

Intrinsic wavelet regression for curves of Hermitian positive definite matrices

Joris Chau* and Rainer von Sachs†

Abstract

In multivariate time series analysis, non-degenerate autocovariance and spectral density matrices are necessarily Hermitian and positive definite and it is important to preserve these properties in any estimation procedure. Our main contribution is the development of intrinsic wavelet transforms and nonparametric wavelet regression for curves in the non-Euclidean space of Hermitian positive definite matrices. The primary focus is on the construction of intrinsic average-interpolation wavelet transforms in the space equipped with a natural invariant Riemannian metric. In addition, we derive the wavelet coefficient decay and linear wavelet thresholding convergence rates of intrinsically smooth curves of Hermitian positive definite matrices. The intrinsic wavelet transforms are computationally fast and nonlinear wavelet shrinkage or thresholding captures localized features, such as cups or kinks, in the matrix-valued curves. In the context of nonparametric spectral estimation, the intrinsic linear or nonlinear wavelet spectral estimator satisfies the important property that it is equivariant under a change of basis of the time series, in contrast to most existing approaches. The finite-sample performance of the intrinsic wavelet spectral estimator based on nonlinear tree-structured trace thresholding is benchmarked against several state-of-the-art nonparametric curve regression procedures in the Riemannian manifold by means of simulated time series data.

Keywords: Riemannian manifold, Hermitian positive definite matrices, Intrinsic wavelet transform, Nonlinear wavelet thresholding, Spectral matrix estimation, Multivariate time series.

1 Introduction

In multivariate time series analysis, the second-order behavior of a multivariate time series is studied by means of its autocovariance matrices in the time domain, or its spectral density matrices in the frequency domain. Non-degenerate spectral density matrices are necessarily curves of Hermitian positive definite (HPD) matrices, and one generally constrains a spectral curve estimator to preserve these properties. This is important for several reasons: i) interpretation of the spectral estimator as the Fourier transform of symmetric positive definite (SPD) autocovariance matrices

*Corresponding author, j.chau@uclouvain.be, Institute of Statistics, Biostatistics, and Actuarial Sciences (ISBA), Université catholique de Louvain, Voie du Roman Pays 20, B-1348, Louvain-la-Neuve, Belgium.

†Institute of Statistics, Biostatistics, and Actuarial Sciences (ISBA), Université catholique de Louvain, Voie du Roman Pays 20, B-1348, Louvain-la-Neuve, Belgium.

in the time domain or as HPD covariance matrices across frequency in the Fourier domain; ii) well-defined transfer functions in the Cramér representation of the time series for the purpose of e.g. simulation of time series and bootstrapping; iii) sufficient regularity to avoid computational problems in subsequent inference procedures (requiring e.g. the inverse of the estimated spectrum). Our main contribution is the development of intrinsic wavelet transforms and nonparametric wavelet regression for curves in the non-Euclidean space of HPD matrices, exploiting the geometric structure of the space as a Riemannian manifold. In this work, the primary focus is on nonparametric spectral density matrix estimation of stationary multivariate time series, but we emphasize that the methodology applies equally to general matrix-valued curve estimation or denoising problems, where the target is a curve of symmetric or Hermitian PD matrices. Examples include; curve denoising of SPD diffusion covariance matrices in diffusion tensor imaging as in e.g. Yuan et al. (2012), or estimation of time-varying autocovariance matrices of a locally stationary time series as in e.g. Dahlhaus (2012).

A first important consideration to perform estimation in the space of HPD matrices is the associated metric in the space. The metric gives the space its curvature and induces a distance between HPD matrices. Traditional nonparametric spectral matrix estimation relies on smoothing the periodogram via e.g. kernel regression as in (Brillinger, 1981, Chapter 5), (Brockwell and Davis, 2006, Chapter 11), or multitaper spectral estimation as in e.g. Walden (2000). These approaches equip the space of HPD matrices with the Euclidean (i.e. Frobenius) metric and view it as a *flat* space. An important disadvantage is that this metric space is incomplete, as the boundary of singular matrices lies at a finite distance. For this reason, flexible nonparametric (e.g. wavelet- or spline-) periodogram smoothing embedded in a Euclidean space cannot guarantee a PD spectral estimate. Exceptions to this rule include inflexible kernel or multitaper periodogram smoothing, which rely on a sufficiently large equivalent smoothing bandwidth for each matrix component. To avoid this issue, Dai and Guo (2004), Rosen and Stoffer (2007) and Krafty and Collinge (2013) among others construct a HPD spectral estimate as the square of an estimated curve of Cholesky square root matrices. This allows for more flexible estimation of the spectrum, such as individual smoothing of Cholesky matrix components, while at the same time guaranteeing a HPD spectral estimate. In this context, the space of HPD matrices is equipped with the Cholesky metric, where the distance between two matrices is given by the Euclidean distance between their Cholesky square roots. Unfortunately, the Cholesky metric and Cholesky-based smoothing are not equivariant to permutations of the components of the underlying time series. That is, if one observes a reordering of the

time series components, the spectral estimate is not necessarily a permuted version of the spectral estimate obtained under the original ordering of the time series components.

In this work, we exploit the geometric structure of the space of HPD matrices as a Riemannian manifold equipped with the natural invariant (Smith (2000)) –also affine-invariant (Pennec et al. (2006)), canonical (Holbrook et al. (2016)), trace (Yuan et al. (2012)), Rao-Fisher (Said et al. (2015))– Riemannian metric, or simply the Riemannian metric (Bhatia, 2009, Chapter 6), Dryden et al. (2009)). The natural invariant Riemannian metric plays an important role in estimation problems in the space of symmetric or Hermitian PD matrices for several reasons: (i) the space of HPD matrices equipped with the Riemannian metric is a complete metric space, (ii) there is no swelling effect as with the Euclidean metric, where interpolating two HPD matrices may yield a matrix with a determinant larger than either of the original matrices (e.g. Pasternak et al. (2010)), and (iii) the induced Riemannian distance is invariant under congruence transformation by any invertible matrix, see Section 2. The first property guarantees a HPD spectral estimate, while allowing for flexible spectral matrix estimation as with Cholesky-based smoothing. The third property ensures that the spectral estimator is –not only– permutation or unitary congruence equivariant, but also *general linear congruence equivariant*, which essentially implies that the estimator does not non-trivially depend on the chosen coordinate system of the time series. In Dryden et al. (2009), the authors list several additional suitable metrics to perform estimation in the space of HPD matrices, such as the Log-Euclidean metric also discussed in e.g. Yuan et al. (2012) or Boumal and Absil (2011b). The Log-Euclidean metric transforms the space of HPD matrices in a complete metric space and is unitary congruence invariant, but not general linear congruence invariant.

Several recent works on nonparametric curve regression in the space of SPD matrices equipped with the Riemannian metric include: intrinsic geodesic and linear regression in Pennec et al. (2006) and Zhu et al. (2009) among others, intrinsic local polynomial regression in Yuan et al. (2012) and intrinsic penalized spline-like regression in Boumal and Absil (2011b). In the context of frequency-specific spectral matrix estimation Holbrook et al. (2016) recently introduced a Bayesian geodesic Lagrangian Monte Carlo (gLMC) approach based on the Riemannian metric. The latter may not be best-suited to estimation of the entire spectral curve, as this requires application of the gLMC to each individual Fourier frequency, which is computationally quite challenging. In this work, we develop fast intrinsic wavelet regression in the manifold of HPD matrices equipped with the Riemannian metric. Wavelet-based estimation of spectral matrices allows us to capture poten-

tially very localized features, such as local peaks or troughs in the spectral matrix at pointwise frequencies or frequency bands, in contrast to the approaches mentioned above, which rely on globally homogeneous smoothness behavior. The intrinsic wavelet transform is a generalization of the average-interpolation (AI) wavelet transform on the real line in Donoho (1993). In this sense it is related to the *midpoint-interpolation* (MI) approach in Rahman et al. (2005) for general symmetric Riemannian manifolds with tractable exponential and logarithmic maps. The fundamental difference lies in the fact that we do not apply a Euclidean refinement scheme to the data projected a set of tangent spaces as in Rahman et al. (2005). Such an approach introduces a certain degree of ambiguity as the base points of the tangent spaces to which the midpoints (i.e. scaling coefficients) are projected need to be specified by the user and different base points lead to different wavelet coefficients. In the introduced intrinsic wavelet transform there is no such ambiguity as we construct an intrinsic AI refinement scheme directly in the manifold equipped with the Riemannian metric. In fact, the construction of the wavelet transform is valid for any of the above-mentioned metrics, and one can substitute the Riemannian metric by e.g. the Log-Euclidean or Cholesky metric. To further illustrate, the intrinsic wavelet transform based on the Euclidean metric simplifies to the matrix version of the AI wavelet transform in Donoho (1993). The second advantage of a truly intrinsic approach is that in contrast to the MI approach in Rahman et al. (2005), the wavelet refinement scheme of order $k \geq 0$ reproduces *intrinsic polynomial* curves up to order k as defined in Hinkle et al. (2014), see Section 2. This allows us to derive wavelet coefficient decay and non-parametric estimation rates for smooth curves of HPD matrices.

The structure of the paper is as follows. In Section 2, we introduce the necessary geometric notions and tools and in Section 3 we describe the intrinsic AI refinement scheme and forward and backward wavelet transform in the Riemannian manifold. In Section 4, we derive wavelet coefficient decay rates of intrinsically smooth curves of HPD matrices and convergence rates of linear wavelet thresholding. In Section 5, we consider intrinsic wavelet regression in the context of spectral matrix estimation. In particular, we examine nonlinear tree-structured thresholding of matrix-valued wavelet coefficients based on their *trace* as it is shown that, asymptotically, the traces of the wavelet coefficients decompose into a scalar additive signal plus noise model with homogeneous variances across wavelet scales. In Section 6, we compare the finite-sample performance of wavelet-based spectral matrix estimation to several benchmark procedures. The technical proofs are deferred to the appendix and a toolbox for data analysis and estimation problems in the space of HPD matrices is publicly available in the R-package `pdSpecEst` on CRAN, Chau (2017).

2 Geometry of HPD matrices

In order to develop an intrinsic wavelet transform for curves in the space of HPD (Hermitian positive definite) matrices, we consider the space as a Riemannian manifold equipped with a specific invariant Riemannian metric, see e.g. Pennec et al. (2006), (Bhatia, 2009, Chapter 6), or Smith (2000) for more details. Below, we introduce the necessary tools and notions for the construction of a proper intrinsic manifold wavelet transform.

Riemannian metric For notational simplicity, let us denote $\mathcal{M} := \mathbb{P}_{d \times d}$ for the space of $(d \times d)$ -dimensional HPD matrices, a d^2 -dimensional smooth Riemannian manifold. The tangent space $\mathcal{T}_p(\mathcal{M})$ at a point, i.e. a matrix, $p \in \mathcal{M}$ can be identified by the real vector space $\mathcal{H} := \mathbb{H}_{d \times d}$ of $(d \times d)$ -dimensional Hermitian matrices, and as detailed in Pennec et al. (2006), the Frobenius inner product on \mathcal{H} induces a natural invariant Riemannian metric on the manifold \mathcal{M} given by the smooth family of inner products:

$$\langle h_1, h_2 \rangle_p = \text{Tr}((p^{-1/2} * h_1)(p^{-1/2} * h_2)), \quad \forall p \in \mathcal{M}, \quad (2.1)$$

with $h_1, h_2 \in \mathcal{T}_p(\mathcal{M})$. Here and throughout this paper, $y^{1/2}$ always denotes the Hermitian square root matrix of $y \in \mathcal{M}$, and we write $y * x := y^* x y$ for matrix congruence transformation, where $*$ denotes the conjugate transpose of a matrix. The Riemannian distance on \mathcal{M} derived from the Riemannian metric is given by:

$$\delta(p_1, p_2) = \|\text{Log}(p_1^{-1/2} * p_2)\|_F, \quad (2.2)$$

where $\|\cdot\|_F$ denotes the matrix Frobenius norm and $\text{Log}(\cdot)$ is the matrix logarithm. The mapping $x \mapsto a * x$ is an isometry for each invertible matrix $a \in \text{GL}(d, \mathbb{C})$, i.e. it is distance-preserving:

$$\delta(p_1, p_2) = \delta(a * p_1, a * p_2), \quad \forall a \in \text{GL}(d, \mathbb{C}),$$

and by (Bhatia, 2009, Theorem 6.1.6 and Prop. 6.2.2), the shortest curve with respect to the Riemannian distance, i.e. the *geodesic* segment, joining any two points $p_1, p_2 \in \mathcal{M}$ is unique and can be parametrized as,

$$\gamma(p_1, p_2, t) = p_1^{1/2} * (p_1^{-1/2} * p_2)^t, \quad 0 \leq t \leq 1. \quad (2.3)$$

Since (\mathcal{M}, δ) is a complete separable metric space, the Hopf-Rinow Theorem implies that every geodesic curve can be extended indefinitely.

Exp- and Log-maps The characterization of the intrinsic manifold wavelet coefficients requires the notions of exponential and logarithmic maps, relating the manifold \mathcal{M} to its tangent spaces $\mathcal{T}_p(\mathcal{M})$ in a one-to-one fashion. By (Pennec et al. (2006)) the exponential maps $\text{Exp}_p : \mathcal{T}_p(\mathcal{M}) \rightarrow \mathcal{M}$ are (locally) diffeomorphic maps from the tangent space at a point $p \in \mathcal{M}$ to the manifold given by,

$$\text{Exp}_p(h) = p^{1/2} * \text{Exp}\left(p^{-1/2} * h\right), \quad \forall h \in \mathcal{T}_p(\mathcal{M}),$$

where $\text{Exp}(\cdot)$ denotes the matrix exponential. Since \mathcal{M} is a geodesically complete manifold and minimizing geodesics are always unique, it follows by (do Carmo, 1992, Chapter 13) that for each $p \in \mathcal{M}$ the image of the exponential map Exp_p is the entire manifold \mathcal{M} , and the exponential maps are in fact global diffeomorphisms. In the other direction, the logarithmic maps $\text{Log}_p : \mathcal{M} \rightarrow \mathcal{T}_p(\mathcal{M})$ are global diffeomorphisms from the manifold to the tangent space at the point $p \in \mathcal{M}$, given by the inverse exponential maps:

$$\text{Log}_p(\tilde{p}) = p^{1/2} * \text{Log}\left(p^{-1/2} * \tilde{p}\right).$$

The Riemannian distance can now also be expressed in terms of the logarithmic map as:

$$\delta(p_1, p_2) = \|\text{Log}_{p_1}(p_2)\|_{p_1} = \|\text{Log}_{p_2}(p_1)\|_{p_2}, \quad \forall p_1, p_2 \in \mathcal{M}, \quad (2.4)$$

where throughout this paper $\|h\|_p := \langle h, h \rangle_p$ denotes the norm of $h \in \mathcal{T}_p(\mathcal{M})$ induced by the Riemannian metric.

Parallel transport The parallel transport is required to define intrinsic polynomials and intrinsic Taylor expansions of smooth curves as in (Lang, 1995, Chapter 9), which allow for the derivation of the wavelet coefficient decay for smooth curves in the manifold.

Let $\gamma : I \rightarrow \mathcal{M}$ for some interval $I \subseteq \mathbb{R}$ be a smooth curve and $P(t)$ be a vector field along γ , such that $P(t) \in \mathcal{T}_{\gamma(t)}(\mathcal{M})$ for each $t \in I$. The parallel transport $\Gamma(\gamma)_{t_0}^{t_1} : \mathcal{T}_{\gamma(t_0)}(\mathcal{M}) \rightarrow \mathcal{T}_{\gamma(t_1)}(\mathcal{M})$ transports a vector in the tangent space at $\gamma(t_0)$ to the tangent space at $\gamma(t_1)$ along the curve γ , such that the inner product between vectors across tangent spaces is preserved, i.e.

$$\langle \Gamma(\gamma)_{t_0}^{t_1}(h_1), \Gamma(\gamma)_{t_0}^{t_1}(h_2) \rangle_{\gamma(t_1)} = \langle h_1, h_2 \rangle_{\gamma(t_0)}, \quad \forall h_1, h_2 \in \mathcal{T}_{\gamma(t_0)}(\mathcal{M}).$$

The *covariant derivative* (or affine connection) ∇ induced by the Riemannian metric can be retrieved from the parallel transport by differentiation as,

$$\nabla_{\dot{\gamma}} P(t) = \lim_{\Delta t \rightarrow 0} \frac{\Gamma(\gamma)_{t+\Delta t}^t(P(t+\Delta t)) - P(t)}{\Delta t}.$$

Here, the time derivative $\dot{\gamma}(t)$ is a tangent vector in $T_{\gamma(t)}(\mathcal{M})$ and can be expressed using the logarithmic map (i.e. in terms of *normal coordinates*) as,

$$\dot{\gamma}(t) := \frac{d}{dt}\gamma(t) = \lim_{\Delta t \rightarrow 0} \frac{\text{Log}_{\gamma(t)}(\gamma(t + \Delta t))}{\Delta t} \in \mathcal{T}_{\gamma(t)}(\mathcal{M}). \quad (2.5)$$

A vector field $P(t)$ is parallel transported along the curve γ if $\nabla_{\dot{\gamma}}P(t) = \mathbf{0}$ for each $t \in I$. In particular, geodesic curves –previously defined as shortest line segments on (\mathcal{M}, δ) – can also be characterized as the curves γ for which $\dot{\gamma}$ is parallel transported along the curve itself, i.e. $\nabla_{\dot{\gamma}}\dot{\gamma}(t) = \mathbf{0}$ for each $t \in I$.

In (\mathcal{M}, δ) , parallel transport of a vector $w \in \mathcal{T}_p(\mathcal{M})$ from a point $p \in \mathcal{M}$ along a geodesic curve in the direction of $v \in T_p(\mathcal{M})$ for time Δt is given explicitly by:

$$\mathfrak{T}(p, \Delta tv, w) = \text{Exp}_p(\Delta tv/2) * p^{-1} * w.$$

Substituting $\Delta tv = \text{Log}_p(\text{Id})$, we obtain the *whitening* transport as a specific case transporting $w \in \mathcal{T}_p(\mathcal{M})$ to the tangent space at the identity $\mathcal{T}_{\text{Id}}(\mathcal{M})$ along a geodesic curve,

$$\mathfrak{T}(p, \text{Log}_p(\text{Id}), w) = p^{-1/2} * w \in \mathcal{T}_{\text{Id}}(\mathcal{M}). \quad (2.6)$$

Intrinsic polynomials Intrinsic polynomials as defined in Hinkle et al. (2014) play a key role in the construction of the intrinsic AI refinement scheme. Essentially, polynomial curves of degree $k \geq 0$ on the manifold are defined as the curves with vanishing k -th and higher order covariant derivatives. Let $\gamma : I \rightarrow \mathcal{M}$ be a smooth curve on the manifold, with existing covariant derivatives of all orders, then it is said to be a polynomial curve of degree k if,

$$\nabla_{\dot{\gamma}}^\ell \dot{\gamma}(t) = \mathbf{0}, \quad \forall \ell \geq k \text{ and } t \in I,$$

where $\nabla_{\dot{\gamma}}^0 \dot{\gamma}(t) := \dot{\gamma}(t)$. A zero degree polynomial is a curve for which $\dot{\gamma}(t) = \mathbf{0}$, i.e. a constant curve. A first-degree polynomial is a curve for which $\nabla_{\dot{\gamma}}\dot{\gamma}(t) = \mathbf{0}$ corresponding to a geodesic curve, i.e. a *straight* line on the manifold. In general, higher degree polynomials are difficult to represent in closed form, but discretized polynomial curves are straightforward to generate via numerical integration as outlined in Hinkle et al. (2014).

