SIAM J. MATRIX ANAL. APPL.
Vol. 42, No. 2, pp. 528–556

GEODESICALLY PARAMETERIZED COVARIANCE ESTIMATION*

ANTONI MUSOLAS†, STEVEN T. SMITH‡, AND YOUSSEF MARZOUK†

Abstract. Statistical modeling of spatiotemporal phenomena often requires selecting a covariance matrix from a covariance class. Yet standard parametric covariance families can be insufficiently flexible for practical applications, while nonparametric approaches may not easily allow certain kinds of prior knowledge to be incorporated. We propose instead to build covariance families out of geodesic curves. These covariances offer more flexibility for problem-specific tailoring than classical parametric families and are preferable to simple convex combinations. Once the covariance family has been chosen, one typically needs to select a representative member by solving an optimization problem, e.g., by maximizing the likelihood of a data set. We consider instead a differential geometric interpretation of this problem: minimizing the geodesic distance to a sample covariance matrix ("natural projection"). Our approach is consistent with the notion of distance employed to build the covariance family and does not require assuming a particular probability distribution for the data. We show that natural projection and maximum likelihood estimation within the covariance family are locally equivalent up to second order. We also demonstrate that natural projection may yield more accurate estimates with noise-corrupted data.

Key words. covariance estimation, geodesic, symmetric positive-definite matrix manifold, natural metric, maximum likelihood, optimization on manifolds, Fisher information, denoising

AMS subject classifications. 53C22, 62J10

DOI. 10.1137/19M1284646

1. Introduction.

Statistical modeling of spatiotemporal phenomena often requires employing and estimating covariance matrices. Classical parametric covariance families (e.g., based on Matérn [13, 40] kernels) can be insufficiently flexible for practical applications. By construction, these approaches describe a high-dimensional object (a symmetric positive semidefinite matrix with $O(n^2)$ degrees of freedom) using only a few generic parameters that are not problem-specific; more broadly, these parametric families may not be rich enough to capture the phenomena of interest. Nonparametric methods (e.g., sparse precision matrix estimation [11, 20, 12], tapering [24, 21], diagonal loading [42], and shrinkage [26, 27, 41]) can be much more flexible. However, neither approach easily allows prior knowledge—for instance, known covariance matrices at related conditions—to be incorporated. Estimation in both settings often involves solving an optimization problem, such as maximizing the likelihood of a data set. Defining this objective function requires prescribing a specific probability distribution for the data, which may not be readily available. Moreover, maximum likelihood is not linked to the natural distance on the manifold of symmetric positive matrices. Also, as we shall show later, the resulting estimates can be sensitive to noise.

*Received by the editors January 7, 2020; accepted for publication (in revised form) by J. Chung December 21, 2020; published electronically April 8, 2021.

https://doi.org/10.1137/19M1284646

Funding: This work was supported by the "la Caixa" Banking Foundation (ID 100010434) under project LCF/BQ/AN13/10280009 and the Air Force Office of Scientific Research Computational Mathematics Program. This material is based upon work supported by the Under Secretary of Defense for Research and Engineering under Air Force contract FA8702-15-D-0001. Any opinions, findings, conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the Under Secretary of Defense for Research and Engineering.

†Center for Computational Science and Engineering, Massachusetts Institute of Technology, Cambridge, MA 02139 USA (musolas@mit.edu, ymarz@mit.edu).

‡MIT Lincoln Laboratory, Lexington, MA 02420 USA (stsmith@ll.mit.edu).

To overcome these obstacles, we propose to build covariance families by connecting *representative* covariance matrices (called "anchors") through geodesics. The resulting covariance classes can thus be tailored to the problem of interest. Second, as an alternative to selecting the most representative parameter by maximizing the likelihood, we advocate for a differential geometric approach to estimation within the family: minimizing the geodesic distance to a sample covariance matrix. Our approach, which we call *natural projection*, is consistent with the notion of distance employed to build the covariance family and does not require assuming a particular probability distribution for the data. Later, in a case study involving observations of a groundwater flow model, we will show that natural projection can outperform maximum likelihood estimation in the presence of noise.

Geodesic interpolation, smoothing, and regression of matrices and related objects have been active topics of research. Several papers [6, 5, 7] use interpolation on matrix manifolds to adapt and construct reduced-order models, while others [37, 29] propose a Riemannian framework for the interpolation and regularization of tensor fields, with broad applications in imaging. Local polynomial regression in the manifold of symmetric positive-definite matrices has been explored in [47], also in the context of computer vision and medical imaging. Other authors have pursued higher-order interpolation of positive-definite [1, 23] and semidefinite [22] matrices using Bézier curves. Additionally, the authors of [34, 35] characterize the Riemannian mean of positive-definite matrices and propose a multivariate geodesic interpolation scheme for such matrices using weights. Our work also relies on geodesic interpolation, but for the purpose of building covariance families. The idea of differential geometric methods for covariance *estimation* has links to the broad field of information geometry [4], which constructs manifolds of probability distributions and analyzes their geometric properties. In this general setting, the likelihood function has been characterized as a notion of distance in [3, 2]. Matrix nearness using Bregman divergences, and its geometric interpretations, have been discussed thoroughly in [15].

As described above, our covariance families will follow from geodesic interpolation of a given set of anchor matrices. The anchors should be representative of known problem instances—e.g., empirical observations or computational simulations of the relevant spatiotemporal process at related conditions. Combining these instances into a parametric family constitutes a hybrid approach to covariance modeling that can yield much richer and more problem-specific covariances than standard kernels. Using geodesics ensures that the entire covariance family lies in the manifold of symmetric positive-definite matrices. These families can also be interpretable: the internal parameters may serve as explicative variables for the problem of interest. Another advantage of this approach is that it harnesses the asymmetry of information between online and offline stages of a problem; that is, each anchor covariance matrix can be computed offline to a desired accuracy and later used for online estimation with limited data.

