- As a second main contribution, with our paper we go beyond existing preliminary *nonlinear* thresholding results by suggesting appropriate threshold rules *and* the choice of their threshold values which are both more flexible than the existing ones of [11] and as such directly taylored to the non-trivial problem of denoising matrix-valued curves. We propose a series of original theoretical results, in particular a new mean-squared error consistency result for general nonlinear wavelet estimation. This includes showing the usefulness of an explicitly derived correction of the bias of our estimator caused by the geometric nature of our log-Euclidean approach. Empirical investigations confirm our theoretical results, via extensive simulation studies and an application of our method to transient EEG signals, which reveals new insights about linear associations between signals at different EEG channels and their evolution over time.

Estimation of covariance matrices which evolve over time (or space) arises in a variety of applied fields such as computer vision [9], medical imaging [17, 18], signal processing [4] and neuroscience [19]. Perhaps the most prominent example (with $d = 3$) are diffusion tensors [29, 16] which describe the shape of diffusion localised to a voxel (human tissue, e.g. brain or heart) by modelling the covariance of the Brownian motion of water at this voxel. In this work we address the analysis of observed multivariate EEG seizure time series, through their TVCM: this is a different approach than [22] which used an estimate of the time-varying spectral density matrix.

Our paper is structured as follows: in the remaining of this introductory section we recall from the literature notation, achievements and results that we need to familiarize the reader with the topic of this

research. In particular we recall in Subsection 1.2.3 existing results on *linear* wavelet estimation of SPD matrix-valued curves – for which appropriate signal-plus-noise models had to be established (see Equation (12)) – and that we want to improve by further developing *nonlinear* thresholding. In particular, we aim at rendering it applicable to adaptive smoothing of empirical TVCMs: to this end, Section 2 introduces local covariance estimation where we present the model and the necessary assumptions on our data generating process, as well as a "pilot" estimator used as input in the first step of our subsequently developed wavelet denoising scheme. Section 3 then is the core section of our paper where we treat genuinely nonlinear wavelet regression for denoising SPD matrices: the first part of this section is devoted to the discussion of a set of newly proposed thresholding rules which adapt to the matrix-nature of our wavelet coefficients, and in the second part we show a new result on integrated mean-squared error (IMSE) consistency of our nonlinear regression scheme for general SPD matrices. We then specify the application to sample covariance matrices, benefiting from the preparatory achievements of the previous section. The performance of our new thresholding rules is then investigated in the next Section 4 first by some simulation studies where we i) generate more generically regression data from a given signal-plus-noise model; but also ii) consider smoothing of empirical time-varying covariance matrices, by simulating both serially independent data as well as time series data from a vector-autoregressive (VAR) model. Finally, in the last Subsection 4.3, we will show the performance of our method on multivariate time series EEG data collected from patients suffering from epilepsy: these data show rather abrupt changes over time – at the onset of the seizure – and it will be interesting to see how these potentially differ from one element of the covariance matrix to another. A discussion section summarizes our achievements and outlines future research whereas a short appendix section presents the central tool of the Baker-Campbell-Hausdorff formula for non-commutative matrices. Finally, the following parts are relegated to a substantial series of Supplement Sections which presents in particular all proofs of our results and some further numerical simulation results.

1.1 Preliminaries

Notation. We denote the set of $d \times d$ symmetric matrices by $\mathrm{Sym}(d)$, the set of $d \times d$ symmetric positive definite matrices by $\mathrm{Sym}^+(d)$, the set of $d \times d$ orthogonal matrices by $\mathbb{O}(d)$. The matrix $A = \mathrm{diag}(a_i)_{i=1}^d$ is a diagonal square matrix with elements $A_{ii} = a_i$. The operation $A * B := ABA^\top$ is the matrix congruence transformation between A and B.

Recall the eigenvalue decomposition (EVD) of a matrix $A \in \mathrm{Sym}^+(d)$, i.e. there exists some $U \in \mathbb{O}(d)$ and $\mathrm{diag}(\lambda_i)_{i=1}^d =: \Lambda$ such that $A = U\Lambda U^\top$. The EVD allows functional calculus for SPD matrices, in the sense that we can extend some univariate map $\varphi : \mathbb{R}^+ \to \mathbb{R}$ to its matrix counterpart, by the concept of $\mathbb{O}(d)$-equivariant univariate maps [26]: i.e. maps $\varphi : \mathrm{Sym}^+(d) \to \mathrm{Sym}(d)$ such that

$$\varphi(U\Lambda U^\top) = U \ \mathrm{diag}(\varphi(\lambda_i))_{i=1}^d \ U^\top \tag{1}$$

for all $U \in \mathbb{O}(d)$ and Λ with positive diagonal entries, and where the inverse φ^{-1} exists. In the remaining, the distinction between the univariate map and the $\mathbb{O}(d)$-equivariant univariate map is often ignored and clear from the context. For example, the principal matrix square root and principal matrix logarithm are defined by setting $\varphi(\cdot) = (\cdot)^{1/2}$ and $\varphi(\cdot) = \log(\cdot)$. The inverse of the latter is $\exp : \mathrm{Sym}(d) \to \mathrm{Sym}^+(d) : A \mapsto \sum_{k=0}^\infty (k!)^{-1} A^k$, the matrix exponential. Both $\exp$ and $\log$ are differentiable: we will use the directional Fréchet derivatives of the matrix exponential and matrix logarithm, which are well defined and denoted by $\mathbb{D}_A\varphi(E)$, with $A, E \in \mathrm{Sym}(d)$ and $A = U \ \mathrm{diag}(\lambda_i)_{i=1}^d \ U^\top$, and can be expressed in closed-form:

$$\mathbb{D}_A\varphi(E) := U(Z \odot (U^\top E U))U^\top \ ,$$

where $\odot$ is the Hadamard product and where Z is the matrix of first divided differences, i.e.

$$Z_{ij} = \begin{cases} \dfrac{\varphi(\lambda_i) - \varphi(\lambda_j)}{\lambda_i - \lambda_j} & \text{if } \lambda_i \neq \lambda_j \\ \varphi'(\lambda_i) & \text{otherwise.} \end{cases}$$

Log-Euclidean manifold. Consider $\mathrm{Sym}^+(d)$ as a differentiable manifold embedded in $\mathrm{Sym}(d)$, which can be equipped with a Riemannian metric g – i.e. a family of inner products for tangent spaces $T_S\mathrm{Sym}^+(d)$ defined at each point S in $\mathrm{Sym}^+(d)$ and varying smoothly in S. The pair $(\mathrm{Sym}^+(d), g)$ is called a Riemannian manifold: here, we consider the log-Euclidean Riemannian metric of [5] defined by $g = \{g_S\}_{S\in\mathrm{Sym}^+(d)}$ where $g_S(U,V) := \langle \mathbb{D}_S \log(U), \mathbb{D}_S \log(V)\rangle_F$ for $U, V \in T_S\mathrm{Sym}^+(d)$, inducing the norm $\|U\|_S = \|\mathbb{D}_S \log(U)\|_F$. This metric also induces the simple *log-Euclidean distance* between points in the SPD manifold:

$$\delta_{\mathrm{L}}(S_1, S_2) = \|\log(S_2) - \log(S_1)\|_F \ , \tag{2}$$

for $S_1, S_2 \in \mathrm{Sym}^+(d)$ – note that this is the length of the shortest curve between the two points, i.e. the *geodesic curve* parameterized by

$$\gamma(t; S_1, S_2) = \exp((1-t)\log(S_1) + t\log(S_2)). \tag{3}$$

We can generalize the concept of addition for points in the manifold by defining the *exponential map* $\mathrm{Exp}_S : T_S\mathrm{Sym}^+(d) \to \mathrm{Sym}^+(d)$ mapping matrices U in the tangent space of S_1 to a point S_2 in $\mathrm{Sym}^+(d)$:

$$\mathrm{Exp}_{S_1}(U) = \exp\big(\log(S_1) + \mathbb{D}_{S_2}\log(U)\big) = S_2 \ .$$

To generalize the concept of subtraction we define its inverse, the *logarithmic map* $\mathrm{Log}_S : \mathrm{Sym}^+(d) \to T_S\mathrm{Sym}^+(d)$ which is unique (because the sectional curvature of $(\mathrm{Sym}^+(d), g)$ is nonpositive) and defined by

$$\mathrm{Log}_{S_1}(S_2) = \mathbb{D}_{\log(S_1)} \exp\big(\log(S_2) - \log(S_1)\big) = U \ .$$

Furthermore, another important tool is the *parallel transport* which maps vectors in tangent spaces to vectors in other tangent spaces, while preserving its intrinsic norm: $\Gamma_{S_1}^{S_2} : T_{S_1}\mathrm{Sym}^+(d) \to T_{S_2}\mathrm{Sym}^+(d)$ such that for $U \in T_{S_1}\mathrm{Sym}^+(d)$, $\|U\|_{S_1} = \|\Gamma_{S_1}^{S_2}(U)\|_{S_2}$. It is defined by

$$\Gamma_{S_1}^{S_2}(U) = \mathbb{D}_{\log(S_2)} \exp\big(\mathbb{D}_{S_1}\log(U)\big) \ . \tag{4}$$

Finally, we consider manifold-valued random variables: as $(\mathrm{Sym}^+(d), \delta_{\mathrm{L}})$ is a complete separable metric space, then $(\mathrm{Sym}^+(d), \mathcal{B}(\mathrm{Sym}^+(d)))$ is a measurable space with $\mathcal{B}(\mathrm{Sym}^+(d)))$ its Borel σ-algebra. Given a probability space $(\Omega, \mathcal{A}, \mathbb{P})$, a random element in $\mathrm{Sym}^+(d)$ is a measurable function $X : (\Omega, \mathcal{A}) \to (\mathrm{Sym}^+(d), \mathcal{B}(\mathrm{Sym}^+(d)))$ with induced probability measure (or distribution) $\nu := \mathbb{P} \circ X^{-1}$, which we denote by $X \sim \nu$. We denote by $\mathcal{P}(\mathrm{Sym}^+(d))$ the set of probability measures on $(\mathrm{Sym}^+(d), \mathcal{B}(\mathrm{Sym}^+(d)))$, and by

$$\mathcal{P}_p(\mathrm{Sym}^+(d)) := \left\{ \nu \in \mathcal{P}(\mathrm{Sym}^+(d)) \;\middle|\; \int_{\mathrm{Sym}^+(d)} \delta_{\mathrm{L}}^p(R, S)\, \mathrm{d}\nu(S) < \infty \right\} \ ,$$

with arbitrary $R \in \mathrm{Sym}^+(d)$, the subset of probability measures having finite p-th moment with respect to the log-Euclidean distance. For $X \sim \nu$, we define the Fréchet mean of X by

$$\mathbb{E}_\nu[X] = \mu := \arg\min_{S_0 \in \mathrm{Sym}^+(d)} \int_{\mathrm{Sym}^+(d)} \delta_{\mathrm{L}}^2(S_0, S)\, \mathrm{d}\nu(S) \ ,$$

which exists whenever $\nu \in \mathcal{P}_2(\mathrm{Sym}^+(d))$ and is unique because $(\mathrm{Sym}^+(d), g)$ has nonpositive sectional curvature. Furthermore, under the log-Euclidean metric, it is characterized by $\mathbb{E}[X] = \exp(\mathbf{E}[\log(X)])$, where $\mathbf{E}[\cdot]$ is the ordinary Euclidean expectation in the vector space $\mathrm{Sym}(d)$.

1.2 Wavelet transform of curves in the log-Euclidean manifold

In order to give the reader a better understanding of how wavelet transforms of curves in Riemannian manifolds can be constructed via *average-interpolation* schemes [11], we first briefly recall classical wavelet smoothing.

1.2.1 Recalling classical wavelet denoising

We begin by (very schematically) recalling the first-generation (linear) wavelet estimator in a (demonstratively simple) nonparametric signal-plus-noise regression model, where we observe $n = 2^J$ noisy observations Y_k from an equidistantly sampled square-integrable curve $f(t)$, $t \in [0,1]$, in the presence of (i.i.d.) homoscedastic (mean zero) noise ξ_k:

$$Y_k = f(k/n) + \xi_k, \quad k = 0, \dots, n-1 \, . \tag{5}$$

Projecting each of the ingredients of this additive model on the scaling functions $\{\Phi_{j,k}\}$ of an $L^2([0,1])$ orthonormal wavelet basis furnishes the corresponding additive i.i.d. model in the coefficient domain

$$M_{j,k} = M^f_{j,k} + M^\xi_{jk}, \quad k = 0, \dots, 2^j - 1, \; j = 0, \dots, J \, , \tag{6}$$

where $M_{j,k}$ and $M^f_{j,k} = \langle f, \Phi_{j,k} \rangle_{L^2}$ denote the empirical and the true scaling coefficients (of the curve f), respectively, on resolution scales $j = 0, \dots, J$, and locations $k = 0, \dots 2^j - 1$. Then a "linearly thresholded" MSE-consistent curve estimator of f is given by

$$\widehat{f}_{J_0}(t) = \sum_k M_{J_0,k} \Phi_{J_0,k}(t) \, , \tag{7}$$

where the smoothing scale J_0 is chosen to be smaller ("coarser") than the sampling scale J, akin the choice of the bandwidth $b \sim 2^{-J_0}$ of the associated (wavelet) kernel estimator to fulfil $b \to 0$ and $nb \to \infty$ with $n \to \infty$.

However, the full potential of wavelet denoising shows only for *nonlinearly* thresholding the empirical wavelet coefficients D_{jk} (see (15)) associated to the set of $\{M_{jk}\}_{jk}$, i.e. setting (and/or shrinking) those D_{jk} to zero that, in absolute value, do not surpass an appropriately chosen threshold (e.g. the famous "universal threshold" being proportional to the standard deviation of $\{D_{Jk}\}_{Jk}$, times a factor $\sqrt{2\log(n)}$ which protects against large deviations in its (asymptotically) Gaussian distribution). These truly "multiscale" estimators are able to denoise signals with locally inhomogeneous smoothness provided the data follow some version of the above-mentioned signal-to-noise model (6) (whereas the "monoscale" estimators (7) cannot).

The following construction of second-generation wavelet schemes, which replaces the above, in Euclidean set-ups more direct, construction for scalar-valued functions, we will use *average-interpolation* [13]: note that although *no* (scaling or wavelet) basis functions $\{\Phi_{j,k}\}$ do exist in the considered non-Euclidean geometry, we will benefit, by equation (12), from a signal-to-noise model similar to (6) which holds with the same properties as for first-generation wavelet estimation of (scalar-valued) curves. In particular, as it is common practice, the scaling coefficients $M_{J,k}$ are identified with the observations Y_k, before a dyadic fast pyramid algorithm (of computational order $O(n)$) is used to obtain coarser scaling coefficients $M_{j,k}$ essentially via successive weighted averaging with filters of finite length.

1.2.2 Average-Interpolation (AI) schemes for SPD matrices in the log-Euclidean manifold

Let now $f : [0,1] \to \mathrm{Sym}^+(d)$ be a square-integrable function from (rescaled) time $u \in [0,1]$ into the space of SPD matrices such that $\int_0^1 \|f(u)\|_F^2 \, \mathrm{d}u < \infty$, observed through dyadic sampling in the presence

of noise: $(f(u_k))_{k=0}^{n-1}$ with $u_k = (k+1)/n$, $n = 2^J$, $J \in \mathbb{N}$. Again, the sampled points are identified with the finest scale *scaling coefficients* of the intrinsic average-interpolation wavelet transform, $(M_{J,k})_{k=0}^{2^J-1}$, now being a sequence (in time) of SPD matrices. This particular wavelet transform, derived by [11] for geodesically complete Riemannian manifolds (for which exponential maps, logarithmic maps and parallel transport are defined) as an extension of [13], is described in the following. As for first-generation wavelets, the *order* $N \in \{1, 3, 5, ...\}$ of the chosen set of filter coefficients corresponds to number of vanishing moments of the associated wavelet kernel (similarly to the those of classical smoothing with convolution kernels of a given order).

The following ingredients make up the AI wavelet scheme (for more details we also refer to [23]).

Midpoint pyramid. The *scaling coefficients* of the wavelet transform are identified by constructing a midpoint pyramid from the finest scale coefficients. Denote the midpoint pyramid by $\{(M_{j,k})_{j,k} \mid j = 0, \dots, J\,; k = 0, \dots, 2^j - 1\}$ where

$$M_{j,k} = \gamma(1/2; M_{j+1,2k}, M_{j+1,2k+1}) \,, \tag{8}$$

with γ defined in (3).

Forward wavelet transform. Consider the j-scale midpoints $(M_{j,k})_k$: the following procedure outputs the j-scale *wavelet coefficients* $(D_{j,k})_k$:

- PREDICTION STEP. Compute the $(j-1)$-scale midpoints $(M_{j-1,k})_k$. Construct for each $k = 0, \dots, 2^j - 1$ the unique *intrinsic polynomial* $\rho : [0,1] \to \mathrm{Sym}^+(d)$ of order $N-1$ (where $N = 2M+1$ for some $M \geq 0$) interpolating the midpoints $M_{j-1,k-L}, \dots, M_{j-1,k+L}$, and average it to obtain the *predict midpoints*: $\widetilde{M}_{j,2k}$ when averaging over the interval $[(2k)/2^j, (2k+1)/2^j)$ and $\widetilde{M}_{j,2k+1}$ when averaging over the interval $[(2k+1)/2^j, (2k+2)/2^j)$. The detailed procedure is found in [11], but in practice this resumes to using weighted intrinsic averages with unique coefficients $(C_{N,i})_{i=0}^{2N}$:

$$\widetilde{M}_{j,2k} = \exp\left(\sum_{m=-M}^{M} C_{N,2m+N-1} \log(M_{j-1,k+m})\right), \; \widetilde{M}_{j,2k+1} = \exp\left(\sum_{m=-M}^{M} C_{N,2m+N} \log(M_{j-1,k-m})\right) . \tag{9}$$

- DIFFERENCING STEP. Define the wavelet coefficients by the intrinsic difference between the actual and predicted midpoints,

$$D_{j,k} = 2^{-j/2} \mathrm{Log}_{\widetilde{M}_{j,2k+1}}(M_{j,2k+1}) \in T_{\widetilde{M}_{j,2k+1}} \mathrm{Sym}^+(d) \,. \tag{10}$$

  Those coefficients live in different tangent spaces: subsequently, to be able to compare them, they should be parallel transported to the tangent space at the identity matrix: the resulting coefficients are *whitened wavelet coefficients*, defined by

$$\mathfrak{D}_{j,k} = \Gamma^{I_d}_{\widetilde{M}_{j,2k+1}}(D_{j,k}) = 2^{-j/2}(\log M_{j,2k+1} - \log \widetilde{M}_{j,2k+1}) \in T_{I_d}\mathrm{Sym}^+(d) = \mathrm{Sym}(d) \,. \tag{11}$$

  where $\|D_{j,k}\|_{\widetilde{M}_{j,2k+1}} = \|\mathfrak{D}_{j,k}\|_F$.

Backward wavelet transform. Consider the $(j-1)$-scale midpoints $(M_{j-1,k})_k$ and j-scale wavelet coefficients $(D_{j,k})_k$: the following procedure outputs the j-scale midpoints $(M_{j,k})_k$.

- PREDICTION STEP. Given the order N, generate for each $k = 0, \dots, 2^{j-1} - 1$ the *odd* predicted midpoints $(\widetilde{M}_{j,2k+1})_k$ based on $(M_{j-1,k})_k$. Then, compute the odd j-scale midpoints $(M_{j,2k+1})_k$ as the intrinsic addition between the predicted midpoints and the wavelet coefficients:

$$M_{j,2k+1} = \mathrm{Exp}_{\widetilde{M}_{j,2k+1}}(2^{j/2} D_{j,k}) \,,$$

which is equivalent to

$$M_{j,2k+1} = \exp\left(\log(\widetilde{M}_{j,2k+1}) + 2^{j/2}\mathfrak{D}_{j,k}\right).$$

- COMPLETING STEP. Compute the *even* j-scale midpoints by reversing the geodesic midpoint relation for $k = 0, \ldots, 2^{j-1} - 1$:

$$M_{j,2k} = \exp\left(2\log(M_{j-1,k}) - \log(M_{j,2k+1})\right).$$

Hence, from the coarsest scaling coefficient $M_{0,0}$ and the wavelet coefficient pyramid $(D_{j,k})_{j,k}$, we can reconstruct the sampled points $(M_{J,k})_k$ by iterating the backward transform up to the finest scale J.

1.2.3 Nonparametric wavelet regression for curves in the log-Euclidean manifold

In the framework of sampled curves corrupted by noise, a nonparametric regression framework is considered. However, we cannot use the standard Euclidean additive signal-to-noise models as in (5) because $\mathrm{Sym}^+(d)$ endowed with the Euclidean metric is not a complete metric space. Instead, as we consider the log-Euclidean manifold which satisfies the properties of a complete separable metric space, we will instead work with an intrinsic signal plus noise model as defined in [23], by replacing the addition with the exponential map $\mathrm{Exp}_\star(\cdot)$, and parallel transporting the noise (which is assumed to live in $T_{I_d}(\mathrm{Sym}^+(d)) = \mathrm{Sym}(d)$) to the right tangent space using the operation $\Gamma^\star_{I_d}(\cdot)$.

Definition 1.1 (Log-Euclidean signal plus noise model)**.** *Suppose $f : [0,1] \to \mathrm{Sym}^+(d)$ is an unknown target signal with $\sup_u \|f(u)\|_F < \infty$, $S_k = f(u_k)$, $u_k = \frac{k+1}{n}$, $k = 0, \ldots, n-1$, and $\xi_k \sim \widetilde{\nu}$ are i.i.d. noise, i.e. mean-zero symmetric random matrices with finite second moment, i.e. $\widetilde{\nu} \in \mathcal{P}_2(\mathrm{Sym}(d))$. The intrinsic signal plus noise model of $(\mathrm{Sym}^+(d), \delta_{\mathrm{L}})$ is defined by*

$$\widetilde{S}_k = \mathrm{Exp}_{S_k}(\Gamma^{S_k}_{I_d}(\xi_k)) = \exp\left(\log(S_k) + \xi_k\right), \quad k = 0, \ldots, n-1. \tag{12}$$

This intrinsic nonparametric regression model is then an additive model in $\mathrm{Sym}(d)$ (or additive model in the log-domain), and it is brought back to $\mathrm{Sym}^+(d)$ with the matrix exponential. This model can then be viewed as a matrix multiplicative model.

Having defined the appropriate model for data $\widetilde{S}_k$ which are noisy versions of the target function $S_k = f(u_k)$, the scheme for estimating the latter as a function over $u \in [0,1]$ via wavelets constructed in $\mathrm{Sym}^+(d)$ is the following.

Remark 1 (Linear wavelet estimator). Note that by choosing $g(\cdot) \equiv 0$ in (13), we include *linear* thresholding, which amounts to a simple linear smoother, constructed from the midpoints $(M_{J_0-1,k})_k$ and lifted back on the sample scale J by successive averaging via (9) In this work, we will give the asymptotic properties of this specific linear estimator (see Theorem 2.4) only for comparisons. Our emphasis will be on the non-linear threshold estimator for which we will present new results in Section 3.

Remark 2 (Non-linear thresholding). A common *non-linear* thresholding scheme that we will adopt in this paper is to choose a keep-or-kill rule of the form

$$g(\widehat{D}_{j,k}) = \widehat{D}_{j,k}\mathbb{1}\left\{h(\widehat{\mathfrak{D}}_{j,k}) > \varepsilon\right\} \tag{15}$$

for some $\varepsilon > 0$ and an appropriate functional $h : \mathrm{Sym}(d) \to [0,\infty)$ to be, importantly, applied onto the empirical *whitened* wavelet coefficients $\widehat{\mathfrak{D}}_{j,k}$ (cf. equation (11)). A particular interesting thresholding scheme (applied and developed in Section 3) is to further split up the thresholded range of scales $J_0, \ldots, J$ into two parts: one below J_1 (with $J_0 < J_1 < J$), where the aforementioned keep-or-kill rule is applied, and one starting from J_1 where all coefficients are systematically set to zero. Section 3 will elaborate on this in all details on some choices, including our new original propositions here.

Algorithm 1: The estimation scheme

1. Set $M_{J,k} = \widetilde{S}_k$ for $k = 0, \dots, n-1$ and let $(M_{j,k})_{j,k}$ denote the *midpoint pyramid* according to (8) .

2. Choose a refinement order $N = 2M + 1$, $M \geq 0$, and apply the *forward wavelet transform* to obtain predicted midpoints $(\widetilde{M}_{j,k})_{j,k}$ (as in (9)) as well as empirical wavelet coefficients $(D_{j,k})_{j,k}$ (from (10)).

3. Choose a *thresholding scale* $J_0 < J$ and define thresholded empirical wavelet coefficients $\widehat{D}_{j,k}$ according to

$$\widehat{D}_{j,k}(J_0) := \begin{cases} D_{j,k} & , \ j = 1, \dots, J_0 - 1 \\ g(D_{j,k}) & , \ j = J_0, \dots, J \end{cases} \quad \text{for } k = 0, \dots, 2^j - 1 \ , \tag{13}$$

   where $g \colon \mathrm{Sym}(d) \to \mathrm{Sym}(d)$ is an appropriate thresholding function (shrinking to zero, or setting to zero its argument, as a function of a threshold to be determined from the data).

4. Given the empirical midpoint pyramid $(M_{j,k})_{j,k}$ and the thresholded empirical wavelet coefficients $(\widehat{D}_{j,k})_{j,k}$, recursively apply the *backward wavelet transform* to obtain estimated midpoints $(\widehat{M}_{j,k}(J_0))_{j,k}$. At the sampling scale J, utilize the *wavelet estimates* $(\widehat{M}_{J,k}(J_0))_k$ to construct the curve estimator

$$\widehat{f}_n(u) = \sum_{k=0}^{n-1} \widehat{M}_{J,k}(J_0) \, \mathbb{1}_{2^{-J}[k,k+1)}(u), \quad u \in [0,1]. \tag{14}$$

2 Time-varying covariance matrix modelling and estimation

In this section we introduce our particular set-up of time-varying covariance matrices to which we will apply our subsequent wavelet thresholding scheme. We recall that both considered population and sample covariance matrices need to be assumed non-degenerate, i.e. with strictly positive determinant, hence SPD (otherwise, the considered geometry of intrinsic estimation would not be the same as employed in this work). For this, we start from a data sample with T observations that can be serially independent or dependent. The relation between the sample size T of the time series and the sample size n of our considered wavelet scheme is determined by the construction of n locally pre-smoothed (or pilot) covariance estimators from the T observations of the time series.

In this work, we consider $(X_t)_{t=1}^T$ to be a sample from a $\mathbb{R}^d$-variate non-stationary process, with $\mathbf{E}[X_t] = 0$ for all t and with $\mathrm{Var}[X_t] = \Sigma_t := \Sigma(t/T)$ where $\Sigma : [0,1] \to \mathrm{Sym}^+(d)$ is a matrix-valued curve. We can write, for $t = 1, \dots, T$,

$$X_t = \Sigma_t^{1/2} \varepsilon_t \ . \tag{16}$$

Assumption 1. *The process $(\varepsilon_t)_t$ is an $\mathbb{R}^d$-valued stationary process with mean $\mathbf{E}[\varepsilon_t] = 0$ and variance-covariance matrix $\mathrm{Var}[\varepsilon_t] = I_d$. We assume existence of all moments of the components $\varepsilon_{t,j}$, $j = 1, \dots, d$. For the serial dependence structure of $(\varepsilon_t)_t$ we assume that*

$$\sum_{t_1, \dots, t_{k-1}}^{\infty} |(\varepsilon_{t_1,a_1}, \dots, \varepsilon_{t_{k-1},a_{k-1}}, \varepsilon_{0,a_k})| < \infty \tag{17}$$

*for all $a_1, \dots, a_k = 1, \dots, d$ and all natural $k \geq 2$ [**?**]with mixed cumulants defined as in¿[Equation (2.3.1)]Brillinger2001. This is a common convenient assumption in time series analysis to control the*

serial dependence, as to be found, e.g., in [8, Assumption 2.6.1].