Intrinsic means A random variable $X : \Omega \rightarrow \mathcal{M}$ is a measurable function from a probability space $(\Omega, \mathcal{A}, \nu)$ to the measurable space $(\mathcal{M}, \mathcal{B}(\mathcal{M}))$, where $\mathcal{B}(\mathcal{M})$ is the Borel algebra in the complete separable metric space (\mathcal{M}, δ) . By $P(\mathcal{M})$, we denote the set of all probability measures

on $(\mathcal{M}, \mathcal{B}(\mathcal{M}))$ and $P_p(\mathcal{M})$ denotes the subset of probability measures in $P(\mathcal{M})$ that have finite moments of order p with respect to the Riemannian distance,

$$P_p(\mathcal{M}) := \left\{ \nu \in P(\mathcal{M}) : \exists y_0 \in \mathcal{M}, \text{ s.t. } \int_{\mathcal{M}} \delta(y_0, x)^p \nu(dx) < \infty \right\}$$

Note that if $\int_{\mathcal{M}} \delta(y_0, x)^p \nu(dx) < \infty$ for some $y_0 \in \mathcal{M}$ and $1 \leq p < \infty$, this is true for any $y \in \mathcal{M}$. This follows by the triangle inequality and the fact that $\delta(p_1, p_2) < \infty$ for any $p_1, p_2 \in \mathcal{M}$, as $\int_{\mathcal{M}} \delta(y, x)^p \nu(dx) \leq 2^p (\delta(y, y_0)^p + \int_{\mathcal{M}} \delta(y_0, x)^p \nu(dx)) < \infty$.

In the intrinsic AI refinement scheme, the center of a random variable $X \sim \nu$ is characterized by its intrinsic (also Karcher or Fréchet) mean, see e.g. Pennec (2006). The set of intrinsic means is given by the points that minimize the second moment with respect to the Riemannian distance,

$$\mu = \mathbb{E}_\nu[X] := \arg \min_{y \in \text{supp}(\nu)} \int_{\mathcal{M}} \delta(y, x)^2 \nu(dx).$$

If $\nu \in P_2(\mathcal{M})$, then at least one intrinsic mean exists and since the manifold \mathcal{M} is a space of non-positive curvature with no cut-locus (see Pennec et al. (2006)), by (Le, 1995, Proposition 1) the intrinsic mean μ is also unique.

3 Intrinsic AI wavelet transforms

The intrinsic average-interpolation (AI) wavelet transform for curves in the space of HPD matrices is a natural unambiguous generalization of the scalar AI wavelet transform on the real line in Donoho (1993). Although closely related to the midpoint-interpolation (MI) wavelet transform in Rahman et al. (2005), an important difference is that the intrinsic AI wavelet transform reproduces higher-order intrinsic polynomial curves in the manifold, whereas the MI wavelet transform only reproduces first-order polynomials, i.e. geodesic curves. As a consequence, this intrinsic polynomial reproduction property allows us to derive wavelet coefficient decay and nonparametric convergence rates for smooth curves in Section 4. We emphasize that the construction of the wavelet transform is based on the idea of lifting transforms, see e.g. Jansen and Oonincx (2005) or Klees and Haagmans (2000) for a general overview of first- and second-generation wavelet transforms using the lifting scheme.

3.1 Intrinsic AI refinement scheme

Midpoint pyramid With in mind the application to spectral matrices, we consider a discretized curve $f(x_\ell) \in \mathcal{M}$ at equidistant points $x_\ell \in \mathcal{I}$ for $\ell = 1, \dots, n$, with $\mathcal{I} \subset \mathbb{R}$. Here, without loss

of generality we set $\mathcal{I} = [0, 1]$ and it is assumed that $n = 2^J$ is dyadic in order to allow for straightforward construction of the wavelet transforms. The latter is not an absolute limitation of the approach, as the lifting wavelet transforms can also be adapted to non-dyadic values of n . Given the sequence $(f(x_\ell))_\ell$, we build a redundant midpoint or scaling coefficient pyramid analogous to Rahman et al. (2005). At the finest scale J , set $M_{J,k} = f(x_{k+1})$ for $k = 0, \dots, 2^J - 1$. At the next coarser scale $j = J - 1$ set,

$$M_{j,k} = \gamma(M_{j+1,2k}, M_{j+1,2k+1}, 1/2), \quad \text{for } k = 0, \dots, 2^j - 1, \quad (3.1)$$

and continue this coarsening operation up to scale $j = 0$, such that each scale j contains a total of 2^j midpoints. Here, $\gamma(p_1, p_2, 1/2)$ is the halfway point or *midpoint* on the geodesic segment connecting $p_1, p_2 \in \mathcal{M}$, which coincides with the intrinsic sample mean of p_1 and p_2 .

Intrinsic polynomial interpolation At scale $j \in \{0, \dots, J - 1\}$, the intrinsic AI refinement scheme takes as input coarse-scale midpoints $(M_{j,k})_k$ and outputs imputed or predicted finer-scale midpoints $(\widetilde{M}_{j+1,k'})_{k'}$. The predicted midpoints are computed as the $(j + 1)$ -scale midpoints of the unique intrinsic polynomial $(\tilde{f}(x_\ell))_\ell$ with j -scale midpoints $(M_{j,k})_k$. In order to reconstruct intrinsic polynomials from a discrete set of points on the manifold, we consider a generalized intrinsic version of Neville's algorithm as in (Ma and Fu, 2012, Chapter 9.2), replacing ordinary linear interpolation by *geodesic interpolation*.

Given $P_0, \dots, P_n \in \mathcal{M}$ and $x_0 < \dots < x_n \in \mathbb{R}$, set $p_{i,i}(x) := P_i$ for all x and $i = 0, \dots, n$. The $p_{i,i}$ are zero-th order polynomials, since $\dot{p}_{i,i}(x) = \mathbf{0}$. Iteratively define,

$$p_{i,j}(x) := \text{Exp}_{p_{i,j-1}(x)} \left(\frac{x - x_i}{x_j - x_i} \text{Log}_{p_{i,j-1}(x)}(p_{i+1,j}(x)) \right), \quad 0 \leq i < j \leq n,$$

where $p_{i+1,j}(x)$ and $p_{i,j-1}(x)$ are the intrinsic polynomials of degree $j - i - 1$ passing through P_{i+1}, \dots, P_j at x_{i+1}, \dots, x_j and through P_i, \dots, P_{j-1} at x_i, \dots, x_{j-1} respectively. Then $p_{i,j}(x)$ is the intrinsic polynomial of degree $j - i$ passing through P_i, \dots, P_j at x_i, \dots, x_j . Continuing the above iterative reconstruction, at the final iteration we obtain the intrinsic polynomial $p_{0,n}(x)$ of order n passing through P_0, \dots, P_n at x_0, \dots, x_n .

To illustrate, $p_{0,1}(x)$ is the geodesic, i.e. first-order intrinsic polynomial, passing through P_0 and P_1 at x_0 and x_1 . In general, since $p_{i,j}(x)$ geodesically interpolates two polynomials of degree $j - i - 1$, $p_{i,j}(x)$ is itself a polynomial of degree $j - i$ introducing one additional higher-degree non-vanishing covariant derivative. This is exactly analogous to the Euclidean setting, where linear interpolation of two polynomials of degree r results in a polynomial of degree $r + 1$.

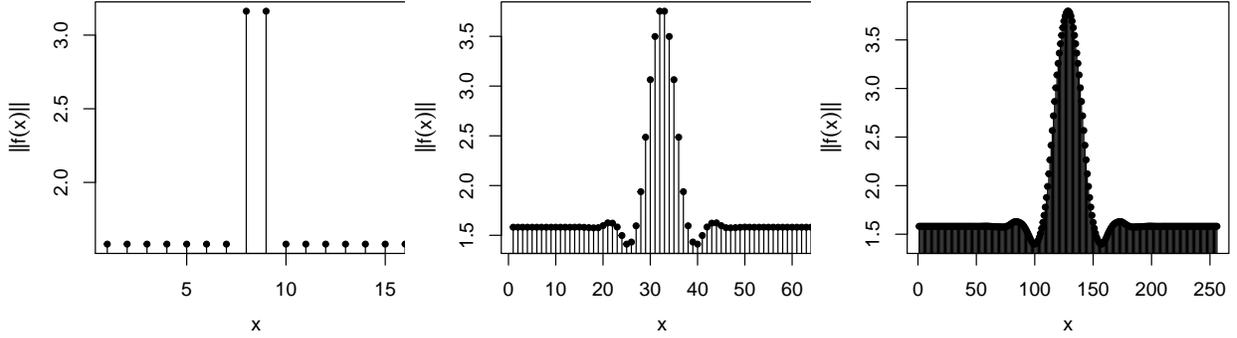


Figure 1: (Away from the boundary) Euclidean norm of $f(x)$ after successive applications of the refinement scheme of order $N = 7$ starting from $n = 2^4$ observations.

3.1.1 Midpoint prediction via intrinsic average interpolation

We emphasize that reconstruction of the intrinsic polynomial $\tilde{f}(x)$ with j -scale midpoints $(M_{j,k})_k$ is *not* the same as reconstructing the intrinsic polynomial passing through the j -scale midpoints. To compute predicted midpoints in the intrinsic AI refinement scheme, instead we consider the cumulative intrinsic mean of $\tilde{f}(x)$, $M_{y_0} : (y_0, 1] \rightarrow \mathcal{M}$, given by:

$$M_{y_0}(y) = \text{Exp}_{M_{y_0}(y)} \left(\int_{y_0}^y \text{Log}_{M_{y_0}(y)}(\tilde{f}(u)) du \right). \quad (3.2)$$

If $\tilde{f}(x)$ is the intrinsic polynomial with j -scale midpoints $(M_{j,k})_{k=0,\dots,2^j-1}$, then $M_0((k+1)2^{-j})$ equals the cumulative intrinsic average of $\{M_{j,0}, \dots, M_{j,k-1}\}$. The key consideration is that the cumulative intrinsic mean of an intrinsic polynomial of order r is again an intrinsic polynomial of order $\leq r$. For instance, given a geodesic segment, i.e. a first-order polynomial, its cumulative intrinsic mean is a geodesic segment moving at half the original speed. Again, this is analogous to the Euclidean setting, where an integrated polynomial remains a polynomial.

Prediction away from the boundary Fix a location $k \in \{L, \dots, 2^{(j-1)} - (L+1)\}$ on scale $j-1$ for some $L \geq 0$. Given the neighboring $j-1$ -scale midpoints $\{M_{j-1,k-L}, \dots, M_{j-1,k}, \dots, M_{j-1,k+L}\}$, the goal is to predict the finer-scale midpoints $\{M_{j,2k}, M_{j,2k+1}\}$. The number of collected neighbors $N := 2L + 1 \geq 0$ is referred to as the *order* or degree of the refinement scheme. First, to predict the midpoint $M_{j,2k+1}$, fit an intrinsic polynomial $\widehat{M}_{(k-L)2^{-(j-1)}}(y)$ of order $N-1$ through the N known points $\{\overline{M}_{j-1,0}, \dots, \overline{M}_{j-1,N-1}\}$ by means of Neville's algorithm, where $\overline{M}_{j-1,\ell}$ denotes the

cumulative intrinsic average:

$$\bar{M}_{j-1,\ell} := M_{(k-L)2^{-(j-1)}}((k-L+\ell)2^{-(j-1)}) = \text{Exp}_{\bar{M}_{j-1,\ell}} \left(\frac{1}{\ell+1} \sum_{i=k-L}^{k-L+\ell} \text{Log}_{\bar{M}_{j-1,\ell}}(M_{j-1,i}) \right). \quad (3.3)$$

By construction of the cumulative intrinsic mean curve, $M_{(k-L)2^{-(j-1)}}((2k+1)2^{-j})$ lies on the geodesic segment connecting the known cumulative average $\bar{M}_{j-1,L}$ and the midpoint $M_{j,2k+1}$. The predicted midpoint $\widetilde{M}_{j,2k+1}$ is then given by:

$$\widetilde{M}_{j,2k+1} = \gamma \left(\bar{M}_{j-1,L}, \widehat{M}_{(k-L)2^{-(j-1)}}((2k+1)2^{-j}), -2L \right),$$

where $M_{(k-L)2^{-(j-1)}}((2k+1)2^{-j})$ is replaced by its estimate $\widehat{M}_{(k-L)2^{-(j-1)}}((2k+1)2^{-j})$. We refer to the proof of Proposition 4.2 for the derivation of the above expression. The value of $\widetilde{M}_{j,2k}$ directly follows from the midpoint relation $\gamma(\widetilde{M}_{j,2k}, \widetilde{M}_{j,2k+1}, 1/2) = M_{j-1,k}$ as:

$$\widetilde{M}_{j,2k} = M_{j-1,k} * \widetilde{M}_{j,2k+1}^{-1}.$$

An important observation is that if the coarse-scale midpoints $\{M_{j-1,k-L}, \dots, M_{j-1,k}, \dots, M_{j-1,k+L}\}$ are generated from an intrinsic polynomial $f(x)$ of degree $\leq N-1$, then the midpoints $\{M_{j,2k}, M_{j,2k+1}\}$ are reproduced without error. This is analogous to the scalar AI refinement scheme in Donoho (1993) and is referred to as the *intrinsic polynomial reproduction* property.

Prediction at the boundary If $k \in \{0, \dots, L-1\} \cup \{2^{(j-1)} - (L-1), \dots, 2^{(j-1)} - 1\}$ is located near or at the boundary, not all symmetric neighbors around $M_{j-1,k}$ are available for prediction of $\{M_{j,2k}, M_{j,2k+1}\}$. Instead, collect the N closest neighbors of $M_{j-1,k}$ either to the left or right and predict the j -scale midpoints as above through $(N-1)$ -th order intrinsic polynomial interpolation based on the non-symmetric neighbors $(M_{j-1,k+\ell})_\ell$. This boundary modification preserves the intrinsic polynomial reproduction property. In Figures 1 and 2, we demonstrate successive applications of the intrinsic AI refinement scheme for an interior and a boundary midpoint starting from a sequence of $n = 2^4$ dummy HPD matrix-valued observations at $x = 1, \dots, n$.

3.1.2 Faster midpoint prediction in practice

In the scalar AI refinement scheme on the real line in Donoho (1993) or (Klees and Haagmans, 2000, pg.95), the predicted j -scale scaling coefficients obtained via polynomial average-interpolation of the $(j-1)$ -scale scaling coefficients are equivalent to weighted linear combinations of the input scaling coefficients, with weights depending on the order of the AI refinement scheme $N =$

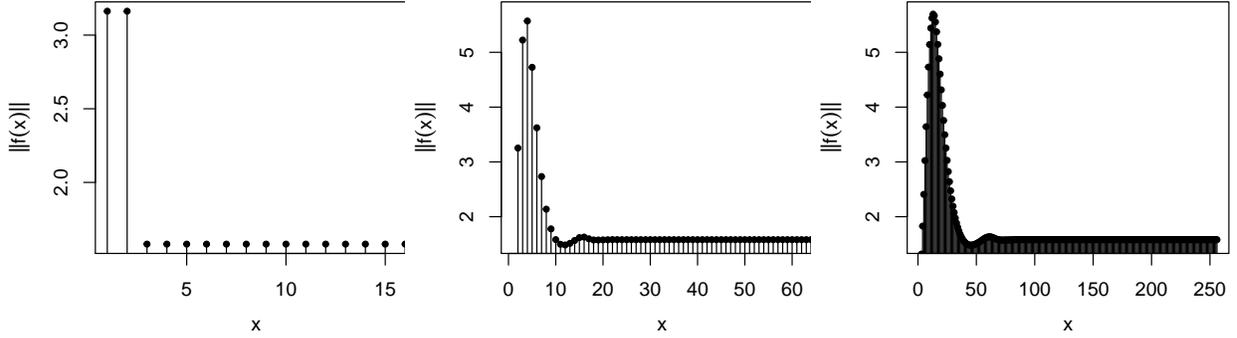


Figure 2: (At the boundary) Euclidean norm of $f(x)$ after successive applications of the refinement scheme of order $N = 7$ starting from $n = 2^4$ observations.

$2L + 1 \geq 0$. In the intrinsic version of Neville's algorithm, the sole change with respect to its Euclidean counterpart is the nature of the interpolation, i.e. linear interpolation is substituted by geodesic interpolation. It turns out that $\{\widetilde{M}_{j,2k}, \widetilde{M}_{j,2k+1}\}$ remain weighted averages of the inputs $\{M_{j-1,k-L}, \dots, M_{j-1,k}, \dots, M_{j-1,k+L}\}$, with the same weights as in the Euclidean case. The difference being that –instead of weighted Euclidean averages– the weighted averages are intrinsic weighted averages in the Riemannian manifold. That is,

$$\begin{aligned} \widetilde{M}_{j,2k} &= \text{Exp}_{\widetilde{M}_{j,2k}} \left(\sum_{\ell=-L}^L C_{N,2\ell+N-1} \text{Log}_{\widetilde{M}_{j,2k}} (M_{j-1,k+\ell}) \right), \\ \widetilde{M}_{j,2k+1} &= \text{Exp}_{\widetilde{M}_{j,2k+1}} \left(\sum_{\ell=-L}^L C_{N,2\ell+N} \text{Log}_{\widetilde{M}_{j,2k+1}} (M_{j-1,k+\ell}) \right), \end{aligned} \quad (3.4)$$

where the weights $\mathbf{C}_N = (C_{N,i})_{i=0,\dots,2N-1}$ depend on the refinement order $N \geq 1$ and sum up to 2. For instance, away from the boundary; for $N = 1$, $\mathbf{C}_1 = (1, 1)$; for $N = 3$, $\mathbf{C}_3 = (1, -1, 8, 8, -1, 1)/8$; for $N = 5$, $\mathbf{C}_5 = (-3, 3, 22, -22, 128, 128, -22, 22, 3, -3)/128$; and for $N = 7$, $\mathbf{C}_7 = (5, -5, -44, 44, 201, -201, 1024, 1024, -201, 201, 44, -44, -5, 5)/1024$. In the `pdSpecEst` package, these prediction weights are pre-determined up to order $N \leq 9$ at locations k both *away of* and *at* the boundary, allowing for faster computation of the predicted midpoints in practice. For higher refinement orders, (i.e. $N > 9$), the midpoints are predicted via Neville's algorithm.