Having constructed a geodesically parameterized covariance family, we also analyze different alternatives for estimation within the family—i.e., given a data set, identifying the most "representative" member. This problem is usually solved by assigning a probability distribution to the data and selecting the parameter values that maximize the resulting likelihood function. Under some conditions, this process is equivalent to minimizing a particular direction of the Kullback–Leibler (KL) divergence, known as reverse information-projection (reverse I-projection) [14]; this is further equivalent to minimizing Stein's loss [44, 45]. Alternatively, one might minimize the opposite direction of KL divergence; this choice is known as I-projection.

Consistent with the construction of the family, we instead propose to use natural projection: selecting the covariance matrix within the family that minimizes the geodesic distance to the sample covariance matrix. We will show that the other methods are essentially linear approximations of natural projection. In particular, we will show that the estimates produced by natural projection and the two forms of I-projection are locally equivalent up to second order. In contrast with other methods, however, the optimality condition for natural projection is equivalent to an orthogonality condition on the tangent space of the matrix manifold. Since natural projection does not require modeling the distribution of the data, it may be easier to apply in practice. We also show that it can yield reduced sensitivity to noise.

Performing natural projection requires that the sample covariance lie in the manifold of symmetric positive matrices, and thus that the size of the sample q generally be greater than the dimension n of the parameters. Even though the sample covariance is itself a consistent estimator of the population covariance as $q/n \to \infty$, it can be improved upon significantly for finite q/n (say $q/n < O(100)$) [25, 21, 16, 43, 32]. This also defines the regime of applicability for many of the estimation (and regularization) methods we propose in this paper. But we also emphasize that the construction of geodesically parameterized covariance families, apart from natural projection, is of independent interest, and that one can perform maximum likelihood estimation within these parametric families for $q < n$. Our goal, also, is not to create a consistent estimator of generic population covariances, but rather to create useful and expressive *parametric models* for covariance matrices and to study estimation *within* these models in the appropriate sample size regime.

To summarize, the original contributions of this paper are (1) to devise a general framework for problem-specific geodesically parameterized covariance families; (2) to propose natural projection as an alternative means of estimation within a covariance family; (3) to analyze the differences between natural projection and other standard estimation techniques; and (4) to demonstrate the advantages of geodesically parameterized families and natural projection in a case study.

The ability to find the closest member of a covariance family has several further applications. First, consider denoising or regularization: if a covariance matrix is well approximated by a given family, one can project it to the family to reduce sampling noise. Second, consider efficient storage using a geodesic basis: given a covariance family and a set of related matrices, one could store only the values of the parameters of the closest matrices in the family. As a consequence, storage is reduced only to the optimal parameters and the anchor matrices.

The plan of the paper is as follows. In section 2, we review the geometry of the manifold of symmetric positive-definite matrices, define the notion of a geodesic covariance family, and introduce natural projection alongside some standard alternatives. Section 3 discusses properties of the optimization problems associated with each of these estimation methods. In section 4, we perform local analyses that compare natural projection with existing alternatives. Section 5 extends the geodesic covariance family construction to general multiparameter settings. In section 6, we demonstrate the performance of geodesic covariance families and natural projection in a case study: characterizing the spatial variations of hydraulic head in an aquifer. Conclusions follow in section 7.

2. Tools for covariance estimation on a geodesic family. In subsection 2.1, we recall some results on the geometry of the symmetric positive-definite cone. In subsection 2.2, we introduce the idea of a geodesic covariance family. In subsection 2.3, we present the loss functions we will consider for estimation. In subsection 2.4, we

introduce natural projection and contrast it with canonical approaches to parametric covariance estimation.

2.1. The geometry of the symmetric positive-definite cone. Let $\mathbf{S}(n)$ be the space of $n \times n$ symmetric matrices, and let $\mathbf{S}_+(n)$ denote the manifold of symmetric positive-definite $n \times n$ matrices. This manifold has been studied extensively in the literature (see, e.g., [9, 18, 43, 17]).

Let $X_A, Y_A \in \mathbf{S}(n)$ be tangent vectors to $\mathbf{S}_+(n)$ at A: $T_A\,\mathbf{S}_+(n)$. The natural metric g_A is defined as the inner product in the tangent space at A:

$$g_A(X_A, Y_A) = \operatorname{tr}(X_A A^{-1} Y_A A^{-1}).$$

We will denote the tangent vector X_A simply as $\underline{X}$ when there is no ambiguity in the choice of tangent space.

The exponential map transports an object in the tangent space to its corresponding element on the manifold and is defined as

$$B = \exp_A(\underline{B}) = A^{\frac{1}{2}} \exp(A^{-\frac{1}{2}} \underline{B} A^{-\frac{1}{2}}) A^{\frac{1}{2}},$$

where $A^{\frac{1}{2}}$ is the symmetric square root. Conversely, the logarithm map transports objects from the manifold to the tangent space:

$$\underline{B} = \log_A(B) = A^{\frac{1}{2}} \log_m(A^{-\frac{1}{2}} B A^{-\frac{1}{2}}) A^{\frac{1}{2}}, \quad A^{-\frac{1}{2}} \underline{B} A^{-\frac{1}{2}} = \log_m(A^{-\frac{1}{2}} B A^{-\frac{1}{2}}).$$