Here, without loss of generality, we consider mean-centred $(\varepsilon_t)_t$, in order to simplify presentation of our results which, with minor modifications, continue to hold as such in the more general case.

Remark 3. Under the additional assumption that the process $(\varepsilon_t)_t$ is Gaussian with componentwise absolutely summable autocovariance matrices, condition (17) is equivalent to

$$\sum_{s\in\mathbb{Z}} |\mathrm{Cov}(\varepsilon_{t,a}, \varepsilon_{t+s,b})| < \infty$$

for all $a, b = 1, \ldots, d$ [8, Section 2.6]. Note that for the derivation of the majority of our results, we do not need any distributional assumption. However, in the particular case of independent Gaussian $(\varepsilon_t)_t$ we can further exploit distributional properties in order to be able to use an explicit bias correction, see Proposition 2.3.

2.1 Pilot estimator for the TVCM

Our goal is to nonparametrically estimate $\Sigma(u)$ as a curve in $u \in (0,1)$. For this we intend to adapt the recently developed wavelet-methodology ([10], [23]) in non-Euclidean set-ups, for correctly treating the multivariate situation under the paradigm that any estimator of $\Sigma(u)$ should remain positive-definite for any time $u \in (0,1)$. As our main interest is to develop a nonlinear wavelet thresholding estimator of $\Sigma(u)$ constructed from observed data following model (16), we need to propose a positive-definite *pilot* estimator to be used as input for our wavelet thresholding scheme presented in Algorithm 1. Note also that this multivariate approach of ours generalises existing linear wavelet estimation of time-varying variances in the univariate situation (e.g. [2]).

Two possible choices of pilot estimating local covariances can be envisaged: (i) a sliding window approach akin kernel smoothing with a rectangular window of length L, i.e. $\check{S}_k = L^{-1}\sum_{t=k-L/2+1}^{k+L/2} X_t X_t^\top$; (ii) local estimation by non-overlapping segmentation: for this we need to chop the observation period T up into n non-overlapping segments of length L, such that $T = nL$. Then a discrete estimator of Σ is given by $(\widetilde{S}_0, \ldots, \widetilde{S}_{n-1})$ with $\widetilde{S}_k$ an estimator of $S_k := \Sigma(u_k)$ at the midpoint $u_k = \frac{2k+1}{2n}$ of the $k-$th segment and defined by

$$\widetilde{S}_k = L^{-1}\sum_{t=kL+1}^{(k+1)L} X_t X_t^\top \ , \ k = 0, \ldots, n-1 \ . \tag{18}$$

As studied a bit in more detail in Section E of our Supplementary Material, in general we favor to use $(\widetilde{S}_k)_{k=0,\ldots,n-1}$ as pilot estimator for our nonlinear wavelet scheme, with $L \geq d$ to guarantee positive-definiteness of this pilot. Although, in comparison with $\check{S}_k$, the segment estimator $\widetilde{S}_k$ shows a somewhat higher bias of approximating $\Sigma(k/n)$ on a reduced resolution scale $n < T$, some illustrations in this Supplement Section E show that in the finite sample setting this will be highly compensated by the reduced variability of the resulting wavelet coefficients based on this pilot estimator $\widetilde{S}_k$. In the sequel, we therefore concentrate on presenting our asymptotic results while using as pilot the segment estimator $\widetilde{S}_k$.

Under the assumption that the series $(X_t)_{t=1}^T$ are generated by model (16), the pilot estimator (18) can be rewritten as $\widetilde{S}_k = S_k^{1/2} W_k S_k^{1/2}$ with

$$W_k = L^{-1}\sum_{t=kL+1}^{(k+1)L} \varepsilon_t \varepsilon_t^\top \ . \tag{19}$$

Proposition 2.1. *Let $(X_t)_{t=1}^T$ be time series generated by model (16) fulfilling Assumption 1. Then $W_k \xrightarrow{\mathbb{P}} I_d$ as $L \to \infty$. Thereby, $\widetilde{S}_k \xrightarrow{\mathbb{P}} S_k$ as $L \to \infty$. Here, convergence in probability holds with respect to the Frobenius norm.*

The reader might ask why not use $\widetilde{S}_k$ beyond being a pilot estimator to solve our smoothing problem, potentially with a rather large smoothing span L. To illustrate this question, we point the reader to Figure 8 in Section 4.3 which shows that for signals with transient behaviour, a pilot estimator such as the shown $\widetilde{S}_k$ with large smoothing span L, would be clearly suboptimal, as not capable to provide spatially adaptive smoothing. Note that in fact, because we estimate the points S_k, i.e. the curve Σ at the midpoint u_k of some segment $[(kL+1)/T, (k+1)L/T]$, we define as population analog of the expected $\widetilde{S}_k$, the average of the curve of the corresponding segment:

$$\overline{\Sigma}_k = \frac{1}{L} \sum_{t=kL+1}^{(k+1)L} \Sigma_t \ .$$

As $\Sigma : [0,1] \to \mathrm{Sym}^+(d)$ is sufficiently regular, we have that $\overline{\Sigma}_k = S_k + O(h)$ where $h = L/T$. Therefore, in practice, one should set $L \geq d$ as small as possible to mitigate the potential approximation bias of the localized estimator $\widetilde{S}_k$. We then need that L grows with the slowest possible rate when T increases, to avoid oversmoothing localized features of the underlying curve. For further considerations related to the growth rate of this window length L, see Remark 6 in the next section.

2.2 Nonparametric regression for sample covariance matrices

Having motivated the need for an appropriate smoothing procedure, we will use our pilot estimator as input for some subsequent, more sophisticated, denoising scheme, based on wavelets. In the following we aim at correctly treating bias, variance and integrated mean squared error (IMSE) of the thresholding estimator presented in Equation (14), now applied on denoising empirical variance-covariance matrices $\widetilde{S}_k$ of data following model (16). Motivated by the nonparametric regression framework in Subsection 1.2.3, we will work with the log-Euclidean signal-plus-noise model as presented in Equation (12):

$$\widetilde{S}_k = \mathrm{Exp}_{S_k}(\Gamma_{I_d}^{S_k}(\xi_k)) = \exp\big(\log(S_k) + \xi_k\big), \quad k = 0, \dots, n-1 \ .$$

Modelling sample covariance matrices with this log-Euclidean approach is justified by the following remark.

Remark 4. This model can actually be seen as an approximation of the "canonical signal-noise model" of empirical covariance matrices, namely $\widetilde{S}_k = S_k^{1/2} W_k S_k^{1/2}$ where W_k, as defined in (19), takes the role of "noise". Using the symmetric Baker-Campbell-Hausdorff (BCH) formula (see Appendix A), this canonical signal-noise model is equivalent to

$$\widetilde{S}_k = \exp\big(\log(S_k) + \log(W_k) + r_k\big) \ , \tag{20}$$

where $\log(W_k) + r_k$ takes the role of ξ_k in our log-Euclidean signal-plus-noise model. Here, r_k – defined by equation (29) – is a sum of vanishing Lie brackets polynomials of $\log S_k$ and $\log W_k$ (because $\log(S_k^{1/2} W_k S_k^{1/2}) \neq \log(S_k) + \log(W_k)$ for non-commuting matrices S_k and W_k, and the deviation from the right-hand side is quantified by the remainder term r_k).

Following from Remark 4, we can show that the pilot estimator has an intrinsic bias in model (12) – for which we recall that we consider the Fréchet (intrinsic) mean $\mathbb{E}[\cdot]$, not the Euclidean mean $\mathbf{E}[\cdot]$ (for which, obviously $\mathbf{E}[\widetilde{S}_k] = S_k$ but $\mathbb{E}[\widetilde{S}_k] \neq S_k$).

Proposition 2.2 (Intrinsic bias of the pilot estimator)**.** *Let $(X_t)_{t=1}^T$ be generated by model (16) and satisfy Assumption 1. Under the log-Euclidean model (12), the pilot estimator $\widetilde{S}_k$ has an intrinsic bias:*

$$\mathbb{E}[\widetilde{S}_k] = \exp\big(\log(S_k) + \mathbf{E}[\log(W_k)] + \mathbf{E}[r_k]\big) = S_k + O(L^{-1}) \ .$$

Remark 5. Recall that we do not work with any distributional assumption on the time series. However, under the assumption of Gaussianity of the multivariate time series $(X_t)_t$ and of independence between

the observations in time (i.e. the special case where $(\varepsilon_t)_{t\in\mathbb{Z}}$ consists of white noise), we can provide a distribution for the matrices W_k – which turns out to be useful for providing a bias correction. Indeed, we have segments consisting of the samples $X_{kL+1}, \dots, X_{(k+1)L}$ with $X_t \sim N_d(0, \Sigma_t)$. Using from the previous subsection that $\overline{\Sigma}_k = S_k + O(h)$ where $h = L/T$, then asymptotically each of the pointwise estimators $\widetilde{S}_k$ has an asymptotic Wishart distribution with L degrees of freedom: $\widetilde{S}_k \sim \mathcal{W}_d(L, L^{-1}S_k)$ which has a well-defined probability density function when $L \geq d$, see [21] for a global review of this distribution. In particular, rewriting $\widetilde{S}_k = S_k^{1/2} W_k S_k^{1/2}$ means that $W_k \sim \mathcal{W}_d(L, L^{-1}I_d)$.

A consequence of the previous remark is that we can derive a correction for the contribution of the "noise" matrices W_k to the intrinsic bias of the sample covariance matrix.

Proposition 2.3 (Intrinsic bias correction)**.** *Let $(X_t)_{t=1}^T$ be generated by model (16), with $(\varepsilon_t)_t$ Gaussian and independent. Under the log-Euclidean model (12), the pilot estimator $\widetilde{S}_k$ has an intrinsic bias. In particular, $\mathbf{E}[\log(W_k)] = c(d, L)I_d$ with $c(d, L) = \log(2/L) + \frac{1}{d}\sum_{i=1}^{d} \psi\left(\frac{L-i+1}{2}\right)$ and $\psi(\cdot)$ the digamma function. If we denote the bias-corrected pilot estimator $\widetilde{S}_k^c = \mathrm{e}^{-c(d,L)}\widetilde{S}_k$, then*

$$\mathbb{E}[\widetilde{S}_k^c] = \exp\left(\log(S_k) + \mathbf{E}[r_k]\right) = S_k + O(L^{-1}) \,,$$

where $\mathbf{E}[r_k] = O(L^{-1})$ is negligible.

Theoretically, this bias correction might not be useful because of the remaining term $\mathbf{E}[r_k] = O(L^{-1})$. However, this term is negligible in practice and most of the bias is removed when applying this correction. While theoretically those results do not hold if the assumption of independence of the series is not satisfied, using this correction nevertheless for serially-dependent time series satisfying Assumption 1 still provides an obvious numerical enhancement, as it can be viewed in Figure 1 where we compare two cases of $(\varepsilon_t)_t$ in model (16): the first is white noise, and the second is a VAR(1) process as in Section 4.2, scenario 2 of the "Dependent observations" paragraph. Additional experiments with non-Gaussian stationary processes showed robustness of this bias correction.

Furthermore, while the "generalized" variance of scaled log-Wishart random matrices $\log(W_k)$ can be shown to be homoscedastic [11, proof of Proposition 4.4], the remainder r_k prevents the noise in (20) from being homoscedastic (as the variance of this term depends on the signal S_k). However, we actually show in the proof of Proposition 2.2 that $r_k = O_{\mathbb{P}}(L^{-1/2})$ – which means that the additive model becomes approximately homoscedastic as the window length L increases.

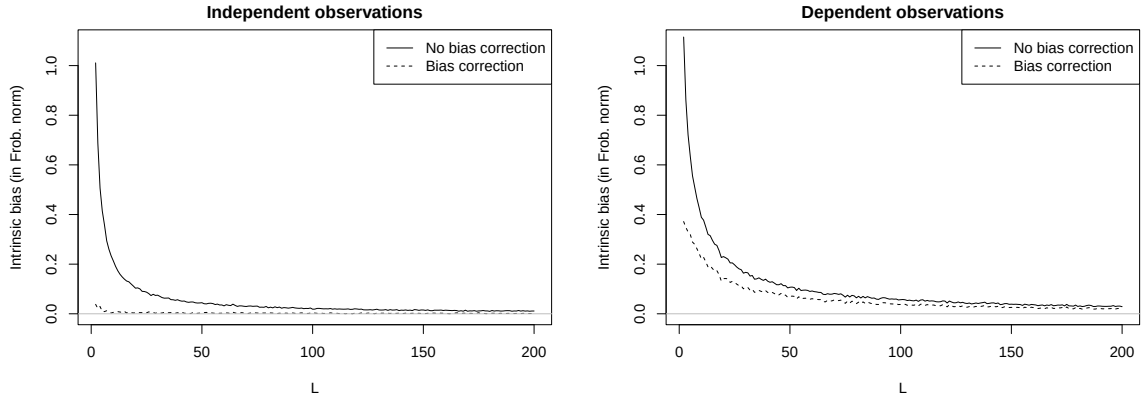


Figure 1: Simulation of empirical intrinsic bias as a function of L. For each L, we average $M = 10000$ sample covariance matrices based on a window of L observations. The bias correction in the dependent case reduces the bias by roughly 60 to 40% in the considered range of L.

We will now present asymptotic results for the linear curve estimator (14) where our observations on the finest wavelet scale $J = \log_2(n)$ are given by $(\log(\widetilde{S}_k^c))_{k=0,\dots,2^J-1}$. Although, in this paper, we

strongly promote the *nonlinear* version of this estimator, to be treated by Theorem 3.5, presenting first asymptotics for the linear estimator will be insightful for the sequel. Note that for these asymptotics, borrowing strength from the existing rates of IMSE-convergence as to be found in [11, Theorem 3.4], we assume that sample size n, finest resolution scale $J = J(n)$ and linear cut-off scale $J_0 = J_0(n)$ (see step 3 of our estimation scheme 1 of Section 1.2.3) tend to infinity such that $2^{J_0}/2^J \to 0$ as $n \to \infty$ (with exact rates for balancing approximation bias and variance of the linear wavelet estimator to be found below in Theorem 2.4). As it becomes apparent from Proposition 2.2 as well as from the simulations in Section 4, using the smallest possible L, i.e. $L = d$ (or rather $L = 2^{\lceil \log_2(d) \rceil}$ under the dyadic sample size constraint), would not be sufficient within our wavelet denoising scheme because of the intrinsic bias of the pilot estimator: we refer to the discussion on the need of a slowly growing L in Theorem 2.4 and Remark 6.

Theorem 2.4 (Consistency of linear estimator with IMSE-rate of convergence)**.** *Consider the wavelet estimator $\widehat{M}_{J,k}(J_0)$, based on a refinement scheme of order N and on linear thresholding on scale J_0. Let further $f(u) = \Sigma(u), u \in [0,1]$ be a smooth curve of time-varying covariance matrices $\Sigma(u) \in \mathrm{Sym}^+(d)$, possessing at least N continuous derivatives in time u. Construct data $\widetilde{S_k}$ according to the PD pilot estimator (18) with $T^{N/(3N+1)} \lesssim L \lesssim T^{(2N+1)/(3N+1)}$ and apply the bias correction of Proposition 2.2. Feed the resulting $\widetilde{S_k^c}$ into Step 1 of our estimation scheme 1 of Section 1.2.3 to get, via linear thresholding above scale J_0 the scaling coefficients $\widehat{M}_{J,k}(J_0)$ which fulfill, as $n \to \infty$,*

$$\mathbf{E}[\| \log \widehat{M}_{J,k}(J_0) - \mathbf{E}[\log \widehat{M}_{J,k}(J_0)]\|_F^2] = O(2^{-(J-J_0)})$$

and

$$\|\mathbf{E}[\log \widehat{M}_{J,k}(J_0)] - \log M_{J,k}^f\|_F = O(2^{-NJ_0}) \, .$$

As usual for linear nonparametric curve estimators, the choice $J_0 = \lfloor J/(2N+1) \rfloor$ balances variance and squared bias of the wavelet estimator, such that

$$\mathbf{E}[\delta_{\mathrm{L}}^2(\widehat{M}_{J,k}(J_0), M_{J,k}^f)] = \mathbf{E}[\| \log \widehat{M}_{J,k}(J_0) - \log M_{J,k}^f\|_F^2] = O(n^{-2N/(2N+1)}) \, ,$$

which implies that $n^{-1} \sum_{k=0}^{n-1} \mathbf{E}[\delta_{\mathrm{L}}^2(\widehat{M}_{J,k}(J_0), M_{J,k}^f)] = O(n^{-2N/(2N+1)}) \, .$

As explained in Remark C.1 of the Supplementary Material, this also implies that the IMSE of $\widehat{\Sigma}_n(u)$ converges with the same rate:

$$\int_0^1 \mathbf{E}[\delta_{\mathrm{L}}^2(\widehat{\Sigma}_n(u), \Sigma(u))] \, \mathrm{d}u = O(n^{-2N/(2N+1)}) \, .$$

Remark 6. By the discussion on the approximation bias from Subsection 2.1, L should grow with the smallest possible rate to keep this bias small. This suggests for the control of the intrinsic bias to choose $L \sim T^{N/(3N+1)}$. In any case, note that we should respect the condition $L \lesssim T^{(2N+1)/(3N+1)})$ such that the approximation bias $\overline{\Sigma}_k(t/T) - S_k(t/T)$ of order L/T is also smaller than the bias of the wavelet estimator of order 2^{-NJ_0}.

In the next section we will treat the nonlinear version of this estimator, see Theorem 3.5.

3 Nonlinear wavelet regression for denoising SPD matrices

3.1 Coefficient noise model

Using the algebraic properties of the log-Euclidean signal plus noise model (12), [27] showed that the intrinsic AI wavelet transform with the log-Euclidean metric actually leads to a model where the whitened wavelet coefficients follow a matrix-valued signal-plus-noise model, which allows many thresholding possibilities.

Proposition 3.1. *Let $(\widetilde{S}_k)_{k=0}^{n-1} = (M_{Jk})_{k=0}^{n-1}$ be defined by equation (12), with $n = 2^J$ for $J > 0$. Let $(\mathfrak{D}_{j,k})_{j,k}$ be the set of empirical whitened wavelet coefficients (as in equation (11)) obtained via the intrinsic forward wavelet transform with the log-Euclidean metric and order $N = 2M + 1$ for some $M \geq 0$. Then, for each scale j and location k, it holds that*

$$\mathfrak{D}_{jk} = \mathfrak{D}^f_{jk} + \mathfrak{D}^\xi_{jk} \ ,$$

where $\mathfrak{D}^f_{jk}$ is the deterministic whitened coefficient based on the signal $(S_k)_{k=0}^{n-1}$ and $\mathfrak{D}^\xi_{jk}$ is the random whitened coefficient based on the i.i.d. noise sequence $(\xi_k)_{k=0}^{n-1}$. Moreover, $\mathbf{E}[\mathfrak{D}_{jk}] = \mathfrak{D}^f_{jk}$ and

$$\mathrm{Var}[\mathfrak{D}_{jk}] = \frac{1}{2n} \sum_{m=-M}^{M} C^2_{N,m} \mathrm{Var}[\xi_0] \ ,$$

where $\mathrm{Var}[\xi_0] = \boldsymbol{\mathcal{C}}$ is a fourth order tensor.

Remark 7. This property of the whitened wavelet coefficients in the log-Euclidean wavelet transform implies that the (non-whitened) wavelet coefficients also follow an additive model. Parallel transporting the coefficients $(\mathfrak{D}_{j,k})_{j,k}$ back to their own tangent space will result in the wavelet coefficients

$$D_{j,k} = \Gamma_{I_d}^{\widetilde{M}_{j,2k+1}}(\mathfrak{D}_{j,k}) = \mathbb{D}_{\log \widetilde{M}_{j,2k+1}} \exp(\mathfrak{D}_{j,k}) \ ,$$

where $\mathbb{D}_{\log S} \exp$ with $S \in \mathrm{Sym}^+(d)$ is a linear map. Consequently, the wavelet coefficients also follow a signal-plus-noise model:

$$D_{j,k} = D^f_{j,k} + D^\xi_{j,k} \ .$$

As a consequence of Proposition 3.1, linear combinations of the elements of the empirical coefficient matrices also satisfy the properties of a signal-plus-noise model, which facilitates derivation of more sophisticated thresholding methodology.

Proposition 3.2. *Let $(\widetilde{S}_k)_{k=0}^{n-1} = (M_{Jk})_{k=0}^{n-1}$ be defined by Equation (12), with $n = 2^J$ for $J > 0$. Let $(\mathfrak{D}_{j,k})_{j,k}$ be the set of empirical whitened wavelet coefficients obtained via the intrinsic forward wavelet transform with the log-Euclidean metric and order $N = 2M + 1$ for some $M \geq 0$. Let $\varphi : \mathrm{Sym}(d) \to \mathbb{R}$ be an affine map such that for $X, Y \in \mathrm{Sym}(d)$ and $a, b \in \mathbb{R}$: $\varphi(aX + bY) = a\varphi(X) + b\varphi(Y)$. Then, for each scale j and location k, it holds that*

$$\varphi(\mathfrak{D}_{jk}) = \varphi(\mathfrak{D}^f_{jk}) + \varphi(\mathfrak{D}^\xi_{jk}) \ .$$

Moreover, $\mathbf{E}[\varphi(\mathfrak{D}_{jk})] = \varphi(\mathfrak{D}^f_{jk})$ and

$$\mathrm{Var}[\varphi(\mathfrak{D}_{jk})] = \left(\frac{1}{2n} \sum_{m=-M}^{M} C^2_{N,m} \right) \mathrm{Var}[\varphi(\xi_0)] \ .$$

3.2 Appropriate choice of thresholds

Assumption 2. *In the following section, we assume that the i.i.d noise matrices ξ_i have a normal distribution on $\mathrm{Sym}(d)$ with mean 0 and covariance operator $\boldsymbol{\mathcal{C}}$:*

$$\langle \xi_i, A \rangle_F \sim \mathcal{N}\big(0, \langle \boldsymbol{\mathcal{C}} A, A \rangle_F\big) \quad \forall A \in \mathrm{Sym}(d) \ ,$$

where

$$[\boldsymbol{\mathcal{C}}]_{abmn} = \begin{cases} \sigma^2_{ab} & a = n, b = m \text{ or } a = m, b = n \\ 0 & \text{else.} \end{cases} \tag{21}$$

In particular, this can be written as

$$\eta(\xi_i) \sim \mathcal{N}_q(0, \boldsymbol{\mathcal{C}}^\eta)$$

where $q = d(d+1)/2$, $\boldsymbol{\mathcal{C}}^\eta = \mathrm{Cov}(\eta(\xi_0))$ *with the stacking operator* $\eta : \mathrm{Sym}(d) \to \mathbb{R}^q$ *a linear isometry defined by*

$$\eta(A) = (A_{11}, \ldots, A_{dd}, \sqrt{2}A_{12}, \ldots, \sqrt{2}A_{1d}, \sqrt{2}A_{23}, \ldots, \sqrt{2}A_{2d}, \ldots, \sqrt{2}A_{d-1,d})^\top \ , \quad A = [A_{ij}]_{i,j} \ .$$

This is equivalent to rewrite the noise matrices as a linear combination of independent univariate Gaussian random variables and matrices which form a basis in $\mathrm{Sym}(d)$*, i.e.*

$$\xi_k = \sum_{1 \le a \le b \le d} z_{a,b}^k E_{a,b} \ ,$$

where $(z_{a,b}^k)_k \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma_{a,b}^2)$*, and the matrices* $(E_{a,b})_{1 \le a \le b \le d}$ *are defined by*

$$[E_{a,b}]_{k,\ell} = \begin{cases} 1 \ , & a = k, b = \ell \text{ or } a = \ell, b = k \\ 0 \ , & \text{else.} \end{cases}$$

3.2.1 Scalar Gaussian thresholding

Our aim with scalar Gaussian rules for nonlinear thresholding is to find a map $g : \mathrm{Sym}(d) \to \mathbb{R}$ such that, under Assumption 2, $g(\mathfrak{D}_{j,k}^\xi) \sim N(0, \sigma_g^2)$ so we can use a universal threshold $\tau_n = \sigma_g^2 \sqrt{2 \log n}$ and our nonlinear estimator of $\mathfrak{D}_{j,k}^f$ is then

$$\widehat{\mathfrak{D}}_{j,k} = \mathbb{1}\left\{ |g(\mathfrak{D}_{j,k})| > \tau_n \right\} \mathfrak{D}_{j,k} \ ,$$

where the coefficients consisting only in noise are asymptotically removed with probability 1. This is motivated by the upper bound on the maximum of standard Gaussian random variables: for $Z_1, \ldots, Z_n \sim \mathcal{N}(0,1)$,

$$\mathbf{E}\left[\max_{1 \le i \le n} |Z_i| \right] \le \sqrt{2 \log n} \ .$$

Proposition 3.1 implies that the trace of the whitened coefficients obtained from the intrinsic AI wavelet transform (with refinement order N) also follows a signal-plus-noise model:

$$\mathrm{Tr}(\mathfrak{D}_{jk}) = \mathrm{Tr}(\mathfrak{D}_{jk}^f) + \mathrm{Tr}(\mathfrak{D}_{jk}^\xi) \ ,$$

which satisfies $\mathbf{E}[\mathrm{Tr}(\mathfrak{D}_{jk})] = \mathrm{Tr}(\mathfrak{D}_{jk}^f)$ and has homogeneous variance across scales and locations:

$$\mathrm{Var}[\mathrm{Tr}(\mathfrak{D}_{j,k})] = \left(\frac{1}{2n} \sum_{m=-M}^{M} C_{N,m}^2 \right) \mathrm{Var}[\mathrm{Tr}(\xi_0)] \ ,$$

where under Assumption 2 we have $\mathrm{Var}[\mathrm{Tr}(\xi_0)] = \sum_{a=1}^d \sigma_{a,a}^2$. This thresholding rule was first developed in the framework described in [11]. Another possibility to exploit further the properties of the whitened coefficients shown in Proposition 3.1, is to take the upper-sum of the elements (which only takes the $d(d+1)/2$ independent elements into account): for some $A \in \mathrm{Sym}^+(d)$, the upper-sum map $\triangle : \mathrm{Sym}^+(d) \to \mathbb{R}$ is defined by

$$\triangle(A) = \sum_{a=1}^{d} \sum_{b=a}^{d} [A]_{a,b} \ .$$

By affinity of the upper-sum, we have the following signal-plus-noise model:

$$\triangle(\mathfrak{D}_{jk}) = \triangle(\mathfrak{D}^f_{jk}) + \triangle(\mathfrak{D}^\xi_{jk}) \ ,$$

which satisfies $\mathbf{E}[\triangle(\mathfrak{D}_{jk})] = \triangle(\mathfrak{D}^f_{jk})$ and has variance

$$\mathrm{Var}[\triangle(\mathfrak{D}_{j,k})] = \left(\frac{1}{2n}\sum_{m=-M}^{M} C^2_{N,m}\right)\mathrm{Var}[\triangle(\xi_1)] \ ,$$

where under Assumption 2 we have $\mathrm{Var}[\triangle(\xi_0)] = \sum_{1\le a\le b\le d}\sigma^2_{a,b}$.