3.2 Intrinsic forward and backward AI wavelet transform

Forward wavelet transform The intrinsic AI refinement scheme leads to an intrinsic AI wavelet transform passing from j -scale midpoints to $(j-1)$ -scale midpoints plus j -scale wavelet coefficients as follows:

1. **Coarsen/Predict:** given the j -scale midpoints $(M_{j,k})_{k=0,\dots,2^j-1}$, compute the $(j-1)$ -scale midpoints $(M_{j-1,k})_{k=0,\dots,2^{j-1}-1}$ through the midpoint relation in eq.(3.1). Select a refinement order $N \geq 1$ and generate the predicted midpoints $(\widetilde{M}_{j,k})_{k=0,\dots,2^j-1}$ based on $(M_{j-1,k})_k$.
2. **Difference:** given the true and predicted j -scale midpoints $M_{j,2k+1}, \widetilde{M}_{j,2k+1}$, define the wavelet coefficients as an *intrinsic difference* according to,

$$D_{j,k} = 2^{-j/2} \text{Log}_{\widetilde{M}_{j,2k+1}}(M_{j,2k+1}) \in \mathcal{T}_{\widetilde{M}_{j,2k+1}}(\mathcal{M}). \quad (3.5)$$

Note that $\|D_{j,k}\|_{\widetilde{M}_{j,2k+1}}^2 = 2^{-j} \delta(M_{j,2k+1}, \widetilde{M}_{j,2k+1})^2$ by definition of the Riemannian distance, giving the wavelet coefficients the interpretation of a (scaled) difference between $M_{j,2k+1}$ and $\widetilde{M}_{j,2k+1}$. In the remainder, we also keep track of the *whitened* wavelet coefficients,

$$\mathfrak{D}_{j,k} = \widetilde{M}_{j,2k+1}^{-1/2} * D_{j,k} \in \mathcal{T}_{\text{Id}}(\mathcal{M}). \quad (3.6)$$

The whitened coefficients correspond to the coefficients in eq.(3.5) transported to the same tangent space (at the identity) via the whitening transport in eq.(2.6). This allows for straightforward comparison of coefficients across scales and locations in Section 4 and 5, since $\|\mathfrak{D}_{j,k}\|_F^2 = \|D_{j,k}\|_{\widetilde{M}_{j,2k+1}}^2$.

Remark In terms of memory usage, the forward wavelet transform is executed in place, i.e. one substitutes the old fine-scale midpoints by new coarse-scale midpoints and wavelet coefficients. In particular, the $(d \times d)$ -dimensional midpoints $\{M_{j,2k}, M_{j,2k+1}\}$ are replaced by the $(d \times d)$ -dimensional midpoint $M_{j-1,k}$ and the $(d \times d)$ -dimensional wavelet coefficient $D_{j,k}$. The coefficients at the even locations $D_{j,k}^{\text{even}} = 2^{-j/2} \text{Log}_{\widetilde{M}_{j,2k}}(M_{j,2k})$ do not need to be stored, as they are uniquely determined by $M_{j,k}$ and $D_{j,k}$. In addition, if $D_{j,k}^{\text{odd}} := D_{j,k} = \mathbf{0}$, then also $D_{j,k}^{\text{even}} = \mathbf{0}$.

Backward wavelet transform The backward wavelet transform passing from coarse $(j-1)$ -scale midpoints plus j -scale wavelet coefficients to finer j -scale midpoints follows from reverting the above operations:

1. **Predict/Refine:** given $(j - 1)$ -scale midpoints $(M_{j-1,k})_{k=0,\dots,2^{j-1}-1}$ and a refinement order $N \geq 1$, generate the predicted midpoints $(\widetilde{M}_{j,2k+1})_{k=0,\dots,2^{j-1}-1}$ and compute the j -scale midpoints at the odd locations $2k + 1$ for $k = 0, \dots, 2^j - 1$ through:

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

2. **Complete:** the j -scale midpoints at the even locations $2k$ for $k = 0, \dots, 2^j - 1$ are retrieved from $M_{j-1,k}$ and $M_{j,2k+1}$ through the midpoint relation in eq.(3.1) as,

$$M_{j,2k} = M_{j-1,k} * M_{j,2k+1}^{-1}.$$

Given the coarsest midpoint $M_{0,0}$ at scale $j = 0$ and the pyramid of wavelet coefficients $(D_{j,k})_{j,k}$, for $j = 0, \dots, J - 1$ and $k = 0, \dots, 2^j - 1$, repeating the reconstruction procedure above up to scale J , we retrieve the original discretized curve $M_{J,k} = f(x_{k+1})$ for $k = 0, \dots, 2^J - 1$.

4 Wavelet regression for smooth HPD curves

In this section, we derive the wavelet coefficient decay and linear wavelet thresholding convergence rates in the context of intrinsically smooth curves of HPD matrices. It turns out that the derived rates coincide with the usual scalar wavelet coefficient decay and linear thresholding convergence rates on the real line. Nonlinear thresholding will not improve the convergence rates in the case of a homogeneous smoothness space. However, nonlinear wavelet thresholding is expected to improve the convergence rates in the case of a non-homogeneous smoothness space. This requires a well-defined intrinsic generalization of e.g. the family of Besov spaces to the Riemannian manifold, which is outside the scope of this work.

Repeated midpoint operator The repeated midpoint operator in eq.(3.1) in the construction of the midpoint pyramid is a valid intrinsic averaging operator in the sense that it converges to the intrinsic mean in the metric space (\mathcal{M}, δ) at a rate conjectured in Rahman et al. (2005) for general Riemannian manifolds. As in Rahman et al. (2005), recursively define,

$$\mu_n := \mu_n(X_1, \dots, X_n) = \gamma(\mu_{n/2}(X_1, \dots, X_{n/2}), \mu_{n/2}(X_{n/2+1}, \dots, X_n), 1/2). \quad (4.1)$$

Proposition 4.1. (Law of large numbers) Let $X_1, \dots, X_n \stackrel{\text{iid}}{\sim} \nu$, such that $\nu \in P_2(\mathcal{M})$ with intrinsic mean $\mathbb{E}_\nu[X] = \mu$, and $n = 2^J$ for some $J > 0$. Then,

$$\mathbf{E}[\delta(\mu_n, \mu)^2] \lesssim n^{-1},$$

with \lesssim smaller or equal up to a constant. In particular, $\mu_n \xrightarrow{P} \mu$ as $n \rightarrow \infty$, where the convergence holds with respect to the Riemannian distance, i.e. for every $\epsilon > 0$, $\Pr(\delta(\mu_n, \mu) > \epsilon) \rightarrow 0$.

Wavelet coefficient decay of smooth curves The derivation of the wavelet coefficient decay of intrinsically smooth curves in the Riemannian manifold relies on the fact that the derivative $\dot{\gamma}(t) \in \mathcal{T}_{\gamma(t)}(\mathcal{M})$ of a smooth curve $\gamma : I \rightarrow \mathcal{M}$ with $I \subseteq \mathbb{R}$ can be Taylor expanded in terms of the parallel transport and covariant derivatives according to (Lang, 1995, Chapter 9, Proposition 5.1) as,

$$\dot{\gamma}(t) = \sum_{k=0}^m \Gamma(\gamma)_{t_0}^t \left(\nabla_{\dot{\gamma}}^k \dot{\gamma}(t_0) \right) \frac{(t - t_0)^k}{k!} + O((t - t_0)^{m+1}), \quad \text{as } t \rightarrow t_0. \quad (4.2)$$

If $\gamma(t)$ is an intrinsic polynomial curve of order $r > 0$. Then, since $\Gamma(\gamma)_{t_0}^t(\mathbf{0}) = \mathbf{0}$, all terms of order higher or equal to r vanish and $\dot{\gamma}(t)$ simplifies to,

$$\dot{\gamma}(t) = \sum_{k=0}^{r-1} \Gamma(\gamma)_{t_0}^t \left(\nabla_{\dot{\gamma}}^k \dot{\gamma}(t_0) \right) \frac{(t - t_0)^k}{k!}.$$

In the specific case of a first-order polynomial, the above expression reduces to $\dot{\gamma}(t) = \Gamma(\gamma)_{t_0}^t(\dot{\gamma}(t_0))$, i.e. $\dot{\gamma}$ is parallel transported along the curve γ itself, or in other words $\gamma(t)$ is a geodesic curve.

Proposition 4.2. *Suppose that $\gamma : [0, 1] \rightarrow \mathcal{M}$ is a smooth curve with existing covariant derivatives up to order $N \geq 1$. Then, for each scale $j \geq 0$ sufficiently large and location k ,*

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

where $\mathfrak{D}_{j,k}$ denotes the whitened wavelet coefficient at scale-location (j, k) as in eq.(3.6), and N is the intrinsic AI refinement order. Here, the finest-scale midpoints are given by the local intrinsic averages $M_{J,k} = \text{Exp}_{M_{J,k}} \left(\int_{t_k}^{t_k+1} \text{Log}_{M_{J,k}}(\gamma(u)) \right)$ with $t_k = k/2^J$ for $k = 0, \dots, 2^J - 1$.

Remark Note that the above decay rates correspond to the usual wavelet coefficient decay rates of smooth real-valued curves in a Euclidean space based on wavelets with N vanishing moments, see e.g. (Walnut, 2002, Theorem 9.5).

Consistency and convergence rates The following results detail the convergence rates of linear wavelet regression of intrinsically smooth curves $\gamma : [0, 1] \rightarrow \mathcal{M}$ corrupted by noise. In particular, suppose that X_1, \dots, X_n is an independent sample, such that $X_i \sim \nu_i$ with $\mathbb{E}_{\nu_i}[X] = \gamma(i/n)$ and $\nu_i \in P_2(\mathcal{M})$ for each $i = 1, \dots, n$. The first proposition gives the estimation error of the

empirical wavelet coefficients based on X_1, \dots, X_n with respect to the true wavelet coefficients of the discretized curve $(\gamma(i/n))_{i=1, \dots, n}$. The proof relies on the rate in Proposition 4.1 above.

Proposition 4.3. (*Estimation error*) *Let $n = 2^J$ for some $J > 0$. For each scale $j \geq 0$ sufficiently small and location $0 \leq k \leq 2^j - 1$, it holds that,*

$$\mathbf{E} \|\widehat{\mathfrak{D}}_{j,k,n} - \mathfrak{D}_{j,k}\|_F^2 \lesssim n^{-1},$$

where $\widehat{\mathfrak{D}}_{j,k,n} = 2^{-j/2} \text{Log}(\widetilde{M}_{j,2k+1,n}^{-1/2} * M_{j,2k+1,n})$ is the empirical whitened wavelet coefficient at scale-location (j, k) , with $M_{j,2k+1,n}$ the estimated repeated midpoint at scale-location $(j, 2k+1)$ based on X_1, \dots, X_n and $\widetilde{M}_{j,2k+1,n}$ the predicted midpoint based on the estimated midpoints $(M_{j-1,k',n})_{k'}$.

Combining Proposition 4.2 and 4.3, the main theorem below gives integrated mean squared error in terms of the Riemannian distance of a linear wavelet estimator $\hat{\gamma}(t)$ of a smooth curve $\gamma(t)$, with existing covariant derivatives up to the intrinsic refinement order N , based on the sample of observations X_1, \dots, X_n . Again, the convergence rates correspond to the usual convergence rates of linear wavelet estimators of smooth real-valued curves in Euclidean space based on wavelets with N vanishing moments, see Antoniadis (1997).

Theorem 4.4. (*Convergence rates linear thresholding*) *Let $n = 2^J$ for some $J \geq 0$, and consider a linear wavelet estimator that thresholds all wavelet coefficients at scales $j \geq J_0$, such that $J_0 \sim \log_2(n)/(2N+1)$, with $N \geq 1$ the intrinsic AI refinement order. The total mean squared error in terms of the wavelet coefficients satisfies:*

$$\sum_{j,k} \mathbf{E} \|\widehat{\mathfrak{D}}_{j,k} - \mathfrak{D}_{j,k}\|_F^2 \lesssim n^{-2N/(2N+1)}, \quad (4.3)$$

where the sum ranges over $0 \leq j \leq J-1$ and $0 \leq k \leq 2^j-1$. Moreover, denote by $\hat{\gamma}(t)$ the linear wavelet thresholded curve on the manifold. Then,

$$\mathbb{E} \left[\int_0^1 \delta(\gamma(t), \hat{\gamma}(t))^2 dt \right] \lesssim n^{-2N/(2N+1)}. \quad (4.4)$$

5 Wavelet-based spectral matrix estimation

In the context of multivariate spectral matrix estimation, consider data observations from a d -dimensional strictly stationary time series with spectral matrix $f(\omega)$ and raw periodogram matrix $I_n(\omega_\ell)$ at the Fourier frequencies $\omega_\ell = 2\pi\ell/n \in (0, \pi]$ for $\ell = 1, \dots, n$. We apply the intrinsic wavelet transform to estimate $f(\omega)$ by denoising the inconsistent spectral estimator $I_n(\omega_\ell)$ through shrinkage or thresholding of coefficients in the intrinsic wavelet domain.

Pre-smoothed periodogram By construction, the raw periodogram matrix $I_n(\omega_\ell)$ is Hermitian, but only positive semi-definite as the rank of $I_n(\omega_\ell)$ is one. The intrinsic wavelet transform acts on curves of HPD matrices and for this reason we pre-smooth the periodogram to guarantee that it is HPD or full rank analogous to e.g. Dai and Guo (2004). By (Dai and Guo, 2004, Lemma 1), for $\omega_\ell \not\equiv 0 \pmod{\pi}$, a multitaper spectral estimate $\bar{I}_n(\omega_\ell)$ of a strictly stationary time series, with a fixed number of tapers B , is asymptotically independent at the Fourier frequencies, and its asymptotic distribution satisfies:

$$\bar{I}_n(\omega_\ell) \xrightarrow{d} W_d^c(B, B^{-1}f(\omega_\ell)), \quad \text{as } n \rightarrow \infty.$$

Here, $W_d^c(B, B^{-1}f(\omega_\ell))$ denotes a complex Wishart distribution of dimension d with B degrees of freedom. If $B \geq d$, then the spectral estimate $\bar{I}_n(\omega_\ell)$ is HPD with probability one, i.e. $\bar{I}_n(\omega_\ell) \in \mathcal{M}$ almost surely. In practice, we choose B as small as possible, preferably $B = d$, so that only the necessary small amount of pre-smoothing is performed to guarantee a HPD noisy spectral estimate.

Asymptotic bias-correction Suppose that $X \sim W_d^c(B, B^{-1}f)$ exactly, the Euclidean mean of X equals f , and if $X_1, \dots, X_n \stackrel{\text{iid}}{\sim} W_d^c(B, B^{-1}f)$, the ordinary arithmetic mean $\frac{1}{n} \sum_{\ell=1}^n X_\ell$ is an unbiased and consistent estimator of f as $n \rightarrow \infty$. Intrinsic averaging in the midpoint pyramid is performed through repeated application of the midpoint operator. From Proposition 4.1, it is clear that if the Euclidean mean $\mathbf{E}[X_\ell] = f$ and the intrinsic mean $\mathbb{E}[X_\ell] = \mu$ do not coincide, the repeated midpoint functional is not a consistent estimator of f , the quantity of interest. By defining the notion of intrinsic bias as in Smith (2000), the repeated midpoint functional of a multitaper spectral estimate is seen to be asymptotically biased with respect to the spectrum f .

Definition 5.1. Given an estimator $\hat{\mu}$ of $\mu \in \mathcal{M}$, define the bias $b(\hat{\mu}, \mu) \in \mathcal{T}_\mu(\mathcal{M})$ of $\hat{\mu}$ as,

$$b(\hat{\mu}, \mu) = \mathbf{E}[\text{Log}_\mu(\hat{\mu})],$$

where $\mathbf{E}[\cdot]$ is the (ordinary) Euclidean expectation. In the context of a Euclidean space, the Exp- and Log-maps reduce to ordinary matrix addition and subtraction, in which case the definition above simplifies to the usual vector space definition of the bias.

Theorem 5.1. (*Bias-correction*) Let $X \sim W_d^c(B, B^{-1}f)$ and $c(d, B) = -\log(B) + \frac{1}{d} \sum_{i=1}^d \psi(B - (d - i))$, with $\psi(\cdot)$ the digamma function. The intrinsic bias of X with respect to f is given by,

$$b(X, f) = \mathbf{E}[\text{Log}_f(X)] = c(d, B) \cdot f.$$

If $(\tilde{X}_\ell)_{\ell=1,\dots,n} := (e^{-c(d,B)} X_\ell)_{\ell=1,\dots,n}$, such that $X_1, \dots, X_n \stackrel{\text{iid}}{\sim} W_d^c(B, B^{-1}f)$ with $n = 2^J$. Then,

$$\mu_n(\tilde{X}_1, \dots, \tilde{X}_n) \xrightarrow{P} f, \quad \text{as } n \rightarrow \infty,$$

where the convergence in probability holds with respect to the Riemannian distance.

Remark Note that if $d = B = 1$, the bias-correction simplifies to multiplication by the scalar $\exp(-c(d, B)) = \exp(-\psi(1))$, i.e. the exponential of the Euler-Mascheroni constant. This corresponds to the asymptotic bias-correction for the ordinary log-periodogram in the context of a univariate time series, see e.g. Wahba (1980).

5.1 Nonlinear intrinsic wavelet thresholding

Given a sequence of d -dimensional time series observations, wavelet-based spectral estimation exploits the sparsity of representations of smooth curves in the intrinsic AI wavelet domain by proceeding along the usual steps:

Step 1: Apply the forward intrinsic wavelet transform to the bias-corrected HPD periodogram.

Step 2: Shrink or threshold the coefficients in the intrinsic wavelet domain.

Step 3: Apply the inverse intrinsic wavelet transform to the modified coefficients.

There are various possible ways to nonlinearly shrink or threshold coefficients in the intrinsic manifold wavelet domain. In particular, expanding the matrix-valued coefficients in a basis of the vector space of Hermitian matrices, nonlinear thresholding or shrinkage of individual components allows to capture inhomogeneous smoothness behavior across components of the spectral matrix, similar to the Cholesky-based smoothing procedures in e.g. Dai and Guo (2004) or Krafty and Collinge (2013). The wavelet-denoised estimator is guaranteed to be HPD, as the inverse wavelet transform always outputs a curve in the manifold of HPD matrices. From the perspective of wavelet coefficients being intrinsic local differences in the manifold, another sensible approach is to shrink or threshold all components of a matrix-valued wavelet coefficient simultaneously, e.g. a kink or cusp in a curve in the manifold likely affects all components of the matrix-valued wavelet coefficients at the corresponding scale-locations instead of a single or only a few components. Here, we pursue the latter approach and consider keep-or-kill thresholding of entire wavelet coefficients.

Congruence equivariance In general, the only requirement that is imposed on the intrinsic wavelet thresholding or shrinkage procedure is that it is *unitary congruence equivariant*. That is,

if D^X is a noisy matrix-valued wavelet coefficient and \widehat{D}^X is its shrunken or thresholded version, then $U * \widehat{D}^X$ should be the shrunken or thresholded version of $U * D^X$ for each $U \in \mathcal{U}_d$, where $\mathcal{U}_d = \{U \in \text{GL}(d, \mathbb{C}) \mid U^*U = \text{Id}\}$. In practice, this property virtually always holds. For instance, if one thresholds or shrinks components of coefficients data-adaptively, the component-specific threshold or shrinkage parameters rotate in the same fashion as the components of the coefficients.