Let A_1 and A_2 belong to this manifold. Associated with the natural metric, there exists a natural distance $d(A_1, A_2)$ that is invariant with respect to matrix inversion,

$$d(A_1, A_2) = d(A_1^{-1}, A_2^{-1}), \tag{2.1}$$

and with respect to congruence via any invertible matrix Z:

$$d(A_1, A_2) = d(Z A_1 Z^\top, Z A_2 Z^\top). \tag{2.2}$$

Moreover, a parameterization of the geodesic, which at any point minimizes the natural distance to A_1 and A_2, is given by

$$\varphi_{A_1 \to A_2}(t) = A_1^{\frac{1}{2}} \exp_m\big(t \log_m(A_1^{-\frac{1}{2}} A_2 A_1^{-\frac{1}{2}})\big) A_1^{\frac{1}{2}},$$

where $\varphi_{A_1 \to A_2}(t) \in \mathbf{S}_+(n)$ for all $t \in \mathbb{R}$. Clearly, $A_1^{-\frac{1}{2}} A_2 A_1^{-\frac{1}{2}}$ admits an orthogonal eigendecomposition of the form $U \Lambda U^\top$. Notice that Λ contains the generalized eigenvalues of the pencil (A_2, A_1), which we denote as $\lambda_k^{(A_2, A_1)}$, $k = 1, \ldots, n$. Therefore, $\varphi_{A_1 \to A_2}(t)$ can be expressed as

$$\varphi_{A_1 \to A_2}(t) = A_1^{\frac{1}{2}} \exp_m\big(t \log_m(U \Lambda U^\top)\big) A_1^{\frac{1}{2}} = A_1^{\frac{1}{2}} U \Lambda^t U^\top A_1^{\frac{1}{2}}. \tag{2.3}$$

Notice that we recover the trivial cases $\varphi_{A_1 \to A_2}(t = 0) = A_1$ and $\varphi_{A_1 \to A_2}(t = 1) = A_2$; we call A_1 and A_2 the *anchor* matrices. The geodesic can also be expressed as

$$\varphi_{A_1 \to A_2}(t) = A_1 (A_1^{-1} A_2)^{t_1} = A_2 (A_2^{-1} A_1)^{1 - t_1}.$$

Additionally, there is a closed-form expression for the natural distance between any two matrices in $\mathbf{S}_+(n)$:

$$d(A_1, A_2) = d(A_1^{-\frac{1}{2}} A_2 A_1^{-\frac{1}{2}}, I) = \| \log_m(A_1^{-\frac{1}{2}} A_2 A_1^{-\frac{1}{2}}) \|_F = \sqrt{\sum_{k=1}^{n} \log^2 \lambda_k^{(A_2, A_1)}}. \tag{2.4}$$

From the above, we have

$$d\big(A_1, \varphi_{A_1 \to A_2}(t)\big) = |t| d(A_1, A_2). \tag{2.5}$$

Equation (2.4) appears extensively in the literature of differential geometry and inference. Up to a constant, it is known as the Fisher information metric or as Rao's distance [8, 39, 38]. When measuring distance between covariance matrices, it is also known as the Förstner metric [19].

Unlike the manifold of symmetric positive-definite matrices, the manifold of symmetric positive *semidefinite* matrices does not enjoy as much structure. As shown clearly in [10] and discussed in [46], the main drawback is that there is no notion of distance that satisfies the invariance properties in (2.1) and (2.2).

2.2. Definition of the geodesic covariance family. As described in section 1, the ability to create tailored parametric covariance *functions* out of two (or more) covariance matrices of interest has several potential benefits. First, via the choice of anchor matrices, the covariance function can be made representative of the particular problem at hand; second, the parameters of the covariance function can become meaningful explicatory variables of the spatiotemporal process.

To begin, we define the notion of a one-parameter covariance family. We will generalize this notion to the multiparameter case in section 5.

DEFINITION 2.1 (covariance function and family). *A one-parameter covariance function is a map $\varphi : \mathbb{R} \to \mathbf{S}_+(n)$; its corresponding covariance family is the image of φ.*

The covariance family structure we will employ in sections 2 to 4 is a geodesic between anchor matrices. Let A_1 and A_2 be two elements in $\mathbf{S}_+(n)$. Then $\varphi_{A_1 \to A_2}(t)$ as defined in (2.3) is a one-parameter covariance function, whose image is a covariance family. This covariance function immediately satisfies the following two properties:

1. $\varphi^{-1}_{A_1 \to A_2}(t) = \varphi_{A_1^{-1} \to A_2^{-1}}(t)$,
2. $\varphi_{A_1 \to A_2}(t) = \varphi_{A_2 \to A_1}(1 - t)$.

Since the parameter $t \in \mathbb{R}$, the covariance family is an infinitely long curve on the manifold rather than a segment between the two anchor matrices. Notice that the first property allows one to work with precision matrices and still obtain the same results. The second guarantees invariance with respect to the order of the matrices. These properties make the geodesic a compelling covariance function to be used for practical applications.

It is often useful to be able to scale the family to adapt to changes of the magnitude of the problem. Within our covariance family framework, this extra degree of freedom is achieved by building a geodesic between any matrix and the same times a constant.

REMARK 2.2 (scaling of a covariance function). *Let A_1 be an element in $\mathbf{S}_+(n)$, and let $\alpha \in \mathbb{R}^+$. Notice that $\varphi_{A_1 \to \alpha A_1}(t)$ as defined in* (2.3) *is a one-parameter covariance function of the form $\varphi_{A_1 \to \alpha A_1}(t) = \alpha^t A_1$. If A_1 is replaced by a one-parameter covariance function, the scaling applies to the whole family.*

The scaling factor α^t in Remark 2.2 is positive for any t in the real line. Clearly, if the scaling is applied to a one-parameter covariance function, we are left with two degrees of freedom: one that moves across the anchor matrices and another that controls the magnitude of the entries.