3.2.2 Intrinsic Elementwise Thresholding

Another possibility following from Proposition 3.1 is to directly threshold the elements of the whitened coefficients. Using the canonical basis of $\mathrm{Sym}(d)$, we can express the whitened coefficients by

$$\mathfrak{D}_{j,k} = \sum_{1\le a\le b\le d}([\mathfrak{D}^f_{j,k}]_{a,b} + [\mathfrak{D}^\xi_{j,k}]_{a,b})E_{a,b} \ , \tag{22}$$

which implies that we only threshold the $d(d+1)/2$ elements while ensuring the thresholded coefficients are still symmetric. Having a matrix valued signal-plus-noise model implies that the elements can also be expressed as a signal-plus-noise model. Let $\varphi_{a,b}(\mathfrak{D}_{j,k}) = [\mathfrak{D}_{j,k}]_{a,b} =: \mathfrak{d}_{j,k,a,b}$ – as we have an affine-map $\varphi_{a,b} : \mathrm{Sym}(d) \to \mathbb{R}$, Proposition 3.2 applies and we have

$$\mathfrak{d}_{j,k,a,b} = \mathfrak{d}^f_{j,k,a,b} + \mathfrak{d}^\xi_{j,k,a,b} \ ,$$

which satisfies $\mathbf{E}[\mathfrak{d}_{j,k,a,b}] = \mathfrak{d}^f_{j,k,a,b}$ and with variance

$$\mathrm{Var}[\mathfrak{d}_{j,k,a,b}] = \left(\frac{1}{2n}\sum_{\ell=0}^{2N-1} C^2_{N,\ell}\right)\mathrm{Var}\left([\xi_0]_{a,b}\right) =: \varsigma^2_{a,b} \ ,$$

where under Assumption 2 we have $\mathrm{Var}\left([\xi_0]_{a,b}\right) = \sigma^2_{a,b}$. We call $(\widehat{\mathfrak{D}}^*_{j,k})_{j,k}$ the intrinsic elementwise (IEWT) nonlinear thresholding estimator of $(\mathfrak{D}^f_{j,k})_{j,k}$, such that

$$\widehat{\mathfrak{D}}^*_{j,k} = W_{j,k} \odot \mathfrak{D}_{j,k} \ ,$$

where $W_{j,k} \in \mathrm{Sym}(d)$ has elements $[W_{j,k}]_{a,b} = \mathbb{1}\{|\mathfrak{d}_{j,k,a,b}| > \tau_{a,b,n}\}$, with thresholds $\tau_{a,b,n} = \varsigma_{a,b}\sqrt{2\log n}$ computed for each of the $d(d+1)/2$ elements.

Additionally, we show in Corollary 3.3 that it is sufficient that wavelet coefficients stay symmetric to ensure the reconstructed signal is SPD. This result is important because it means that we have a lot more flexibility to remove some component specific noise, which is more adaptive compared to the case where we use a scalar threshold only taking the overall regularity of the signal into account.

Corollary 3.3. *Let $M_{0,0}$ be the coarsest scale midpoints, $(\widetilde{M}_{j,k})_{j,k}$ be the pyramid of predicted midpoints, and $(D_{j,k})_{j,k}$ be the pyramid of wavelet coefficients; all obtained from the intrinsic AI wavelet transform. Let $(\widehat{D}^*_{j,k})_{j,k}$ be the pyramid of thresholding estimates of the wavelet coefficients. The reconstruction of a signal based on those thresholded wavelet coefficients is necessarily SPD, because $\widehat{D}_{j,k}$ is indeed symmetric for all (j,k) when thresholding in the basis $(E_{a,b})_{a\le b}$.*

Remark 8. The fact that the thresholding scheme presented here is elementwise does not mean that estimation of the curve can be done elementwise with Euclidean tools: using univariate smoothing on

the different elements of the matrix curve does not guarantee that the resulting estimator is SPD, and can cause other artifacts such as the swelling effect. On the other hand, when using *intrinsic* elementwise thresholding (IEWT), the resulting coefficients of the thresholding scheme are still in $\mathrm{Sym}(d)$. Thus, they can be used in the backward transform using intrinsic operations such as the parallel transport and exponential map, hence ensuring the resulting estimator is SPD and does not suffer from well-known problems in Euclidean smoothing for curves of covariance matrices.

3.2.3 Chi-squared thresholding

A conceptual weakness of the scalar thresholding methodology presented in Section 3.2.1 is that it might not catch some important signal of relatively small magnitude on some element of the observed matrix if the noise contribution of other elements annihilates the contribution of the signal. Note that due to the affinity requirement of scalar Gaussian thresholding, we cannot use the trace and upper-sum of the absolute value of the whitened wavelet coefficients. Using the Frobenius norm would also be a conceptually more intuitive and stronger solution, but this would not satisfy the properties of the Gaussian wavelet thresholding theory either. Instead, consider the following random variable:

$$X_0 = \sum_{1 \le a \le b \le d} \left\{ \frac{[\xi_0]_{a,b}}{\sigma_{a,b}} \right\}^2 .$$

Under Assumption 2, we have $X_0 \sim \chi^2_p$ with $p = d(d+1)/2$, which implies that

$$h(\mathfrak{D}_{jk}) := \sum_{1 \le a \le b \le d} \left\{ \frac{\mathfrak{d}_{j,k,a,b}}{\varsigma_{a,b}} \right\}^2 \sim \chi^2_p ,$$

where $\varsigma^2_{a,b} = \mathrm{Var}[\mathfrak{d}_{j,k,a,b}]$. Parallel to the case of normally distributed functionals g of the whitened wavelet coefficients, we need to find a universal threshold motivated by asymptotic removal of the noisy coefficients. Instead of using the threshold for asymptotic Gaussian noise removal, we then consider a new thresholding rule for χ^2 distributed random variables: using that centered Chi-squared random variables belong the the class of sub-gamma random variables on the right tail, it can be shown (see [6, Example 2.7]) that with iid $X_1, \dots, X_n \sim \chi^2_p$,

$$\mathbf{E}\left[\max_{1 \le i \le n} X_i - p \right] \le 2\sqrt{p \log n} + 2 \log n .$$

Hence, the Chi-squared thresholding estimator of $\mathfrak{D}^f_{j,k}$ is defined by $\widehat{\mathfrak{D}}^*_{jk} = w_{j,k} \mathfrak{D}_{jk}$ with, for some $\varepsilon > 0$,

$$w_{j,k} = \mathbb{1}\left\{ h(\mathfrak{D}_{jk}) - p > 2(1+\varepsilon)\sqrt{p \log n} + 2 \log n \right\} .$$

Proposition 3.4 (Asymptotic noise removal)**.** *Let $X_1, \dots, X_n \sim \chi^2_p$ with $p > 1$. Then, for $\varepsilon > 0$*

$$\mathbb{P}\left\{ \max_{1 \le i \le n} X_i - p \ge 2(1+\varepsilon)\sqrt{p \log n} + 2 \log n \right\} \to 0 \quad \text{as } n \to \infty .$$

Remark 9. The presence of the factor $(1+\varepsilon)$ with positive ε is essential for theoretical purpose. However, it turns out that in finite sample situations, choosing the threshold $\tau_n = 2\sqrt{p \log n} + 2 \log n$ is working sufficiently well, whereas the slightly larger one of Proposition 3.4 solely serves to show the asymptotic noise removal property.

3.3 IMSE-consistency of nonlinear threshold estimator

Nonlinear thresholding estimators are motivated from an intuitive point of view by capturing localized high frequencies and jumps that might not be captured by linear thresholding. One question that naturally

arises is whether the nonlinear estimator is L_2 consistent – i.e. if the integrated mean squared error (IMSE) converges to 0 when $n \to \infty$. Deriving such a result in the context of AI-schemes with filters that do not correspond to orthogonal wavelets is the main contribution of the following Theorem 3.5, which includes stating rates of convergence of the IMSE for sufficiently smooth curves.

Given a refinement order $N \geq 1$, suppose $f : [0,1] \to \mathrm{Sym}^+(d)$ is a smooth curve with existing covariant derivatives of order N or higher, and let $(\mathfrak{D}_{jk})_{jk}$ be the empirical whitened wavelet coefficients obtained from the intrinsic AI wavelet transform of order N. Now let $(\widehat{\mathfrak{D}}_{jk})_{jk}$ be the nonlinear thresholding estimates of the whitened wavelet coefficients. We will denote by $(\widehat{\mathfrak{D}}^*_{jk})_{jk}$ the hybrid thresholding estimates of the whitened wavelet coefficients, namely

$$\widehat{\mathfrak{D}}^*_{jk} = \begin{cases} \mathfrak{D}_{jk} & j = 1, \dots, J_0 - 1 \, ; \\ \widehat{\mathfrak{D}}_{jk} & j = J_0, \dots, J_1 - 1 \, ; \\ 0 & j = J_1, \dots, J \, . \end{cases} \tag{23}$$

Theorem 3.5 (Convergence rates nonlinear thresholding)**.** *Given a refinement order $N \geq 1$, suppose $f : [0,1] \to \mathrm{Sym}^+(d)$ is a smooth curve with existing covariant derivatives of order N or higher. Then, as $n \to \infty$,*

$$\sum_{jk} \mathbf{E} \|\widehat{\mathfrak{D}}^*_{jk} - \mathfrak{D}^f_{jk}\|_F^2 = O(n^{-2N/(2N+1)}) \, , \tag{24}$$

where $\widehat{\mathfrak{D}}^*_{j,k}$ denotes the hybrid nonlinear thresholding estimator of the whitened wavelet coefficients as defined in (23), with $J_0 = \lfloor J/(2N+1) \rfloor$ and $J_0 < J_1 < J$ such that $2^{J_0} \sim 2^{J_1}$ and $2^{J_1 - J} \to 0$ as $n \to \infty$. Moreover,

$$\frac{1}{n} \sum_{k=0}^{n-1} \mathbf{E}[\delta_{\mathrm{L}}^2(\widehat{M}^*_{J,k}, M^f_{J,k})] = O(n^{-2N/(2N+1)}) \, , \tag{25}$$

where $\widehat{M}^*_{J,k}$ denotes the nonlinear thresholding estimator of $M^f_{J,k}$ reconstructed from $(\widehat{\mathfrak{D}}^*_{j,k})_{j,k}$.

Remark 10. This Theorem 3.5 is stated in a very generic way, holding for all the rules of how to choose the appropriate threshold value as presented in Section 3.2. Hence it calls for two important remarks to be added:

(1) It can directly be applied to the situation of Theorem 2.4 on smoothing the (bias-corrected) pilot estimator $\widetilde{S}^c_k$ with window length $L \sim T^{N/(3N+1)}$, and hence shows IMSE-consistency of the nonlinear threshold version of our estimation scheme of Section 1.2.3 applied to denoising TVCM. Note that the results of Proposition 3.1 remain valid asymptotically for this situation.

(2) Inspection of its proof suggests that the employed techniques are not that specific to the use of the log-Euclidean metric but would work also for more general AI-schemes adapted to the geometry of the underlying Riemannian manifold. In particular, by comparison with the tools used in [11] on the affine-invariant Riemannian metric, it can be seen that this new Theorem 3.5 also holds for nonlinear thresholding based on the use of the "trace-rule" (of Section 3.2.1) when using this latter metric. As a consequence it can be applied for denoising Hermitian positive-definite matrices (such as spectral density matrices), as in [11] and [12].

3.4 Denoising of sample covariance matrices

In this section, we address possible issues concerning the effectiveness of our nonlinear thresholding scheme in the TVCM framework.

Gaussianity of the coefficients. In practice, Assumption 2 is not verified when working with sample covariance matrices, because the nature of the noise in this framework is not Gaussian on $\mathrm{Sym}(d)$. This assumption of Gaussianity is important because the nonlinear thresholding estimator based on the

"universal threshold" is asymptotically noise-free if the noise is Gaussian. However, as $L, T \to \infty$ and $L/T \to 0$, the non-Gaussian additive model of the whitened wavelet coefficients is asymptotically equivalent to a Gaussian additive model as the empirical wavelet noise coefficients $(\mathfrak{D}^{\xi}_{j,k})_{j,k}$ are essentially differences of intrinsic local averages of sample covariance matrices: from Proposition C.5 in the Supplementary Material, asymptotic normality (on $\mathrm{Sym}(d)$) of the empirical whitened wavelet coefficients is fulfilled – this suggests to still work in practice under the assumption of normally distributed empirical wavelet coefficients.

Heteroscedasticity of the noise. As discussed in Subsection 2.2, the remainder term r_k is heteroscedastic with respect to k. However as we let the window length grow as $L \sim T^{N/(3N+1)}$ to reduce the intrinsic bias of the pilot estimator, the heteroscedasticity becomes negligible as well because $r_k = O_{\mathbb{P}}(L^{-1/2})$. To treat this problem in the non-asymptotic setting, we use scale-dependent thresholding which empirically enhances the performance of our nonlinear scheme and which had initially been developed for dealing with denoising in presence of stationary serially dependent noise, as in [20]. That is, instead of setting a global threshold τ_n proportional to $\sqrt{2\log n}$ (or $2\sqrt{p\log n} + 2\log n$ for Chi-squared thresholding), we set for each scale j a threshold of the form $\tau_{j,n} \sim \sqrt{2\log 2^j}$ (and accordingly replacing n by 2^j for Chi-squared thresholding).

Signal-noise ratio. Because $r_k = O_{\mathbb{P}}(L^{-1/2})$ and (by the proof of Proposition 2.2) $\log(W_k) = O_{\mathbb{P}}(L^{-1/2})$ too, we get by letting the window length grow as $L \sim T^{N/(3N+1)}$ that the noise amplitude will vanish as $L \to \infty$, which facilitates the nonlinear thresholding procedure by reducing the noise level relative to the signal.

4 Numerical examples

In this section, we assess by simulation study the performance of the multiple thresholding rules considered in Section 3.2 by Monte-Carlo simulations with different data generating processes (DGPs). For each curve in each scenario, we will simulate $M = 500$ realisations of the considered DGPs.

We will consider multiple SPD curves, each being studied under six scenarios for the noise structure. In particular, consider the following matrix-valued function:

$$f(t) = \begin{pmatrix} 10 + (5 - 10t)^2 & 15 - 10t \\ 15 - 10t & 20 \end{pmatrix} .$$

Let $c_1(t), c_2(t), c_3(t), c_4(t)$ be respectively the Blocks, Bumps, HeaviSine and Doppler functions defined in [15] and recalled in the Supplementary Material. We will then study the following SPD curves, and focus on the Bumps SPD curve f_2 while the same analysis for each of the four curves is to be found in the Supplementary Material:

$$\begin{aligned} f_1(t) &= (10 + c_1) \circ f(t) , \\ f_2(t) &= (1 + c_2) \circ f(t) , \\ f_3(t) &= (7 + c_3) \circ f(t) , \\ f_4(t) &= (1 + c_4) \circ f(t) . \end{aligned}$$

4.1 Log-normal noise model

We study multiple possibilities for the noise structure. We simulate data following model (12) as in Definition 1.1, where the i.i.d. noise matrices $(\xi_k)_{k=0}^n$ are constructed as described in Assumption 2; i.e. we simulate log-normal noise. We will consider the following scenarios:

$$\sigma_{a,b} = 0.01 \quad \forall a \leq b \, ; \tag{Homog1}$$

$$\sigma_{a,b} = 0.1 \quad \forall a \leq b \; ; \tag{Homog2}$$
$$\sigma_{a,b} = 0.3 \quad \forall a \leq b \; ; \tag{Homog3}$$
$$\sigma_{1,1} = 0.01, \; \sigma_{1,1} = 0.1, \; \sigma_{1,2} = 0.001 \; ; \tag{Inhomog1}$$
$$\sigma_{1,1} = 0.01, \; \sigma_{1,1} = 0.1, \; \sigma_{1,2} = 0.1 \; ; \tag{Inhomog2}$$
$$\sigma_{1,1} = 0.01, \; \sigma_{1,1} = 0.1, \; \sigma_{1,2} = 0.3 \; ; \tag{Inhomog3}$$

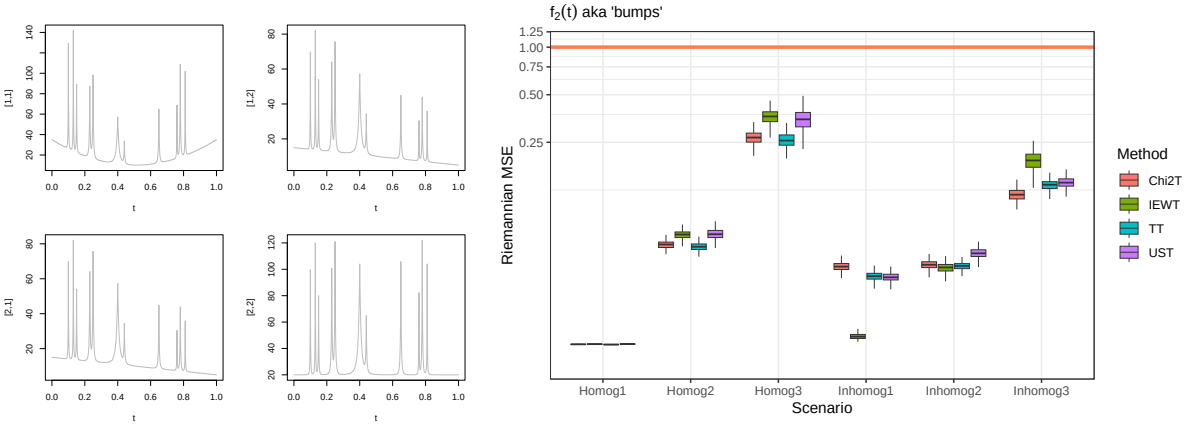


Figure 2: Left: Bumps SPD curve $f_2(t)$. Right: performance of the nonlinear thresholding estimators (relative to the performance of linear thresholding).

For each simulation, we estimate the underlying curve using linear thresholding, trace thresholding (`TT`), upper-sum thresholding (`UST`), Chi-square thresholding (`Chi2T`) and intrinsic elementwise thresholding (`IEWT`). We will use the log-Euclidean AI wavelet transform with $N = 3$, and define our nonlinear wavelet thresholding estimator as in equation (23) with $J_0 = 3$ and $J_1 = 10$. Additionally, we impose a tree structure constraint in the nonlinear thresholding as described in [14], where the goal is to exploit the inter-scale correlation between the wavelet coefficients to obtain a better reconstruction in the presence of singularities in the signal. Finally, we estimate the variances $\varsigma^2_{a,b}$ (necessary to compute the thresholds) by the usual robust mean-absolute deviation - scaled to match (approximately) Gaussian data - on the finest scale: $\widehat{\varsigma}^2_{a,b} = \mathrm{MAD}[\mathfrak{d}_{J,k,a,b}]/0.6745$. For the linear thresholding estimator, we set $J_0 = 6$ to be able to adapt to the less regular parts of the functions, without incorporating too many noisy coefficients. We summarize the results of our simulation study in Figure 2, using as performance metric the Riemannian mean-squared error with respect to the log-Euclidean distance:

$$\mathrm{RieMSE}(\widehat{f}^*, f) = \frac{1}{n} \sum_{k=0}^{n-1} \delta^2_{\mathrm{L}}(\widehat{S}^*_k, S_k) \; .$$

The RieMSE of the nonlinear thresholding estimators are standardized by the RieMSE of the linear estimator in each simulation: hence, values above 1 indicate worse performance while values below 1 indicate that linear thresholding is outperformed by the nonlinear estimator: we immediately observe that nonlinear rules globally outperform the linear estimator. Comparing the rules, we observe that scalar rules do not perform significantly differently (we note that `UST` is generally worse than `TT` and `Chi2T`), while `IEWT` does not seem to work well except in the Inhomog1 and Inhomog2 scenarios.

More interestingly, we simulate a SPD curve with $d = 5$ (which is defined in Section F of the Supplementary Material), which consists of 15 curves over elements which satisfy the positive-definite constraint (5 on the diagonal and 10 off-diagonal). In this framework, the more flexible `Chi2T` and `IEWT` definitely outperform the other two rules. This confirms that in particular the trace is not sufficient to describe the overall behaviour of the wavelet coefficients – and that our new thresholding rules are better suited to more realistic, higher-dimensional frameworks.

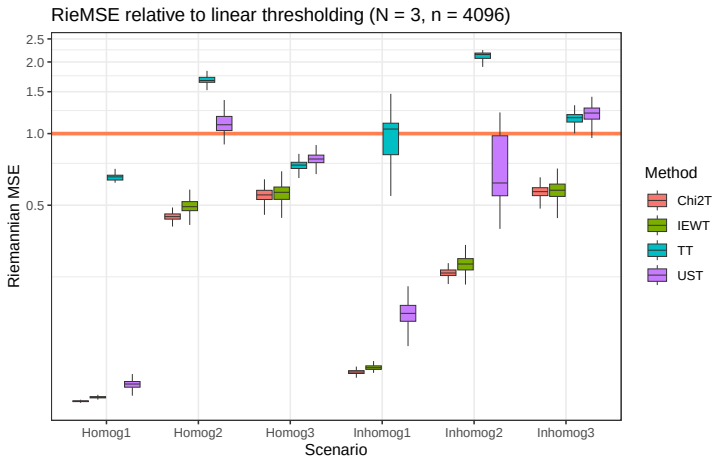


Figure 3: Performance of the nonlinear thresholding estimators (relative to the performance of linear thresholding) of the higher-dimensional SPD curve (with $d = 5$).

4.2 Wavelet denoising of empirical TVCM

In the following, we apply our wavelet denoising scheme to estimation of the TVCM of some series $(X_t)_{t=1}^T$ with $X_t \sim N(0, \Sigma(t))$ where $\Sigma(t) = f_2(t)$. We will first study the performance of our methodology for denoising TVCMs of time series constructed from white noise, and then we assess those performances on time series constructed from VAR(1) processes. In this framework, we will only use linear and Chi-squared (scale-dependent) thresholding; we chose the latter compared to the other nonlinear thresholds because it was shown in the previous simulations to perform satisfyingly well in many situations and because of its enhanced originality. We observe in the two following scenarios that it indeed outperforms the linear estimator, especially when choosing the optimal window length $L = 2^{\lceil \log_2(T^{N/(3N+1)}) \rceil}$.

Serially independent observations. We simulate time series $X_t = \Sigma_t^{1/2} \varepsilon_t$ where $\varepsilon_t \overset{\text{i.i.d.}}{\sim} N(0, I_d)$, for $t = 1, \ldots, T$ where $T = 2^{15}$. We use the pilot estimator (18), and increase the window length L from 2 to 256.

Serially dependent observations. We simulate time series $X_t = \Sigma_t^{1/2} \varepsilon_t$ where $\varepsilon_t = A\, \varepsilon_{t-1} + \upsilon_t$, with $\upsilon_t \sim N(0, \Psi)$ and Ψ such that $\mathrm{Var}[\varepsilon_t] = I_d$ for all t, as the solution of

$$\mathrm{vec}(\Psi) = (I_{d^2} - A \otimes A)\, \mathrm{vec}(I_d)\ .$$

We consider two scenarios for A, one for weaker and one for stronger dependence – respectively,

$$A^{(1)} = \begin{pmatrix} 0.3 & 0.15 \\ -0.05 & 0.1 \end{pmatrix} \quad \text{and} \quad A^{(2)} = \begin{pmatrix} 0.8 & 0.2 \\ -0.05 & 0.5 \end{pmatrix}\ .$$

4.3 Application to electroencephalography data

In this section, we present an analysis of an electroencephalogram (EEG), measuring the electrical brain activity of a female patient suffering from left temporal lobe epilepsy, diagnosed by Dr. Beth Malow at the Epilepsy Disorder Laboratory (University of Michigan). We consider 3 channels (i.e. electrodes on particular locations of the scalp, see Figure 6) T3, P3 and T4, for which we are interested in obtaining a denoised (i.e. smooth but adaptive) estimate of the variance-covariance structure between those channels during the epileptic seizure. We will compare the seizure onsets by inspecting the curves of estimated variances, and use the time-varying covariance matrix to compute a smooth estimate of the time-varying correlation matrix and evaluate the linear relationship between the considered electrodes. We use the last

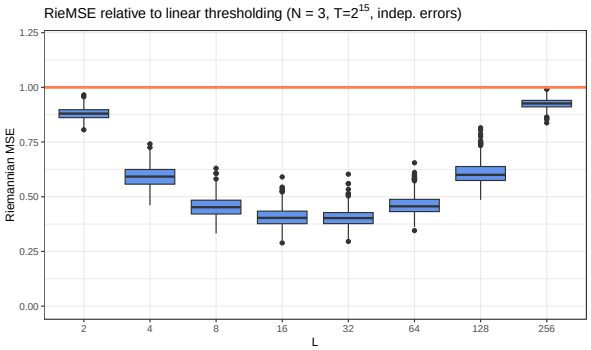


Figure 4: Performance of the Chi-squared thresholding estimator relative to linear thresholding, on empirical TVCMs as a function of window length L.

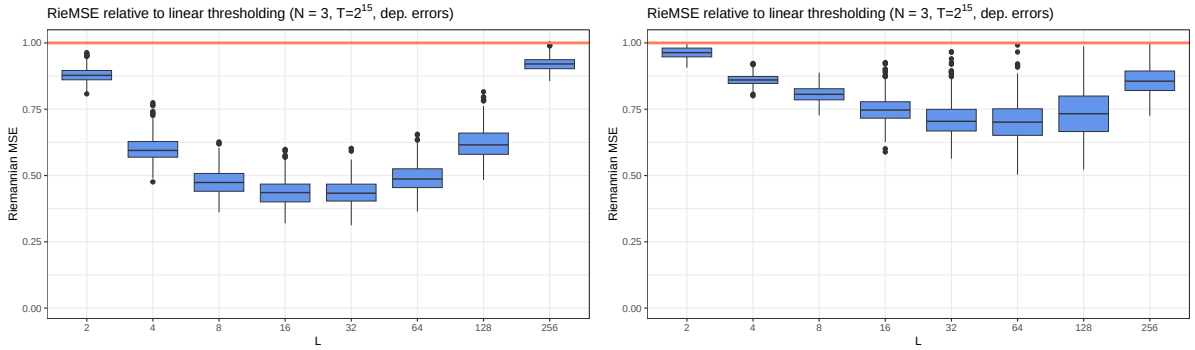


Figure 5: Performance of the Chi-squared thresholding estimator relative to linear thresholding, on empirical TVCMs as a function of window length L. Left: $A^{(1)}$, right: $A^{(2)}$.

$T = 2^{15} = 32,768$ observations, scaled to the unit interval and being displayed in Figure 6.

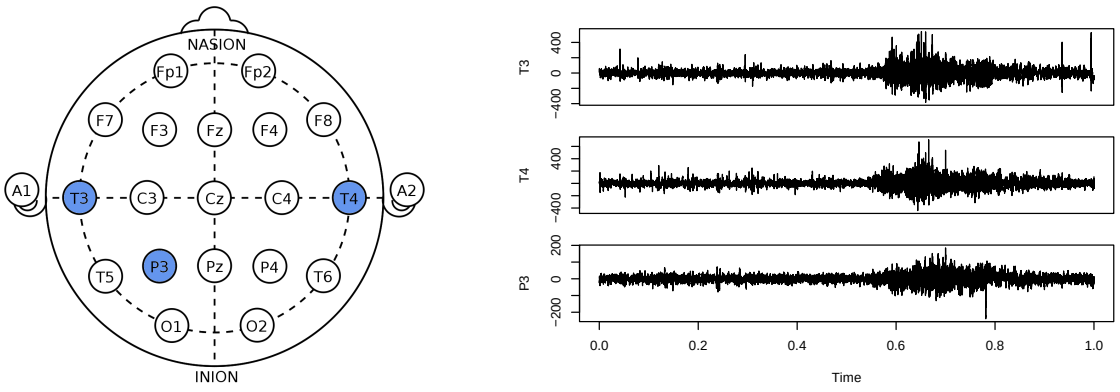


Figure 6: Left: layout of electrodes on the scalp, with considered channels in blue. Right: time series of considered channels.