Proposition 5.2. (*Unitary congruence equivariance*) *Let $(X_\ell)_\ell$ be a sequence of HPD matrices and $(\hat{f}_\ell)_\ell$ its wavelet-denoised estimate. If the wavelet thresholding or shrinkage procedure is unitary congruence equivariant, then the same is true for the wavelet estimator, i.e. the wavelet-denoised estimate of $(U * X_\ell)_\ell$ is $(U * \hat{f}_\ell)_\ell$ for each $U \in \mathcal{U}_d$.*

This is an important property in the context of spectral estimation of multivariate time series. Rotation of the observed time series, e.g. a permutation of the time series components, results in a congruence transformation $U * f(\omega)$ of the underlying spectral matrix, with $U \in \mathcal{U}_d$. Such rotations should not have a nontrivial effect on the spectral estimator, otherwise it is not clear which rotation of the time series to consider for estimation. The spectral estimation methods based on smoothing the Cholesky decomposition of an initial noisy spectral estimator (Dai and Guo (2004), Rosen and Stoffer (2007), or Krafty and Collinge (2013)) do not necessarily satisfy this property. This is due to the fact that Cholesky square root matrices are generally not congruence-equivariant, i.e. $\text{Chol}(U * f(\omega)) \neq U * \text{Chol}(f(\omega))$ for a non-trivial unitary matrix $U \in \mathcal{U}_d$. To circumvent this problem, Zheng et al. (2017) propose to average many Cholesky-based estimates based on random permutations of the data. The main drawback of such an approach is the significant increase in computational effort.

Trace thresholding of coefficients A method that is particularly *tractable* is thresholding or shrinkage based on the trace of the whitened wavelet coefficients. For a sequence of independent complex Wishart matrices, the trace of the noisy whitened coefficients decomposes into an additive signal plus mean-zero noise sequence model. Moreover, the variance of the trace of the noisy whitened coefficients is constant across wavelet scales, and since the trace operator outputs a scalar, one can directly apply ordinary scalar thresholding or shrinkage procedures to the matrix-valued coefficients. Thresholding or shrinkage of the trace of the whitened coefficients is equivariant under unitary congruence transformations as in Proposition 5.2. In fact, it is equivariant under congruence transformation by any invertible matrix, i.e. *general linear* congruence equivariant. In the context of spectral estimation of multivariate time series, this means that the estimator does

not nontrivially depend on the chosen basis or coordinate system of the time series, as the spectral estimator is equivariant under a change of basis of the time series.

Lemma 5.3. *(General linear congruence equivariance) Let $(X_\ell)_\ell$ be a sequence of HPD matrices and $(\hat{f}_\ell)_\ell$ its wavelet-denoised estimate based on thresholding or shrinkage of the trace of the whitened wavelet coefficients. The estimator is equivariant under general linear congruence transformation in the sense that the wavelet-denoised estimate $(\hat{f}_{A,\ell})_\ell$ of $(A * X_\ell)_\ell$ equals $(A * \hat{f}_\ell)_\ell$ for each $A \in \text{GL}(d, \mathbb{C})$.*

Below, we write \tilde{P}_f for the probability distribution corresponding to a bias-corrected complex Wishart distribution $e^{-(d,B)}W_d^c(B, B^{-1}f)$ as in Theorem 5.1, with $B \geq d$ to ensure positive-definiteness of the random variable. Here, $\tilde{P}_f \in P_2(\mathcal{M})$ is understood to be the distribution of a random variable $X = f^{1/2} * W$, where W is a HPD complex Wishart matrix, with B degrees of freedom, not depending on f , and with intrinsic mean equal to the identity matrix Id . Note that the latter directly implies that $f^{1/2} * W$ has intrinsic mean equal to f .

Proposition 5.4. *(Trace properties) Let $X_\ell \sim \tilde{P}_{f_\ell}$, independently distributed for $\ell = 1, \dots, n$, with $n = 2^J$. For each scale-location (j, k) , the whitened wavelet coefficients obtained from the intrinsic AI wavelet transform of order $N = 2L + 1 \geq 1$ satisfy:*

$$\text{Tr}(\mathfrak{D}_{j,k}^X) = \text{Tr}(\mathfrak{D}_{j,k}^f) + \text{Tr}(\mathfrak{D}_{j,k}^W),$$

where $\mathfrak{D}_{j,k}^X$ is the random whitened coefficient based on the sequence $(X_\ell)_{\ell=1}^n$, $\mathfrak{D}_{j,k}^f$ is the deterministic whitened coefficient based on the sequence of intrinsic means $(f_\ell)_{\ell=1}^n$, and $\mathfrak{D}_{j,k}^W$ is the random whitened coefficient based on an iid sequence of Wishart matrices $(W_\ell)_{\ell=1}^n$, with intrinsic mean equal to the identity, independent of $(f_\ell)_{\ell=1}^n$.

Moreover, $\mathbf{E}[\text{Tr}(\mathfrak{D}_{j,k}^X)] = \text{Tr}(\mathfrak{D}_{j,k}^f)$, and,

$$\text{Var}(\text{Tr}(\mathfrak{D}_{j,k}^X)) = \left(2^{-J} \sum_{i=0}^{2N-1} \mathbf{C}_{L,i}^2 \right) \left(\sum_{i=1}^d \psi'(B - (d - i)) \right), \quad (5.1)$$

where $\psi'(\cdot)$ denotes the trigamma function, and $(\mathbf{C}_{L,i})_i$ are the filter coefficients as in Section 3.1.2. In particular, $\text{Var}(\text{Tr}(\mathfrak{D}_{j,k}^X))$ is independent of the scale-location (j, k) and whenever $\text{Tr}(\mathfrak{D}_{j,k}^f)$ vanishes $\mathbf{E}[\text{Tr}(\mathfrak{D}_{j,k}^X)] = 0$, e.g. when $(f_\ell)_\ell$ is an intrinsic polynomial of order smaller than N .

Corollary 5.5. *(Centered noise) With the same notation as in Proposition 5.4, the random whitened wavelet coefficients $\mathfrak{D}_{j,k}^W$ based on a sequence of i.i.d. Wishart matrices $(W_\ell)_{\ell=1}^n$, with*

identity intrinsic mean satisfy,

$$\mathbf{E}[\mathfrak{D}_{j,k}^W] = \mathbf{0},$$

where $\mathbf{E}[\cdot]$ denotes the (ordinary) Euclidean expectation.

Based on the trace of the whitened coefficients, by Proposition 5.4, in the context of a curve of approximate complex random Wishart matrices, we can apply any preferred classical wavelet thresholding or shrinkage procedure well-suited to scalar additive signal plus noise sequence models, with homogeneous variances across coefficient scales.

Tree-structured thresholding In the simulated data examples below, because of its good empirical performance and computational tractability, we consider dyadic tree-structured thresholding of wavelet coefficients as in Donoho (1997). For each scale-location (j, k) , denote $d_{j,k} = \text{Tr}(\mathfrak{D}_{j,k}^X)$ for the trace of the noisy whitened wavelet coefficient and let $x_{j,k} \in \{0, 1\}$ be a binary label. Given a penalty parameter λ , we optimize the *complexity penalized residual sum of squares* (CPRESS) criterion:

$$\underset{\mathbf{x}}{\text{argmin}} \text{CPRESS}(\mathbf{x}) = \underset{\mathbf{x}}{\text{argmin}} \|\mathbf{d}\mathbf{x} - \mathbf{d}\|^2 + \lambda^2 \|\mathbf{x}\|,$$

under the constraint that the nonzero labels $\{x_{j,k} \mid x_{j,k} = 1\}$ form a dyadic rooted tree, i.e. for each nonzero label $x_{j+1,2k+1}$ or $x_{j+1,2k}$, the label $x_{j,k}$ also has to be nonzero. This minimization problem can be solved in $O(N)$ computations, with N the total number of coefficients via the fast dyadic tree-pruning algorithm in Donoho (1997). The optimal tree in the CPRESS criterion is a good compromise between goodness-of-fit and sparsity and the imposed tree-structure captures singularities in the signal, as coefficients of singularities typically persist across wavelet scales at the location of the singularity. On the other hand, isolated large coefficients, usually arising from noise and not signal, are set to zero. For sufficiently large n , the scalar coefficients $d_{j,k}$ are approximately normally distributed for scales j away from the finest wavelet scale $J - 1$, as the scalar coefficients $d_{j,k}$ are essentially based on local weighted averages, see e.g. Moulin (1994), Neumann (1996) or Chau and von Sachs (2016). For normally distributed coefficients, as detailed in Donoho (1997), a natural choice for the smoothing parameter is the universal threshold $\lambda \sim \sigma_w \sqrt{2 \log(N)}$, with σ_w^2 the noise variance determined either via eq.(5.1) or from the data itself. Instead, the smoothing parameter λ can also be determined in a data-adaptive fashion through two-fold cross-validation as in Nason (1996). The only necessary modifications to perform two-fold cross-validation in the

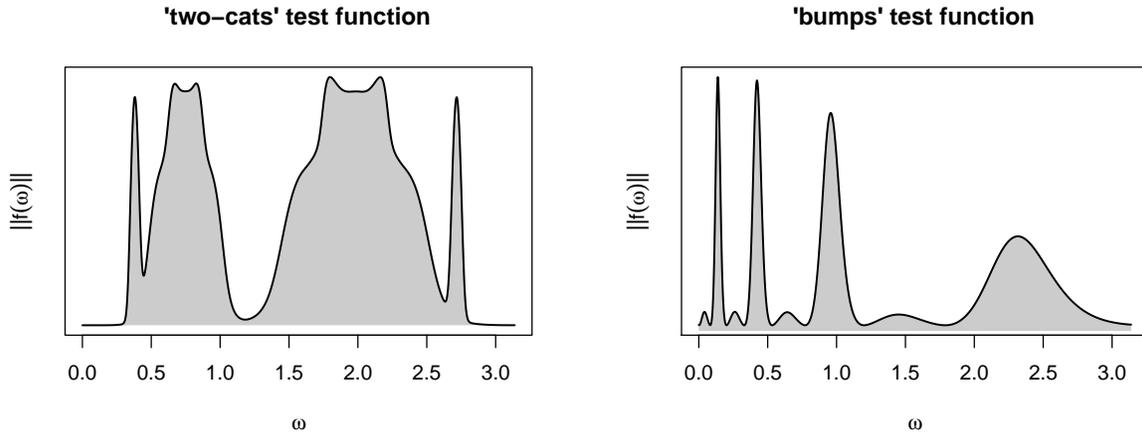


Figure 3: Euclidean norms of (3×3) -dimensional HPD test spectra generated with `rExamples()`.

Riemannian manifold are: (i) replace linear interpolation of the curve estimator of half the data at the left-out data points by geodesic interpolation, and (ii) replace the Euclidean distance in the approximate MISE by the Riemannian distance.

6 Illustrative data examples

6.1 Finite-sample performance

In the figures below, we assess the finite-sample performance of intrinsic wavelet-based spectral estimation in the space of HPD matrices equipped with several different metrics and benchmark the performance against a number of alternative nonparametric curve estimation procedures in the Riemannian manifold. In particular, we consider two HPD test functions (i.e. spectral matrix curves) displaying locally varying smoothness behavior, such as local peaks or troughs.

Test functions As test functions, we construct two (3×3) -dimensional HPD-matrix valued curves available through the function `rExamples()` in the R-package `pdSpecEst` by specifying the argument `example = "two-cats"` or `example = "bumps"`. The `two-cats` function visualizes the contours of two cats and consists of relatively smooth parts combined with local peaks and troughs. The `bumps` function is a curve of HPD matrices containing local bumps of various degrees of smoothness. Figure 3 displays the Euclidean norm of the HPD matrix-valued curves along the frequency domain. We refer to the package documentation for more details and additional test functions.

Benchmark procedures As mentioned before, the intrinsic AI wavelet transform in Section 3 does not fundamentally rely on the Riemannian metric. For instance, substituting the Riemannian metric by the Euclidean metric, the intrinsic AI wavelet transform reduces to the matrix-version of the scalar AI wavelet transform in Donoho (1993). In the simulation studies below, we consider nonlinear tree-structured wavelet thresholding (as in Section 5.1) in the space of HPD matrices equipped with three different metrics: (i) the Riemannian metric, (ii) the Log-Euclidean metric and (iii) the Cholesky metric. We point out that wavelet regression based on the Log-Euclidean or Cholesky metric is exactly equivalent to: (i) wavelet regression applied to the matrix logarithms or Cholesky decompositions of the HPD-matrix valued curves under the Euclidean metric, combined with (ii) projection back to the manifold via the matrix exponential (for the Log-Euclidean metric) or multiplication by the conjugate transpose (for the Cholesky metric). In the context of spectral matrix estimation, it turns out that the asymptotic bias-correction for the periodogram when smoothing with respect to the Log-Euclidean metric coincides with the bias-correction for the Riemannian metric as in Theorem 5.1. For the Cholesky metric, we apply the asymptotic bias-correction in (Dai and Guo, 2004, Theorem 1), which is the analogue of Theorem 5.1 when smoothing the periodogram with respect to the Cholesky metric. Intrinsic nonlinear tree-structured wavelet spectral matrix estimation is directly available in the package `pdSpecEst` through `pdSpecEst1D()` and specification of the appropriate metric. In addition to intrinsic wavelet-based curve estimation, we have implemented the following benchmark estimation procedures for curves of HPD matrices:

- **Intrinsic NN-regression:** we consider intrinsic nearest-neighbor regression in the Riemannian manifold by replacing the ordinary local Euclidean averages by their intrinsic counterparts based on the Riemannian metric. The local intrinsic averages are calculated efficiently by the gradient descent algorithm in Pennec (2006).
- **Intrinsic cubic spline regression:** we consider the penalized regression approach in Boumal and Absil (2011a) or Boumal and Absil (2011b) based on the Riemannian metric, with penalty parameters $(\lambda = 0, \mu > 0)$, such that the minimizers of the objective function are approximating cubic splines in the manifold of HPD matrices. The Riemannian conjugate gradient descent method in Boumal and Absil (2011b) to compute the estimator is available through the function `pdSplineReg()` in the package `pdSpecEst`. Here, we use a backtracking line search based on the Armijo-Goldstein condition.

Table 1: Implemented wavelet and benchmark estimators and their properties.

Abbreviation	Method	Metric	U -equiv.*	A -equiv.†	Always PD
$W_{av-\delta_R}$	Tree-structured manifold wavelet regression, refinement order $N = 5$	Riemannian	✓	✓	✓
$W_{av-\delta_L}$	" "	Log-Euclidean	✓	✗	✓
$W_{av-\delta_C}$	" "	Cholesky	✗	✗	✓
$NN-\delta_R$	Intrinsic nearest-neighbor regression	Riemannian	✓	✓	✓
$CS-\delta_R$	Intrinsic cubic spline regression	Riemannian	✓	✓	✓
$LP_0-\delta_R$	Intrinsic local polynomial regression, degree $p = 0$	Riemannian	✓	✓	✓
$LP_3-\delta_L$	Intrinsic local polynomial regression, degree $p = 3$	Log-Euclidean	✓	✗	✓

*, †: U -equiv. and A -equiv. respectively denote whether the estimator is equivariant under congruence transformation by a unitary matrix $U \in \mathcal{U}_d$ or a general linear matrix $A \in GL(\mathbb{C}, d)$, see Section 5.1.

- **Intrinsic local polynomial regression:** we consider local polynomial regression in the space of HPD matrices as in Yuan et al. (2012) based on: (i) the Riemannian metric, and (ii) the Log-Euclidean metric. For the Riemannian metric, we have only implemented the locally constant estimator, i.e. degree $p = 0$, as local polynomial regression under the Riemannian metric for $p > 0$ requires the optimization of a non-convex objective function and is computationally quite challenging, see Yuan et al. (2012) for additional details.

Simulation setup and results Given the `two-cats` and `bumps` test functions as target spectral curves, we apply the intrinsic wavelet and benchmark estimation procedures in the context of spectral matrix estimation as follows: (i) generate a 3-dimensional time series trace of length $2n$ with underlying spectral matrix f via the Cramér representation based on complex normal random variates as in e.g. (Brillinger, 1981, Section 4.6); (ii) compute an initial HPD multitaper periodogram based on three Slepian (discrete prolate spheroidal) taper functions with `pdPgram()`;

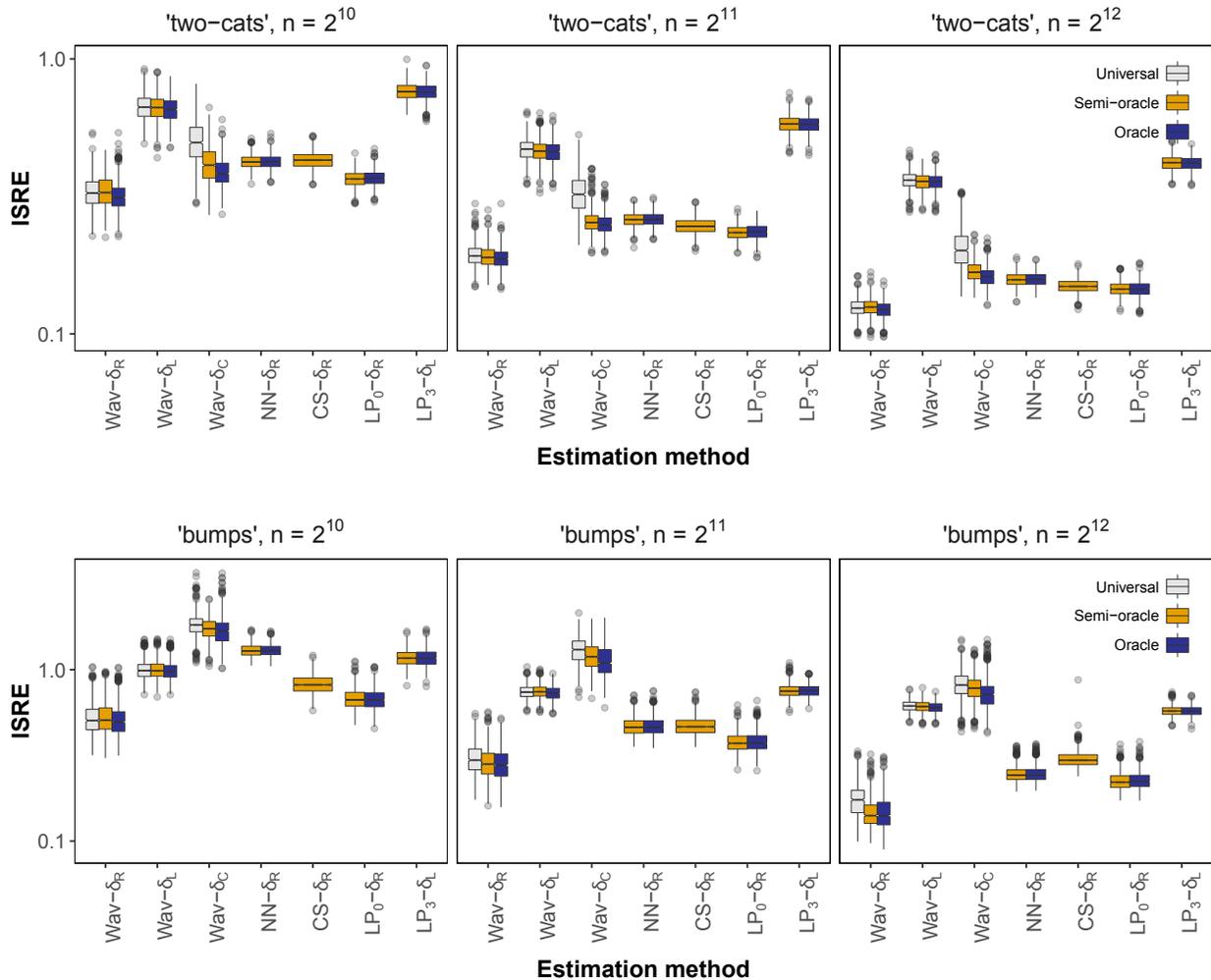


Figure 4: Integrated squared Riemannian errors (ISRE) of the wavelet and benchmark spectral estimation procedures for time series traces generated from the `two-cats` and `bumps` test spectra.