2.3. Spectral functions. In section 3, we will be interested in selecting the "most representative" member of a covariance family given a data set. Doing so will

entail minimizing certain loss functions: distances or divergences between distributions. All the loss functions we employ are *spectral functions*.

DEFINITION 2.3 (spectral function). *Let A_1 be a matrix in $\mathbf{S}_+(n)$. A function $F(A_1)$ is a* spectral function *if it is a differentiable and symmetric map from the eigenvalues of A_1 to the reals. The function F can be understood as a composition of the eigenvalue function λ and a differentiable and symmetric map f; that is, $F(A_1) = f \circ \lambda(A_1)$.*

Closed-form expressions for some spectral functions that we shall use later are presented below.

REMARK 2.4. *The following notions of distance or divergence can be expressed as functions of the generalized eigenvalues $(\lambda_k)_{k=1}^n$ of the pencil (A_2, A_1):*

- *Natural distance in $\mathbf{S}_+(n)$:*

$$d(A_1, A_2) = \sqrt{\sum_{k=1}^{n} \log^2 \lambda_k}. \tag{2.6}$$

- *KL divergence between multivariate normals:*

$$D_{KL}\big(N(0, A_1) \,\big\|\, N(0, A_2)\big) = \sum_{k=1}^{n} \frac{\lambda_k^{-1} + \log \lambda_k - 1}{2}. \tag{2.7}$$

- *KL divergence between multivariate normals, swapping the order:*

$$D_{KL}\big(N(0, A_2) \,\big\|\, N(0, A_1)\big) = \sum_{k=1}^{n} \frac{\lambda_k - \log \lambda_k - 1}{2}. \tag{2.8}$$

2.4. Covariance estimation in a geodesic family. Now we define several alternative optimization problems that each describe estimation in a geodesic covariance family. We will formally contrast these problems in the next sections. Figure 1 illustrates the covariance function as a geodesic from A_1 and A_2 on the manifold of symmetric positive-definite matrices, along with multiple projections (i.e., estimates within the family) of the sample covariance matrix $\widehat{C}$.

Let $y_1, \ldots, y_q$ be independent and identically distributed observations from some distribution with density function $p_Y\big(\cdot\,; \varphi_{A_1 \to A_2}(t)\big)$, parameterized by φ. In general, the maximum likelihood estimate of t is then

$$t^{\mathrm{ML}} \in \operatorname*{argmax}_{t \in (-\infty, \infty)} \sum_{i=1}^{q} \log p_Y(y_i; \varphi_{A_1 \to A_2}(t)). \tag{2.9}$$

For comparison with other techniques, we consider the specific case where p_Y is multivariate Gaussian and, without loss of generality, zero mean. In this case, the maximum likelihood estimate is as follows.

PROBLEM 2.5 (maximum likelihood with a Gaussian model).

$$y_i \overset{iid}{\sim} N\big(0, \varphi_{A_1 \to A_2}(t)\big),$$

$$\hat{t} \in \operatorname*{argmax}_{t \in (-\infty, \infty)} \sum_{i=1}^{q} \log N\big(y_i; 0, \varphi_{A_1 \to A_2}(t)\big).$$

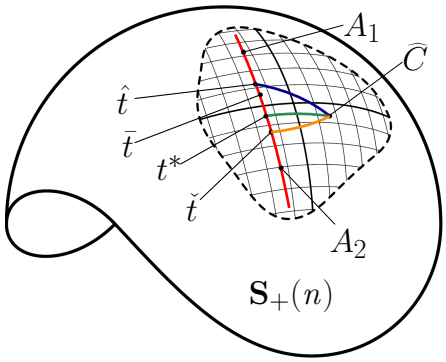

FIG. 1. *Representation of the geodesic between anchors A_1 and A_2 and projection of the sample covariance matrix $\widehat{C}$ as solution of Problems* 2.5 *to* 2.8. *We denote $\hat{t}$ as the value of t that maximizes the likelihood (reverse I-projection), $\check{t}$ the I-projection, and t^* the natural projection (cf. Lemma* 3.2*).*

Note that (2.9) and Problem 2.5 can in general be solved for $q < n$. In other words, maximum likelihood estimation within the geodesic covariance family, under some distributional assumption on the data, does not require a full rank sample covariance $\widehat{C} \coloneqq \frac{1}{q-1}\sum_{i=1}^{q}(y_i - \bar{y})(y_i - \bar{y})^\top$.

Now, suppose that the sample covariance matrix $\widehat{C}$ of zero-mean Gaussian data $(y_i)_{i=1}^q$ is full rank, which is typically the case when $q \geq n$. In this case, the solution to Problem 2.5 is equivalent to that obtained by reverse I-projection, which consists in minimizing the KL divergence as follows. This equivalence is shown in Lemma A.2.

PROBLEM 2.6 (reverse I-projection).

$$\hat{t} \in \operatorname*{argmin}_{t\in(-\infty,\infty)} D_{KL}\Big(N(0,\widehat{C}) \,\Big\|\, N\big(0,\varphi_{A_1\to A_2}(t)\big)\Big). \tag{2.10}$$

Since Problems 2.5 and 2.6 are equivalent, we will only refer to Problem 2.6 going forward. We will also consider *I-projection*, which minimizes the KL divergence but in the opposite order.