Our analysis is based on the segmented (and bias-corrected) pilot estimator $\widetilde{S}_k^c$ with $L = 4$, hence we have a sample size of $n = T/L = 8192$, note that this $(\widetilde{S}_k^c)_{k=0}^{n-1}$ is displayed in Figure 7. We can observe that this raw curve estimator is very noisy: it is not evident, from a visual point of view, to extract patterns out of this data neither to compare the different elements. Increasing L, as in Figure 8, smoothes out the noise, however, it clearly also smoothes out some important local peaks which could carry relevant information for parts of the signal. In addition, it is quite clear that the variability behaves

differently before and during the seizure (which occurs roughly right before $u = 0.6$). Therefore, although smoothing with a large L seems to work reasonably well to approximate the behaviour before the seizure, it does not reflect well the rapid but steady evolution towards a highly oscillating behaviour of the brain activity during an epileptic seizure.

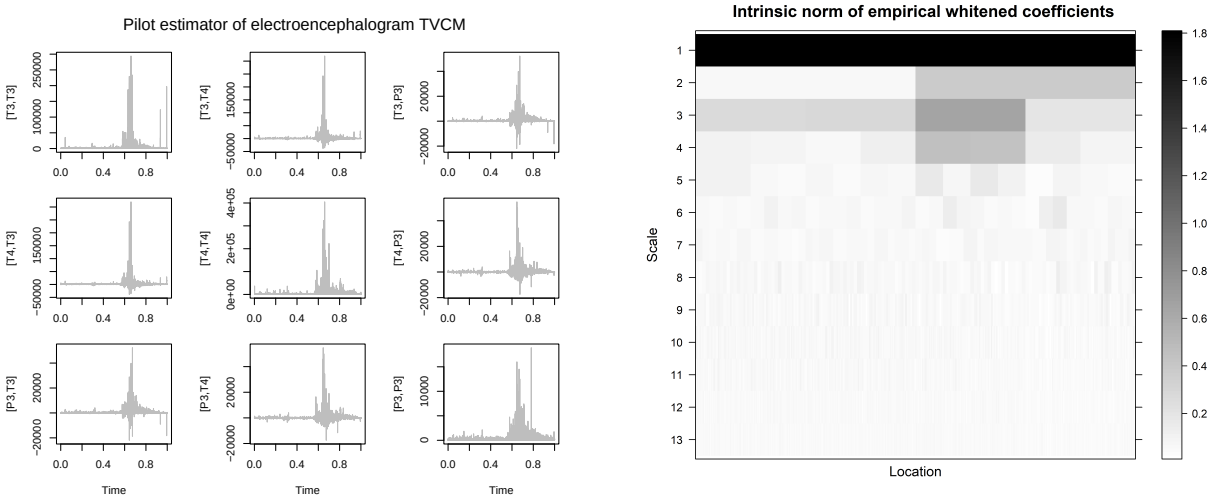


Figure 7: Left: pilot estimators $(\widetilde{S}_k^c)_{k=0}^{n-1}$. Right: pyramid of $(\|\mathfrak{D}_{j,k}\|_F)_{j,k}$.

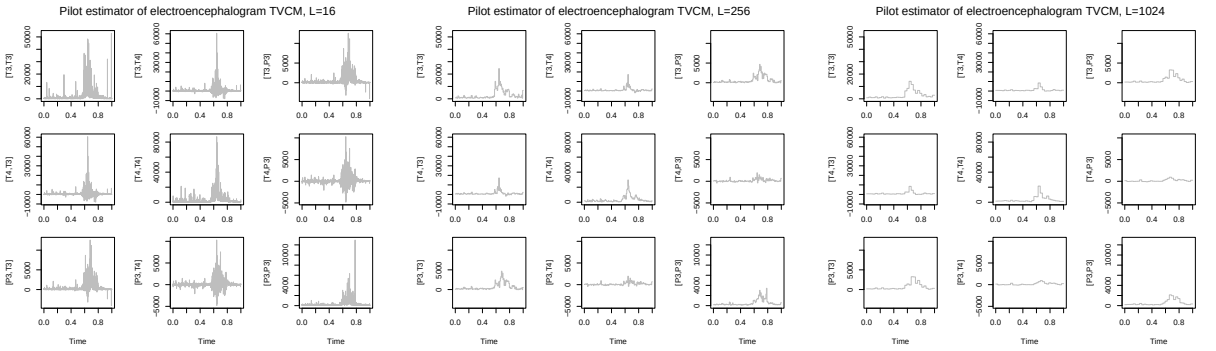


Figure 8: Example of pilot estimators $(\widetilde{S}_k^c)_{k=0}^{T/L-1}$ with fixed T and increasing L.

Once this inhomogeneous behaviour of the electroencephalogram's TVCM is established, it is quite natural to consider to construct the pilot estimator $\widetilde{S}_k^c$ with the smallest possible L in order to capture the possibly important local features of the signal and to smooth out the noise by means of nonlinear wavelet thresholding: comparing Figure 9 with Figure 10, our estimator $\widehat{S}_k^*$ has the advantage over the linear estimator that it is able to adequately estimate the essentially constant behaviour of the EEG before the seizure while being able to capture the local phenomena occurring during epilepsy. For its construction, in particular, we choose the transform order $N = 3$ and the χ^2 thresholding rule (of Section 3.2.3) to threshold coefficients within scales $\{J_0 + 1, \dots, J_1\}$ where $J_0 = 1$ and $J_1 = 6$, and allow some fine tuning of the threshold: we use $\tau = \alpha\, 2(\sqrt{p \log 2^j} + \log 2^j)$ where $\alpha = 1.2$ and $p = d(d+1)/2$ for $d = 3$ here. In the right hand part of Figure 10, we visualize the Frobenius norm of the empirical whitened wavelet coefficients after the nonlinear thresholding procedure. We remark that in particular, the pre-seizure period has globally low frequency content while during the seizure, higher frequencies occur in a localized and inhomogeneous manner (which confirms the first visual analysis of the pilot estimator). In particular, shortly before the onset of the seizure, there is a small amplitude oscillation at a frequency higher than for most of the pre-attack period which could indicate the beginning of the perturbation: this information is not at all easy detectable when looking just at the pilot estimator.

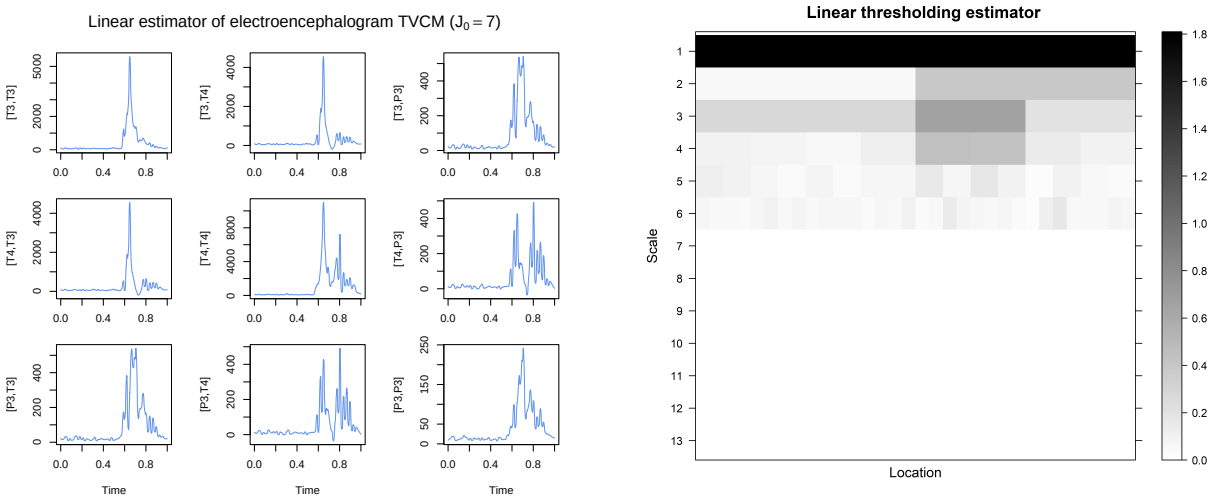


Figure 9: Left: linear estimator of the TVCM $(\widehat{S}_k)_{k=0}^{n-1}$. Right: pyramid of linearly thresholded whitened coefficients $(\|\widehat{\mathfrak{D}}_{j,k}\|_F)_{j,k}$.

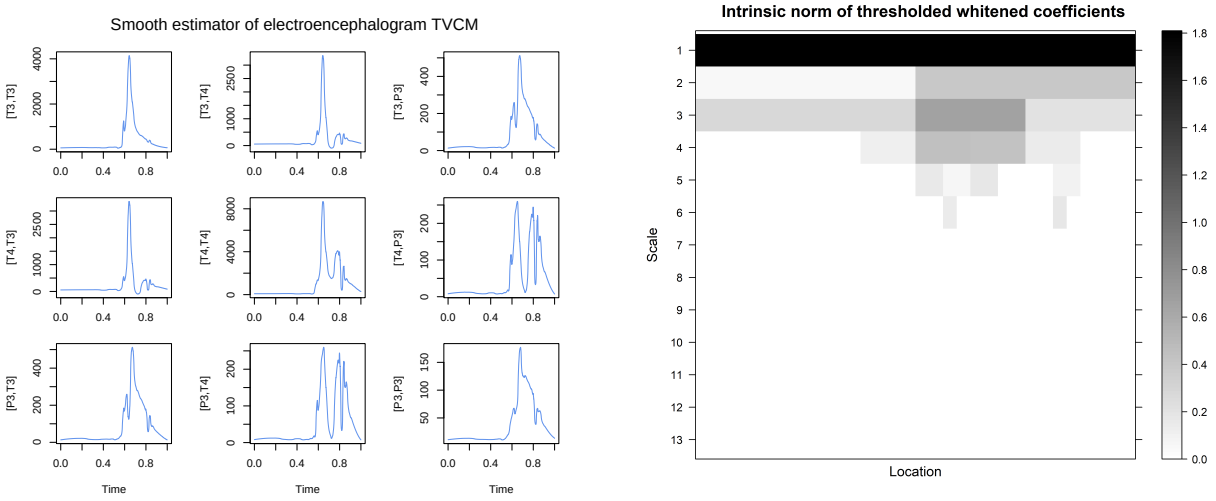


Figure 10: Left: nonlinear estimator of the TVCM $(\widehat{S}_k^*)_{k=0}^{n-1}$. Right: pyramid of nonlinearly thresholded whitened coefficients $(\|\widehat{\mathfrak{D}}_{j,k}^*\|_F)_{j,k}$.

In order to analyze the linear dependence between the channels, we can use the TVCM estimator $(\widehat{S}_k^*)_{k=0}^{n-1}$ to compute an estimator of the time-varying *correlation* matrix (TVcM) $(\widehat{R}_k^*)_{k=0}^{n-1}$, where for all k we have $\widehat{R}_k^* = \mathrm{Diag}^{-1/2}(\widehat{S}_k^*)\ \widehat{S}_k^*\ \mathrm{Diag}^{-1/2}(\widehat{S}_k^*)$: here Diag denotes the operation that maps the square matrix A to a diagonal matrix with the same diagonal elements as A. This TVcM is displayed in Figure 11. In addition, we display in Figure 12 the variances of each element together, both in their original scales and then on a comparable scale by normalizing them into $[0, 1]$. This will be useful to compare the amplitude of the channels variability, and to compare the onsets and possible lags between each other. Finally, in order to visualize better by which channels the covariances are dominated, we display pairwise variance-covariance plots in Figure 13.

What we can first observe in the analysis of the TVCM is that T4 has dominating variability compared to T3 and P3 – hence, it is dominating in the covariance with other channels. We observe that, while the onset of the seizure is almost perfectly synchronized between T3 and T4 – which are on opposite sides of the scalp but on the same latitude – it seems that the signal of P3, although closer to T3 on the scalp, is a lagged version of the other two. Interestingly, while its variability has the smallest amplitude

compared to other channels, P3 dominates the covariance with T3. Then, turning to the TvCM, it appears that T3 and T4 are positively correlated in the pre-seizure period – but as soon as epilepsy strikes, this correlation has a tendency to vanish (and even becomes negative during the peak of variability of those channels), and increases back when the attack stops. This effect is less visible when looking at the correlation between P3 and T4, which are less correlated in the first part. Regarding the correlation between P3 and T3, it has a tendency to increase during the seizure. We stress that this analysis on the correlations would be far from trivial when using only the pilot estimator to compute the TVcM, as it would be so noisy that not clear interpretation would be possible. Considering directly smoothing the empirical *correlation* matrices with an intrinsic wavelet transform would lead to a very different methodological problem because the geometry of these matrices is very different from that of $\mathrm{Sym}^+(d)$, with different tools that are more difficult to obtain in closed-form (see [25]).

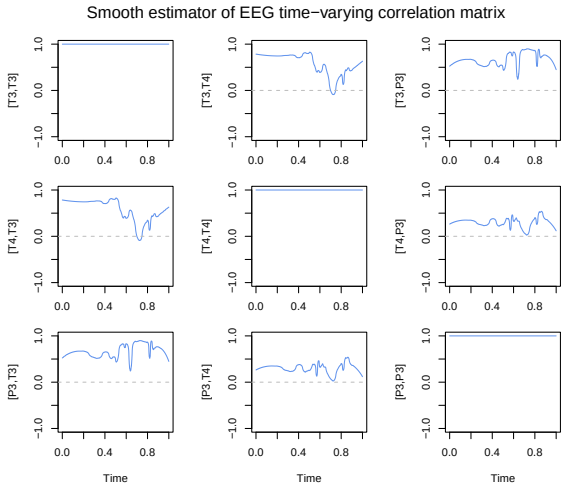


Figure 11: Smooth nonlinear wavelet estimator of the TVcM: $(\widehat{R}_k^*)_{k=0}^{n-1}$.

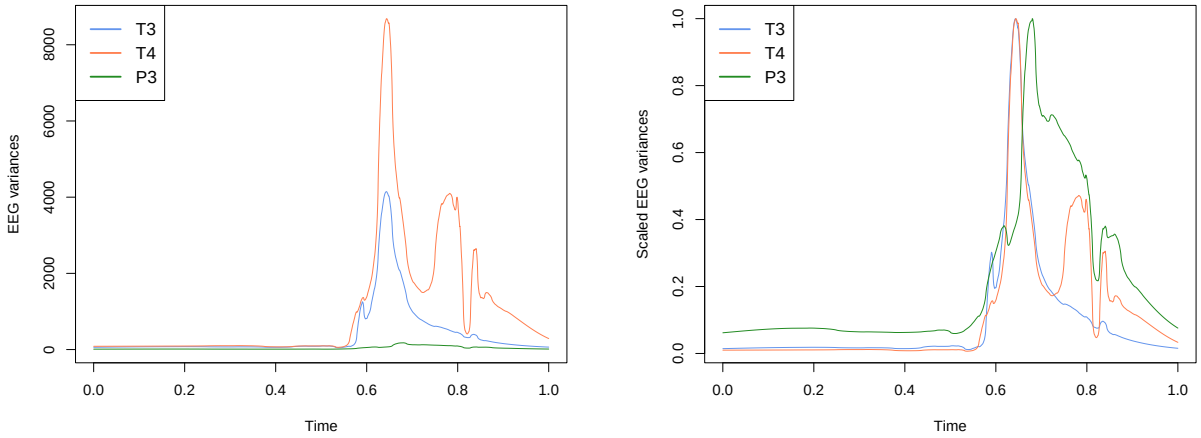


Figure 12: Left: estimated time-varying variances. Right: estimated time-varying variances, on a comparable scale $[0, 1]$.

Compared to previous investigations of this EEG signal based on the time-varying spectral density matrix [22, 12], our approach benefits from the possibility of an increased time resolution. Indeed, comparing the onset of the seizure is visually easier with a very fine time grid: we recall that (finite-sample) time-frequency analyses inherently cannot provide simultaneously both high time resolution and stable estimates over frequency. Furthermore, detection of the onset is more direct because with

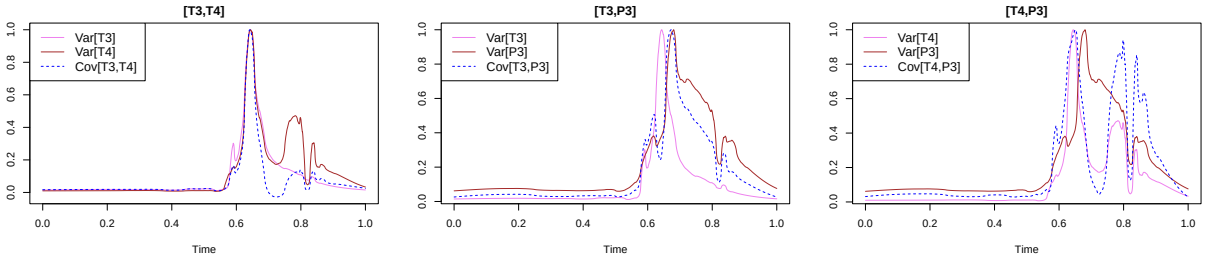


Figure 13: Pairwise comparisons of the variances, with their corresponding covariance (all rescaled to allow for visual evidence).

TVCM the variability of the series is more compactly represented in the time domain than if split onto multiple frequencies in the spectral domain.

We conclude that not only is the intrinsic wavelet transform for SPD matrices an efficient tool for denoising TVCMs, but it also provides a more reliable visual interpretation on the variability and linear dependence of time series, by inspecting the elements of the smooth curve (in a global or pairwise manner), as well as the underlying TVcM. Additionally, inspecting the scalogram representing the norm of the activated coefficients is a convenient way to get an overview of the "frequency" content of the curve at different locations. This last tool can be seen as a compression of the information in the underlying curves, and might be helpful for experts in the domain of interest to investigate previously unobserved features of the signal at specific moments in time.

5 Conclusion

In this paper, we studied estimation of time-varying covariance matrices by denoising sequences of sample covariance matrices of time series over short time windows. This is done by nonlinear thresholding the coefficients of the intrinsic average-interpolation wavelet transform for curves in the log-Euclidean SPD manifold: in particular, we developed new nonlinear thresholding rules which go beyond the trace thresholding of [11], exploiting the algebraic structure of the intrinsic signal plus noise model in the log-Euclidean setting. On the theoretical side, we contribute by a new result on IMSE-consistency, including rates of convergence, for a class of non-linear threshold estimators for SPD matrix-valued curves in non-Euclidean geometries. We complement this generic result by studying the asymptotic behaviour of both linearly and non-linearly denoising the aforementioned sample covariance matrices. We propose a bias-correction which we show to be theoretically justified and empirically very satisfying. Extensive simulation studies for data generated from a variety of pertinent signal-plus-noise models confirm our theoretical findings. An application to a transient EEG shows the interest of our proposed nonlinear estimation method to simultaneously provide for: (i) increased time-resolution compared to linear smoothing (and to existing time-frequency analyses); (ii) preserving positive-definiteness and invariance with respect to unitary congruent transformations, despite its non-linear nature; (iii) reproducing all the desirable features of existing first-generation wavelet thresholding methods (including fast implementation).

In future investigations it will be interesting to study *statistical inference* for our *nonlinear* wavelet estimator: [23] have successfully developed confidence regions for SPD matrix-valued curves in a log-Euclidean manifold, which – for the *linear* estimator only – are based on both asymptotic normality and some well-working wild bootstrap approach. Whereas proving the former is already a quite difficult task for first generation nonlinear thresholding, we believe that proposing bootstrap-based inference for our estimator could be an achievable and interesting contribution. Another though quite elaborated investigation concerns the question to what extent our newly proposed thresholding schemes which go beyond trace thresholding could be worked out for other Riemannian geometries (such as the affine-invariant of [11]), to allow, among others, an extension of our approach here to denoising multivariate

spectral densities: in particular from an applied point of view (of, e.g., denoising brain data) we strongly believe that most of our technology can be transferred to this situation.

*Acknowledgments

The authors would like to thank Stéphane Lhaut, Daniel Rademacher, Christian Hafner and Estelle Massart for their useful comments on this work. Further, we thank Hernando Ombao for having made available the EEG data collected by Dr. Beth Malow at the Epilepsy Disorder Laboratory (University of Michigan).

References

[1] ACHILLES, R., AND BONFIGLIOLI, A. The early proofs of the theorem of Campbell, Baker, Hausdorff, and Dynkin. *Archive for History of Exact Sciences 66*, 3 (2012), 295–358.

[2] ANTONIADIS, A., AND LAVERGNE, C. *Variance Function Estimation in Regression by Wavelet Methods*. Springer New York, 1995, pp. 31–42.

[3] ARNAL, A., CASAS, F., AND CHIRALT, C. A note on the Baker–Campbell–Hausdorff series in terms of right-nested commutators. *Mediterranean Journal of Mathematics 18*, 2 (2021).

[4] ARNAUDON, M., BARBARESCO, F., AND YANG, L. Riemannian medians and means with applications to radar signal processing. *IEEE Journal of Selected Topics in Signal Processing 7*, 4 (2013), 595–604.

[5] ARSIGNY, V., FILLARD, P., PENNEC, X., AND AYACHE, N. Geometric means in a novel vector space structure on symmetric positive-definite matrices. *SIAM Journal on Matrix Analysis and Applications 29*, 1 (2007), 328–347.

[6] BOUCHERON, S., LUGOSI, G., AND MASSART, P. *Concentration Inequalities*. Oxford University Press, 2013.

[7] BOUMAL, N., AND ABSIL, P.-A. Discrete regression methods on the cone of positive-definite matrices. In *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (2011), pp. 4232–4235.

[8] BRILLINGER, D. R. *Time Series: Data Analysis and Theory*. Society for Industrial and Applied Mathematics, 2001.

[9] CASEIRO, R., MARTINS, P., HENRIQUES, J. F., AND BATISTA, J. A nonparametric Riemannian framework on tensor field with application to foreground segmentation. *Pattern Recognition 45*, 11 (2012), 3997–4017.

[10] CHAU, J. *Advances in Spectral Analysis for Multivariate, Nonstationary and Replicated Time Series*. PhD thesis, Faculté des sciences, UCLouvain, 2018.

[11] CHAU, J., AND VON SACHS, R. Intrinsic wavelet regression for curves of Hermitian positive definite matrices. *Journal of the American Statistical Association 116*, 534 (2021), 819–832.

[12] CHAU, J., AND VON SACHS, R. Time-varying spectral matrix estimation via intrinsic wavelet regression for surfaces of Hermitian positive definite matrices. *Computational Statistics and Data Analysis 174* (2022), 107477.

[13] DONOHO, D. L. Smooth wavelet decompositions with blocky coefficient kernels. In *Recent Advances in Wavelet Analysis*. Academic Press, New-York, 1993, pp. 259–308.

[14] DONOHO, D. L. CART and best-ortho-basis: a connection. *The Annals of Statistics 25*, 5 (1997).

[15] DONOHO, D. L., AND JOHNSTONE, I. M. Ideal spatial adaptation by wavelet shrinkage. *Biometrika 81*, 3 (1994), 425–455.

[16] DRYDEN, I. L., KOLOYDENKO, A., AND ZHOU, D. Non-Euclidean statistics for covariance matrices, with applications to diffusion tensor imaging. *The Annals of Applied Statistics 3*, 3 (2009), 1102 – 1123.

[17] FILLARD, P., ARSIGNY, V., PENNEC, X., HAYASHI, K. M., THOMPSON, P. M., AND AYACHE, N. Measuring brain variability by extrapolating sparse tensor fields measured on sulcal lines. *NeuroImage 34*, 2 (2007), 639–650.

[18] FLETCHER, P. T., AND JOSHI, S. Riemannian geometry for the statistical analysis of diffusion tensor data. *Signal Processing 87*, 2 (Feb. 2007), 250–262.

[19] FRISTON, K. J. Functional and effective connectivity: A review. *Brain Connectivity 1*, 1 (2011), 13–36.

[20] JOHNSTONE, I. M., AND SILVERMAN, B. W. Wavelet threshold estimators for data with correlated noise. *Journal of the Royal Statistical Society Series B: Statistical Methodology 59*, 2 (July 1997), 319–351.

[21] MUIRHEAD, R. J. *Aspects of Multivariate Statistical Theory*. John Wiley & Sons, Inc., 1982.

[22] OMBAO, H., VON SACHS, R., AND GUO, W. SLEX analysis of multivariate nonstationary time series. *JASA 100*, 470 (2005), 519–531.

[23] RADEMACHER, D., KREBS, J., AND VON SACHS, R. Statistical inference for wavelet curve estimators of symmetric positive definite matrices. *J. Statistical Planing and Inference* (2023). *published online: JSPI 104160*.

[24] RAHMAN, I. U., DRORI, I., STODDEN, V. C., DONOHO, D. L., AND SCHRÖDER, P. Multiscale representations for manifold-valued data. *Multiscale Modeling & Simulation 4*, 4 (2005), 1201–1232.

[25] THANWERDAS, Y., AND PENNEC, X. Theoretically and computationally convenient geometries on full-rank correlation matrices. *SIAM Journal on Matrix Analysis and Applications 43*, 4 (2022), 1851–1872.

[26] THANWERDAS, Y., AND PENNEC, X. O(n)-invariant Riemannian metrics on SPD matrices. *Linear Algebra and its Applications 661* (2023), 163–201.

[27] WACHEL, Y. Wavelet Thresholding for Symmetric Positive-Definite Matrix-Valued Curves in a Log-Euclidean Manifold. Master's thesis, Université catholique de Louvain, supervisor: Rainer von Sachs, 2022.

[28] YUAN, Y., ZHU, H., LIN, W., AND MARRON, J. S. Local polynomial regression for symmetric positive definite matrices. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 74*, 4 (2012), 697–719.

[29] ZHU, H., CHEN, Y., IBRAHIM, J. G., LI, Y., HALL, C., AND LIN, W. Intrinsic regression models for positive-definite matrices with applications to diffusion tensor imaging. *Journal of the American Statistical Association 104*, 487 (2009), 1203–1212.

A Symmetric Baker-Campbell-Hausdorff formula

This short section develops the background of how to derive our important equation (20) on rewriting the canonical model of empirical covariance matrices $\widetilde{S}_k = S_k^{1/2} W_k S_k^{1/2}$.

Define the Lie bracket (often called commutator) between two symmetric matrices: $[X, Y] = XY - YX$, which satisfies the following properties:

- $[A + B, C] = [A, C] + [B, C]$ (bilinearity),
- $[A, A] = 0$,
- $[A, B] = -[B, A]$ (antisymmetry),
- $[A, [B, C]] + [B, [C, A]] + [C, [A, B]] = 0$ (Jacobi identity).