(iii) apply the intrinsic wavelet and benchmark manifold regression techniques to the initial HPD periodogram.

Figure 4 displays the integrated squared Riemannian errors (ISRE) over the entire frequency domain for each of the nonparametric curve estimation procedures as the result of $M = 1000$ simulations. Each estimation procedure depends on a single tuning parameter: for the tree-structured wavelet and CS-regression this is the penalty parameter; for the NN-regression this is the number of nearest neighbors; and for the LP-regression the bandwidth parameter. For the tree-structured wavelet estimators, we include the ISRE based on the following choices: (i) λ equal to the universal threshold, (ii) λ a semi-oracular penalty and (iii) λ the oracular penalty. The oracular penalty min-

imizes the ISRE with respect to the true target at each individual simulation. The semi-oracular penalty is a fixed penalty pre-determined by averaging the oracular penalty over a number of simulation runs. The reason we also include the semi-oracular penalty is that it is computationally infeasible to compute the oracular penalty at each individual simulation for the intrinsic manifold CS estimator. For this estimator, the results are displayed only for the semi-oracular penalty. For each of the other estimators, the displayed results include both a semi-oracular and oracular choice of the tuning parameter. Note that, in addition, the appropriate (asymptotic) bias-corrections are applied for each of the benchmark estimators.

The intrinsic nonlinear wavelet spectral estimator based on the Riemannian metric outperforms the benchmark spectral estimators in terms of the ISRE in each of the simulated scenarios. This is attributed to the fact that, in contrast to the benchmark procedures, the wavelet-based estimator is able to capture locally varying or inhomogeneous smoothness behavior. Moreover, and of considerable importance in practical applications, the estimation error of the data-adaptive nonlinear tree-structured wavelet estimator based on a simple universal threshold is close to the optimal finite-sample estimation error. For each of the benchmark procedures there is no simple heuristic choice for the penalty or bandwidth parameter, and in practice one needs to resort to computationally expensive cross-validation methods. For test spectral matrix curves displaying globally homogeneous behavior (e.g. generated with `rExamples()` by specifying `example = "gaussian"`) the performance of the wavelet and benchmark estimators is roughly similar, and we did not include the results here. It is emphasized that, although the wavelet estimator may not significantly improve upon the estimation error for globally smooth curves, from a computational perspective the wavelet-based estimator remains the preferred option, as it provides a fast heuristic choice of the penalty parameter in contrast to the benchmark procedures.

7 Conclusion

The primary contribution of this work is the development of intrinsic average-interpolation wavelet transforms and intrinsic wavelet regression for curves in the Riemannian manifold of HPD matrices equipped with the natural invariant Riemannian metric. As discussed in Section 6, the intrinsic AI wavelet transforms are constructed independently of the chosen metric and although the wavelet coefficient decay and nonparametric convergence rates in Section 4 are derived under the Riemannian metric, similar arguments apply for other metrics. For instance, in the Euclidean case,

the intrinsic Taylor expansions reduce to the usual Euclidean Taylor expansions, as the parallel transport is the identity map and the covariant derivatives are ordinary matrix derivatives.

Additional material Given the wavelet spectral matrix estimator, we can assess its variability with the function `pdConfInt1D()` in the R-package `pdSpecEst`, which uses a parametric bootstrap procedure to construct depth-based confidence regions based on the intrinsic manifold data depths developed in Chau et al. (2017). The bootstrap procedure, which exploits the data generating process of a stationary time series via its Cramér representation, is equivalent to Dai and Guo (2004) and Fiecas and Ombao (2016) among others. In addition to spectral matrix estimation, the intrinsic wavelet transforms are also useful for fast clustering of spectral matrices based on sparse representations in the intrinsic wavelet domain. Wavelet-based spectral matrix clustering is available through `pdSpecClust1D()`. In the intrinsic wavelet domain, we combine both: (i) denoising by thresholding wavelet coefficients, and (ii) clustering of spectral matrices based on the non-zero wavelet coefficients. Such an approach allows for significantly reduced computational effort in comparison to a more naive approach, in which we first estimate the individual spectra and subsequently cluster spectra based on integrated inter-spectra distances in the frequency domain.

Outlook We aim to extend the developed intrinsic wavelet methods for curves of HPD matrices to intrinsic 2D wavelet transforms and 2D wavelet regression for surfaces of HPD matrices, with in mind the application to time-varying spectral matrix estimation for nonstationary multivariate time series. Furthermore, Hermitian or symmetric positive definite matrices are encountered as autocovariance matrices or spectral density matrices in multivariate time series analysis, but also play an important role in the fields of medical imaging, computer vision or radar signal processing (e.g. Pennec et al. (2006)), and it is of interest to apply the intrinsic wavelet methods for the purpose of compression or denoising in other settings than spectral matrix estimation. For instance, applied to diffusion tensor imaging, intrinsic wavelet shrinkage or thresholding shows potential for fast denoising of large collections of non-smoothly varying diffusion tensors.

Acknowledgements

This research is supported by IAP research network P7/06 of the Belgian government (Belgian Science Policy), and by the contract ‘Projet d’Actions de Recherche Concertées’ (ARC) No. 12/17-045 of the ‘Communauté française de Belgique’ granted by the Académie universitaire Louvain.

The first author gratefully acknowledges funding from the Belgian Fund for Scientific Research (FRIA/FRS-FNRS).

References

- Antoniadis, A. (1997). Wavelets in statistics: a review. *Statistical methods & applications* 6(2), 97–130.
- Bhatia, R. (2009). *Positive Definite Matrices*. New Jersey: Princeton University Press.
- Boumal, N. and P.-A. Absil (2011a). A discrete regression method on manifolds and its application to data on $SO(n)$. *IFAC Proceedings Volumes* 44(1), 2284–2289.
- Boumal, N. and P.-A. Absil (2011b). Discrete regression methods on the cone of positive-definite matrices. In *IEEE ICASSP, 2011*, pp. 4232–4235.
- Brillinger, D. (1981). *Time Series: Data Analysis and Theory*. San Francisco: Holden-Day.
- Brockwell, P. and R. Davis (2006). *Time Series: Theory and Methods*. New York: Springer.
- Chau, J. (2017). *pdSpecEst: Positive-Definite Wavelet-Based Multivariate Spectral Analysis*. R package version 1.2.1.
- Chau, J., H. Ombao, and R. von Sachs (2017). Data depth and rank-based tests for covariance and spectral density matrices. *ArXiv preprint 1706.08289*.
- Chau, J. and R. von Sachs (2016). Functional mixed effects wavelet estimation for spectra of replicated time series. *Electronic Journal of Statistics* 10(2), 2461–2510.
- Dahlhaus, R. (2012). *Locally stationary processes*, Chapter in Time Series Analysis: Methods and Applications, Vol. 30, pp. 351–413. Amsterdam: Elsevier B.V.
- Dai, M. and W. Guo (2004). Multivariate spectral analysis using Cholesky decomposition. *Biometrika* 91(3), 629–643.
- do Carmo, M. (1992). *Riemannian Geometry*. Boston: Birkhäuser.
- Donoho, D. (1993). *Smooth wavelet decompositions with blocky coefficient kernels*, Chapter in Recent Advances in Wavelet Analysis, pp. 259–308. New York: Academic Press.
- Donoho, D. (1997). Cart and best-ortho-basis: a connection. *The Annals of Statistics* 25(5), 1870–1911.
- Dryden, I., A. Koloydenko, and D. Zhou (2009). Non-Euclidean statistics for covariance matrices, with applications to diffusion tensor imaging. *The Annals of Applied Statistics* 3(3), 1102–1123.
- Fiecas, M. and H. Ombao (2016). Modeling the evolution of dynamic brain processes during an associative learning experiment. *Journal of the American Statistical Association* 111(516), 1440–1453.
- Higham, N. J. (2008). *Functions of Matrices: Theory and Computation*. Philadelphia: Siam.
- Hinkle, J., P. Fletcher, and S. Joshi (2014). Intrinsic polynomials for regression on Riemannian manifolds. *Journal of Mathematical Imaging and Vision* 50(1-2), 32–52.
- Ho, J., G. Cheng, H. Salehian, and B. Vemuri (2013). Recursive karcher expectation estimators and recursive law of large numbers. *AISTATS*, 325–332.
- Holbrook, A., S. Lan, A. Vandenberg-Rodes, and B. Shahbaba (2016). Geodesic Lagrangian Monte Carlo over the space of positive definite matrices: with application to Bayesian spectral density estimation. *ArXiv preprint 1612.08224*.
- Jansen, M. and P. Oonincx (2005). *Second Generation Wavelets and Applications*. London: Springer-Verlag.
- Klees, R. and R. Haagmans (2000). *Wavelets in the Geosciences*. Berlin: Springer-Verlag.
- Krafty, R. and W. Collinge (2013). Penalized multivariate Whittle likelihood for power spectrum estimation. *Biometrika*, 1–12.
- Lang, S. (1995). *Differential and Riemannian Manifolds*.

- New York: Springer-Verlag.
- Le, H. (1995). Mean size-and-shapes and mean shapes: a geometric point of view. *Advances in Applied Probability* 27(1), 44–55.
- Ma, Y. and Y. Fu (2012). *Manifold Learning Theory and Applications*. CRC Press, Taylor & Francis.
- Moulin, P. (1994). Wavelet thresholding techniques for power spectrum estimation. *IEEE Transactions on Signal Processing* 42(11), 3126–3136.
- Muirhead, R. (1982). *Aspects of Multivariate Statistical Theory*. New Jersey: John Wiley & Sons.
- Nason, G. (1996). Wavelet shrinkage using cross-validation. *Journal of the Royal Statistical Society: Series B* 58, 463–479.
- Neumann, M. (1996). Spectral density estimation via nonlinear wavelet methods for stationary non-gaussian time series. *Journal of Time Series Analysis* 17(6), 601–633.
- Pasternak, O., N. Sochen, and P. Basser (2010). The effect of metric selection on the analysis of diffusion tensor MRI data. *NeuroImage* 49(3), 2190–2204.
- Pennec, X. (2006). Intrinsic statistics on Riemannian manifolds: Basic tools for geometric measurements. *Journal of Mathematical Imaging and Vision* 25(1), 127–154.
- Pennec, X., P. Fillard, and N. Ayache (2006). A Riemannian framework for tensor computing. *International Journal of Computer Vision* 66(1), 41–66.
- Rahman, I., I. Drori, V. Stodden, D. Donoho, and P. Schröder (2005). Multiscale representations for manifold-valued data. *Multiscale Modeling & Simulation* 4(4), 1201–1232.
- Rosen, O. and D. Stoffer (2007). Automatic estimation of multivariate spectra via smoothing splines. *Biometrika* 94(2), 335–345.
- Said, S., L. Bombrun, Y. Berthoumieu, and J. Manton (2015). Riemannian Gaussian distributions on the space of symmetric positive definite matrices. *ArXiv preprint 1507.01760*.
- Smith, S. (2000). Intrinsic Cramér-Rao bounds and subspace estimation accuracy. In *Proceedings of the IEEE Sensor Array and Multichannel Signal Processing Workshop*, pp. 489–493. IEEE.
- Tulino, A. and S. Verdú (2004). *Random Matrix Theory and Wireless Communications*. Hanover: Now Publishers Inc.
- Wahba, G. (1980). Automatic smoothing of the log periodogram. *Journal of the American Statistical Association* 75(369), 122–132.
- Walden, A. (2000). A unified view of multitaper multivariate spectral estimation. *Biometrika* 87(4), 767–788.
- Walnut, D. (2002). *An Introduction to Wavelet Analysis*. Boston: Birkhäuser.
- Yuan, Y., H. Zhu, W. Lin, and J. Marron (2012). Local polynomial regression for symmetric positive definite matrices. *Journal of the Royal Statistical Society: Series B* 74(4), 697–719.
- Zheng, H., K.-W. Tsui, X. Kang, and X. Deng (2017). Cholesky-based model averaging for covariance matrix estimation. *Statistical Theory and Related Fields* 1(1), 48–58.
- Zhu, H., Y. Chen, J. Ibrahim, Y. Li, C. Hall, and W. Lin (2009). Intrinsic regression models for positive-definite matrices with applications to diffusion tensor imaging. *Journal of the American Statistical Association* 104(487), 1203–1212.

8 Appendix I: Proofs

8.1 Proof of Proposition 4.1

Proof. Denote the distribution of $\mu_n := \mu_n(X_1, \dots, X_n)$ by ν_n , we show recursively that:

$$\mathbf{E}[\delta(\mu_n, \mu)^2] = \int_{\mathcal{M}} \delta(x, \mu) d\nu_n(x) \leq \frac{1}{n} \text{Var}(X_1)$$

with $\text{Var}(X_1) = \mathbf{E}[\delta(X_1, \mu)^2]$.

By (Bhatia, 2009, Theorem 6.1.9), if $X_1, X_2, X_3 \in \mathcal{M}$, then:

$$\delta(\gamma(X_1, X_2, t), X_3)^2 \leq (1-t)\delta(X_1, X_3)^2 + t\delta(X_2, X_3)^2 - t(1-t)\delta(X_1, X_2)^2, \quad \text{for all } t \in [0, 1]$$

Substituting $X_3 = \mu$ and $t = 1/2$, (note that $\mu_2 = \gamma(X_1, X_2, 1/2)$), and taking expectations on both sides yields:

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

Using that $X_1, X_2 \stackrel{\text{iid}}{\sim} \nu$ we obtain,

$$\mathbf{E}[\delta(\mu_2, \mu)^2] \leq \text{Var}(X_1) - \frac{1}{4} \mathbf{E}_{X_1} \mathbf{E}_{X_2}[\delta(X_1, X_2)^2] \tag{8.1}$$

From the semi-parallelogram law above, (Ho et al., 2013, Proposition 1) derive:

$$\int_{\mathcal{M}} [\delta(x, y)^2 - \delta(x, \mu)^2] d\nu(x) \geq \delta(y, \mu)^2, \quad \text{for any } y \in \mathcal{M}$$

By the above inequality (and independence of X_1, X_2),

$$\begin{aligned} \mathbf{E}_{X_2}[\delta(X_1, X_2)^2 \mid X_1 = x_1] &= \int_{\mathcal{M}} \delta(x_1, X_2)^2 d\nu(X_2) \\ &\geq \delta(x_1, \mu)^2 + \int_{\mathcal{M}} \delta(X_2, \mu)^2 d\nu(X_2) \\ &= \delta(x_1, \mu)^2 + \text{Var}(X_2) \end{aligned}$$

and consequently,

$$\begin{aligned} \mathbf{E}_{X_1} \mathbf{E}_{X_2}[\delta(X_1, X_2)^2] &\geq \int_{\mathcal{M}} \delta(X_1, \mu)^2 d\nu(X_1) + \text{Var}(X_2) \\ &= 2\text{Var}(X_1) \end{aligned}$$

Returning to eq.(8.1),

$$\mathbf{E}[\delta(\mu_2, \mu)^2] \leq \frac{1}{2} \text{Var}(X_1)$$

Repeating the same argument, using independence of $\gamma(X_1, X_2, 1/2)$ and $\gamma(X_3, X_4, 1/2)$,

$$\mathbf{E}[\delta(\mu_4, \mu)^2] \leq \frac{1}{2} \mathbf{E}[\delta(\mu_2, \mu)^2] \leq \frac{1}{4} \text{Var}(X_1)$$

Continuing this iteration up to μ_n , we find the upper bound:

$$\mathbf{E}[\delta(\mu_n, \mu)^2] \leq \frac{1}{2} \mathbf{E}_{n/2}[\delta(\mu_{n/2}, \mu)^2] \leq \dots \leq \frac{1}{n} \text{Var}(X_1)$$

By Markov's inequality, $P(\delta(\mu_n, \mu) > \epsilon) \rightarrow 0$ for each $\epsilon > 0$ as $n \rightarrow \infty$, since the distribution ν is assumed to have finite second moment. \square

8.2 Proof of Proposition 4.2

Proof. Denote $L := (N-1)/2$, with $L \geq 0$, and fix $j \geq 1$ sufficiently large and $k \in [L, 2^{j-1} - (L+1)]$ away from the boundary, such that the neighboring $(j-1)$ -midpoints $M_{j-1, k-L}, \dots, M_{j-1, k+L}$ exist.

Remark: For $k < L$ or $k > 2^{j-1} - (L+1)$ near the boundary, we collect the N available closest neighbors of $M_{j-1, k}$ (either to the left or right). The remainder of the proof for the boundary case is exactly analogous to the non-boundary case and follows directly by mimicking the arguments outlined below.

We predict $M_{j, 2k+1}$ from $M_{j-1, k-L}, \dots, M_{j-1, k+L}$ via intrinsic polynomial interpolation of degree $N-1$ passing through the N points $\bar{M}_{j-1, 0}, \dots, \bar{M}_{j-1, N-1}$, where $\bar{M}_{j-1, k}$ denotes the cumulative intrinsic average as in eq.(3.3) in the main document. The predicted midpoint $\widetilde{M}_{j, 2k+1}$ is then a weighted intrinsic average of the estimated polynomial at $(2k+1)2^{-j}$, i.e. $\widehat{M}_{(k-L)2^{-(j-1)}}((2k+1)2^{-j})$, and the given midpoint $\bar{M}_{j-1, L} = M_{(k-L)2^{-(j-1)}}(2k2^{-j})$, (with notation as in eq.(3.2) in the main document).