PROBLEM 2.7 (I-projection).

$$\check{t} \in \operatorname*{argmin}_{t\in(-\infty,\infty)} D_{KL}\Big(N\big(0,\varphi_{A_1\to A_2}(t)\big) \,\Big\|\, N(0,\widehat{C})\Big).$$

Note that the KL divergence is proportional to Stein's loss [28]. Therefore, Problems 2.6 and 2.7 can also be cast as minimizing Stein's loss with the appropriate order of the arguments.

Finally, we explore the possibility of covariance estimation by minimizing the geodesic distance $d(\cdot,\cdot)$ defined in (2.4).

PROBLEM 2.8 (natural projection).

$$t^* \in \operatorname*{argmin}_{t\in(-\infty,\infty)} d\big(\varphi_{A_1\to A_2}(t), \widehat{C}\big).$$

2.5. Choosing the anchor matrices. Before proceeding with further analysis, we offer a few comments on the choice of the anchor covariance matrices A_1 and A_2

defining the parametric covariance family. (These comments are equally applicable to the multiparameter case of section 5, with more than two anchor matrices.) In general, we view the choice of anchors as a *modeling* issue, tied to the purpose of the family and the availability of relevant information. Yet it can be shaped by the following principles. Clearly, anchors should be chosen such that they are representative of the spatiotemporal processes that the covariance family will use to describe. How can this be done? In general, we suggest that the anchors should be chosen and connected in a way that reflects any *possible* latent or explanatory variables that describe the problem. For instance, in the example of section 6, we choose the anchors to span a range of parameters describing statistics of the input to a stochastic PDE. The resulting geodesic parameters (connecting covariances of the PDE *solution* field) do not necessarily correspond directly to the input's parameters, but they do encompass a continuum of values and thus relevant behaviors.

While in some problems the explanatory variables will be clear, there are other problems where some explanatory variables may not be apparent. In this case, a strategy would be to collect anchors that capture all "qualitatively different" patterns of covariance. The resulting covariance family will, by construction, be able to interpolate continuously between these different regimes. For example, in mathematical finance, modeling the covariance of multiple asset prices is essential to mitigating volatility through diversification [33]. To construct a useful model, one can connect multiple covariance matrices corresponding to different market conditions (e.g., time of the year, prevailing Federal Reserve interest rates) through geodesics to create a richer family of covariances suitable for any condition.

3. Properties of the covariance estimation problem, one-parameter case. In this section, we characterize various properties of Problems 2.6 to 2.8. Our main results are the optimality conditions and their corresponding geometric interpretations. These results are presented in Propositions 3.3 to 3.5, the proofs of which are deferred to Appendix A. First, however, we establish uniqueness of the optima and idempotence of the associated projections. Supporting results (Lemmas A.1 to A.4) and their proofs are also deferred to Appendix A.

LEMMA 3.1 (uniqueness of the solution). *Each of Problems* 2.6 *to* 2.8 *has a unique solution.*

Proof. Since the optimization problems are unconstrained, uniqueness follows immediately from the convexity of Problem 2.8 (shown in Lemma A.3) and of Problems 2.6 and 2.7 (shown in Lemma A.4). ∎

Since the solutions are unique, though in general distinct (see below), we will use $\hat{t}$ to denote the result of reverse I-projection, $\check{t}$ that of I-projection, and t^* that of natural projection. Uniqueness also allows defining the distance between the sample covariance matrix and the geodesic as $d\big(\varphi_{A_1\to A_2}(t^*), \widehat{C}\big)$.

If the sample covariance matrix already belongs to the covariance family, the most representative member of the family ought to be the sample covariance matrix itself. This result holds true for all three problems.

LEMMA 3.2 (idempotence of projections). *If $\widehat{C} \in \varphi_{A_1\to A_2}(t)$, then there exists a unique $\bar{t}$ such that $(\lambda_k^{(A_2,A_1)})^{\bar{t}} = \lambda_k^{(\widehat{C},A_1)}$, $k = 1,\dots,n$, where $\lambda_k^{(A_2,A_1)}$ and $\lambda_k^{(\widehat{C},A_1)}$ are the kth eigenvalues of the pencils (A_2, A_1) and $(\widehat{C}, A_1)$, respectively. Moreover, under this condition, $\bar{t} = t^* = \hat{t} = \check{t}$ (cf. Figure* 1*), $\widehat{C} = \varphi_{A_1\to A_2}(\bar{t})$ with*

$$\bar{t} = \frac{\sum_{k=1}^n \log \lambda_k^{\widehat{C}} - \sum_{k=1}^n \log \lambda_k^{A_1}}{\sum_{k=1}^n \log \lambda_k^{A_2} - \sum_{k=1}^n \log \lambda_k^{A_1}},$$

where $\lambda_k^{A_1}$, $\lambda_k^{A_2}$, and $\lambda_k^{\widehat{C}}$ are the kth eigenvalues of A_1, A_2, and $\widehat{C}$, respectively. This expression also holds when $A_1 = \alpha A_2$ for any $\widehat{C}$ and $\alpha > 0$.

Proof. If $\widehat{C} \in \varphi_{A_1 \to A_2}(t)$, then there exists a $\bar{t}$ such that $\widehat{C} = A_1^{\frac{1}{2}} U \Lambda^{\bar{t}} U^\top A_1^{\frac{1}{2}}$. Rearranging the terms, we obtain $A_1^{-\frac{1}{2}} \widehat{C} A_1^{-\frac{1}{2}} = U \Lambda^{\bar{t}} U^\top$, which is an eigendecomposition of $A_1^{-\frac{1}{2}} \widehat{C} A_1^{-\frac{1}{2}}$. This is equivalent to saying that $\Lambda^{\bar{t}}$ contains the $\lambda_k^{(\widehat{C}, A_1)}$. But also, from our notation in (2.3) we knew that Λ contains the $\lambda_k^{(A_2, A_1)}$.