Recall we have $\widetilde{S}_k = S_k^{1/2} W_k S_k^{1/2}$. In order to exploit simultaneously asymptotic properties of W_k and the assumed log-additive model, we can use the symmetric Baker-Campbell-Hausdorff (BCH) formula. Let X and Y be symmetric matrices, then the symmetric BCH formula can be defined as in [3]:

$$\exp(1/2\ X)\exp(Y)\exp(1/2\ X) = \exp\left(\sum_{h=0}^{\infty} \frac{(-1)^h}{2^h h!} \mathrm{ad}_X^h\big(\Phi(X,Y)\big)\right) , \tag{26}$$

where

$$\mathrm{ad}_X^0(Y) = Y\ , \quad \mathrm{ad}_X(Y) = [X, Y]\ , \quad \mathrm{ad}_X^j(Y) = [X, \mathrm{ad}_X^{j-1}(Y)]\ ;$$

and where we use Dynkin's formula of the BCH series [1]:

$$\Phi(X,Y) = \sum_{n=1}^{\infty} \frac{(-1)^{n-1}}{n} \sum_{\substack{r_1+s_1>0 \\ \vdots \\ r_n+s_n>0}} \frac{[X^{r_1}Y^{s_1}X^{r_2}Y^{s_2}\cdots X^{r_n}Y^{s_n}]}{\left(\sum_{j=1}^n (r_j+s_j)\right)\cdot \prod_{i=1}^n r_i! s_i!} \tag{27}$$

where the sum is performed over all nonnegative values of s_i and r_i, and the following notation has been used:

$$[X^{r_1}Y^{s_1}\cdots X^{r_n}Y^{s_n}] = [\underbrace{X,[X,\cdots[X}_{r_1},\underbrace{[Y,[Y,\cdots[Y}_{s_1},\cdots\underbrace{[X,[X,\cdots[X}_{r_n},\underbrace{[Y,[Y,\cdots Y}_{s_n}]]\cdots]]$$

with the understanding that $[X] = X$. The first few terms of this complicated combinatorial formula are as follows:

$$\Phi(X,Y) = X + Y + \frac{1}{2}[X,Y] + \frac{1}{12}([X,[X,Y]] + [Y,[Y,X]]) - \frac{1}{24}[Y,[X,[X,Y]]] + \ldots$$

Denote $\widetilde{\Phi}(X,Y) = \Phi(X,Y) - (X+Y)$, which therefore only consists of (nested) Lie brackets. Consequently, identifying X with $\log(S_k)$ and Y with $\log(W_k)$, the canonical model for sample covariance matrices can be rewritten using Equation (26) as

$$\begin{aligned} \widetilde{S}_k &= \exp(1/2\log(S_k))\exp(\log(W_k))\exp(1/2\log(S_k)) \\ &= \exp\big(\log(S_k) + \log(W_k) + r_k\big)\ , \end{aligned} \tag{28}$$

where we let

$$r_k := \widetilde{\Phi}(\log(S_k), \log(W_k)) + \sum_{h\geq 1} \frac{(-1)^h}{2^h h!} \mathrm{ad}_{\log S_k}^h\big(\Phi(\log S_k, \log W_k)\big)\ . \tag{29}$$

Note that whenever $\log(S_k)$ and $\log(W_k)$ commute, $r_k = 0$; however in the finite sample size setting, this is generally not the case – it is then a quantitative measure of the deviation from commutativity between the signal and the noise. In the Supplementary Material, proof of Proposition 2.2, we treat the asymptotic behaviour of r_k.

Supplementary Material: Nonlinear wavelet thresholding estimation of time-varying covariance matrices in a log-Euclidean manifold

Gabriel Bailly Rainer von Sachs

Institute of Statistics, Biostatistics and Actuarial Sciences (ISBA), LIDAM
UCLouvain

Contents

B Moments of the pilot estimator **2**

C Asymptotic properties of intrinsic average-interpolating wavelets **2**
C.1 Serially independent case 2
C.2 Serially dependent case 4

D Proofs of main results **5**
D.1 Proof of Proposition 1 6
D.2 Proof of Proposition 2 7
D.3 Proof of Proposition 3 8
D.4 Proof of Theorem 1 8
D.5 Proof of Proposition 4 8
D.6 Proof of Proposition 5 11
D.7 Proof of Proposition 6 11
D.8 Proof of Theorem 2 12
D.9 Proof of Corollary 1 16

E Our pilot estimator for subsequent non-linear wavelet thresholding **17**

F Further simulation results **19**
F.1 Log-normal noise 20
F.2 Wavelet denoising of empirical TVCM 22

For notational convenience, for a matrix $A \in \mathbb{R}^{d\times d}$, we use the notation $A = O(\lambda_n)$ to mean that $A = \lambda_n E$ with a constant λ_n and a deterministic matrix $E \in \mathbb{R}^{d\times d}$. Note that this is equivalent as meaning that $\mathrm{Tr}(A) = \lambda_n \mathrm{Tr}(E)$, i.e. $\mathrm{Tr}(A) = O(\lambda) \iff A = O(\lambda)$. Additionally, for a random matrix $A \in \mathbb{R}^{d\times d}$, we denote $A = O_{\mathbb{P}}(\lambda_n)$ to mean $A = \lambda_n E_n$ with a constant λ_n and $E_n = O_{\mathbb{P}}(1)$ a random error matrix which is bounded in probability, i.e.

$$\forall \epsilon > 0,\ \exists a > 0,\ \exists N \in \mathbb{N}, \quad \inf_{n \geq N} \mathbb{P}(\|E_n\|_F \leq a) \geq 1 - \epsilon\ .$$

B Moments of the pilot estimator

Recall that our pilot estimators $\widetilde{S}_k$ write as $\widetilde{S}_k = S_k^{1/2} W_k S_k^{1/2}$, hence to bound their central moments of arbitrary order, it is sufficient to investigate those of W_k.

Proposition B.1. *We assume that the series $(X_t)_{t=1}^T$ are generated by model* (16) *and satisfy Assumption 1. Consequently, for all j,*

$$\mathbf{E}[(W_k - I_d)^j] = O(L^{-1}) \ . \tag{B1}$$

Proof. Omitting the index k to ease the notation and letting $w_{a,b} := [W]_{a,b} = L^{-1} \sum_{i=1}^L \varepsilon_{t_i,a} \varepsilon_{t_i,b}$ without loss of generality, we find that the elements of $\mathbf{E}[(W - I_d)^j]$ satisfy the convergence rate because, using the notation

$${}^c w_{a,b}^j := [(W - I_d)^j]_{a,b} = \sum_{i_1=1}^d \sum_{i_2=1}^d \cdots \sum_{i_{j-1}=1}^d (w_{a,i_1} - \delta_{a,i_1})(w_{i_1,i_2} - \delta_{i_1,i_2}) \cdots (w_{i_{j-1},b} - \delta_{i_{j-1},b}) \ ,$$

we get that $\mathbf{E}[{}^c w_{a,b}^j]$ involves mixed moments of ε order 2 up to order j. Then we show that for mixed moments of order h,

$$\begin{aligned}
\mathbf{E}[w_{a,i_1} w_{i_1,i_2} \cdots w_{i_{h-1},b}] &= L^{-h} \sum_{s_1=1}^L \sum_{s_2=1}^L \cdots \sum_{s_h=1}^L \mathbf{E}[\varepsilon_{t_{s_1},a} \varepsilon_{t_{s_1},i_1} \varepsilon_{t_{s_2},i_1} \varepsilon_{t_{s_2},i_2} \cdots \varepsilon_{t_{s_h},b}] \\
&= L^{-h} \sum_{s_1=1}^L \sum_{s_2=1}^L \cdots \sum_{s_{h-1}=1}^L O(1) \\
&= O(L^{-1}) \ ,
\end{aligned}$$

where we used Assumption 1 such that at least one of the h sums is bounded as $L \to \infty$. This holds for all moments involved in $\mathbf{E}[{}^c w_{a,b}^j]$, hence indeed the result holds for the matrix $\mathbf{E}[(W - I_d)^j]$. □

This result will be used in the proof of Proposition 2 to show that $\mathbf{E}[\log^j(W_k)] = O(L^{-1})$, which is necessary to treat the intrinsic bias of $\widetilde{S}_k$ in estimating S_k in the log-Euclidean metric, i.e the bias of $\log \widetilde{S}_k$ in estimating $\log(S_k)$ in the Euclidean metric. Motivated from the situation of Gaussian independent $(X_t)_{t=1}^T$ we reckon that the decay rate of the j-th central moment in Equation (B1) is faster, presumably of order $O(L^{-\lceil j/2 \rceil})$. However, the rate deay of Equation (B1) will be sufficient for proving Proposition 2.

C Asymptotic properties of intrinsic average-interpolating wavelets

We observe a sequence $(\widetilde{S}_k)_k$ of random SPD matrices and we naturally assume, throughout this section, that model

$$\widetilde{S}_k = \exp\left(\log(S_k) + \xi_k\right)$$

holds. We present in this section the asymptotic properties of average-interpolation wavelets in this framework: in the first place, we recall the results from (Chau and von Sachs, 2021, Section 3) assuming the matrices are independent over time. In order to relax this assumption and thereby allowing for dependence, we prove in the second subsection that these results also hold when the observed random matrices are not serially independent.

C.1 Serially independent case

We recall the results from (Chau and von Sachs, 2021, Section 3), which were derived under the assumption of independence. While the original proofs were done for $\mathrm{Sym}^+(d)$ equipped with the affine-

invariant Riemannian metric, they naturally extend to the log-Euclidean metric.

Let $\mathrm{Ave}(\{X_i\}_i)$ be the intrinsic average of a sequence $X_1, \ldots, X_n$ with values in $(\mathrm{Sym}^+(d), \delta_{\mathrm{L}})$:

$$\mathrm{Ave}(\{X_i\}_{i=1}^n) = \frac{1}{n}\sum_{i=1}^n \log(X_i) \ .$$

We recall the midpoint operator of this sequence, defined recursively by

$$\mu_n := \mu_n(X_1, \ldots, X_n) = \mathrm{Ave}(\{\mu_{n/2}(X_1, \ldots, X_{n/2}), \mu_{n/2}(X_{n/2+1}, \ldots, X_n)\}) \ .$$

Proposition C.1 (Convergence of midpoint operator)**.** *Let $X_1, \ldots, X_n \overset{i.i.d.}{\sim} \nu$ where $\nu \in \mathcal{P}_2(\mathrm{Sym}(d))$, with $\mathbb{E}_\nu[X_i] = \mu$ and $n = 2^J$ for some $J > 0$. Then,*

$$\mathbf{E}[\delta_{\mathrm{L}}^2(\mu_n, \mu)] = O(n^{-1}) \ .$$

In particular, $\mu_n \overset{\mathbb{P}}{\to} \mu$ as $n \to \infty$, where convergence holds with respect to the log-Euclidean distance.

The following results will then be intensely used in the proof of Theorem 2. We first begin by recalling that the wavelet coefficients of smooth curves decay over scales with a known rate.

Proposition C.2 (Coefficient decay of wavelet coefficients for smooth curves)**.** *Suppose $f\colon [0,1] \to Sym^+(d)$ is a smooth curve and set $M_{J,k}^f = \mathrm{Ave}_{I_{J,k}}(f)$ with $I_{J,k} = 2^{-J}[k, k+1]$ for $k = 0, \ldots, 2^J - 1$ as finest scale midpoints. Denote $(\mathfrak{D}_{j,k}^f)_{j,k}$ the whitened wavelet coefficients (of the curve f) as in (11) obtained from repeating the forward wavelet transform with refinement order $N = 2M + 1$ up the scale $j = 0$.*

If f has existing covariant derivatives of order N or higher, then, for each scale $j > 0$ and location $k \in \{0, \ldots, 2^{j-1} - 1\}$,

$$\|\mathfrak{D}_{j,k}^f\|_F \lesssim 2^{-j/2}2^{-jN}.$$

The next two results recall consistency (with rates) first for a given empirical whitened wavelet coefficients to converge to $\mathfrak{D}_{j,k}^f$, and second, for the linear wavelet estimator of the curve f – as to be found in (Chau and von Sachs, 2021, Proposition 3.3 and Theorem 3.4), even for more general Riemannian metrics.

Proposition C.3 (Consistency of whitened wavelet coefficients)**.** *Given a refinement order N, denote $(\mathfrak{D}_{j,k}^f)$ and $(\mathfrak{D}_{j,k})$ the true and empirical whitened wavelet coefficients based on an observed sequence $(M_{J,k})_k$ on scale $J = \log_2(n)$. Then, for each fixed scale $j > 0$ and location k, it holds that*

$$\mathbf{E}[\|\mathfrak{D}_{j,k} - \mathfrak{D}_{j,k}^f\|_F^2] = O(n^{-1}) \ .$$

The linear estimator attains the optimal nonparametric rate, which is given in terms of the number of (covariant) derivatives of the SPD curve f, being equal to the filter parameter N.

Proposition C.4 (Consistency of linear estimator with IMSE-rate of convergence)**.** *Consider the wavelet estimator $\widehat{M}_{J,k}(J_0)$, based on a refinement scheme of order N and on (linear) thresholding on scale J_0. Let further $f(t)$ be a smooth curve of SPD matrices, possessing at least N continuous derivatives. Then,*

$$\mathbf{E}[\|\log \widehat{M}_{J,k}(J_0) - \mathbf{E}[\log \widehat{M}_{J,k}(J_0)]\|_F^2] = O(2^{-(J-J_0)})$$

and

$$\|\mathbf{E}[\log \widehat{M}_{J,k}(J_0)] - \log M_{J,k}^f\|_F = O(2^{-NJ_0}) \ .$$

In particular, the choice $J_0 = \lfloor J/(2N+1) \rfloor$ balances variance and squared bias, so

$$\mathbf{E}[\delta_{\mathrm{L}}(\widehat{M}_{J,k}(J_0), M_{J,k}^f)^2] = \mathbf{E}[\| \log \widehat{M}_{J,k}(J_0) - \log M_{J,k}^f \|_F^2] = O(n^{-2N/(2N+1)}) \,,$$

which implies that $n^{-1} \sum_{k=0}^{n-1} \mathbf{E}[\delta_{\mathrm{L}}(\widehat{M}_{J,k}(J_0), M_{J,k}^f)^2] = O(n^{-2N/(2N+1)})$.

Remark C.1. In order to derive a consistency result for the original SPD curve $f(t)$, we need to assume that the approximation by the averages $M_{J,k}^f$ is sufficiently close to the true curve f. For this purpose, we first set $f_n = \sum_{k=0}^{n-1} M_{J,k}^f \mathbb{1}_{I_{J,k}}$. Then we have for $t \in I_{J,k}$

$$\begin{aligned} \delta_{\mathrm{L}}(f_n(t), f(t)) &= \Big\| \frac{1}{2^{-J}} \int_{I_{J,k}} \log f(s) \, \mathrm{d}s - \log f(t) \Big\|_F \\ &\leq \sup_{s,t \in I_{J,k}} \Big\| \log f(s) - \log f(t) \Big\|_F = O(n^{-1}) \,, \end{aligned}$$

because f is smooth. Consequently, we obtain that

$$\int_0^1 \mathbf{E}[\delta_{\mathrm{L}}(\widehat{f}_n(t), f(t))^2] \, \mathrm{d}t = O(n^{-2N/(2N+1)}) \,.$$

The last result on asymptotic normality of the whitened wavelet coefficients (useful for motivating the universal thresholding) is taken from (Rademacher et al., 2022, Theorem 4.5):

Proposition C.5 (Asymptotic normality of the wavelet coefficients)**.** *Let the assumptions of Proposition C.3 be satisfied. Let (j,k) be arbitrary but fixed and let $L \leq k \leq 2^J - 1 - L$. Then*

$$2^{-j/2} n^{1/2} (\widehat{\mathfrak{D}}_{j,k} - \mathbf{E}[\widehat{\mathfrak{D}}_{j,k}])$$

converges to a zero mean normal distribution $\mathcal{N}(0, \Xi_{j,k})$ with asymptotic variance-covariance (operator) $\Xi_{j,k}$ (which essentially depends on the covariance tensor $\boldsymbol{\mathcal{C}}$ defined in (21)*).*

C.2 Serially dependent case

The previous results of this section were derived under the hypothesis that the averaged observations in $\mathrm{Sym}^+(d)$ are serially independent. Under the hypothesis of weak serial dependence, they however remain true thanks to the following proposition which is an adaptation of Proposition C.1. In particular, we use that the cumulant mixing condition (17) on the data generating process transfer from the time series observations to the sequence of pilot estimators $(\widetilde{S}_k)_k$ – which in the following proposition we abstractly denote by $(X_k)_k$.

Proposition C.6 (Convergence of midpoint operator for weakly dependent observations)**.** *Let $X_1, \ldots, X_n$ be weakly dependent observations where $X_k \sim \nu_{X_k} \in \mathcal{P}_2(\mathrm{Sym}^+(d))$, with $n = 2^J$ for some $J > 0$. Then,*

$$\mathbf{E}[\delta_{\mathrm{L}}^2(\mu_n, \mu)] = O(n^{-1}) \,.$$

In particular, $\mu_n \xrightarrow{\mathbb{P}} \mu$ as $n \to \infty$, where convergence holds with respect to the log-Euclidean distance.

Proof. Let $X_k \sim \nu_{X_k} \in \mathcal{P}_2(\mathrm{Sym}(d))$. By (Bhatia, 2007, Theorem 6.1.9), if $X_1, X_2, \mu \in \mathrm{Sym}^+(d)$ then for $t \in [0,1]$,

$$\delta_{\mathrm{L}}^2(\gamma(t; X_1, X_2), \mu) \leq (1-t)\delta_{\mathrm{L}}^2(X_1, \mu) + t\delta_{\mathrm{L}}^2(X_2, \mu) - t(1-t)\delta_{\mathrm{L}}^2(X_1, X_2) \,.$$

Setting $t = 1/2$, $\mu_2 = \gamma(1/2; X_1, X_2)$ and taking expectations (with respect to the joint distribution) on both sides,

$$\mathbf{E}_{X_1 X_2}[\delta_{\mathrm{L}}^2(\mu_2, \mu)] \leq \frac{1}{2}\mathbf{E}_{X_1}[\delta_{\mathrm{L}}^2(X_1, \mu)] + \frac{1}{2}\mathbf{E}_{X_2}[\delta_{\mathrm{L}}^2(X_2, \mu)] - \frac{1}{4}\mathbf{E}_{X_1 X_2}[\delta_{\mathrm{L}}^2(X_1, X_2)]$$

$$\leq \max_{i\in\{1,2\}} \mathbf{E}_{X_i}[\delta_{\mathrm{L}}^2(X_i,\mu)] - \frac{1}{4}\mathbf{E}_{X_1X_2}[\delta_{\mathrm{L}}^2(X_1,X_2)] \ . \tag{C2}$$

We can upper-bound the second term of this last inequality by showing

$$\begin{aligned}\mathbf{E}_{X_1X_2}[\delta_{\mathrm{L}}^2(X_1,X_2)] &= \mathbf{E}_{X_1}\left[\mathbf{E}_{X_2|X_1}[\delta_{\mathrm{L}}^2(X_1,X_2)|X_1]\right] \\ &\geq \mathbf{E}_{X_1}[\delta_{\mathrm{L}}^2(X_1,\mu)] + \mathbf{E}_{X_2}[\delta_{\mathrm{L}}^2(X_2,\mu)] \ ,\end{aligned}$$

where the inequality follows from the semi-parallelogram law Ho et al. (2013):

$$\int_{\mathrm{Sym}^+(d)} \left(\int_{\mathrm{Sym}^+(d)} \left(\delta_{\mathrm{L}}^2(x_2,x_1) - \delta_{\mathrm{L}}^2(x_2,\mu)\right) \mathrm{d}\nu_{X_2|X_1=x_1}(x_2) \right) \mathrm{d}\nu_{X_1}(x_1) \geq \int_{\mathrm{Sym}^+(d)} \delta_{\mathrm{L}}^2(x_1,\mu)\,\mathrm{d}\nu_{X_1}(x_1) \ ,$$

which implies that

$$\begin{aligned}\mathbf{E}_{X_1}\left[\mathbf{E}_{X_2|X_1}[\delta_{\mathrm{L}}^2(X_1,X_2) - \delta_{\mathrm{L}}^2(X_2,\mu)|X_1]\right] \geq \mathbf{E}_{X_1}[\delta_{\mathrm{L}}(X_1,\mu)] \quad &\iff \\ \mathbf{E}_{X_1}\left[\mathbf{E}_{X_2|X_1}[\delta_{\mathrm{L}}^2(X_1,X_2)|X_1]\right] - \mathbf{E}_{X_2}[\delta_{\mathrm{L}}^2(X_2,\mu)] \geq \mathbf{E}_{X_1}[\delta_{\mathrm{L}}(X_1,\mu)] \quad &\iff \\ \mathbf{E}_{X_1}\left[\mathbf{E}_{X_2|X_1}[\delta_{\mathrm{L}}^2(X_1,X_2)|X_1]\right] \geq \mathbf{E}_{X_1}[\delta_{\mathrm{L}}(X_1,\mu)] + \mathbf{E}_{X_2}[\delta_{\mathrm{L}}^2(X_2,\mu)] \ .&\end{aligned}$$

Consequently,

$$\begin{aligned}-\frac{1}{4}\mathbf{E}_{X_1X_2}[\delta_{\mathrm{L}}^2(X_1,X_2)] &\leq -\frac{1}{4}\mathbf{E}_{X_1}[\delta_{\mathrm{L}}^2(X_1,\mu)] - \frac{1}{4}\mathbf{E}_{X_2}[\delta_{\mathrm{L}}^2(X_2,\mu)] \\ &\leq -\frac{1}{2}\max_{i\in\{1,2\}} \mathbf{E}_{X_i}[\delta_{\mathrm{L}}^2(X_i,\mu)] \ .\end{aligned}$$

Plugging this result into Equation (C2), we obtain

$$\mathbf{E}_{X_1X_2}[\delta_{\mathrm{L}}^2(\mu_2,\mu)] \leq \frac{1}{2}\max_{i\in\{1,2\}} \mathbf{E}_{X_i}[\delta_{\mathrm{L}}^2(X_i,\mu)] \ .$$

Repeating the same argument for μ_4 we obtain

$$\mathbf{E}_{X_1X_2X_3X_4}[\delta_{\mathrm{L}}^2(\mu_4,\mu)] \leq \frac{1}{2}\mathbf{E}_{X_1X_2X_3X_4}[\delta_{\mathrm{L}}^2(\mu_2,\mu)] \leq \frac{1}{4}\max_{i\in\{1,\ldots,4\}} \mathbf{E}_{X_i}[\delta_{\mathrm{L}}^2(X_i,\mu)]$$

Denoting $\nu_{X_1X_2\cdots X_m} =: \nu_m$ and $\mathbf{E}_{X_1X_2\cdots X_m} =: \mathbf{E}_m$, we can repeat the argument over successive midpoints to find the upper bound

$$\mathbf{E}_n[\delta_{\mathrm{L}}^2(\mu_n,\mu)] \leq \frac{1}{2}\mathbf{E}_{n/2}[\delta_{\mathrm{L}}^2(\mu_{n/2},\mu)] \leq \ldots \leq \frac{1}{n}\max_{1\leq i\leq n} \mathbf{E}_{X_i}[\delta_{\mathrm{L}}^2(X_i,\mu)] \ ,$$

where under the assumption that $\nu_n \in \mathcal{P}_2(\mathrm{Sym}^+(d))$, the upper-bound is finite. By Markov's inequality, we can finally rewrite $\mu_n \xrightarrow{\mathbb{P}} \mu$ because $\forall\epsilon > 0$,

$$\mathbb{P}(\delta_{\mathrm{L}}(\mu_n,\mu) > \epsilon) \leq \frac{\mathbf{E}_n[\delta_{\mathrm{L}}^2(\mu_n,\mu)]}{\epsilon^2} \leq \frac{\max_{1\leq i\leq n} \mathbf{E}_{X_i}[\delta_{\mathrm{L}}^2(X_i,\mu)]}{n\epsilon^2} \to 0 \quad \text{as } n\to\infty \ . \qquad \square$$

D Proofs of main results

Before going into the proofs, let us recall a few useful results for asymptotic boundedness in the presence of matrix exponential, the matrix logarithm and its Fréchet derivative. The following lemmas also apply if we consider *stochastic* boundedness.

Lemma D.1 (Chau and von Sachs (2021))**.** *If $\lambda \to 0$ sufficiently small, $M_1 = \exp(M_2 + O(\lambda))$ implies*

$M_1 = \exp(M_2) + O(\lambda)$.

Proof. The result follows from the Taylor expansion of the matrix exponential:

$$\begin{aligned} M_1 &= \exp(M_2 + O(\lambda)) \\ &= \sum_{k=0}^{\infty} \frac{\big(M_2 + O(\lambda)\big)^k}{k!} \\ &= \sum_{k=0}^{\infty} \frac{(M_2)^k + O(\lambda)}{k!} \\ &= \sum_{k=0}^{\infty} \frac{(M_2)^k}{k!} + O(\lambda) \sum_{k=0}^{\infty} \frac{1}{k!} \\ &= \exp(M_2) + O(\lambda) \ . \end{aligned}$$

□

Lemma D.2 (Chau and von Sachs (2021))**.** *If $\lambda \to 0$ sufficiently small, then* $\log(M+O(\lambda)) = \log(M)+O(\lambda)$.

Proof. Let $\lambda E = O(\lambda)$ with $E \in \mathbb{R}^{d\times d}$ an arbitrary deterministic matrix. Rewrite $\log(M + \lambda E) = \log(M(\mathbf{I} + \lambda M^{-1}E))$. Using the Baker-Campbell-Hausdorff series (refer to Appendix A of the main document) with $X = \log(M)$ and $Y = \log(\mathbf{I} + \lambda M^{-1}E)$,

$$\log(M(\mathbf{I} + \lambda M^{-1}E)) = X + Y + \frac{1}{2}[X,Y] + \frac{1}{12}\big([X,[X,Y]] + [Y,[Y,X]]\big) + \dots$$

In particular,

$$\begin{aligned} [X,Y] &= [\log(M), \log(\mathbf{I} + \lambda M^{-1}E)] \\ &= \log(M)\log(\mathbf{I} + \lambda M^{-1}E) - \log(\mathbf{I} + \lambda M^{-1}E)\log(M) \\ &= \log(M)(\lambda M^{-1}E + O(\lambda^2)) - (\lambda M^{-1}E + O(\lambda^2))\log(M) \\ &= O(\lambda) \ , \end{aligned}$$

where we used the Mercator series expansion $\log(\mathbf{I} + \lambda M^{-1}E) = \lambda M^{-1}E + O(\lambda^2)$ for spectral radius $\rho(\lambda M^{-1}E) < 1$ and $\lambda \to 0$ sufficiently small (Higham, 2008, Section 11.3). Iterating the same argument over the higher-order commutators which are of order $O(\lambda)$, the desired result follows

$$\log(M + \lambda E) = \log(M) + \log(\mathbf{I} + \lambda M^{-1}E) + O(\lambda) = \log(M) + O(\lambda) \ .$$

□

Lemma D.3. *If $\lambda \to 0$ sufficiently small, then* $\mathbb{D}_{M_1}\log(M_2 + O(\lambda)) = \mathbb{D}_{M_1}\log(M_2) + O(\lambda)$.