For notational simplicity, write $M(t) := M_{(k-L)2^{-(j-1)}}(t)$ and $\widehat{M}(t) := \widehat{M}_{(k-L)2^{-(j-1)}}(t)$ for the true and estimated intrinsic cumulative mean curves respectively, where the latter is a polynomial of order $N-1$. $M(t)$ is a smooth curve with existing covariant derivatives up to order N and we can Taylor expand its derivative $\frac{d}{dt}M(t)$ at $t = 2k2^{-j}$ as in eq.(4.2) in the main document. Choose $t_0 \in (2k2^{-j}, (2k+1)2^{-j})$ such that $|t - t_0| \leq 2^{-j}$ and $t \rightarrow t_0$ as $j \rightarrow \infty$, then,

$$\frac{d}{dt}M|_{t=2k2^{-j}} = \sum_{\ell=0}^{N-1} \Gamma(M)_{t_0}^{2k2^{-j}} \left(\nabla_{\frac{d}{dt}M}^{\ell} \frac{d}{dt}M|_{t=t_0} \right) \frac{(2k2^{-j} - t_0)^{\ell}}{\ell!} + O(2^{-jN}) \quad (8.2)$$

Similarly, Taylor expanding the derivative $\frac{d}{dt}\widehat{M}|_{t=2k2^{-j}}$ results in terms up to order $N-1$ matching the terms in the expansion of $\frac{d}{dt}M|_{t=2k2^{-j}}$ and vanishing higher order terms. This follows from the

fact that $\widehat{M}(t)$ is the intrinsic polynomial passing through N different points on the curve $M(t)$, thus matching the covariant derivatives up to order $N - 1$. In particular, matching terms of the intrinsic Taylor expansions up to order $N - 1$, by the definition of the derivative:

$$\begin{aligned} \lim_{\Delta t \rightarrow 0} \frac{1}{\Delta t} \left(\text{Log}_{\widehat{M}(2k2^{-j})}(\widehat{M}(2k2^{-j} + \Delta t)) - \text{Log}_{M(2k2^{-j})}(M(2k2^{-j} + \Delta t)) \right) &= \frac{d}{dt} \widehat{M} \Big|_{t=2k2^{-j}} - \frac{d}{dt} M \Big|_{t=2k2^{-j}} \\ &= O(2^{-jN}) \end{aligned}$$

By construction $M(2k2^{-j}) = \widehat{M}(2k2^{-j})$, therefore we can rewrite,

$$\frac{1}{\Delta t} (1 + o(1)) \left(\text{Log}_{M(2k2^{-j})}(\widehat{M}(2k2^{-j} + \Delta t)) - \text{Log}_{M(2k2^{-j})}(M(2k2^{-j} + \Delta t)) \right) = O(2^{-jN})$$

Setting $\Delta t = 2^{-j}$, since $2^{-j}/(1 + o(1)) \rightarrow 0$ as $j \rightarrow \infty$, also,

$$\text{Log}_{M(2k2^{-j})}(\widehat{M}((2k+1)2^{-j})) - \text{Log}_{M(2k2^{-j})}(M((2k+1)2^{-j})) = O(2^{-jN}) \quad (8.3)$$

For notational convenience, in the remainder of this proof, we write $\Lambda = \lambda E$ for some arbitrary (not necessarily fixed) deterministic matrix $E \in \mathbb{C}^{d \times d}$ and constant $\lambda \lesssim 2^{-jN}$, i.e. $\Lambda = O(2^{-jN})$.

Let $M, M_1, M_2 \in \mathcal{M}$ be deterministic matrices, we verify the following implication:

Claim *If $\text{Log}_M(M_1) - \text{Log}_M(M_2) = O(\lambda)$, then also $M_1 = M_2 + O(\lambda)$.*

Proof. Reverting the steps in the claimed implication $\text{Log}(M + \Lambda) = \text{Log}(M) + O(\lambda)$ in the proof of Proposition 4.3, (note that this is the deterministic version). It can also be checked that the reverse implication $\text{Log}(M) + \Lambda = \text{Log}(M + O(\lambda))$ holds. Using this relation, starting from $\text{Log}_M(M_1) - \text{Log}_M(M_2) = O(\lambda)$, by the definition of the logarithmic map, we write out,

$$\begin{aligned} M^{1/2} * \text{Log}(M^{-1/2} * M_1) &= M^{1/2} * \text{Log}(M^{-1/2} * M_2) + O(\lambda) &\Rightarrow \\ \text{Log}(M^{-1/2} * M_1) &= \text{Log}(M^{-1/2} * M_2) + O(\lambda) &\Rightarrow \\ \text{Log}(M^{-1/2} * M_1) &= \text{Log}(M^{-1/2} * M_2 + O(\lambda)) &\Rightarrow \\ M^{-1/2} * M_1 &= M^{-1/2} * M_2 + O(\lambda) &\Rightarrow \\ M^{-1/2} * (M_1 - M_2) &= O(\lambda) &\Rightarrow \\ M_1 &= M_2 + O(\lambda) \end{aligned}$$

□

Applying the above implication to eq.(8.3) yields,

$$\widehat{M}((2k+1)2^{-j}) = M((2k+1)2^{-j}) + O(2^{-jN}) \quad (8.4)$$

The predicted midpoint $\widetilde{M}_{j,2k+1}$ is reconstructed from $\widehat{M}((2k+1)2^{-j})$ and $M(2k2^{-j})$ as follows. By definition of $M(t)$ as the cumulative intrinsic mean curve, we can write $M((2k+1)2^{-j})$ as a weighted intrinsic average between $\overline{M}_{j-1,L} = M(2k2^{-j})$ and $M_{j,2k+1}$ according to:

$$M((2k+1)2^{-j}) = \text{Exp}_{M((2k+1)2^{-j})} \left(\frac{(N-1)2^{-j}}{N2^{-j}} \text{Log}_{M((2k+1)2^{-j})}(\overline{M}_{j-1,L}) + \frac{2^{-j}}{N2^{-j}} \text{Log}_{M((2k+1)2^{-j})}(M_{j,2k+1}) \right)$$

Application of the logarithmic map $\text{Log}_{M((2k+1)2^{-j})}(\cdot)$ to both sides and rearranging terms (substitute $N-1=2L$), gives,

$$\frac{-2L}{N} \text{Log}_{M((2k+1)2^{-j})}(\overline{M}_{j-1,L}) = \frac{1}{N} \text{Log}_{M((2k+1)2^{-j})}(M_{j,2k+1})$$

Or in terms of $M_{j,2k+1}$,

$$\begin{aligned} M_{j,2k+1} &= \text{Exp}_{M((2k+1)2^{-j})} \left(-2L \cdot \text{Log}_{M((2k+1)2^{-j})}(\overline{M}_{j-1,L}) \right) \\ &= \gamma \left(M((2k+1)2^{-j}), \overline{M}_{j-1,L}, -2L \right) \end{aligned}$$

The predicted midpoint $\widetilde{M}_{j,2k+1}$ is given by replacing the true point $M((2k+1)2^{-j})$ by the estimated point $\widehat{M}((2k+1)2^{-j})$, ($\overline{M}_{j-1,L}$ is known), i.e.

$$\widetilde{M}_{j,2k+1} = \gamma \left(\widehat{M}((2k+1)2^{-j}), \overline{M}_{j-1,L}, -2L \right) \quad (8.5)$$

Below, we use that $(M+\Lambda)^a = M^a + O(\lambda)$ for $a \in \mathbb{N}$, $(M+\Lambda)^{1/2} = M^{1/2} + O(\lambda)$ and $(M+\Lambda)^{-1} = M^{-1} + O(\lambda)$ for $M \in \mathcal{M}$ and $\lambda \rightarrow 0$ sufficiently small, as verified in the proof of Proposition 4.3, (note that this is the deterministic version), combined with eq.(8.4) and the definition of the geodesic in eq.(2.3) in the main document. Writing out eq.(8.5) gives,

$$\begin{aligned} \widetilde{M}_{j,2k+1} &= \left(M((2k+1)2^{-j})^{1/2} + \Lambda \right) * \left(\left(M((2k+1)2^{-j})^{-1/2} + \Lambda \right) * \overline{M}_{j-1,L} \right)^{-2L} \\ &= \left(M((2k+1)2^{-j})^{1/2} + \Lambda \right) * \left(\left(M((2k+1)2^{-j})^{-1/2} * \overline{M}_{j-1,L} \right)^{-1} + \Lambda \right)^{2L} \\ &= \left(M((2k+1)2^{-j})^{1/2} + \Lambda \right) * \left(\left(M((2k+1)2^{-j})^{-1/2} * \overline{M}_{j-1,L} \right)^{-2L} + \Lambda \right) \\ &= M_{j,2k+1} + O(2^{-jN}) \end{aligned} \quad (8.6)$$

Substituting the above result in the whitened wavelet coefficient $\mathfrak{D}_{j,k} = 2^{-j/2} \text{Log}(\widetilde{M}_{j,2k+1}^{-1/2} * M_{j,2k+1})$, by the same identities as used above combined with $\text{Log}(M+\Lambda) = \text{Log}(M) + O(\lambda)$, (verified in

the proof of Proposition 4.3), it follows that for $j \geq 1$ sufficiently large,

$$\begin{aligned}
\|\mathfrak{D}_{j,k}\|_F &= \left\| 2^{-j/2} \text{Log}((M_{j,2k+1} + \Lambda)^{-1/2} * M_{j,2k+1}) \right\|_F \\
&= 2^{-j/2} \left\| \text{Log}((M_{j,2k+1}^{-1/2} + \Lambda) * M_{j,2k+1}) \right\|_F \\
&= 2^{-j/2} \left\| \text{Log}(\text{Id} + \Lambda) \right\|_F \\
&= 2^{-j/2} (\|\mathbf{0}\|_F + O(2^{-jN})) = O\left(2^{-j/2} 2^{-jN}\right)
\end{aligned}$$

□

8.3 Proof of Proposition 4.3

Proof. By the proof of Proposition 4.1, $\mathbf{E}[\delta(M_{j,k,n}, M_{j,k})^2] = O(2^{-(J-j)})$ for each $j \geq 0$ and $0 \leq k \leq 2^j - 1$. For notational convenience, in the remainder of this proof $\epsilon_{j,n}$ denotes a general (not necessarily the same) random error matrix that satisfies $\mathbf{E}\|\epsilon_{j,n}\|_F^2 = O(2^{-(J-j)})$. Furthermore, we can appropriately write $M_{j,k,n} = \text{Exp}_{M_{j,k}}(\epsilon_{j,n})$, such that $M_{j,k,n} \xrightarrow{P} M_{j,k}$ as $J \rightarrow \infty$ at the correct rate since,

$$\begin{aligned}
\mathbf{E}[\delta(\text{Exp}_{M_{j,k}}(\epsilon_{j,n}), M_{j,k})^2] &= \mathbb{E}\|\text{Log}(M_{j,k}^{-1/2} * \text{Exp}_{M_{j,k}}(\epsilon_{j,n}))\|_F^2 \\
&= \mathbb{E}\|M_{j,k}^{-1/2} * \epsilon_{j,n}\|_F^2 \\
&= O(2^{-(J-j)})
\end{aligned}$$

using the definitions of the Riemannian distance function and the logarithmic and exponential maps in Section 2. In particular, by a first-order Taylor expansion of the matrix exponential, (abusing notation of $\epsilon_{j-1,n}$), $M_{j-1,k,n} = M_{j-1,k}^{1/2} * \text{Exp}(\epsilon_{j-1,n}) = M_{j-1,k}^{1/2} * (\text{Id} + \epsilon_{j-1,n} + \dots) = M_{j-1,k} + \epsilon_{j-1,n}$.

By eq.(3.4) in the main document, the predicted midpoint $\widetilde{M}_{j,2k+1,n}$ is a weighted intrinsic mean of N coarse-scale midpoints $(M_{j-1,k,n})_k$ with weights summing up to 1. The rate of $\widetilde{M}_{j,2k+1,n}$ is therefore upper bounded by the (worst) convergence rate of the individual midpoints $(M_{j-1,k,n})_k$, and we can also write $\widetilde{M}_{j,2k+1,n} = \widetilde{M}_{j,2k+1} + \epsilon_{j-1,n}$.

Below, we verify several implications that are needed to finish the proof. let $M \in \mathcal{M}$ be a deterministic matrix and $\lambda E = O_p(\lambda)$ a random error matrix, such that $\mathbf{E}\|\lambda E\|_F = O(\lambda)$.

Claim If $\lambda \rightarrow 0$ sufficiently small, then $\text{Log}(M + \lambda E) = \text{Log}(M) + O_p(\lambda)$.

Proof. Rewrite $\text{Log}(M + \lambda E) = \text{Log}(M(\text{Id} + \lambda M^{-1}E))$. By the Baker-Campbell-Hausdorff

formula (e.g. (Higham, 2008, Theorem 10.4)), with $X = \text{Log}(M)$ and $Y = \text{Log}(\text{Id} + \lambda M^{-1}E)$,

$$\text{Log}(M + \lambda E) = X + Y + \frac{1}{2}[X, Y] + \frac{1}{12}([X, [X, Y]] + [Y, [Y, X]]) + \frac{1}{24}[Y, [X, [X, Y]]] - \dots$$

where $[X, Y] = XY - YX$ denotes the commutator of X and Y . In particular,

$$\begin{aligned} [X, Y] &= [\text{Log}(M), \text{Log}(\text{Id} + \lambda M^{-1}E)] \\ &= \text{Log}(M)\text{Log}(\text{Id} + \lambda M^{-1}E) - \text{Log}(\text{Id} + \lambda M^{-1}E)\text{Log}(M) \\ &= \text{Log}(M)(\lambda M^{-1}E + O_p(\lambda^2)) - (\lambda M^{-1}E + O_p(\lambda^2))\text{Log}(M) \\ &= O_p(\lambda) \end{aligned}$$

Here, we expanded $\text{Log}(\text{Id} + \lambda M^{-1}E) = \lambda M^{-1}E + O_p(\lambda^2)$ via its Mercator series (e.g. (Higham, 2008, Section 11.3)), using that the spectral radius $\rho(\lambda M^{-1}E) = \lambda \rho(M^{-1}E) < 1$ almost surely for $\lambda \rightarrow 0$ sufficiently small.

Iterating the above argument, it follows that all the nested (higher-order) commutators are of the order $O_p(\lambda)$ as well, and we rewrite:

$$\text{Log}(M + \lambda E) = \text{Log}(M) + \text{Log}(\text{Id} + \lambda M^{-1}E) + O_p(\lambda)$$

Expanding again $\text{Log}(\text{Id} + \lambda M^{-1}E) = \lambda M^{-1}E + O_p(\lambda^2) = O_p(\lambda)$, (for λ sufficiently small), the claim follows. \square

Claim If $\lambda \rightarrow 0$ sufficiently small, then $(M + \lambda E)^{1/2} = M^{1/2} + O_p(\lambda)$ and $(M + \lambda E)^{-1} = M^{-1} + O_p(\lambda)$.

Proof. For the first claim, Taylor expanding the matrix exponential,

$$\begin{aligned} (M + \lambda E)^{1/2} &= \text{Exp}\left(\frac{1}{2}\text{Log}(M + \lambda E)\right) = \sum_{k=0}^{\infty} \frac{(\text{Log}(M + \lambda E))^k}{2^k k!} \\ &= \sum_{k=0}^{\infty} \frac{(\text{Log}(M) + O_p(\lambda))^k}{2^k k!} = \sum_{k=0}^{\infty} \frac{(\text{Log}(M))^k}{2^k k!} + O_p(\lambda) = M^{1/2} + O_p(\lambda) \end{aligned}$$

using the previous claim $\text{Log}(M + \lambda E) = \text{Log}(M) + O_p(\lambda)$ for $\lambda \rightarrow 0$ sufficiently small.

For the second claim, rewrite, (for λ sufficiently small),

$$\begin{aligned} (M + \lambda E)^{-1} &= (M(\text{Id} + \lambda M^{-1}E))^{-1} \\ &= (\text{Id} + \lambda M^{-1}E)^{-1}M^{-1} \\ &= (\text{Id} - \lambda M^{-1}E + (\lambda M^{-1}E)^2 - \dots)M^{-1} = M^{-1} + O_p(\lambda) \end{aligned}$$

applying a binomial series expansion of the matrix inverse $(\text{Id} + \lambda M^{-1}E)^{-1}$, using that the spectral radius $\rho(\lambda M^{-1}E) = \lambda \rho(M^{-1}E) < 1$ almost surely for $\lambda \rightarrow 0$ sufficiently small. Combining the two claims, we find in particular also that $(M + \lambda E)^{-1/2} = M^{-1/2} + O_p(\lambda)$. \square

Combining the above results, with some abuse of notation for $\epsilon_{j,n}$, we write out for the empirical whitened wavelet coefficient $\widehat{\mathfrak{D}}_{j,k,n}$,

$$\begin{aligned}\widehat{\mathfrak{D}}_{j,k,n} &= 2^{-j/2} \text{Log} \left((\widetilde{M}_{j,2k+1} + \epsilon_{j-1,n})^{-1/2} * (M_{j,2k+1} + \epsilon_{j,n}) \right) \\ &= 2^{-j/2} \text{Log} \left((\widetilde{M}_{j,2k+1}^{-1/2} + \epsilon_{j-1,n}) * (M_{j,2k+1} + \epsilon_{j,n}) \right) \\ &= 2^{-j/2} \text{Log} \left(\widetilde{M}_{j,2k+1}^{-1/2} * M_{j,2k+1} + \epsilon_{j,n} + \dots \right) \\ &= 2^{-j/2} \text{Log} \left(\widetilde{M}_{j,2k+1}^{-1/2} * M_{j,2k+1} \right) + 2^{-j/2} O_p(2^{-(J-j)/2}) \\ &= \mathfrak{D}_{j,k} + 2^{-j/2} O_p(2^{-(J-j)/2})\end{aligned}$$

Plugging in the above result, it follows that for $j < J$ sufficiently small,

$$\mathbf{E} \|\widehat{\mathfrak{D}}_{j,k,n} - \mathfrak{D}_{j,k}\|_F^2 = O(2^{-j} 2^{-(J-j)}) = O(n^{-1})$$

\square

8.4 Proof of Theorem 4.4

Proof. For the first part of the theorem, suppose that $J_0 = \log_2(n)/(2N + 1) \gg 1$ is sufficiently large such that the rates in Propositions 4.2 and 4.3 hold. Then,

$$\begin{aligned}\sum_{j,k} \mathbf{E} \|\widehat{\mathfrak{D}}_{j,k} - \mathfrak{D}_{j,k}\|_F^2 &= \sum_{j \geq J_0} \|\mathfrak{D}_{j,k}\|_F^2 + \sum_{j < J_0} \mathbf{E} \|\widehat{\mathfrak{D}}_{j,k} - \mathfrak{D}_{j,k}\|_F^2 \\ &\lesssim \sum_{j \geq J_0} 2^j (2^{-j} 2^{-2jN}) + \sum_{j < J_0} 2^j n^{-1} \\ &= \left(\sum_{j=0}^J (2^{-2N})^j - \sum_{j=0}^{J_0-1} (2^{-2N})^j \right) + n^{-1} \sum_{j=1}^{J_0-1} 2^j \\ &= \frac{(2^{-2N})^{J_0} - (2^{-2N})^{(J+1)}}{1 - 2^{-2N}} + n^{-1} (2^{J_0} - 2) \\ &\lesssim (2^{-2N})^{J_0} + n^{-1} 2^{J_0} + n^{-1} \\ &\lesssim n^{-2N/(2N+1)}\end{aligned} \tag{8.7}$$

where the last step follows from substituting $J_0 = \log_2(n)/(2N + 1)$ since,

$$\begin{aligned}(2^{-2N})^{J_0} &= \exp(-2N J_0 \log(2)) = \exp\left(\frac{-2N}{2N+1} \log(n)\right) = n^{-2N/(2N+1)} \\ n^{-1} 2^{J_0} &= \exp(-\log(n) + J_0 \log(2)) = \exp\left(\frac{-2N}{2N+1} \log(n)\right) = n^{-2N/(2N+1)}\end{aligned}$$

For the second part of the theorem, first approximate:

$$\mathbf{E} \left[\int_0^1 \delta(\gamma(t), \hat{\gamma}(t))^2 dt \right] = \frac{1}{n} \sum_{k=0}^{n-1} \mathbf{E} \left[\delta(M_{J,k}, \widehat{M}_{J,k,n})^2 \right] + O(n^{-1}) \quad (8.8)$$

where $M_{J,k}$ is the true midpoint at scale-location (J, k) and $\widehat{M}_{J,k,n}$ is its linear wavelet thresholded counterpart. If for each $k = 0, \dots, n-1$, we can verify that $\mathbf{E}[\delta(M_{J,k}, \widehat{M}_{J,k,n})^2] \lesssim n^{-2N/(2N+1)}$ the proof is finished.