Notice that $\bar{t}$ satisfies the general form (cf. (A.2)) of distance minimization, and similarly for the reverse I-projection (cf. (A.3)) and the I-projection (cf. (A.4)). Using uniqueness in Lemma 3.1, we conclude that $\bar{t} = t^* = \hat{t} = \check{t}$. The last $\widehat{C} = \varphi_{A_1 \to A_2}(\bar{t})$ follows by the definition of $\bar{t}$. For the closed-form solution, set $\widehat{C} = \varphi_{A_1 \to A_2}(t^*) = A_1^{\frac{1}{2}} U \Lambda^{t^*} U^\top A_1^{\frac{1}{2}} = A_1^{\frac{1}{2}} (A_1^{-\frac{1}{2}} A_2 A_1^{-\frac{1}{2}})^{t^*} A^{\frac{1}{2}}$. Now, take the determinant of both sides and apply its properties to have $\det(\widehat{C}) = \det(A_1) \det(A_1^{-\frac{1}{2}} A_2 A_1^{-\frac{1}{2}})^{t^*}$. Applying logarithms on both sides, we obtain the desired result. Clearly, it solves Problem 2.8 since distance in this case is zero. Finally, if $A_1 = \alpha A_2$, then $\Lambda = \alpha I$ and we note that the general expression simplifies to $\operatorname{tr}\big(\log_m(\Lambda^{t^*} \widehat{C}^{-\frac{1}{2}} A_1 \widehat{C}^{-\frac{1}{2}})\big) = 0$. Using Jacobi's formula, it simplifies further to $\det(\Lambda^{t^*} \widehat{C}^{-\frac{1}{2}} A_1 \widehat{C}^{-\frac{1}{2}}) = 1$ and we come back to the same above expression with determinants. ∎

The following propositions characterize the optimal solutions of Problems 2.6 to 2.8.

PROPOSITION 3.3 (natural projection for covariance estimation). *The optimal parameter t^* satisfies*

$$\operatorname{tr}\big(\log_m(Z \Lambda^{-t^*}) \log_m(\Lambda)\big) = 0, \tag{3.1}$$

where $Z = U^\top A_1^{-\frac{1}{2}} \widehat{C} A_1^{-\frac{1}{2}} U$. This optimality equation can also be rewritten as an orthogonality condition on the tangent space:

$$g_{R_{A_1 \to A_2}(t^*)}\big(\underline{A_1^{-\frac{1}{2}} \widehat{C} A_1^{-\frac{1}{2}}}, \underline{R_{A_1 \to A_2}(1 + t^*)}\big) = 0, \tag{3.2}$$

where $R_{A_1 \to A_2}(t) = (A_1^{-\frac{1}{2}} A_2 A_1^{-\frac{1}{2}})^t$ is the whitened geodesic $A_1^{-\frac{1}{2}} \varphi_{A_1 \to A_2} A_1^{-\frac{1}{2}}$.

The natural projection consists in minimizing the natural distance to a certain matrix over a curve. This is similar to what one would do in an Euclidean space, but on a manifold. On the tangent space, this operation looks like finding a point at which the projected geodesic is orthogonal to the direction of the outside matrix (3.2). Figure 2 illustrates this relationship.

As mentioned in section 2, the Fisher information metric between two normal distributions with known mean is proportional to the natural distance, so the former is also minimized when the latter is. Therefore, the aforementioned t^* minimizes the Fisher information metric between $N\big(0, \varphi_{A_1 \to A_2}(t)\big)$ and $N(0, \widehat{C})$.

PROPOSITION 3.4 (reverse I-projection for covariance estimation). *The optimal parameter $\hat{t}$ satisfies*

$$\operatorname{tr}\big((Z \Lambda^{-\hat{t}} - \mathrm{Id}) \log_m(\Lambda)\big) = 0, \tag{3.3}$$

where $Z = U^\top A_1^{-\frac{1}{2}} \widehat{C} A_1^{-\frac{1}{2}} U$. This optimality equation can also be rewritten as an orthogonality condition:

$$g_{R_{A_1 \to A_2}(\hat{t})}\Big(\underline{\exp_{R_{A_1 \to A_2}(\hat{t})}\big(A_1^{-\frac{1}{2}} \widehat{C} A_1^{-\frac{1}{2}} - R_{A_1 \to A_2}(\hat{t})\big)}, \underline{R_{A_1 \to A_2}(1 + \hat{t})}\Big) = 0. \tag{3.4}$$

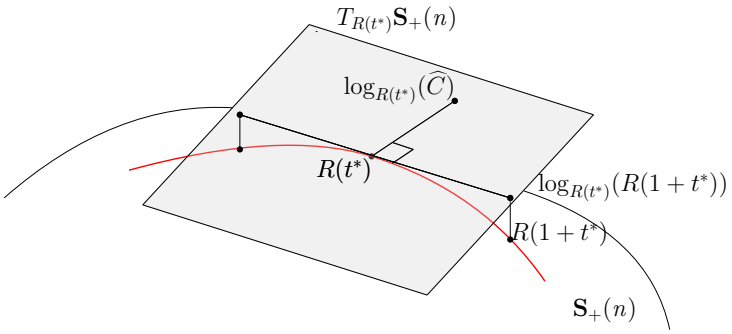

FIG. 2. *Illustration of the orthogonality condition in Proposition* 3.3. *Solving Problem* 2.8 *is equivalent to finding a* t^* *such that* $\widehat{C}$ *is orthogonal to the unitary vector* $R(1+t^*)$.