Proof. We use the closed-form expression of the Fréchet derivative (recall its definition from Section 1.1):

$$\begin{aligned} \mathbb{D}_{M_1}\log(M_2 + O(\lambda)) &= U^\top * (Z \odot (U * (M_2 + O(\lambda)))) \\ &= U^\top * (Z \odot (U * (M_2))) + O(\lambda) \\ &= \mathbb{D}_{M_1}\log(M_2) + O(\lambda) \ . \end{aligned}$$

□

D.1 Proof of Proposition 1

Proof. By definition of X_t in model (16), such that $\widetilde{S}_k = S_k^{1/2} W_k S_k^{1/2}$. In the sequel, we write $A_L \xrightarrow{\mathbb{P}} A$ as $L \to \infty$ to denote that $\forall a > 0$, $\lim_{L\to\infty} \mathbb{P}(\|A_L - A\|_F > a) = 0$. Under Assumption 1,

$$\mathbb{P}(\|W_k - I_d\|_F > a) = \mathbb{P}(\mathrm{Tr}^{1/2}\big((W_k - I_d)^2\big) > a)$$

$$
\begin{aligned}
&\leq \frac{\mathbf{E}\left[\mathrm{Tr}\big((W_k - I_d)^2\big)\right]}{a^2} \\
&= \frac{\mathrm{Tr}\left(\mathbf{E}[(W_k - I_d)^2]\right)}{a^2} \\
&\lesssim L^{-1} \,,
\end{aligned}
$$

where the last step follows from the results of Proposition B.1. Taking the limit as $L \to \infty$ yields $W_k \xrightarrow{\mathbb{P}} I_d$; using that $W_k \mapsto S_k^{1/2} W_k S_k^{1/2}$ is a continuous map, by the continuous mapping theorem $\widetilde{S}_k \xrightarrow{\mathbb{P}} S_k$. □

D.2 Proof of Proposition 2

Proof. By definition of the Fréchet mean in the log-Euclidean manifold, and the development of equations (A1) and (A2) in Appendix A of the main document, we can write

$$
\begin{aligned}
\mathbb{E}[\widetilde{S}_k] &= \exp\left(\mathbf{E}[\log(\widetilde{S}_k)]\,\right) \\
&= \exp\left(\mathbf{E}[\log(S_k) + \log(W_k) + r_k]\,\right) \\
&= \exp\left(\log(S_k) + \mathbf{E}[\log(W_k)] + \mathbf{E}[r_k]\,\right) \,.
\end{aligned}
$$

The matrix logarithm of a SPD matrix A has a Taylor expansion as defined in (Higham, 2008, Theorem 4.7) which converges only if $0 < \lambda_{\max}(A) \leq 2$. By Proposition 1, $W_k \xrightarrow{\mathbb{P}} I_d$ as $L \to \infty$, which implies that $\lambda_{\max}(W_k) \xrightarrow{\mathbb{P}} 1$ by the continuous mapping theorem. Then, the following Taylor expansions of $\log(W_k)$ and $\log^2(W_k)$, respectively, converge when $L \to \infty$ (here, we understand convergence of a series of random variables to its limit in the mean-square sense).

$$
\log(W_k) = \sum_{h \geq 1} \frac{(-1)^{h+1}}{h} (W_k - I_d)^h \quad \text{and} \quad \log^2(W_k) = \sum_{h \geq 2} \frac{(\log^2)^{(h)}(1)}{h!} (W_k - I_d)^h
$$

By Proposition B.1, $\mathbf{E}[(W_k - I_d)^h] = O(L^{-1})$ and

$$
\begin{aligned}
\mathbf{E}[\log(W_k)] &= -\mathbf{E}[(W_k - I_d)^2] + \sum_{h \geq 3} \frac{(-1)^{h+1}}{h} \mathbf{E}[(W_k - I_d)^h] = O(L^{-1}) \,, \\
\mathbf{E}[\log^2(W_k)] &= \mathbf{E}[(W_k - I_d)^2] + \sum_{h \geq 3} \frac{(\log^2)^{(h)}(1)}{h!} \mathbf{E}[(W_k - I_d)^h] = O(L^{-1}) \,.
\end{aligned}
$$

More generally, because $(\log^j)^{(h)}(1) = 0$ for $h < j$ and $= 1$ for $h = j$, we have that $\mathbf{E}[\log^j(W_k)] = O(L^{-1})$. We can use these results to obtain

$$
\mathbf{E}\big[(\log(W_k) - \mathbf{E}[\log(W_k)])^2\big] = O(L^{-1}) \,.
$$

Hence, $\log(W_k) = \mathbf{E}[\log(W_k)] + \varepsilon_k$ where $\varepsilon_k = O_{\mathbb{P}}(L^{-1/2})$, $\mathbf{E}[\varepsilon_k] = 0$ and $\mathbf{E}[\varepsilon_k^2] = O(L^{-1})$.

We use that Lie brackets are bilinear and that $\mathbf{E}[\log(W_k)] = O(L^{-1})$ to show that

$$
\begin{aligned}
\big[\log(S_k), \log(W_k)\big] &= \big[\log(S_k), \mathbf{E}[\log(W_k)] + \varepsilon_k\big] \\
&= \big[\log(S_k), \mathbf{E}[\log(W_k)]\big] + \big[\log(S_k), \varepsilon_k\big] \\
&= \big[\log(S_k), O(L^{-1})\big] + \big[\log(S_k), O_{\mathbb{P}}(L^{-1/2})\big] \\
&= O_{\mathbb{P}}(L^{-1/2}) \,.
\end{aligned}
$$

Plugging this result into nested Lie brackets yields that they are all $O_{\mathbb{P}}(L^{-1/2})$, and consequently $r_k = O_{\mathbb{P}}(L^{-1/2})$.

We now show that expectations of (nested) Lie brackets also converge as $L \to \infty$. We exploit that they consist of terms which are products between $\log(S_k)$ and $\log(W_k)$, and denote such a product by P. We have that $\mathbf{E}[\mathrm{Tr}(P)] = O(L^{-1}) \iff \mathbf{E}[P] = O(L^{-1})$, hence we can show by some tedious but elementary calculations – using in particular properties and inequalities of the trace and Cauchy-Schwarz – that all expectations of nested Lie brackets are indeed $O(L^{-1})$. Plugging this result into $\mathbf{E}[r_k]$ implies that $\mathbf{E}[r_k] = O(L^{-1})$.

We finally conclude by showing that the intrinsic bias vanishes asymptotically:

$$\mathbb{E}[\widetilde{S}_k] = \exp\big(\log(S_k) + \mathbf{E}[\log(W_k)] + \mathbf{E}[r_k]\big) = S_k + O(L^{-1}) \,. \qquad \square$$

D.3 Proof of Proposition 3

Proof. Under the presented assumptions, indeed $W_k \sim \mathcal{W}_d(L, L^{-1}I_d)$. It was proved in (Smith, 2005, Theorem 7) that, with $\psi(\cdot)$ the digamma function,

$$\mathbf{E}[\log(W_k)] = \left(\log(2/L) + \frac{1}{d}\sum_{i=1}^{d}\psi\left(\frac{L-i+1}{2}\right)\right) I_d = c(d,L)I_d \,.$$

Plugging this result into $\mathbb{E}[\widetilde{S}_k^c]$ yields a bias-corrected version of the pilot estimator which for finite-sample L removes the essential contribution of $\log(W_k)$ to the bias of $\log(\widetilde{S}_k)$ (shown in Proposition 2 to be asymptotically of order $O(L^{-1})$). $\square$

D.4 Proof of Theorem 1

Proof. By Proposition C.6, the convergence rates of the linear thresholding estimator (Proposition C.4) hold if the additional bias of the pilot estimator from Proposition 2 is not of larger order than 2^{-NJ_0}, i.e. the bias of the linear wavelet estimator. We find that $L \sim T^{N/(3N+1)}$ by setting

$$\begin{aligned} L^{-1} \lesssim 2^{-NJ_0} = 2^{-NJ/(2N+1)} = \left(\frac{L}{T}\right)^{N/(2N+1)} \qquad &\iff \\ L^{-(3N+1)/(2N+1)} \lesssim T^{-N/(2N+1)} \qquad &\iff \\ L \gtrsim T^{N/(3N+1)} \,, \end{aligned}$$

and by recalling from Subsection 2.1 that L should grow with the smallest possible rate in order to adapt to localized features of the underlying curve. $\square$

D.5 Proof of Proposition 4

Proof. Most of the proof comes from the results of Wachel (2022). Let the finest-scale midpoints $(M_{Jk})_k$ be generated by the Riemannian log-additive model (12). Under such a model, we have (with $M_{j,k}$ the noisy scaling coefficients, superscript f indicating the true scaling coefficients, and superscript ξ indicating noise midpoints) for $j = 0, \ldots, J; k = 0, \ldots, 2^j - 1$:

$$\begin{aligned} \log M_{j,k} &= \log\big(\gamma(1/2; M_{j+1,2k}, M_{j+1,2k+1})\big) \\ &= \log\big(\exp\big\{\frac{1}{2}\log(M_{j+1,2k}) + \frac{1}{2}\log(M_{j+1,2k+1})\big\}\big) \\ &= \frac{1}{2}\log(M_{j+1,2k}) + \frac{1}{2}\log(M_{j+1,2k+1}) \end{aligned}$$

$$
\begin{aligned}
&\vdots \\
&= \frac{1}{2^{J-j}} \sum_{\ell=0}^{2^{J-j}-1} \log(M_{J,k(2^{J-j})+\ell}) \\
&= \frac{1}{2^{J-j}} \sum_{\ell=0}^{2^{J-j}-1} \Big(\log(S_{k(2^{J-j})+\ell}) + \xi_{k(2^{J-j})+\ell} \Big) \\
&=: \log(M^f_{j,k}) + \log(M^\xi_{j,k}) \ , \qquad \text{(D3)}
\end{aligned}
$$

as well as

$$
\begin{aligned}
\log \widetilde{M}_{j,2k+1} &= \sum_{m=-M}^{M} C_{N,2m+N} \log(M_{j-1,k+m}) \\
&= \sum_{m=-M}^{M} C_{N,2m+N} \log(M^f_{j-1,k+m}) + \sum_{m=-M}^{M} C_{N,2m+N} \log(M^\xi_{j-1,k+m}) \\
&= \log(\widetilde{M}^f_{j,2k+1}) + \log(\widetilde{M}^\xi_{j,2k+1}) \ . \qquad \text{(D4)}
\end{aligned}
$$

Using relations (D3) and (D4) in the following yields a matrix signal-plus-noise model for the whitened wavelet coefficients:

$$
\begin{aligned}
\mathfrak{D}_{jk} &= 2^{-j/2}\big(\log(M_{j,2k+1}) - \log(\widetilde{M}_{j,2k+1}) \big) \\
&= 2^{-j/2}\big(\log(M^f_{j,2k+1}) - \log(\widetilde{M}^f_{j,2k+1}) \big) + 2^{-j/2}\big(\log(M^\xi_{j,2k+1}) - \log(\widetilde{M}^\xi_{j,2k+1}) \big) \\
&= \mathfrak{D}^f_{jk} + \mathfrak{D}^\xi_{jk} \ . \qquad \text{(D5)}
\end{aligned}
$$

Following from this result, we get that the expectation of the empirical whitened coefficients is the corresponding true whitened coefficient:

$$
\begin{aligned}
\mathbf{E}[\mathfrak{D}_{jk}] &= \mathbf{E}[\mathfrak{D}^f_{jk}] + \mathbf{E}[\mathfrak{D}^\xi_{jk}] \\
&= \mathfrak{D}^f_{jk} + \mathbf{E}\big[2^{-j/2}(\log M^\xi_{j,2k+1} - \log \widetilde{M}^\xi_{j,2k+1})\big] \\
&= \mathfrak{D}^f_{jk} + 2^{-j/2}\mathbf{E}\left[\left(\frac{1}{2^{J-j}} \sum_{\ell=0}^{2^{J-j}-1} \xi_{k(2^{J-j})+\ell} \right) - \left(\sum_{m=-M}^{M} C_{N,2m+N} \log M^\xi_{j-1,k+m} \right) \right] \\
&= \mathfrak{D}^f_{jk} + \left(\frac{2^{-j/2}}{2^{J-j}} \sum_{\ell=0}^{2^{J-j}-1} \mathbf{E}\left[\xi_{k(2^{J-j})+\ell} \right] \right) - \left(2^{-j/2} \sum_{m=-M}^{M} C_{N,2m+N} \mathbf{E}\left[\log M^\xi_{j-1,k+m} \right] \right) \\
&= \mathfrak{D}^f_{jk} \ ,
\end{aligned}
$$

where in the last step we used for the second term that $\mathbf{E}[\xi_k] = \mathbf{0}$ for all k, and for the third term we use the same argument of the repeated noise midpoints, also yielding $\mathbf{E}[\log M^\xi_{j-1,k}] = \mathbf{0}$ for all k.

For the variance, we have

$$
\begin{aligned}
\mathrm{Var}[\mathfrak{D}_{j,k}] &= \mathrm{Var}\left[2^{-j/2}\Big(\log M_{j,2k+1} - \log \widetilde{M}_{j,2k+1} \Big) \right] \\
&= 2^{-j}\mathrm{Var}\left[\log M^f_{j,2k+1} + \log M^\xi_{j,2k+1} - \log \widetilde{M}^f_{j,2k+1} - \log \widetilde{M}^\xi_{j,2k+1} \right] \\
&= 2^{-j}\mathrm{Var}\left[\log M^\xi_{j,2k+1} - \log \widetilde{M}^\xi_{j,2k+1} \right]
\end{aligned}
$$

$$
\begin{aligned}
&= 2^{-j}\mathrm{Var}\left[\log M^{\xi}_{j,2k+1} - \sum_{m=-M}^{M} C_{N,2m+N}\log M^{\xi}_{j-1,k+m}\right] \\
&\overset{(*)}{=} 2^{-j}\mathrm{Var}\left[\log M^{\xi}_{j,2k+1} - \log M^{\xi}_{j-1,k} - \sum_{\substack{-M\le m\le M;\\ m\neq 0}} C_{N,2m+N}\log M^{\xi}_{j-1,k+m}\right] \\
&= 2^{-j}\mathrm{Var}\Big[\log M^{\xi}_{j,2k+1} - \log M^{\xi}_{j-1,k}\Big] + 2^{-j}\sum_{\substack{-M\le m\le M;\\ m\neq 0}} C^2_{N,2m+N}\mathrm{Var}[\log M^{\xi}_{j-1,k+m}]\ ,
\end{aligned}
$$

where in step $(*)$ we used that $C_{N,N} = 1$, and in the last step we used that the covariance across scales between noise midpoints is zero. We can rewrite the first term using the midpoint relation:

$$
\begin{aligned}
2^{-j}\mathrm{Var}[\log M^{\xi}_{j,2k+1} - \log M^{\xi}_{j-1,k}] &= 2^{-j}\mathrm{Var}[\log M^{\xi}_{j,2k+1} - \frac{1}{2}\log M^{\xi}_{j,2k+1} - \frac{1}{2}\log M^{\xi}_{j,2k}] \\
&= 2^{-(j+2)}\mathrm{Var}[\log M^{\xi}_{j,2k+1} - \log M^{\xi}_{j,2k}] \\
&= 2^{-(j+2)}2\mathrm{Var}[\log M^{\xi}_{j,0}] \\
&= 2^{-(j+1)}\mathrm{Var}[\log M^{\xi}_{j,0}]\ ,
\end{aligned}
$$

where using $C_{N,N} = 1$ in step $(*)$ earlier was necessary to obtain the factor $2^{-(j+1)}$. We can use the same reasoning for the second term:

$$
\begin{aligned}
2^{-j}\sum_{\substack{-M\le m\le M;\\ m\neq 0}} C^2_{N,2m+N}\mathrm{Var}[\log M^{\xi}_{j-1,k+m}] &= 2^{-j}\left(-1 + \sum_{\substack{-M\le m\le M;\\ m\neq 0}} C^2_{N,2m+N}\right)\mathrm{Var}[\log M^{\xi}_{j-1,0}] \\
&= 2^{-j}\left(-1 + \sum_{m=-M}^{M} C^2_{N,2m+N}\right)\frac{1}{2}\mathrm{Var}[\log M^{\xi}_{j,0}] \\
&= 2^{-(j+1)}\left(-1 + \sum_{m=-M}^{M} C^2_{N,2m+N}\right)\mathrm{Var}[\log M^{\xi}_{j,0}]\ .
\end{aligned}
$$

Combining the two terms, we obtain

$$
\mathrm{Var}[\mathfrak{D}_{j,k}] = 2^{-(j+1)}\sum_{m=-M}^{M} C^2_{N,2m+N}\mathrm{Var}[\log M^{\xi}_{j,0}]\ .
$$

Using recursively the midpoint relation, we obtain

$$
\mathrm{Var}[\log M^{\xi}_{j,0}] = \frac{1}{2}\mathrm{Var}[\log M^{\xi}_{j+1,0}] = \ldots = \frac{1}{2^{J-j}}\mathrm{Var}[\xi_0]\ ,
$$

plugging this result yields an expression that does not depend on the scale j:

$$
\begin{aligned}
\mathrm{Var}[\mathfrak{D}_{j,k}] &= 2^{-(j+1)}2^{-(J-j)}\sum_{m=-M}^{M} C^2_{N,2m+N}\mathrm{Var}[\xi_1] \\
&= 2^{-(J+1)}\sum_{m=-M}^{M} C^2_{N,2m+N}\mathrm{Var}[\xi_0]\ .
\end{aligned}
$$

where $\mathrm{Var}[\xi_0] =: \boldsymbol{\mathcal{C}}$. □

D.6 Proof of Proposition 5

Proof. Under the assumption of a log-additive Riemannian model and the intrinsic AI wavelet transform with the log-Euclidean metric, by Proposition 4, the whitened wavelet coefficients follow a matrix signal-plus-noise model. By affinity of φ,

$$\varphi(\mathfrak{D}_{jk}) = \varphi(\mathfrak{D}^f_{jk} + \mathfrak{D}^\xi_{jk}) = \varphi(\mathfrak{D}^f_{jk}) + \varphi(\mathfrak{D}^\xi_{jk}) \; .$$

By linearity of the expectation, it follows that

$$\mathbf{E}[\varphi(\mathfrak{D}_{jk})] = \varphi(\mathbf{E}[\mathfrak{D}_{jk}]) = \varphi(\mathfrak{D}^f_{jk}) \; .$$

For the variance, we have

$$\varphi(\mathfrak{D}_{jk}) = 2^{-j/2}\Big(\varphi(\log M_{j,2k+1}) - \varphi(\log \widetilde{M}_{j,2k+1})\Big) \; ,$$

which implies that we can use the same reasoning as the proof of Proposition 4 to obtain the desired result:

$$\mathrm{Var}[\varphi(\mathfrak{D}_{j,k})] = 2^{-(J+1)} \sum_{m=-M}^{M} C^2_{N,2m+N} \mathrm{Var}[\varphi(\xi_0)] \; . \qquad \square$$

D.7 Proof of Proposition 6

Proof. Before prooving the result, we recall a few results from the theory of concentration inequalities, as presented in Boucheron et al. (2013). A centered random variable X is said to be sub-gamma on the right tail with variance factor ν and scale parameter c, denoted by $\Gamma_+(\nu, c)$, if the logarithm of its moment generating function – denoted by $\psi_X(\lambda)$ – is bounded as follows:

$$\psi_X(\lambda) \leq \frac{\lambda^2 \nu}{2(1 - c\lambda)} \qquad 0 < \lambda < 1/c \; .$$

For example, if $Y \sim \mathcal{G}(a, b)$ and $X := Y - \mathbf{E}[Y]$, then $X \sim \Gamma_+(ab^2, b)$. Additionally, as $Y \sim \chi^2_p = \mathcal{G}(p/2, 2)$ we have that $X \sim \Gamma_+(2p, 2)$ with $X := Y - \mathbf{E}[Y] = Y - p$. The Cramér transform of $X \sim \Gamma_+(2p, 2)$ is

$$\psi^*_X(t) = \frac{p}{2} h_1\left(\frac{t}{p}\right) \; ,$$

with $h_1(u) = 1 + u - \sqrt{1 + 2u}$ for $u > 0$. We can apply the Cramér-Chernoff bounding method to derive a concentration inequality on the maximum of centered χ^2_p random variables:

$$\begin{aligned}
\mathbb{P}\left\{\max_{1 \leq i \leq n} X_i - p \geq t\right\} &\leq \sum_{i=1}^{n} \mathbf{E}[\mathbb{1}\{X_i - p > t\}] = n\mathbb{P}(X_1 - p > t) \\
&\leq n \exp\big(-\psi^*_X(t)\big) = n \exp\left(-\frac{p}{2} h_1\left(\frac{t}{p}\right)\right) \\
&= n \exp\left(-\frac{1}{2}\left(p + t - p\sqrt{1 + \frac{2t}{p}}\right)\right) \; .
\end{aligned}$$

Plugging-in $t_n = 2\sqrt{(1+\epsilon)^2 p \log n} + 2 \log n$ for some $\epsilon > 0$, we obtain

$$\mathbb{P}\left\{\max_{1 \leq i \leq n} X_i - p \geq t_n\right\}$$

$$
\begin{aligned}
&\leq n \exp\left(-\frac{1}{2}\left(p + 2\left(\sqrt{(1+\epsilon)^2 p \log n} + \log n\right) - p\sqrt{1 + \frac{4\left(\sqrt{(1+\epsilon)^2 p \log n} + \log n\right)}{p}}\right)\right) \\
&= n \exp\left(-\frac{p}{2} - \sqrt{(1+\epsilon)^2 p \log n} - \log n + \frac{p}{2}\sqrt{1 + \frac{4\left(\sqrt{(1+\epsilon)^2 p \log n} + \log n\right)}{p}}\right) \\
&= \exp\left(-\frac{p}{2} - \sqrt{(1+\epsilon)^2 p \log n} + \frac{p}{2}\sqrt{1 + \frac{4\left(\sqrt{(1+\epsilon)^2 p \log n} + \log n\right)}{p}}\right) \\
&\leq \exp\left(-\frac{p}{2} - \sqrt{(1+\epsilon)^2 p \log n} + \frac{p}{2} + \frac{p}{2}\sqrt{\frac{4\left(\sqrt{(1+\epsilon)^2 p \log n} + \log n\right)}{p}}\right) \\
&= \exp\left(-\sqrt{(1+\epsilon)^2 p \log n} + p^{1/2}\sqrt{\sqrt{(1+\epsilon)^2 p \log n} + \log n}\right) \\
&\leq \exp\left(-(1+\epsilon)p^{1/2}(\log n)^{1/2} + \frac{(1+\epsilon)}{2}p + p^{1/2}(\log n)^{1/2}\right) \\
&= \exp\left(\frac{(1+\epsilon)}{2}p - \epsilon\sqrt{p \log n}\right) \ ,
\end{aligned}
$$

where we used in the sixth step a Taylor approximation such that $(1+x)^{1/2} \leq 1 + \frac{x}{2}$,

$$
\begin{aligned}
\left(\log(n) + \sqrt{(1+\epsilon)^2 p \log(n)}\right)^{1/2} &= (\log(n))^{1/2}\left(1 + \sqrt{(1+\epsilon)^2 \frac{p}{\log(n)}}\right)^{1/2} \\
&\leq (\log(n))^{1/2}\left(1 + \frac{\sqrt{(1+\epsilon)^2 \frac{p}{\log(n)}}}{2}\right) \\
&= (\log(n))^{1/2} + \frac{(1+\epsilon)}{2}\sqrt{p} \ .
\end{aligned}
$$

Taking the limit $n \to \infty$, we observe that $\exp\left(-\epsilon\sqrt{p \log n}\right) \to 0$, and we obtain the desired result:

$$
\lim_{n\to\infty} \mathbb{P}\left\{\max_{1\leq i\leq n} X_i - p \geq 2\left(\sqrt{(1+\epsilon)^2 p \log n} + \log n\right)\right\} = 0 \ . \qquad \square
$$

D.8 Proof of Theorem 2

Proof. For the first part of the proof, we show the convergence rate (24) by decomposing the sum of nonlinear thresholding whitened coefficients over the following scales:

$$
\begin{aligned}
\sum_{jk} \mathbf{E}\|\widehat{\mathfrak{D}}^*_{jk} - \mathfrak{D}^f_{jk}\|_F^2 = \sum_{j<J_0}\sum_{k=0}^{2^j-1} \mathbf{E}\|\mathfrak{D}_{jk} - \mathfrak{D}^f_{jk}\|_F^2 + \sum_{j=J_0}^{J_1-1}\sum_{k=0}^{2^j-1} \mathbf{E}\|w_{jk}\mathfrak{D}_{jk} - \mathfrak{D}^f_{jk}\|_F^2 \\
+ \sum_{j\geq J_1}\sum_{k=0}^{2^j-1} \|\mathfrak{D}^f_{jk}\|_F^2 \ .
\end{aligned}
$$

We work with scales $J_0, J_1 \gg 1$ sufficiently large such that Proposition C.2 and C.3 hold and such that $2^{J_0} \sim 2^{J_1}$. Then, we get for the first and third terms the classical linear rates:

$$\sum_{j<J_0}\sum_{k=0}^{2^j-1}\mathbf{E}\|\mathfrak{D}_{jk}-\mathfrak{D}^f_{jk}\|_F^2 \lesssim \sum_{j<J_0} 2^j n^{-1} = (2^{J_0}-2)n^{-1} \lesssim \frac{2^{J_0}}{n}\ ,$$

and

$$\sum_{j\geq J_1}\sum_{k=0}^{2^j-1}\|\mathfrak{D}^f_{jk}\|_F^2 \lesssim \sum_{j\geq J_1} 2^j(2^{-j}2^{-2Nj}) \lesssim 2^{-2NJ_1} \sim 2^{-2NJ_0}\ .$$

For the second term involving the nonlinear thresholding shrinkage weights, we can also use that $2^{J_0} \sim 2^{J_1}$ to obtain the following rate:

$$\begin{aligned}
\sum_{j=J_0}^{J_1-1}\sum_{k=0}^{2^j-1}\mathbf{E}\|w_{jk}\mathfrak{D}_{jk}-\mathfrak{D}^f_{jk}\|_F^2 &= \sum_{j=J_0}^{J_1-1}\sum_{k=0}^{2^j-1}\mathbf{E}\|w_{jk}\mathfrak{D}_{jk}-w_{jk}\mathfrak{D}^f_{jk}+w_{jk}\mathfrak{D}^f_{jk}-\mathfrak{D}^f_{jk}\|_F^2\\
&\leq 2\sum_{j=J_0}^{J_1-1}\sum_{k=0}^{2^j-1} w_{jk}^2\mathbf{E}\|\mathfrak{D}_{jk}-\mathfrak{D}^f_{jk}\|_F^2 + 2\sum_{j=J_0}^{J_1-1}\sum_{k=0}^{2^j-1}(1-w_{jk})^2\|\mathfrak{D}^f_{jk}\|_F^2\\
&\leq 2\sum_{j=J_0}^{J_1-1}\sum_{k=0}^{2^j-1}\mathbf{E}\|\mathfrak{D}_{jk}-\mathfrak{D}^f_{jk}\|_F^2 + 2\sum_{j=J_0}^{J_1-1}\sum_{k=0}^{2^j-1}\|\mathfrak{D}^f_{jk}\|_F^2\\
&\lesssim 2\sum_{j=J_0}^{J_1-1}2^j n^{-1} + 2\sum_{j=J_0}^{J_1-1}(2^{-2N})^j\\
&\overset{(*)}{=} \frac{2}{n}(2^{J_1}-2^{J_0}) + 2\left(\frac{2^{-2NJ_0}-2^{-2NJ_1}}{1-2^{-2N}}\right)\\
&\lesssim \frac{2^{J_1}}{n} \sim \frac{2^{J_0}}{n}\ ,
\end{aligned}$$

where in step $(*)$ we used that

$$\begin{aligned}
\left(1-2^{-2N}\right)\left(\sum_{j=J_0}^{J_1-1}(2^{-2N})^j\right) &= \sum_{j=J_0}^{J_1-1}2^{-2Nj} - 2^{-2N}\sum_{j=J_0}^{J_1-1}2^{-2Nj}\\
&= \sum_{j=J_0}^{J_1-1}2^{-2Nj} - \sum_{j=J_0+1}^{J_1}2^{-2Nj}\\
&= 2^{-2NJ_0}-2^{-2NJ_1}\ .
\end{aligned}$$

Combining the three terms, setting $J_0 = J/(2N+1)$ such that $2^{J_0}n^{-1} = 2^{-2NJ_0}$, we get

$$\sum_{jk}\mathbf{E}\|\widehat{\mathfrak{D}}^*_{jk}-\mathfrak{D}^f_{jk}\|_F^2 \lesssim \frac{2^{J_0}}{n}+\frac{2^{J_0}}{n}+2^{-2NJ_0} \sim 2^{-2NJ_0} = n^{-2N/(2N+1)}\ .$$

For the second part of the theorem, we are interested in deriving the rate of convergence (25), i.e. of the hard-thresholding estimator $(\widehat{M}^*_{J,k})_k$, based on the same paradigm (i.e. keeping every coefficient up to scale J_0-1, then nonlinear thresholding the wavelet coefficients up to scale J_1-1, and finally forcing them to zero from scale J_1). We will work recursively as in Chau and von Sachs (2021), from the coarsest to finest scale.