At scales $j = 1, \dots, J$, based on the estimated midpoints $(\widehat{M}_{j-1,k',n})_{k'}$ and the estimated wavelet coefficient $\widehat{D}_{j,k,n}$, in the inverse wavelet transform, the finer-scale midpoint $\widehat{M}_{j,k,n}$ is estimated through,

$$\widehat{M}_{j,k,n} = \text{Exp}_{\widehat{M}_{j,k,n}} \left(2^{j/2} \widehat{D}_{j,k,n} \right)$$

where $\widehat{M}_{j,k,n}$ is the predicted midpoint at scale-location (j, k) based on $(\widehat{M}_{j-1,k',n})_{k'}$.

At scales $j = 1, \dots, J_0 - 1$, we do not alter the wavelet coefficients. Assuming that $j \ll J$ is sufficiently small, such that the rate in Proposition 4.3 holds, we write $\widehat{\mathcal{D}}_{j,k,n} = \mathcal{D}_{j,k} + \eta_n$, with η_n a general (not always the same) random error matrix satisfying $\mathbf{E}\|\eta_n\|_F = O(n^{-1/2})$. Also, by the proof of Proposition 4.3, we can write $\widetilde{M}_{j,k,n} = \widetilde{M}_{j,k} + \epsilon_{j,n}$, where $\epsilon_{j,n}$ is a general (not always the same) random error matrix satisfying $\mathbf{E}\|\epsilon_{j,n}\|_F = O(2^{-(J-j)/2})$.

In particular, at scale $j = 1$,

$$\begin{aligned} \widehat{M}_{1,k,n} &= \text{Exp}_{\widehat{M}_{1,k,n}} \left(2^{1/2} \widehat{D}_{1,k,n} \right) \\ &= \widetilde{M}_{1,k,n}^{1/2} * \text{Exp} \left(2^{1/2} \widetilde{M}_{1,k,n}^{-1/2} * \widehat{D}_{1,k,n} \right) \\ &= \widetilde{M}_{1,k,n}^{1/2} * \text{Exp} \left(2^{1/2} \widehat{\mathcal{D}}_{1,k,n} \right) \\ &= \left(\widetilde{M}_{1,k} + \epsilon_{1,n} \right)^{1/2} * \text{Exp} \left(2^{1/2} \mathcal{D}_{1,k} + \eta_n \right) \\ &= \left(\widetilde{M}_{1,k}^{1/2} + \epsilon_{1,n} \right) * \left(\text{Exp} \left(2^{1/2} \mathcal{D}_{1,k} \right) + \eta_n \right) \\ &= M_{1,k} + O_p(2^{-(J-1)/2}) \end{aligned} \quad (8.9)$$

Here, we used that $(M + \lambda E)^{1/2} = M^{1/2} + O_p(\lambda)$ for $\lambda \rightarrow 0$ sufficiently small as in the proof of Proposition 4.3, and a Taylor expansion of the matrix exponential:

$$\text{Exp}(D + \eta_n) = \sum_{k=0}^{\infty} \frac{(D + \eta_n)^k}{k!} = \sum_{k=0}^{\infty} \frac{D^k}{k!} + O_p(n^{-1/2}) = \text{Exp}(D) + O_p(n^{-1/2})$$

Iterating this same argument for each scale $j = 2, \dots, J_0 - 1$, we find that:

$$\widehat{M}_{J_0-1,k,n} = M_{J_0-1,k} + \sum_{j=1}^{J_0-1} O_p(n^{-1/2}2^{j/2}) = M_{J_0,k} + O_p(2^{-(J-J_0)/2})$$

At scales $j = J_0, \dots, J$, we set $\widehat{D}_{j,k,n} = \mathbf{0}$ for each k . Assuming that $j \gg 1$ is sufficiently large, such that the rate in Proposition 4.2 holds, we can write $\widehat{D}_{j,k,n} = \mathbf{0} = \mathfrak{D}_{j,k} + \zeta_{j,N}$, with $\zeta_{j,N}$ a general (not always the same) deterministic error matrix satisfying $\|\zeta_{j,N}\|_F = O(2^{-j/2}2^{-jN})$.

In particular, at scale $j = J_0$,

$$\begin{aligned} \widehat{M}_{J_0,k,n} &= \text{Exp}_{\widehat{M}_{J_0,k,n}}(2^{J_0/2}\widehat{D}_{J_0,k,n}) \\ &= (\widetilde{M}_{J_0,k} + \epsilon_{J_0,n})^{1/2} * \text{Exp}\left((\widetilde{M}_{J_0,k} + \epsilon_{J_0,n})^{-1/2} * 2^{J_0/2}(\mathfrak{D}_{J_0,k} + \zeta_{J_0,n})\right) \\ &= (\widetilde{M}_{J_0,k}^{1/2} + \epsilon_{J_0,n}) * \text{Exp}\left((\widetilde{M}_{J_0,k}^{-1/2} + \epsilon_{J_0,n}) * (2^{J_0/2}\mathfrak{D}_{J_0,k} + 2^{J_0/2}\zeta_{J_0,n})\right) \\ &= (\widetilde{M}_{J_0,k}^{1/2} + \epsilon_{J_0,n}) * \left(\text{Exp}(2^{J_0/2}D_{J_0,k}) + 2^{J_0/2}\epsilon_{J_0,n}\mathfrak{D}_{J_0,k} + 2^{J_0/2}\zeta_{J_0,n}\right) \\ &= (\widetilde{M}_{J_0,k}^{1/2} + \epsilon_{J_0,n}) * \left(\text{Exp}(2^{J_0/2}D_{J_0,k}) + O_p(2^{-J_0N})\right) \\ &= M_{J_0,k} + O_p(n^{-1/2}2^{J_0/2}) + O_p(2^{-J_0N}) \end{aligned}$$

which follows in the same way as in eq.(8.9) above, combined with the observation that $2^{J_0/2}\epsilon_{J_0,n}\mathfrak{D}_{J_0,k} = O_p(2^{-J_0N})$, since $\|2^{J_0/2}\epsilon_{J_0,n}\mathfrak{D}_{J_0,k}\|_F = O_p(2^{-(J-J_0)/2}2^{-J_0N}) = O_p(2^{-J_0N})$ by Proposition 4.2. Iterating this same argument for each scale $j = J_0 + 1, \dots, J$ yields,

$$\widehat{M}_{J,k,n} = M_{J,k} + O_p(n^{-1/2}2^{J_0/2}) + \sum_{j=J_0}^J O_p(2^{-jN}) = M_{J,k} + O_p(2^{-J_0N}) + O_p(n^{-1/2}2^{J_0/2})$$

The second step follows in the same way as in eq.(8.7). Plugging in $J_0 \sim \log_2(n)/(2N + 1)$, as previously demonstrated, the above expression reduces to:

$$\widehat{M}_{J,k,n} = M_{J,k} + O_p(n^{-N/(2N+1)}), \quad \text{for each } k = 0, \dots, n-1$$

For notational convenience, denote by $\xi_{n,N}$ a general (not always the same) random error matrix such that $\mathbf{E}\|\xi_{n,N}\|_F = O(n^{-N/(2N+1)})$. For each $k = 0, \dots, n-1$, by the previous result:

$$\begin{aligned} \mathbf{E} \left[\delta(M_{J,k}, \widehat{M}_{J,k,n})^2 \right] &= \mathbf{E} \left[\delta(M_{J,k}, M_{J,k} + \xi_{n,N})^2 \right] \\ &= \mathbf{E} \left\| \text{Log} \left(M_{J,k}^{1/2} * (M_{J,k} + \xi_{n,N}) \right) \right\|_F^2 \\ &= \mathbf{E} \left\| \text{Log}(\text{Id} + \xi_{n,N}) \right\|_F^2 = O(n^{-2N/(2N+1)}) \end{aligned}$$

where in the final step we expanded $\text{Log}(\text{Id} + \xi_{n,N}) = O_p(n^{-N/(2N+1)})$ via its Mercator series (e.g. (Higham, 2008, Section 11.3)), using that the spectral radius of $\xi_{n,N}$ is smaller than 1 almost surely for n sufficiently large. Plugging the above rates back into eq.(8.8) yields the claimed result. \square

8.5 Proof of Theorem 5.1

Proof. First, we derive the bias $b(X, f) = c(d, B) \cdot f$. By linearity of the (ordinary) expectation:

$$b(X, f) = \mathbf{E}[\text{Log}_f(X)] = f^{1/2} * \mathbf{E}[\text{Log}(f^{-1/2} * X)] \quad (8.10)$$

using that $g * \text{Log}_{X_1}(X_2) = \text{Log}_{g * X_1}(g * X_2)$ for any $g \in \text{GL}(d, \mathbb{C})$. The transformed random variable $Y := f^{-1/2} * X$ is distributed as $Y \sim W_d^c(B, B^{-1}\text{Id})$, which is unitarily invariant (see e.g. (Muirhead, 1982, Section 3.2)). By (Tulino and Verdú, 2004, Section 2.1.5), taking the eigendecomposition of a unitarily invariant matrix $Y = Q * \Lambda$, the matrix of eigenvectors Q is distributed according to the Haar measure, i.e. the uniform distribution on the set of unitary matrices $\mathcal{U}_d = \{U \in \text{GL}(d, \mathbb{C}) \mid U^*U = \text{Id}\}$, implying that the eigenvectors $(\vec{q}_i)_{i=1, \dots, d}$ (the columns of Q) are identically distributed. Furthermore, Q is independent of the diagonal eigenvalue-matrix Λ , therefore (see also Smith (2000)):

$$\mathbf{E}[\text{Log}(Y)] = \mathbf{E} \left[\sum_{i=1}^d \log(\lambda_i) \vec{q}_i \vec{q}_i^* \right] = \mathbf{E}[\vec{q}_i \vec{q}_i^*] \mathbf{E}[\log(\det(\Lambda))] \quad (8.11)$$

Since Y is Hermitian, $Q \in \mathcal{U}_d$, and therefore $\mathbf{E}[\log(\det(\Lambda))] = \mathbf{E}[\log(\det(Y))]$. By (Muirhead, 1982, Theorem 3.2.15),

$$\log(\det(Y)) \sim -d \log(2B) + \sum_{i=1}^d \log \left(\chi_{2(B-(d-i))}^2 \right)$$

with $\chi_{2(B-(d-i))}^2$ mutually independent chi-squared distributions, with $2(B - (d - i))$ degrees of freedom. Using that $\mathbf{E}[\log(\chi_\nu^2)] = \log(2) + \psi(\nu/2)$, it follows that:

$$\mathbf{E}[\log(\det(\Lambda))] = -d \log(B) + \sum_{i=1}^d \psi(B - (d - i))$$

Following Smith (2000), $\mathbf{E}[\vec{q}_i \vec{q}_i^*] = d^{-1} \text{Id}$, thus by eq.(8.11):

$$\mathbf{E}[\text{Log}(Y)] = \left(-\log(B) + \frac{1}{d} \sum_{i=1}^d \psi(B - (d - i)) \right) \text{Id} = c(d, B) \cdot \text{Id}$$

Plugging this back into eq.(8.10) yields $b(X, f) = c(d, B) \cdot f$.

For the second part of the theorem, observe that \tilde{X}_ℓ ($1 \leq \ell \leq n$) is unbiased with respect to f , since:

$$\begin{aligned} b(\tilde{X}_\ell, f) &= f^{1/2} * \mathbf{E}[\text{Log}(f^{-1/2} * \tilde{X}_\ell)] \\ &= f^{1/2} * \mathbf{E}[\text{Log}(e^{-c(d, B)} \text{Id}) + \text{Log}(f^{-1/2} * X_\ell)] \\ &= f^{1/2} * (-c(d, B) \text{Id} + c(d, B) \text{Id}) = \mathbf{0} \end{aligned}$$

using that $\text{Log}(AB) = \text{Log}(A) + \text{Log}(B)$ for commuting matrices A, B , and $\mathbf{E}[\text{Log}(f^{-1/2} * X_\ell)] = c(d, B) \cdot \text{Id}$ as shown above. By Pennek (2006), the unique intrinsic mean of \tilde{X}_ℓ on \mathcal{M} is characterized by μ such that $b(\tilde{X}_\ell, \mu) = \mathbf{E}[\text{Log}_\mu(\tilde{X}_\ell)] = \mathbf{0}$, i.e. f is the unique intrinsic mean of \tilde{X}_ℓ for each $\ell = 1, \dots, n$. Since the distribution of \tilde{X}_ℓ has finite second moment (rescaled complex Wishart distribution), the convergence in probability follows by Proposition 4.1. \square

8.6 Proofs of Proposition 5.2 and Lemma 5.3

Proof. In this proof, we directly derive the stronger general linear congruence equivariance property in Lemma 5.3. The weaker unitary congruence equivariance property in Proposition 5.2 then follows directly by substituting wavelet thresholding or shrinkage of coefficients that is only equivariant under unitary congruence transformation, (instead of trace thresholding as in Lemma 5.3, which is equivariant under general linear congruence transformation of the coefficients).

Let $M_{j,k}^X, M_{j,k}^{\hat{f}}, D_{j,k}^X$ and $D_{j,k}^{\hat{f}}$ be the midpoints and wavelet coefficients at scale-location (j, k) based on the observations $(X_\ell)_\ell$ and the estimator $(\hat{f}_\ell)_\ell$ respectively. Analogously, let $M_{j,k}^{X,A}, M_{j,k}^{\hat{f},A}, D_{j,k}^{X,A}$ and $D_{j,k}^{\hat{f},A}$ be the midpoints and wavelet coefficients based on the observations $(A * X_\ell)_\ell$ and the estimator $(A * \hat{f}_\ell)_\ell$ respectively, where here and throughout this proof $A \in \text{GL}(d, \mathbb{C})$. Below, we repeatedly make use of the identities $A * \text{Exp}_M(H) = \text{Exp}_{A * M_1}(A * H)$ and $A * \text{Log}_{M_1}(M_2) = \text{Log}_{A * M_1}(A * M_2)$ for $M_1, M_2 \in \mathcal{M}$ and $H \in \mathcal{H}$. In particular, denoting $\text{Mid}(M_1, M_2) := \gamma(M_1, M_2, 1/2)$ for the geodesic midpoint, also,

$$\begin{aligned} A * \text{Mid}(M_1, M_2) &= A * \text{Exp}_{M_1} \left(\frac{1}{2} \text{Log}_{M_1}(M_2) \right) = \\ &= \text{Exp}_{A * M_1} \left(\frac{1}{2} \text{Log}_{A * M_1}(A * M_2) \right) = \text{Mid}(A * M_1, A * M_2) \end{aligned}$$

By construction, the finest-scale midpoints satisfy $M_{j,k}^{X,A} = A * M_{j,k}^X$. Repeated application of the above identity then implies,

$$M_{j,k}^{X,A} = A * M_{j,k}^X \quad \text{for all } j, k \quad (8.12)$$

Furthermore, since the predicted midpoints $\tilde{M}_{j,k}^{X,A}$ are weighted intrinsic means of $(M_{j-1,k'}^{X,A})_{k'}$ according to eq.(3.4) in the main document, the same relation holds for the predicted midpoints, i.e. $\tilde{M}_{j,k}^{X,A} = A * \tilde{M}_{j,k}^X$ for all j, k . Consequently, for the wavelet coefficients at each scale-location (j, k) ,

$$D_{j,k}^{X,A} = 2^{-j/2} \text{Log}_{A * \tilde{M}_{j,2k+1}^X} (A * M_{j,2k+1}^X) = A * D_{j,k}^X \quad (8.13)$$

In Lemma 5.3, we threshold or shrink the wavelet coefficients based on the trace of the whitened coefficients, for which:

$$\begin{aligned}
\text{Tr}(\mathfrak{D}_{j,k}^{X,A}) &= 2^{-j/2} \text{Tr} \left(\text{Log} \left((A * \widetilde{M}_{j,2k+1}^X)^{-1/2} * (A * M_{j,2k+1}^X) \right) \right) \\
&= 2^{-j/2} \left(\text{Tr}(\text{Log}(A * M_{j,2k+1}^X)) - \text{Tr}(\text{Log}(A * \widetilde{M}_{j,2k+1}^X)) \right) \\
&= 2^{-j/2} \left(\text{Tr}(\text{Log}(M_{j,2k+1}^X)) - \text{Tr}(\text{Log}(\widetilde{M}_{j,2k+1}^X)) \right) \\
&= \text{Tr}(\mathfrak{D}_{j,k}^X)
\end{aligned} \tag{8.14}$$

using that $\text{Tr}(\text{Log}(A * X)) = \text{Tr}(\text{Log}(X)) + \text{Tr}(\text{Log}(AA^*))$ and $\text{Tr}(\text{Log}(X^t)) = t\text{Tr}(\text{Log}(X))$ for $X \in \mathcal{M}$ and $t \in \mathbb{R}$, which follows from the fact that $\text{Tr}(\text{Log}(X)) = \log(\det(X))$ and the properties of the determinant and ordinary logarithm. Let $g(\text{Tr}(\mathfrak{D}_{j,k}^X)) \in \mathbb{R}$ be a thresholding or shrinkage constant depending on $\text{Tr}(\mathfrak{D}_{j,k}^X)$, such that $D_{j,k}^{\hat{f}} = g(\text{Tr}(\mathfrak{D}_{j,k}^X))D_{j,k}^X$. Due to the invariance in eq.(8.14) combined with eq.(8.13), it immediately follows that:

$$D_{j,k}^{\hat{f},A} = g(\text{Tr}(\mathfrak{D}_{j,k}^{X,A}))D_{j,k}^{X,A} = A * (g(\text{Tr}(\mathfrak{D}_{j,k}^X))D_{j,k}^X) = A * D_{j,k}^{\hat{f}} \quad \text{for all } j, k$$