Notice that the matrix $A_1^{-\frac{1}{2}}\widehat{C}A_1^{-\frac{1}{2}} - R_{A_1\to A_2}(\hat{t})$ lives in a Euclidean space and not necessarily in $\mathbf{S}_+(n)$. Indeed, such a subtraction is the usual way of obtaining a vector between two points in a "flat" space, but it does not necessarily preserve positive definiteness. Intuitively, the likelihood (which yields the first argument of (3.4)) seems to be a flat notion, whereas the geodesic (yielding the second argument) is in the manifold; thus one could argue that reverse I-projection is actually inconsistent. Instead, one should either maximize the likelihood over a family produced by a convex combination of anchors (i.e., both the family and the divergence we are minimizing are in a flat space) or minimize the natural distance over a proper geodesic (as in Problem 2.8). Indeed, the orthogonality condition for natural projection in Proposition 3.3 is far more direct.

We can develop similar results for Problem 2.7.

PROPOSITION 3.5 (I-projection for covariance estimation). *The optimal parameter* $\check{t}$ *satisfies*

$$\operatorname{tr}\big((\Lambda^{\check{t}}Z^{-1} - \mathrm{Id})\log_m(\Lambda)\big) = 0, \tag{3.5}$$

where $Z = U^\top A_1^{-\frac{1}{2}}\widehat{C}A_1^{-\frac{1}{2}}U$. *This optimality equation can also be rewritten as an orthogonality condition:*

$$g_{R_{A_1\to A_2}(-\check{t})}\Big(\underline{\exp_{R_{A_1\to A_2}(-\check{t})}\big(A_1^{\frac{1}{2}}\widehat{C}^{-1}A_1^{\frac{1}{2}} - R_{A_1\to A_2}(-\check{t})\big)}, \underline{R_{A_1\to A_2}(1-\check{t})}\Big) = 0. \tag{3.6}$$

Notice that (3.3) is the first-order Taylor expansion of (3.1) around the identity matrix if we use $Z\Lambda^{-t}$ as a variable. In this sense, maximizing the likelihood (reverse I-projection) corresponds to solving a linearized version of the natural distance minimization problem. The same can be observed with the I-projection if we Taylor expand in $\Lambda^t Z^{-1}$ (cf. (3.1) and (3.5)). It suffices to adapt (3.1) using $\log_m(Z\Lambda^{-t}) = -\log_m(\Lambda^t Z^{-1})$.

4. Local analysis and comparison of the projections. We now compare the solutions of Problems 2.6 to 2.8 when $\widehat{C}$ is very close to the geodesic (in terms of natural distance).

LEMMA 4.1 (equality of the limit). *Let* A_1, A_2, *and* C *be matrices in* $\mathbf{S}_+(n)$. *Assume that* $d\big(\varphi_{A_1\to A_2}(t), C\big) = \epsilon$, $\epsilon > 0$. *In the limit of* $\epsilon \to 0$, *Problems* 2.6 *to* 2.8 *have the same solution.*

Proof. Let t^* be the minimizer of $d\big(\varphi_{A_1\to A_2}(t), C\big)$, and let $A^* = \varphi_{A_1\to A_2}(t^*)$. Without loss of generality, define $\widehat{C}$ such that $d(A^*, \widehat{C}) = 1$ and $\varphi_{A^*\to\widehat{C}}(\epsilon) = C$. By the properties of the natural distance, we know that $d(A^*, C) = \epsilon$. Define $Z_\epsilon = U^\top A_1^{-\frac{1}{2}} \varphi_{A^*\to\widehat{C}}(\epsilon) A_1^{-\frac{1}{2}} U$, and notice that

$$Z = \lim_{\epsilon\to 0} Z_\epsilon = \lim_{\epsilon\to 0} U^\top A_1^{-\frac{1}{2}} \varphi_{A^*\to\widehat{C}}(\epsilon) A_1^{-\frac{1}{2}} U = U^\top A_1^{-\frac{1}{2}} A^* A_1^{-\frac{1}{2}} U.$$

By definition, $A^* = \varphi_{A_1\to A_2}(t^*) = A_1^{\frac{1}{2}} U \Lambda^{t^*} U^\top A_1^{\frac{1}{2}}$, and thus $Z = \Lambda^{t^*}$. The proof is concluded after realizing that, with this value of Z, (3.3) and (3.5) hold. ∎

Together with idempotence of the projection (Lemma 3.2), Lemma 4.1 implies continuity of $\hat{\Delta}t := \hat{t} - t^*$ at $\epsilon = 0$. Indeed, idempotence means pointwise equivalence at $\epsilon = 0$, and at the limit $\epsilon \to 0$, we also see $\hat{\Delta}t = 0$. Thus, $\hat{\Delta}t$ is continuous at that point. The same is also true for $\check{\Delta}t := \check{t} - t^*$.

THEOREM 4.2 (natural projection versus I-projection and reverse I-projection). *Let A_1, A_2, $C \in \mathbf{S}_+(n)$, and without loss of generality suppose that A_1 is the matrix in $\varphi_{A_1\to A_2}$ that minimizes the distance $d\big(\varphi_{A_1\to A_2}(t), C\big)$. The difference between the solutions of Problems* 2.6 *and* 2.8 *as a function of $\epsilon = d\big(\varphi_{A_1\to A_2}(t), C\big)$ is $\hat{\Delta}t(\epsilon)$, defined implicitly as*

$$\operatorname{tr}\left(\left(U^\top V \Sigma^\epsilon V^\top U \Lambda^{\hat{\Delta}t(\epsilon)} - \mathrm{Id}\right)\log_m(\Lambda)\right) = 0, \tag{4.1}$$

where $V\Sigma V^\top = A_1^{-\frac{1}{2}} C A_1^{-\frac{1}{2}}$ and $U\Lambda U^\top = A_1^{-\frac{1}{2}} A_2 A_1^{-\frac{1}{2}}$ are orthogonal eigendecompositions.