Before going to the derivation of the rates, we recall some technical details. Let $(M_{j,k})_{j,k}$ denote the empirical midpoint pyramid, based on the observed (noisy) signal. The random variable $M_{j,k}$ is constructed from repetition of the midpoint relation, on 2^{J-j} observed samples: as a consequence of Proposition C.1, for j sufficiently small,

$$\mathbf{E}[\delta_{\mathrm{L}}^2(M_{jk}, M_{jk}^f)] = O(2^{-(J-j)}) \quad \forall k = 0, \dots, 2^{j-1} \,, \tag{D6}$$

where M_{jk}^f denotes the associated scaling coefficients of the true underlying signal f. Using this result, we rewrite $M_{j,k} = M_{j,k}^f + \epsilon_{j,n}$ with $\epsilon_{j,n}$ a general (not always the same) random error matrix such that $\mathbf{E}\|\epsilon_{j,n}\|_F^2 = O(2^{-(J-j)})$:

$$\begin{aligned}
M_{j,k} &= \mathrm{Exp}_{M_{j,k}^f}(\epsilon_{j,n}) \\
&= \exp\big(\log(M_{j,k}^f) + \mathbb{D}_{M_{j,k}^f}\log(\epsilon_{j,n})\big) \\
&= \exp\big(\log(M_{j,k}^f) + O_{\mathbb{P}}(2^{-(J-j)/2})\big) \\
&= M_{j,k}^f + O_{\mathbb{P}}(2^{-(J-j)/2}) \,.
\end{aligned}$$

We denote by $(\widetilde{M}_{j,k})_{j,k}$ the empirical predicted midpoints, computed using a weighted average of the empirical midpoints. We then get

$$\begin{aligned}
\widetilde{M}_{j,k} &= \sum_{m=-M}^{M} C_{N,2m+N}\log(M_{j-1,k+m}) \\
&= \sum_{m=-M}^{M} C_{N,2m+N}\log(M_{j-1,k+m}^f + \epsilon_{j-1,n}) \\
&= \sum_{m=-M}^{M} C_{N,2m+N}\big(\log(M_{j-1,k+m}^f) + \epsilon_{j-1,n}\big) \\
&= \sum_{m=-M}^{M} C_{N,2m+N}\log(M_{j-1,k+m}^f) + \epsilon_{j-1,n}\sum_{m=-M}^{M} C_{N,2m+N} \\
&= \widetilde{M}_{j,k}^f + O_{\mathbb{P}}(2^{-(J-j+1)/2}) \\
&= \widetilde{M}_{j,k}^f + O_{\mathbb{P}}(2^{-(J-j)/2}) \,.
\end{aligned} \tag{D7}$$

We denote by $(\widehat{M}_{j,k}^*)_k$, without loss of generality, the estimated midpoints – i.e. the midpoints obtain by reconstruction from the thresholding estimates $(\widehat{D}_{j,k}^*)_k$, not from the empirical estimates $(D_{j,k})_k$. Finally, we denote by $(\widehat{\widetilde{M}}_{j,k})_k$ the estimated predicted midpoints: those are weighted averages of the estimated midpoints at scale $j-1$. Hence, $(\widehat{\widetilde{M}}_{j,k})_k$ inherit the estimation error from the previous scale. Finally, recall that we reconstruct the midpoints with the exponential map:

$$\widehat{M}_{j,2k+1}^* = \mathrm{Exp}_{\widehat{\widetilde{M}}_{j,2k+1}}(2^{j/2}\widehat{D}_{j,k}^*) = \exp\big(\log\widehat{\widetilde{M}}_{j,2k+1} + 2^{j/2}\widehat{\mathfrak{D}}_{j,k}^*\big) \,.$$

For scales $j = 1, \dots, J_0 - 1$: we do not alter the wavelet coefficients, i.e. $\widehat{D}_{j,k}^* = D_{j,k}$: consequently,

we get (for all $k = 0, \ldots, 2^j - 1$) $\widehat{M}^*_{j,k} = M_{j,k}$ and thus $\widehat{\widetilde{M}}_{j,k} = \widetilde{M}_{j,k}$, because

$$\begin{aligned}
\widehat{\widetilde{M}}_{j,2k+1} &= \exp\left(\sum_{m=-M}^{M} C_{N,2m+N} \log(\widehat{M}^*_{j-1,k+m})\right) \\
&= \exp\left(\sum_{m=-M}^{M} C_{N,2m+N} \log(M_{j-1,k+m})\right) \\
&= \widetilde{M}_{j,2k+1} \ . \qquad \text{(D8)}
\end{aligned}$$

Hence, combining (D7) and (D8), we can write $\widehat{\widetilde{M}}_{j,k} = \widetilde{M}^f_{j,k} + \epsilon_{j,n}$ with $\epsilon_{j,n}$ a general random error matrix satisfying $\mathbf{E}\|\epsilon_{j,n}\|_F = O(2^{-(J-j)/2})$. Furthermore, as we do not alter the wavelet coefficients, it follows that $\widehat{\mathfrak{D}}^*_{j,k} = \mathfrak{D}_{j,k}$ and by Proposition C.4 stating $\mathbf{E}\|\mathfrak{D}_{j,k} - \mathfrak{D}^f_{j,k}\|^2_F = O(n^{-1})$, we can rewrite $\mathfrak{D}_{j,k} = \mathfrak{D}^f_{j,k} + \eta_n$ with η_n a general random error matrix satisfying $\mathbf{E}\|\eta_n\|_F = O(n^{-1/2})$. Combining these results with the midpoint estimations, we get for each of the considered scales the following relation:

$$\begin{aligned}
\widehat{M}^*_{j,2k+1} &= \mathrm{Exp}_{\widehat{\widetilde{M}}_{j,2k+1}}(2^{j/2}\widehat{D}^*_{j,k}) \\
&= \exp\big(\log \widehat{\widetilde{M}}_{j,2k+1} + \mathbb{D}_{\widehat{\widetilde{M}}_{j,2k+1}} \log(2^{j/2}\widehat{D}^*_{j,k})\big) \\
&= \exp\big(\log \widehat{\widetilde{M}}_{j,2k+1} + 2^{j/2}\widehat{\mathfrak{D}}^*_{j,k}\big) \\
&= \exp\big(\log(\widetilde{M}^f_{j,2k+1} + \epsilon_{j,n}) + 2^{j/2}(\mathfrak{D}^f_{j,k} + \eta_n)\big) \\
&= \exp\big(\log(\widetilde{M}^f_{j,2k+1}) + \epsilon_{j,n} + 2^{j/2}\mathfrak{D}^f_{j,k} + 2^{j/2}\eta_n\big) \\
&= M^f_{j,2k+1} + O_{\mathbb{P}}(2^{-(J-j)/2}) + O_{\mathbb{P}}(2^{j/2}n^{-1/2}) \\
&= M^f_{j,2k+1} + O_{\mathbb{P}}(2^{-(J-j)/2}) \ .
\end{aligned}$$

For scales $j = J_0, \ldots, J_1 - 1$: we still obtain the estimated predicted midpoints from the empirical midpoints because we did not alter the wavelet coefficients until $J_0 - 1$. However, this time we will use nonlinear thresholding on the wavelet coefficients using shrinkage weights. We estimate $\widehat{\mathfrak{D}}^*_{J_0,k} = w_{J_0,k}\mathfrak{D}_{J_0,k}$, and by Proposition C.4 (because $2^{J_0-J} \to 0$), we have $\widehat{\mathfrak{D}}^*_{J_0,k} = w_{J_0,k}(\mathfrak{D}^f_{J_0,k} + \eta_n)$ with $\mathbf{E}\|\eta_n\|_F = O(n^{-1/2})$. Suppose $J_0 = J/(2N+1)$ is sufficiently large such that Proposition C.2 applies, then

$$\begin{aligned}
\widehat{M}^*_{J_0,2k+1} &= \exp(\log \widehat{\widetilde{M}}_{J_0,2k+1} + 2^{J_0/2}\widehat{\mathfrak{D}}^*_{J_0,k}) \\
&= \exp\big(\log(\widetilde{M}^f_{J_0,2k+1} + \epsilon_{J_0,n}) + 2^{J_0/2}w_{J_0,k}(\mathfrak{D}^f_{J_0,k} + \eta_n)\big) \\
&= \exp\big(\log(\widetilde{M}^f_{J_0,2k+1}) + \epsilon_{J_0,n} + 2^{J_0/2}w_{J_0,k}\mathfrak{D}^f_{J_0,k} + 2^{J_0/2}w_{J_0,k}\eta_n\big) \\
&= M^f_{J_0,2k+1} + O_{\mathbb{P}}(2^{-(J-J_0)/2}) + O_{\mathbb{P}}(2^{-J_0N}) \\
&= M^f_{J_0,2k+1} + O_{\mathbb{P}}(2^{-J_0N}) \ ,
\end{aligned}$$

where we used that, with $w_{J_0,k}$ a binary random variable, $\|2^{J_0/2}w_{J_0,k}\mathfrak{D}^f_{J_0,k}\|_F = O_{\mathbb{P}}(2^{J_0/2}2^{-J_0/2}2^{-J_0N}) = O_{\mathbb{P}}(2^{-J_0N})$. Iterating over the following scales, and using that by $2^{J_1} \sim 2^{J_0}$ we have $2^j \sim 2^{J_0}$ for $j \leq J_1$, we obtain up to $j = J_1 - 1$

$$\begin{aligned}
\widehat{M}^*_{j,2k+1} &= \exp\big(\log(\widehat{\widetilde{M}}_{j,2k+1}) + 2^{j/2}\widehat{\mathfrak{D}}^*_{j,k}\big) \\
&= \exp\big(\log(\widetilde{M}^f_{j,2k+1}) + \epsilon_{j,n} + 2^{j/2}w_{j,k}\mathfrak{D}^f_{j,k} + 2^{j/2}w_{j,k}\eta_n\big)
\end{aligned}$$

$$= M^f_{j,2k+1} + O_{\mathbb{P}}(2^{j/2} n^{-1/2}) + O_{\mathbb{P}}(2^{-jN})$$
$$= M^f_{j,2k+1} + O_{\mathbb{P}}(2^{-J_0 N}) \, .$$

For scales $J_1, \ldots, J$: we force every wavelet coefficients to zero, i.e. $\widehat{D}^*_{j,k} = \widehat{\mathfrak{D}}^*_{j,k} = 0$. Consequently, we use Proposition C.2 to rewrite $\widehat{\mathfrak{D}}^*_{j,k} = \mathfrak{D}^f_{j,k} + \rho_{j,N}$, where $\rho_{j,N}$ is a deterministic error matrix with $\|\rho_{j,N}\|_F = O(2^{-j/2} 2^{-jN})$. Additionally, we define a general interpolation error matrix $\zeta_{J_0,N}$, with $\|\zeta_{J_0,N}\|_F = O_{\mathbb{P}}(2^{-J_0 N})$ – indeed for $j = J_1$ we have

$$\begin{aligned}
\widehat{\widetilde{M}}_{J_1,2k+1} &= \exp\left(\sum_{m=-M}^{M} C_{N,2m+N} \log(\widehat{M}^*_{J_1-1,k+m}) \right) \\
&= \exp\left(\sum_{m=-M}^{M} C_{N,2m+N} \log(M^f_{J_1-1,k+m}) + O_{\mathbb{P}}(2^{-J_0 N}) \right) \\
&= \widetilde{M}^f_{J_1,2k+1} + O_{\mathbb{P}}(2^{-J_0 N}) \, .
\end{aligned}$$

Hence, we obtain

$$\begin{aligned}
\widehat{M}^*_{J_1,2k+1} &= \exp\big(\log \widehat{\widetilde{M}}_{J_1,2k+1} + 2^{J_1/2} \widehat{\mathfrak{D}}^*_{J_1,k} \big) \\
&= \exp\big(\log \widetilde{M}^f_{J_1,2k+1} + \zeta_{J_0,N} + 2^{J_1/2} (\mathfrak{D}_{J_1,k} + \rho_{J_1,N}) \big) \\
&= \exp\big(\log \widetilde{M}^f_{J_1,2k+1} + \zeta_{J_0,N} + 2^{J_1/2} \mathfrak{D}^f_{J_1,k} + 2^{J_1/2} \rho_{J_1,N} \big) \\
&= \exp\big(\log \widetilde{M}^f_{J_1,2k+1} + 2^{J_1/2} \mathfrak{D}^f_{J_1,k} \big) + O_{\mathbb{P}}(2^{-J_0 N}) + O_{\mathbb{P}}(2^{-J_1 N}) \\
&= M^f_{J_1,2k+1} + O_{\mathbb{P}}(2^{-J_0 N}) \, .
\end{aligned}$$

Iterating this argument over the following scales $j = J_1 + 1, \ldots, J$ we obtain

$$\begin{aligned}
\widehat{M}^*_{j,2k+1} &= \exp\big(\log \widehat{\widetilde{M}}_{j,2k+1} + 2^{j/2} \widehat{\mathfrak{D}}^*_{j,k} \big) \\
&= \exp\big(\log \widetilde{M}^f_{j,2k+1} + \zeta_{J_0,N} + 2^{j/2} (\mathfrak{D}^f_{j,k} + \rho_{j,N}) \big) \\
&= \exp\big(\log \widetilde{M}^f_{j,2k+1} + 2^{j/2} \mathfrak{D}^f_{j,k} \big) + O_{\mathbb{P}}(2^{-J_0 N}) + O_{\mathbb{P}}(2^{-jN}) \\
&= M^f_{j,2k+1} + O_{\mathbb{P}}(2^{-J_0 N}) \, .
\end{aligned}$$

This last result implies that at the finest scale J, using $J_0 = J/(2N+1)$,

$$\begin{aligned}
\mathbf{E}[\delta^2_{\mathrm{L}}(\widehat{M}^*_{J,k}, M^f_{J,k})] &= \mathbf{E}\| \log(\widehat{M}^*_{J,k}) - \log(M^f_{J,k}) \|^2_F = \mathbf{E}\| \log(M^f_{J,k}) + \zeta_{J_0,N} - \log(M^f_{J,k}) \|^2_F \\
&= \mathbf{E}\|\zeta_{J_0,N}\|^2_F = O(2^{-2NJ_0}) \\
&= O(n^{-2N/(2N+1)}) \, .
\end{aligned}$$

□

D.9 Proof of Corollary 1

Proof. Recall that the coarsest scaling coefficient $M_{0,0}$ and the predicted midpoints $(\widetilde{M}_{j,k})_{j,k}$ are SPD by construction and do not depend on the wavelet coefficients. Also recall that reconstruction of the scaling coefficients implies first the exponential map for odd-indexed coefficients:

$$M_{j,2k+1} = \mathrm{Exp}_{\widetilde{M}_{j,2k+1}}(D_{j,k}) \, ,$$

and secondly the matrix exponential and logarithm for even-indexed coefficients:

$$M_{j,2k} = \exp(2 \log M_{j-1,k} - \log M_{j,2k+1}) \, .$$

Now consider the limiting case where we reconstruct $(M_{1,k})_{k=0,1}$ from $M_{0,0}$ and $D_{1,0}$. In the case of the odd-indexed midpoint, using the fact that $(\mathrm{Sym}^+(d), \delta_L)$ is a geodesically complete manifold Thanwerdas and Pennec (2023), we have that the exponential mapping is a bijective map from $T_p(\mathrm{Sym}^+(d)) \cong \mathrm{Sym}(d) \to \mathrm{Sym}^+(d)$. Therefore, the reconstruction for $M_{1,1}$ is necessarily SPD if $D_{1,0}$ is symmetric. In the case of the even-indexed midpoint, recall that the matrix exponential of a symmetric matrix is a bijective map $\exp : \mathrm{Sym}(d) \to \mathrm{Sym}^+(d)$ Arsigny et al. (2007). Additionally, using the fact that the matrix logarithm of SPD matrices is unique and defined as the inverse of the matrix exponential, then $\log : \mathrm{Sym}^+(d) \to \mathrm{Sym}(d)$ is also a bijective map. Therefore, the reconstruction for $M_{1,0}$ is SPD, as it involves the matrix exponential of the difference $(2 \log M_{0,0} - \log M_{1,1}) \in \mathrm{Sym}(d)$.

Thus, the sequence of reconstructed midpoints at scale $j = 1$ is indeed a sequence of SPD matrices. We can use the same arguments by recurrence to state that as long as the wavelet coefficients are all in $\mathrm{Sym}(d)$, the reconstructed signal will consist of SPD matrices. □

E Our pilot estimator for subsequent non-linear wavelet thresholding

As indicated in Section 2.1 of the main document, the first basic approach to construct a pre-smoothing positive-definite pilot estimator would consist in local averaging observations akin kernel smoothing with a rectangular window of length L (called "smoothing span" hereafter) with $L \geq d$: Let, for $k = L/2, L/2+1, \ldots, T - L/2$,

$$\check{S}_k = L^{-1} \sum_{t=k-L/2+1}^{k+L/2} X_t X_t^\top \ ,$$

for $L \geq 2$ being even (with the obvious adaptation for an odd natural number L to smooth over the interval $[k-(L-1)/2, k+(L-1)/2]$). Choosing a simple boxcar kernel is sufficient here for shedding light on the behaviour of this moving window smoother only used as a pilot estimator for subsequent wavelet smoothing. We observe that, as $T \to \infty, L \to \infty, L/T \to 0$, the $(\check{S}_k)_k$ provide a consistent estimator for $(\Sigma(\frac{k}{T}))_k$, where asymptotically each $\Sigma(\frac{k}{T})$ tends to $\Sigma(u)$ with u being the limit of the sequence of neighbouring grid points $(\frac{k}{T})_k$ - as it is usual in classical nonparametric curve estimation theory of estimating a continuous curve in time $u \in (0,1)$ by discretely sampled values on an equidistant grid $\{k/T\}_{k=1,\ldots,T}$ of $(0,1)$.

Another possibility to localise our estimation problem is to chop the observation period T up into n non-overlapping segments of length L, such that $T = nL$. Then a discrete estimator of Σ is given by $(\widetilde{S}_0, \ldots, \widetilde{S}_{n-1})$ with $\widetilde{S}_k$ an estimator of $S_k := \Sigma(u_k)$ at the midpoint $u_k = \frac{2k+1}{2n}$ of the $k-$th segment as defined in (18)

$$\widetilde{S}_k = L^{-1} \sum_{t=kL+1}^{(k+1)L} X_t X_t^\top \ , \ k = 0, \ldots, n-1 \ .$$

In this framework, as previously for the sliding window approach, increasing T means that we operate finer sampling on the same time period $(0,1)$. As $T \to \infty$, we let $L, n \to \infty$. To obtain again that $\widetilde{S}_k$ is a consistent estimator of $S_k = \Sigma(u_k)$, and in fine the associated $\Sigma(u)$, with u being the limit of the grid values $(u_k)_k$ in its neighbourhood, here one needs the conditions that $L, n \to \infty$ as $T \to \infty$ such that $L/n \to 0$.

When comparing the two approaches now, we need to take two different aspects into account: 1. high enough time resolution of our sampled data (to feed into the finest scale of our wavelet algorithm) - which translates into reasonable small approximation bias; 2. small enough variability of the resulting wavelet coefficients based on the chosen pilot estimator, in particular when allowing for serially dependent time series data $(X)_{t+1}^T$.

Regarding the first point, it appears to be advantageous to use the sliding window approach which delivers, as starting point, $T - L + 1$ estimators $\check{S}_k, k = L/2, \ldots, T - L/2$. However, especially for

higher dimensions d, calling for a smoothing span $L \geq d$, we need to face an increase of the (asymptotic) variance of $\check{S}_k$, in particular when the data are serially correlated. Moreover, when using $\check{S}_k$ as input to a wavelet thresholding scheme, even in an independent data situation, the correlatedness of the $\check{S}_k$ over time k translates into higher variance of the wavelet coefficients on coarser scales - calling for higher level-dependent thresholds to choose. This effect is, again, of course the more pronounced the higher the dimension d (calling for higher smoothing spans L).

When using, alternatively, the (non-overlapping) segment estimator, the resulting $\widetilde{S}_k$ remain exactly uncorrelated for the independent data situation (and less correlated than the $\check{S}_k$ for correlated data $(X)_{t+1}^{T}$). This advantage, on the side of smaller variability of the resulting wavelet threshold estimator, admittedly goes with the introduction of a higher approximation bias, due to the reduced time resolution n^{-1} (compared to essentially T^{-1} for the sliding-window estimator) with $n = T/L$, and this of course, the more, the higher d (again due to $L \geq d$). Nevertheless we generally recommend to use the segment estimator $\widetilde{S}_k$ as its increase in approximation bias will be highly compensated by the reduced variability of the resulting wavelet coefficients based on this pilot estimator.

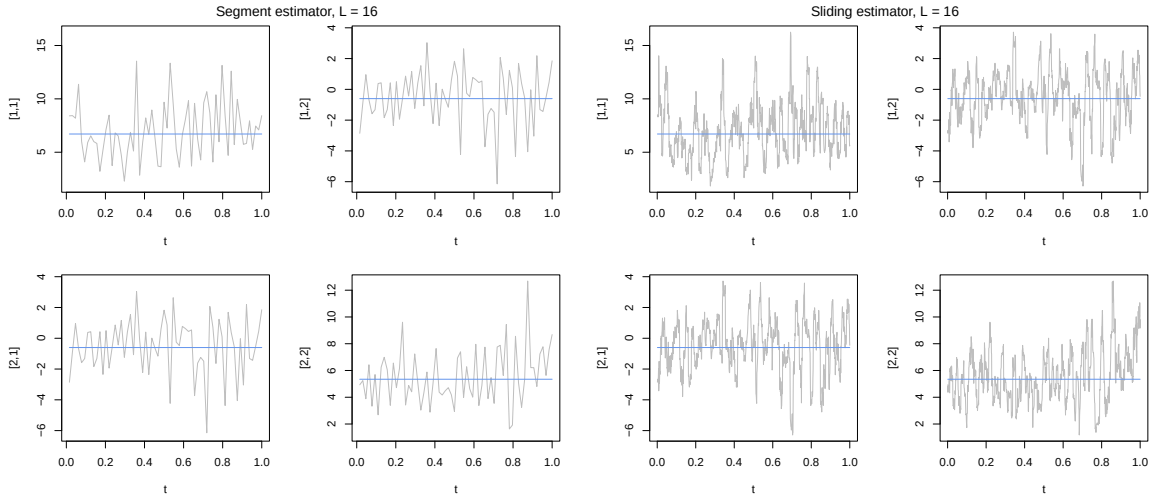


Figure E1: Blue: constant SPD curve. Gray: segmented (left) and sliding (right) estimators of the curve, with $L = 16$.

Indeed, for simulated bivariate independent time series (with constant TVCM), we observe in Figure E1 their segmented and sliding estimators of the TVCM with $L = 16$. We can visually observe that although benefiting from a higher resolution, the sliding estimator appears to feature serial correlation (which visually translates as a noisier version of the segmented estimator with lower resolution, and is confirmed by the ACF plots in Figure E2). This feature is threatening for the performance of our nonlinear thresholding scheme: in fact, the wavelet transform of a constant curve should yield $D_{j,k} = 0$ for all (j, k), so the transform of the noisy version of the curve should feature wavelet coefficients with small amplitude. Inspecting the whitened wavelet coefficients of the pilot estimators in Figure E3, we see that the segmented estimator is mainly composed of noisy coefficients with relatively small amplitude over the scales (which indeed corresponds to the aforementioned situation of a noisy constant curve). The sliding estimator, on the other hand, has a lot of low frequency content with large amplitude induced by the serial correlation while the high frequency content has a very small amplitude. Consequently, the nonlinear thresholding estimators will not be able to catch the constant nature of the curve in this setting. Note that these undesirable properties of the sliding estimator get worse as L increases.

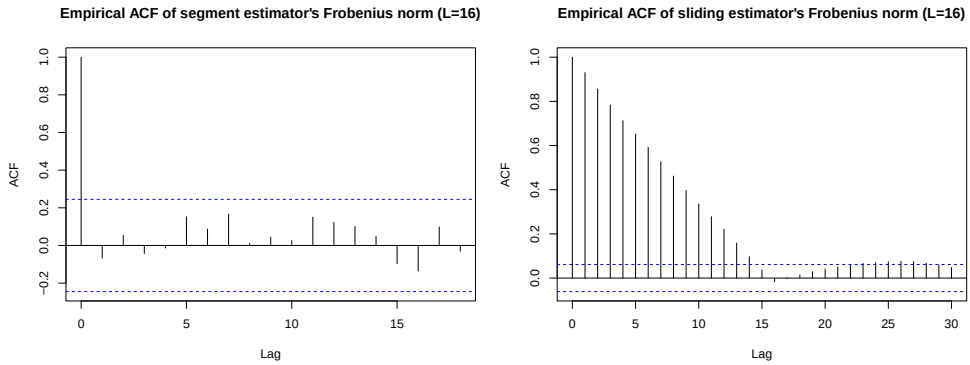


Figure E2: Empirical autocorrelation function of the Frobenius norm of the segment (left) and sliding (right) pilot estimators from Figure E1.

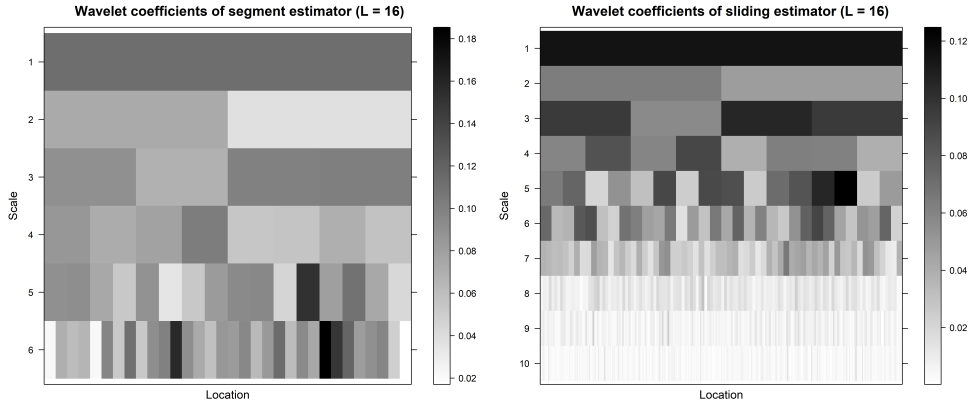


Figure E3: Wavelet decomposition (using the Frobenius norm of the whitened coefficients) of pilot estimators from Figure E1.