The wavelet-thresholded estimator $(\hat{f}_\ell)_\ell$ is retrieved via the inverse wavelet transform applied to the set of thresholded wavelet coefficients (and coarse-scale midpoints). At scale $j = 0$, by eq.(8.12), $M_{0,k}^{\hat{f},A} = M_{0,k}^{X,A} = A * M_{0,k}^X = A * M_{0,k}^{\hat{f}}$. At the odd locations $2k + 1$ at the next coarser scale $j = 1$,

$$\begin{aligned}
M_{1,2k+1}^{\hat{f},A} &= \text{Exp}_{\widetilde{M}_{1,2k+1}^{\hat{f},A}} \left(2^{1/2} D_{j,k}^{\hat{f},A} \right) \\
&= \text{Exp}_{A * \widetilde{M}_{1,2k+1}^{\hat{f}}} \left(A * (2^{1/2} D_{j,k}^{\hat{f}}) \right) \\
&= A * \text{Exp}_{\widetilde{M}_{1,2k+1}^{\hat{f}}} \left(2^{1/2} D_{j,k}^{\hat{f}} \right) \\
&= A * M_{1,2k+1}^{\hat{f}}
\end{aligned}$$

using that $\widetilde{M}_{1,2k+1}^{\hat{f},A} = A * \widetilde{M}_{1,2k+1}^{\hat{f}}$, since the same relation holds for $(M_{0,k'}^{\hat{f},A})_{k'}$ and the predicted midpoints are weighted intrinsic means of $(M_{0,k'}^{\hat{f},A})_{k'}$. Also, at the even locations $2k$,

$$\begin{aligned}
M_{1,2k}^{\hat{f},A} &= M_{0,k}^{\hat{f},A} * (M_{1,2k+1}^{\hat{f},A})^{-1} \\
&= (A * M_{0,k}^{\hat{f}}) * (A * M_{1,2k+1}^{\hat{f}})^{-1} \\
&= A * \left(M_{0,k}^{\hat{f}} * (M_{1,2k+1}^{\hat{f}})^{-1} \right) \\
&= A * M_{1,2k}^{\hat{f}}
\end{aligned}$$

Iterating the same argument up to the finest scale $j = J$ yields the desired result $\hat{f}_{A,\ell} = A * \hat{f}_\ell$ for each $\ell = 1, \dots, 2^J$. \square

8.7 Proof of Proposition 5.4

Proof. Let us write $M_{J,k-1}^X := X_k = f_k^{1/2} * W_k$ for $k = 1, \dots, n$, where the distribution of W_k does not depend on f_k , and the intrinsic mean of W_k is the identity Id. The latter follows from the fact that X_k has intrinsic mean f_k , since:

$$\begin{aligned} \mathbf{E}[\text{Log}_{\text{Id}}(W_k)] &= \mathbf{E}[f_k^{-1/2} * \text{Log}_{f_k}(f_k^{1/2} * W_k)] \\ &= f_k^{-1/2} * \mathbf{E}[\text{Log}_{f_k}(X_k)] \\ &= f_k^{-1/2} * \mathbf{0} = \mathbf{0} \end{aligned}$$

and the intrinsic mean μ of W_k is uniquely characterized by $\mathbf{E}[\text{Log}_\mu(W_k)] = \mathbf{0}$. First, we verify that:

$$\text{Tr}(\text{Log}(M_{j,k}^X)) = \text{Tr}(\text{Log}(M_{j,k}^f)) + \text{Tr}(\text{Log}(M_{j,k}^W)) \quad \text{for all } j, k \quad (8.15)$$

where $M_{j,k}^X$, $M_{j,k}^f$, and $M_{j,k}^W$ are the midpoints at scale-location (j, k) based on the sequences $(X_\ell)_\ell$, $(f_\ell)_\ell$, and $(W_\ell)_\ell$ respectively. Using that $\text{Tr}(\text{Log}(AB)) = \text{Tr}(\text{Log}(A)) + \text{Tr}(\text{Log}(B))$ and $\text{Log}(A^t) = t\text{Log}(A)$ for any $A, B \in \mathcal{M}$, decompose:

$$\begin{aligned} \text{Tr}(\text{Log}(M_{j,k}^X)) &= \text{Tr}(\text{Log}(\text{Mid}(M_{j+1,2k}^X, M_{j+1,2k+1}^X))) \\ &= \text{Tr}(\text{Log}((M_{j+1,2k}^X)^{1/2} * ((M_{j+1,2k}^X)^{-1/2} * M_{j+1,2k+1}^X)^{1/2})) \\ &= \frac{1}{2} \text{Tr}(\text{Log}(M_{j+1,2k}^X)) + \frac{1}{2} \text{Tr}(\text{Log}(M_{j+1,2k+1}^X)) \\ &\vdots \\ &= \frac{1}{2^{J-j}} \sum_{\ell=0}^{2^{J-j-1}} \text{Tr}(\text{Log}(M_{J,(2k)^{J-j-1}+\ell}^X)) \\ &= \frac{1}{2^{J-j}} \sum_{\ell=0}^{2^{J-j-1}} \text{Tr}(\text{Log}(f_{(2k)^{J-j-1}+\ell+1})) + \frac{1}{2^{J-j}} \sum_{\ell=0}^{2^{J-j-1}-1} \text{Tr}(\text{Log}(W_{(2k)^{J-j-1}+\ell+1})) \\ &\vdots \\ &= \text{Tr}(\text{Log}(\text{Mid}(M_{j+1,2k}^f, M_{j+1,2k+1}^f))) + \text{Tr}(\text{Log}(\text{Mid}(M_{j+1,2k}^W, M_{j+1,2k+1}^W))) \\ &= \text{Tr}(\text{Log}(M_{j,k}^f)) + \text{Tr}(\text{Log}(M_{j,k}^W)) \end{aligned}$$

Second, we also verify that:

$$\text{Tr}(\text{Log}(\widetilde{M}_{j,2k+1}^X)) = \text{Tr}(\text{Log}(\widetilde{M}_{j,2k+1}^f)) + \text{Tr}(\text{Log}(\widetilde{M}_{j,2k+1}^W)) \quad \text{for all } j, k \quad (8.16)$$

where $\widetilde{M}_{j,k'}^X$, $\widetilde{M}_{j,k'}^f$, and $\widetilde{M}_{j,k'}^W$ are the imputed midpoints at scale-location (j, k') based on the sequences $(X_\ell)_\ell$, $(f_\ell)_\ell$, and $(W_\ell)_\ell$ respectively. By eq.(3.4) in the main document, the predicted

midpoints at the odd locations $2k + 1$ satisfy:

$$\widetilde{M}_{j,2k+1}^X = \text{Exp}_{\widetilde{M}_{j,2k+1}^X} \left(\sum_{\ell=-L}^L C_{N,2\ell+N} \text{Log}_{\widetilde{M}_{j,2k+1}^X} (M_{j-1,k+\ell}^X) \right)$$

with weights $\mathbf{C}_N = (C_{N,i})_{i=0,\dots,2N-1}$ as in Section 3.1.2. Here, without loss of generality we consider prediction away from the boundary, (at the boundary the sum runs over the $N = 2L + 1$ closest available neighbors to $M_{j,k}$). By eq.(8.15), we decompose,

$$\begin{aligned} \text{Tr}(\text{Log}(\widetilde{M}_{j,2k+1}^X)) &= \text{Tr} \left(\text{Log} \left(\text{Exp}_{\widetilde{M}_{j,2k+1}^X} \left(\sum_{\ell} C_{N,2\ell+N} \text{Log}_{\widetilde{M}_{j,2k+1}^X} (M_{j-1,k+\ell}^X) \right) \right) \right) \\ &= \text{Tr}(\text{Log}(\widetilde{M}_{j,2k+1}^X)) + \text{Tr} \left((\widetilde{M}_{j,2k+1}^X)^{-1/2} * \left(\sum_{\ell} C_{N,2\ell+N} \text{Log}_{\widetilde{M}_{j,2k+1}^X} (M_{j-1,k+\ell}^X) \right) \right) \\ &= \text{Tr}(\text{Log}(\widetilde{M}_{j,2k+1}^X)) + \text{Tr} \left(\sum_{\ell} C_{N,2\ell+N} \text{Log} \left((\widetilde{M}_{j,2k+1}^X)^{-1/2} * M_{j-1,k+\ell}^X \right) \right) \\ &= \text{Tr}(\text{Log}(\widetilde{M}_{j,2k+1}^X)) + \sum_{\ell} C_{N,2\ell+N} \left(\text{Tr}(\text{Log}(M_{j-1,k+\ell}^X)) - \text{Tr}(\text{Log}(\widetilde{M}_{j,2k+1}^X)) \right) \\ &= \sum_{\ell} C_{N,2\ell+N} \text{Tr}(\text{Log}(M_{j-1,k+\ell}^X)) \\ &= \sum_{\ell} C_{N,2\ell+N} \text{Tr}(\text{Log}(M_{j-1,k+\ell}^f)) + \sum_{\ell} C_{N,2\ell+N} \text{Tr}(\text{Log}(M_{j-1,k+\ell}^W)) \\ &\vdots \\ &= \text{Tr}(\text{Log}(\widetilde{M}_{j,2k+1}^f)) + \text{Tr}(\text{Log}(\widetilde{M}_{j,2k+1}^W)) \end{aligned}$$

where we used in particular $g * \text{Log}_{X_1}(X_2) = \text{Log}_{g * X_1}(g * X_2)$ and $g * \text{Exp}_{X_1}(X_2) = \text{Exp}_{g * X_1}(g * X_2)$ for any $g \in \text{GL}(d, \mathbb{C})$, and the fact that $\sum_{\ell} C_{N,2\ell+N} = 1$.

The first claim in the Proposition now follows from eq.(8.15) and eq.(8.16) through:

$$\begin{aligned} \text{Tr}(\mathfrak{D}_{j,k}^X) &= 2^{-j/2} \text{Tr} \left(\text{Log} \left((\widetilde{M}_{j,2k+1}^X)^{-1/2} * M_{j,2k+1}^X \right) \right) \\ &= 2^{-j/2} \text{Tr}(\text{Log}(M_{j,2k+1}^X)) - \text{Tr}(\text{Log}(\widetilde{M}_{j,2k+1}^X)) \\ &= 2^{-j/2} \text{Tr}(\text{Log}(M_{j,2k+1}^f)) + 2^{-j/2} \text{Tr}(\text{Log}(M_{j,2k+1}^W)) \\ &\quad - 2^{-j/2} \left(\text{Tr}(\text{Log}(\widetilde{M}_{j,2k+1}^f)) + \text{Tr}(\text{Log}(\widetilde{M}_{j,2k+1}^W)) \right) \\ &= \text{Tr}(\mathfrak{D}_{j,k}^f) + \text{Tr}(\mathfrak{D}_{j,k}^W) \end{aligned} \tag{8.17}$$

For the second claim in the Proposition, first observe:

$$\mathbf{E}[\text{Tr}(\text{Log}(M_{j,k}^W))] = \frac{1}{2^{J-j}} \sum_{\ell=0}^{2^{J-j}-1} \mathbf{E}[\text{Tr}(\text{Log}(W_{(2k)^{J-j-1+\ell+1}}))] = 0, \quad \text{for each } j, k$$

using that $\mathbf{E}[\text{Tr}(\text{Log}(W_\ell))] = 0$ for each $\ell = 1, \dots, n$, which is implied by $\mathbf{E}[\text{Log}_{\text{Id}}(W_\ell)] = \mathbf{0}$. As a consequence, also,

$$\mathbf{E}[\text{Tr}(\text{Log}(\widetilde{M}_{j,2k+1}^W))] = \sum_{\ell} C_{N,2\ell+N} \text{Tr}(\text{Log}(M_{j-1,k+\ell}^W)) = 0 \quad \text{for each } j, k$$

and therefore,

$$\begin{aligned} \mathbf{E}[\text{Tr}(\mathfrak{D}_{j,k}^X)] &= \text{Tr}(\mathfrak{D}_{j,k}^f) + \mathbf{E}[\text{Tr}(\mathfrak{D}_{j,k}^W)] \\ &= \text{Tr}(\mathfrak{D}_{j,k}^f) + 2^{-j/2} \mathbf{E} \left[\text{Tr}(\text{Log}(M_{j,2k+1}^W)) - \text{Tr}(\text{Log}(\widetilde{M}_{j,2k+1}^W)) \right] \\ &= \text{Tr}(\mathfrak{D}_{j,k}^f) \end{aligned}$$

For the variance of $\text{Tr}(\mathfrak{D}_{j,k}^X)$, we first note that the random variables $(W_\ell)_{\ell=1,\dots,n}$ are i.i.d., implying that the random variables $(\text{Tr}(\text{Log}(M_{j,k}^W))_{k=0,\dots,2^j-1}$ on scale j are independent with equal variance.

We write out:

$$\begin{aligned} \text{Var}(\text{Tr}(\mathfrak{D}_{j,k}^X)) &= 2^{-j} \text{Var} \left(\text{Tr}(\text{Log}(M_{j,2k+1}^W)) - \text{Tr}(\text{Log}(\widetilde{M}_{j,2k+1}^W)) \right) \\ &= 2^{-j} \text{Var} \left(\text{Tr}(\text{Log}(M_{j,2k+1}^W)) - \sum_{\ell} C_{L,2\ell+N} \text{Tr}(\text{Log}(M_{j-1,k+\ell}^X)) \right) \\ &= 2^{-j} \text{Var} \left(\text{Tr}(\text{Log}(M_{j,2k+1}^W)) - C_{N,N} \text{Tr}(\text{Log}(M_{j-1,k}^W)) \right) \\ &\quad + 2^{-j} \sum_{-L \leq \ell \leq L; \ell \neq 0} C_{N,2\ell+N}^2 \text{Var}(\text{Tr}(\text{Log}(M_{j-1,k+\ell}^W))) \\ &= 2^{-(j+1)} \text{Var}(\text{Tr}(\text{Log}(M_{j,2k}^W))) + 2^{-j} \left(\sum_{\ell} C_{N,2\ell+N}^2 - 1 \right) \text{Var}(\text{Tr}(\text{Log}(M_{j-1,k+\ell}^W))) \\ &= 2^{-(j+1)} \sum_{\ell} C_{N,2\ell+N}^2 \text{Var}(\text{Tr}(\text{Log}(M_{j,0}^W))) \end{aligned} \tag{8.18}$$

where in the final two steps we used that $C_{N,N} = 1$, and by the independence of the midpoints within each scale, for each k ,

$$\begin{aligned} \text{Var}(\text{Tr}(\text{Log}(M_{j-1,k}^W))) &= \text{Var} \left(\frac{1}{2} \text{Tr}(\text{Log}(M_{j,2k}^W)) + \frac{1}{2} \text{Tr}(\text{Log}(M_{j,2k+1}^W)) \right) \\ &= \frac{1}{2} \text{Var}(\text{Tr}(\text{Log}(M_{j,0}^W))) \end{aligned}$$

It remains to derive an expression for $\text{Var}(\text{Tr}(\text{Log}(M_{j,0}^W)))$. By repeated application of the above argument,

$$\begin{aligned} \text{Var}(\text{Tr}(\text{Log}(M_{j,0}^W))) &= \frac{1}{2^{J-j}} \text{Var}(\text{Tr}(\text{Log}(M_{J,0}^W))) \\ &= \frac{1}{2^{J-j}} \text{Var}(\text{Tr}(\text{Log}(W_1))) \end{aligned} \tag{8.19}$$

with $W_1 \sim W_d^c(B, B^{-1}e^{-c(d,B)}\text{Id})$. As in the proof of Theorem 5.1,

$$\text{Tr}(\text{Log}(W_1)) \sim -d \log(2e^{c(d,B)}B) + \sum_{i=1}^d \log\left(\chi_{2(B-(d-i))}^2\right)$$

The variance of a $\log(\chi_\nu^2)$ distribution equals $\psi'(\nu/2)$, (with $\psi'(\cdot)$ the trigamma function), therefore:

$$\text{Var}(\text{Tr}(\text{Log}(W_1))) = \sum_{i=1}^d \psi'(B - (d - i))$$

Combining the above result with eq.(8.18) and eq.(8.19) finishes the proof. \square

8.8 Proof of Corollary 5.5

Proof. Analogous to the proof of Theorem 5.1, the random variables $W_1, \dots, W_n \stackrel{\text{iid}}{\sim} W_d^c(B, B^{-1}e^{-c(d,B)}\text{Id})$ are unitarily invariant, see (Muirhead, 1982, Section 3.2). By the same argument as in eq.(8.12) the repeated midpoints based on unitarily invariant random variables satisfy $U * M_{j,k}^W \stackrel{d}{=} M_{j,k}^W$ for each j, k and $U \in \mathcal{U}_d$. It follows that the predicted midpoints $\widetilde{M}_{j,2k+1}^W$ are unitarily invariant as well, as they can be expressed as weighted intrinsic averages of the midpoints $(M_{j-1,k}^W)_k$, which are unitarily invariant themselves. That is, $U * \widetilde{M}_{j,2k+1}^W \stackrel{d}{=} \widetilde{M}_{j,2k+1}^W$ for each j, k and $U \in \mathcal{U}_d$. Combining the above results, it follows that the random whitened coefficient $\mathfrak{D}_{j,k}^W$ is unitarily invariant, as for each $U \in \mathcal{U}_d$,

$$\begin{aligned} U * \mathfrak{D}_{j,k}^W &= U * \text{Log}\left((\widetilde{M}_{j,2k+1}^W)^{-1/2} * M_{j,2k+1}\right) \\ &= \text{Log}\left((U * \widetilde{M}_{j,2k+1}^W)^{-1/2} * (U * M_{j,2k+1})\right) \\ &\stackrel{d}{=} \text{Log}\left((\widetilde{M}_{j,2k+1}^W)^{-1/2} * M_{j,2k+1}\right) \\ &= \mathfrak{D}_{j,k}^W \end{aligned}$$

using that $U * \text{Log}(X) = \text{Log}(U * X)$ for $U \in \mathcal{U}_d$. By the same argument as in Theorem 5.1, if we write the eigendecomposition $\mathfrak{D}_{j,k}^W = Q * \Lambda$, then for a unitarily invariant random matrix $\mathfrak{D}_{j,k}^W$,

$$\begin{aligned} \mathbf{E}[\mathfrak{D}_{j,k}^W] &= \mathbf{E}\left[\sum_{i=1}^d \lambda_i \vec{q}_i \vec{q}_i^*\right] \\ &= \mathbf{E}[\vec{q}_i \vec{q}_i^*] \mathbf{E}[\text{Tr}(\Lambda)] \\ &= \mathbf{E}[\vec{q}_i \vec{q}_i^*] \mathbf{E}[\text{Tr}(\mathfrak{D}_{j,k}^W)] = \mathbf{0} \end{aligned}$$

Here we used that $\text{Tr}(Q * \Lambda) = \text{Tr}(\Lambda)$, since Q is a unitary matrix ($\mathfrak{D}_{j,k}^W$ is Hermitian), combined with the result $\mathbf{E}[\text{Tr}(\mathfrak{D}_{j,k}^W)] = 0$ in Proposition 5.4. \square