F Further simulation results

In this section, we assess by simulation study the performance of the multiple thresholding rules considered in Section 3.2 by Monte-Carlo simulations with different data generating processes (DGPs). For each curve in each scenario, we will simulate $M = 500$ realisations of the considered DGPs.

We will consider multiple SPD curves, each being studied under different scenarios. In particular, consider the following matrix-valued function:

$$f(t) = \begin{pmatrix} 10 + (5 - 10t)^2 & 15 - 10t \\ 15 - 10t & 20 \end{pmatrix} .$$

Let $c_1(t), c_2(t), c_3(t), c_4(t)$ be respectively the Blocks, Bumps, HeaviSine and Doppler functions defined in Donoho and Johnstone (1994), i.e.

$$c_1(t) = \sum_j h_j K(t - t_j) \ , \quad K(t) = (1 + \text{sign}(t))/2$$
$$(t_j) = (0.1, 0.13, 0.15, 0.23, 0.25, 0.40, 0.44, 0.65, 0.76, 0.78, 0.81)$$
$$(h_j) = (4, -5, 3, -4, 5, -4.2, 2.1, 4.3, -3.1, 2.1, -4.2)$$

$$c_2(t) = \sum_j h_j K((t - t_j)/w_j) \ , \quad K(t) = (1 + |t|)^{-4}$$
$$(t_j) = (0.1, 0.13, 0.15, 0.23, 0.25, 0.40, 0.44, 0.65, 0.76, 0.78, 0.81)$$

$$(h_j) = (4, 5, 3, 4, 5, 4.2, 2.1, 4.3, 3.1, 5.1, 4.2)$$
$$(w_j) = (0.005, 0.005, 0.006, 0.01, 0.01, 0.03, 0.01, 0.01, 0.005, 0.008, 0.005)$$

$$c_3(t) = 4\sin(4\pi t) - \mathrm{sign}(t - 0.3) - \mathrm{sign}(0.72 - t)$$

$$c_4(t) = \sqrt{t(1-t)}\sin(2\pi(1 + 0.05)/(t + 0.05)) \ .$$

We will then study the following SPD curves, which can be visualized in Figure F4:

$$f_1(t) = (10 + c_1) \circ f(t) \ , \qquad f_2(t) = (1 + c_2) \circ f(t) \ ,$$
$$f_3(t) = (7 + c_3) \circ f(t) \ , \qquad f_4(t) = (1 + c_4) \circ f(t) \ .$$

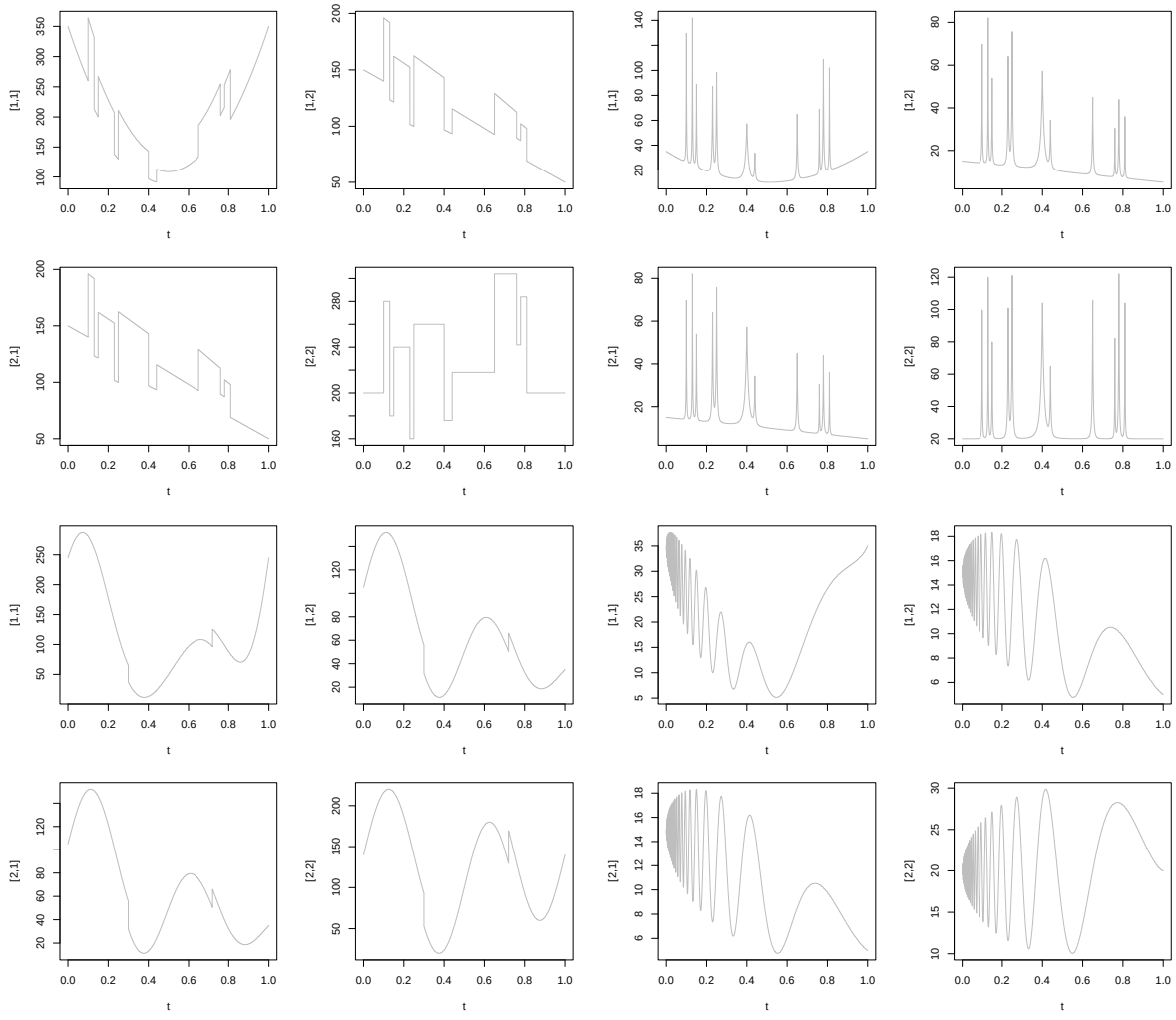


Figure F4: Left to right, top to bottom: f_1, f_2, f_3, f_4.

F.1 Log-normal noise

We study multiple possibilities for the noise structure. We simulate data following model (12) as in Definition 1, where the i.i.d. noise matrices $(\xi_k)_{k=0}^n$ are constructed as described in Assumption 2; i.e. we simulate log-normal noise. We will consider the following scenarios:

$$\sigma_{a,b} = 0.01 \quad \forall a \leq b \ ; \tag{Homog1}$$

$$\sigma_{a,b} = 0.1 \quad \forall a \leq b \; ; \tag{Homog2}$$
$$\sigma_{a,b} = 0.3 \quad \forall a \leq b \; ; \tag{Homog3}$$
$$\sigma_{1,1} = 0.01, \; \sigma_{1,1} = 0.1, \; \sigma_{1,2} = 0.001 \; ; \tag{Inhomog1}$$
$$\sigma_{1,1} = 0.01, \; \sigma_{1,1} = 0.1, \; \sigma_{1,2} = 0.1 \; ; \tag{Inhomog2}$$
$$\sigma_{1,1} = 0.01, \; \sigma_{1,1} = 0.1, \; \sigma_{1,2} = 0.3 \; ; \tag{Inhomog3}$$

For each simulation, we estimate the underlying curve using linear thresholding, trace thresholding (`TT`), upper-sum thresholding (`UST`), Chi-square thresholding (`Chi2T`) and intrinsic elementwise thresholding (`IEWT`). We will use the log-Euclidean AI wavelet transform with $N = 3$, and define our nonlinear wavelet thresholding estimator as in equation (23) with $J_0 = 3$ and $J_1 = 10$. Additionally, we impose a tree structure constraint in the nonlinear thresholding as described in Donoho (1997), where the goal is to exploit the inter-scale correlation between the wavelet coefficients to obtain a better reconstruction in the presence of singularities in the signal. Finally, we estimate the variances $\sigma^2_{a,b}$ (necessary to compute the thresholds) by the usual robust mean-absolute deviation - scaled to match (approximately) Gaussian data - on the finest scale: $\widehat{\sigma}^2_{a,b} = \text{MAD}[\mathfrak{d}_{J,k,a,b}]/0.6745$. For the linear thresholding estimator, we set $J_0 = 6$ to be able to adapt to the less regular parts of the functions, without incorporating too many noisy coefficients. We summarize the results of our simulation study in Figure F5, using as performance metric the Riemannian mean-squared error with respect to the log-Euclidean distance:

$$\text{RieMSE}(\widehat{f}^*, f) = \frac{1}{n} \sum_{k=0}^{n-1} \delta^2_{\text{L}}(\widehat{S}^*_k, S_k) \; .$$

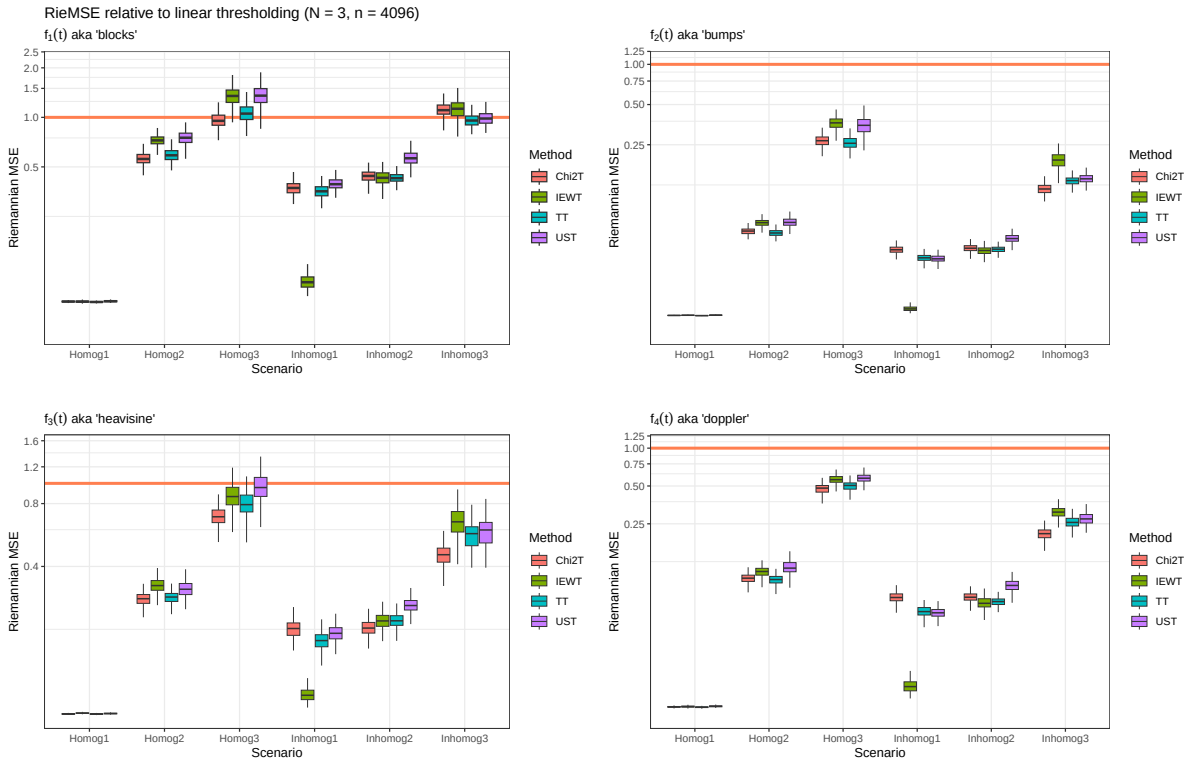


Figure F5: Boxplots, on $M = 500$ simulations for each scenario, of the nonlinear thresholding estimators (relative to the performance of linear thresholding) for (left to right, top to bottom) f_1, f_2, f_3, f_4.

The RieMSE of the nonlinear thresholding estimators are standardized by the RieMSE of the linear estimator in each simulation: hence, values above 1 indicate worse performance while values below 1 indicate that linear thresholding is outperformed by the nonlinear estimator: we immediately observe

that nonlinear rules globally outperform the linear estimator. Comparing the rules, we observe that scalar rules are not significantly different (we note that `UST` is generally worse than `TT` and `Chi2T`), while `IEWT` does not seem to work well except in the Inhomog1 and Inhomog2 scenarios.

More interestingly, we simulate a SPD curve with $d = 5$, which consists of 15 curves over elements which satisfy the positive-definite constraint (5 on the diagonal and 10 off-diagonal) defined by

$$f^{(HD)}(t) = \vartheta(t) \sum_{1 \le a \le b \le d} f_{ab}(t) E_{a,b}$$

where $\vartheta(t) = 10\, \mathbb{1}\{t \le 0.6\} + 5\, \mathbb{1}\{t > 0.6\}$, $(E_{a,b})_{1 \le a \le b \le d}$ is the canonical basis of $\mathrm{Sym}(d)$ and

$$\begin{aligned}
&f_{11}(t) = 10 + (5 - 10t)^2 && f_{12}(t) = 5 - 10t && f_{24}(t) = (1 - t^2)^2 \\
&f_{22}(t) = 10 && f_{13}(t) = \sin(10\pi t) && f_{25}(t) = t \\
&f_{33}(t) = 5 + 10t && f_{14}(t) = \log(t) - 6 && f_{34}(t) = \sin(20\pi t) \\
&f_{44}(t) = 13 && f_{15}(t) = -2t && f_{35}(t) = 0 \\
&f_{55}(t) = 200 - (5 - 12t)^2) && f_{23}(t) = \exp(t) - 5 && f_{45}(t) = 1 - |t - 0.5|
\end{aligned}$$

For the homogeneous noise variances cases, we use the same values as the $d = 2$ case. For the inhomogeneous ones, we use

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Inhomog1	0.025	0.010	0.021	0.021	0.014	0.021	0.013	0.010	0.026	0.004	0.027	0.029	0.021	0.016	0.014
Inhomog2	0.086	0.020	0.088	0.022	0.037	0.014	0.020	0.071	0.086	0.082	0.036	0.065	0.083	0.010	0.012
Inhomog3	0.133	0.088	0.054	0.292	0.203	0.205	0.017	0.290	0.008	0.195	0.262	0.273	0.188	0.118	0.256

In this framework, the more flexible `Chi2T` and `IEWT` definitely outperform the other two rules. This confirms that in particular the trace is not sufficient to describe the overall behaviour of the wavelet coefficients – and that our new thresholding rules are better suited to more realistic, higher-dimensional frameworks.

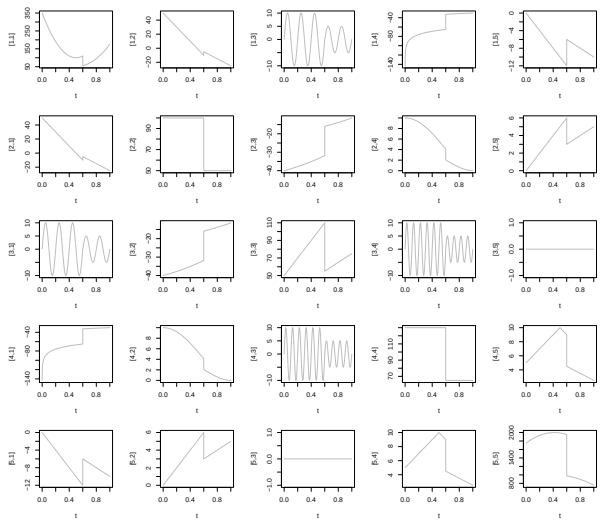


Figure F6: Curve in $\mathrm{Sym}^+(5)$ with very inhomogeneous smoothness between elements.

F.2 Wavelet denoising of empirical TVCM

In the following, we apply our wavelet denoising scheme to estimation of the TVCM of some series $(X_t)_{t=1}^T$ generated by model (16) and such that $X_t \sim N(0, \Sigma(t))$ where $\Sigma(t)$ is chosen among the 2-

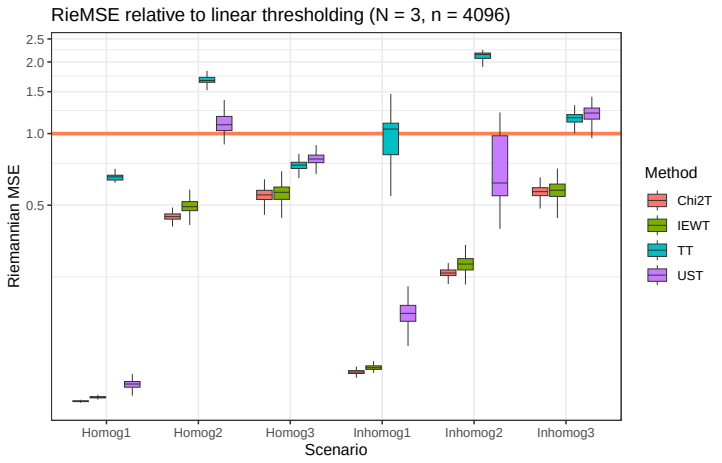


Figure F7: Boxplots, on $M = 500$ simulations for each scenario, of the nonlinear thresholding estimators (relative to the performance of linear thresholding) of the higher-dimensional SPD curve.

dimensional curves defined earlier. We will first study the performance of our methodology for denoising TVCMs of time series constructed from white noise, and then we assess those performances on time series constructed from VAR(1) processes. In this framework, we will only use linear and Chi-squared thresholding; we chose the latter compared to the other nonlinear thresholds because it was shown to be quite adaptive in many situations and because of its enhanced originality. We observe in the two following scenarios that it indeed outperforms the linear estimator, especially when choosing the optimal window length $L = 2^{\lceil \log_2(T^{N/(3N+1)}) \rceil}$.

Independent observations. We simulate time series $X_t = \Sigma_t^{1/2} \varepsilon_t$ where $\varepsilon_t \overset{\text{i.i.d.}}{\sim} N(0, I_d)$, for $t = 1, \ldots, T$ where $T = 2^{15}$. We use the pilot estimator (18), and increase the window length L from 2 to 256.

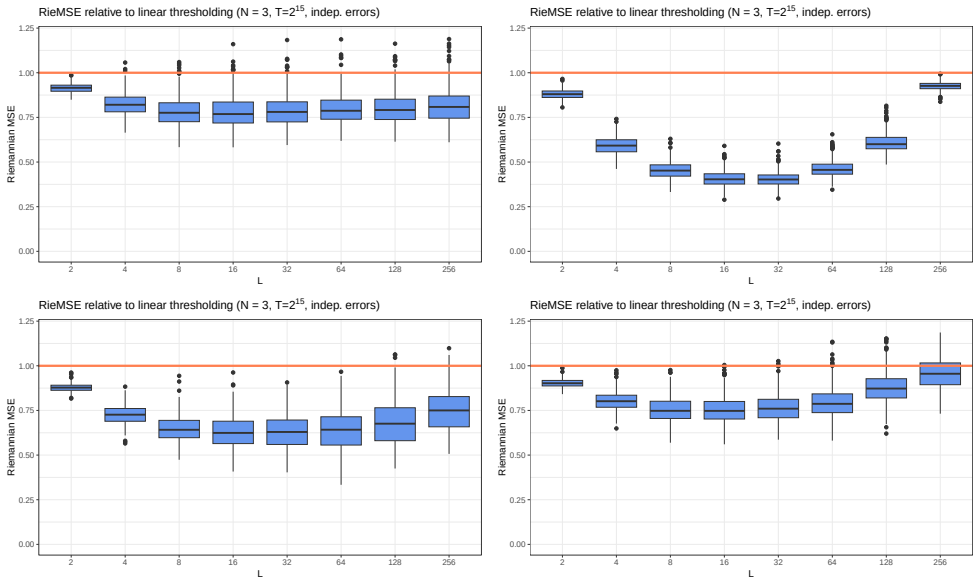


Figure F8: Performance (RieMSE) of the Chi-squared thresholding estimator relative to linear thresholding, and as a function of window length L.

Dependent observations. We simulate time series $X_t = \Sigma_t^{1/2} \varepsilon_t$ where $\varepsilon_t = A\, \varepsilon_{t-1} + \upsilon_t$, with $\upsilon_t \sim N(0, \Upsilon)$ and Υ such that $\mathrm{Var}[\varepsilon_t] = I_d$ for all t, as the solution to

$$\mathrm{vec}(\Upsilon) = (I_{d^2} - A \otimes A)\, \mathrm{vec}(I_d)\ .$$

We consider two scenarios for A, one for weaker and one for stronger dependence – respectively,

$$A^{(1)} = \begin{pmatrix} 0.3 & 0.15 \\ -0.05 & 0.1 \end{pmatrix} \quad \text{and} \quad A^{(2)} = \begin{pmatrix} 0.8 & 0.2 \\ -0.05 & 0.5 \end{pmatrix}\ .$$

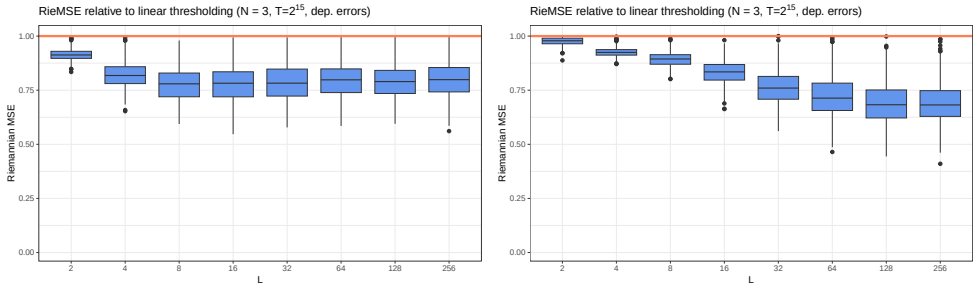


Figure F9: Performance of the Chi-squared thresholding estimator relative to linear thresholding, on the curve $f_1(t)$ and as a function of window length L. Left: $A^{(1)}$, right: $A^{(2)}$.

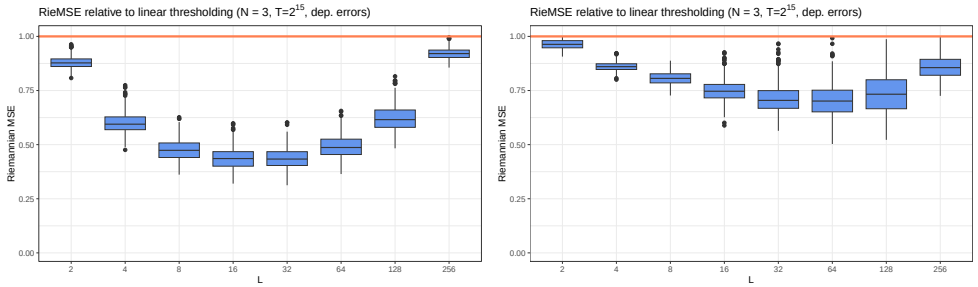


Figure F10: Performance of the Chi-squared thresholding estimator relative to linear thresholding, on the curve $f_2(t)$ and as a function of window length L. Left: $A^{(1)}$, right: $A^{(2)}$.

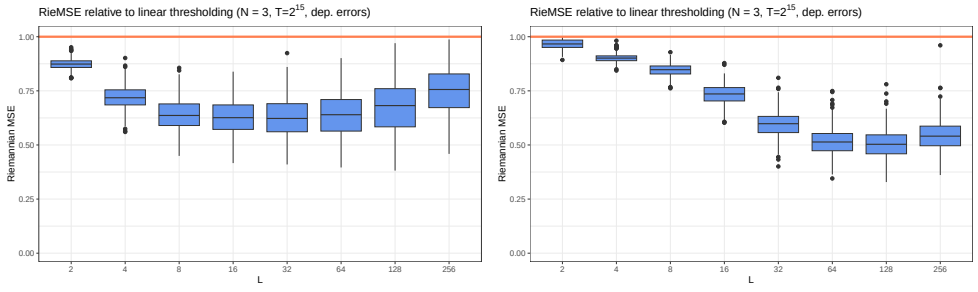


Figure F11: Performance of the Chi-squared thresholding estimator relative to linear thresholding, on the curve $f_3(t)$ and as a function of window length L. Left: $A^{(1)}$, right: $A^{(2)}$.

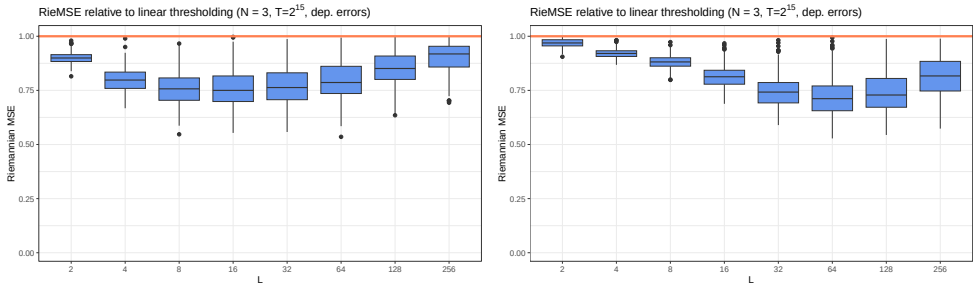


Figure F12: Performance of the Chi-squared thresholding estimator relative to linear thresholding, on the curve $f_4(t)$ and as a function of window length L. Left: $A^{(1)}$, right: $A^{(2)}$.

References

Arsigny, V., Fillard, P., Pennec, X., and Ayache, N. (2007). Geometric means in a novel vector space structure on symmetric positive-definite matrices. *SIAM Journal on Matrix Analysis and Applications*, 29(1):328–347.

Bhatia, R. (2007). *Positive Definite Matrices*. Princeton Series in Applied Mathematics. Princeton University Press, Princeton, NJ, USA.

Boucheron, S., Lugosi, G., and Massart, P. (2013). *Concentration Inequalities*. Oxford University Press.

Chau, J. and von Sachs, R. (2021). Intrinsic wavelet regression for curves of Hermitian positive definite matrices. *Journal of the American Statistical Association*, 116(534):819–832.

Donoho, D. L. (1997). CART and best-ortho-basis: a connection. *The Annals of Statistics*, 25(5).

Donoho, D. L. and Johnstone, I. M. (1994). Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, 81(3):425–455.

Higham, N. J. (2008). *Functions of Matrices: Theory and Computation*. Society for Industrial and Applied Mathematics.

Ho, J., Cheng, G., Salehian, H., and Vemuri, B. (2013). Recursive Karcher expectation estimators and geometric law of large numbers. In Carvalho, C. M. and Ravikumar, P., editors, *Proceedings of the Sixteenth International Conference on Artificial Intelligence and Statistics*, volume 31 of *Proceedings of Machine Learning Research*, pages 325–332, Scottsdale, Arizona, USA. PMLR.

Rademacher, D., Krebs, J., and von Sachs, R. (2022). Statistical inference for wavelet curve estimators of SPD covariance matrices in a log-Euclidean manifold. Technical report. Unpublished manuscript.

Smith, S. (2005). Covariance, subspace, and intrinsic Cramér-Rao bounds. *IEEE Transactions on Signal Processing*, 53(5):1610–1630.

Thanwerdas, Y. and Pennec, X. (2023). O(n)-invariant Riemannian metrics on SPD matrices. *Linear Algebra and its Applications*, 661:163–201.

Wachel, Y. (2022). Wavelet Thresholding for Symmetric Positive-Definite Matrix-Valued Curves in a Log-Euclidean Manifold. Master's thesis, Université catholique de Louvain, supervisor: Rainer von Sachs.