Similarly, the difference in the solutions of Problems 2.7 *and* 2.8 *as a function of ϵ is $\check{\Delta}t(\epsilon)$, defined implicitly as*

$$\operatorname{tr}\left(\left(\Lambda^{-\check{\Delta}t(\epsilon)} U^\top V \Sigma^{-\epsilon} V^\top U - \mathrm{Id}\right)\log_m(\Lambda)\right) = 0. \tag{4.2}$$

Moreover, the functions $\hat{\Delta}t(\epsilon)$ and $\check{\Delta}t(\epsilon)$ are continuous at $\epsilon = 0$, and

$$\hat{\Delta}t'(0) = \check{\Delta}t'(0) = -\frac{g_{A_1}(\underline{C}, \underline{A_2})}{d(A_1, A_2)} = 0. \tag{4.3}$$

Proof. Refer to the construction of the proof in Lemma 4.1, and notice that $A^* = A_1$, $t^* = 0$ for any ϵ, and

$$Z_\epsilon = U^\top A_1^{-\frac{1}{2}} \varphi_{A^*\to\widehat{C}}(\epsilon) A_1^{-\frac{1}{2}} U = U^\top V \Sigma^\epsilon V^\top U.$$

Then (4.1) follows immediately from (3.3), and (4.2) follows from (3.5). Continuity at $\epsilon = 0$ follows from Lemmas 3.2 and 4.1. Now, we can take derivatives of (4.1) and obtain the following:

$$\operatorname{tr}\left(\left(U^\top V \Sigma^\epsilon \log_m(\Sigma) V^\top U \Lambda^{\hat{\Delta}t(\epsilon)} + U^\top V \Sigma^\epsilon V^\top U \log_m(\Lambda) \Lambda^{\hat{\Delta}t(\epsilon)} \hat{\Delta}t'(\epsilon)\right)\log_m(\Lambda)\right) = 0, \tag{4.4}$$

which evaluated at $\epsilon = 0$ results in

$$\hat{\Delta}t'(0) = -\frac{\operatorname{tr}\left(\left(U^\top V \log_m(\Sigma) V^\top U\right)\log_m(\Lambda)\right)}{\operatorname{tr}\left(\log_m^2(\Lambda)\right)}, \tag{4.5}$$

which can be rewritten as

$$\hat{\Delta}t'(0) = -\frac{\operatorname{tr}\left(\log_m(A_1^{-\frac{1}{2}} C A_1^{-\frac{1}{2}}) \log_m(A_1^{-\frac{1}{2}} A_2 A_1^{-\frac{1}{2}})\right)}{d(A_1, A_2)} = 0.$$

Notice that $\hat{\Delta}t'(0)$ vanishes since the numerator is (3.2). An analogous derivation for $\check{\Delta}t'(0)$ provides the same result. ∎

From (4.3), note that $\hat{\Delta}t'(0)$ can be understood as the inner product of the tangent vectors at A_1 pointing to C and A_2, normalized by the distance from the reference point to the latter matrix. The expression is analogous to the classic form of the inner product as a product of the modulus and the angle. In our setting, the modulus is the $d(A_1, A_2)$ and the angle is $\hat{\Delta}t'(0)$.

Since $\hat{\Delta}t(0) = 0$ and $\hat{\Delta}t'(0) = 0$, a second-order Taylor series expansion around $\epsilon = 0$ would be

$$\hat{\Delta}t(\epsilon) = \frac{\hat{\Delta}t''(0)}{2}\epsilon^2 + \mathcal{O}(\epsilon^3),$$

and the same expansion applies for $\check{\Delta}t$.

Finally, we compare I-projection and reverse I-projection, summarizing the results below.

THEOREM 4.3 (I-projection versus reverse I-projection). *Refer to the notation in Theorem* 4.2. *The ith derivatives of* $\hat{\Delta}t$ *and* $\check{\Delta}t$ *satisfy the following:*

$$\hat{\Delta}t^{(i)}(0) = \check{\Delta}t^{(i)}(0),\ i = 1, 3, 5, \ldots;$$
$$\hat{\Delta}t^{(i)}(0) = -\check{\Delta}t^{(i)}(0),\ i = 0, 2, 4, 6, \ldots.$$

Thus, the Taylor expansion of the difference in the solutions of Problems 2.6 *and* 2.7 *as a function of* $\epsilon = d(\varphi_{A_1 \to A_2}(t), C)$, *in a neighborhood of* $\epsilon = 0$, *always attains one additional order of accuracy. In particular,*

$$\hat{t}(\epsilon) - \check{t}(\epsilon) = \hat{\Delta}t''(0)\epsilon^2 + \mathcal{O}(\epsilon^4),$$

where

$$\hat{\Delta}t''(0) = \frac{\operatorname{tr}\left(\log_m^2(A_1^{-\frac{1}{2}} C A_1^{-\frac{1}{2}}) \log_m(A_1^{-\frac{1}{2}} A_2 A_1^{-\frac{1}{2}})\right)}{d(A_1, A_2)}.$$

Proof. Comparing the implicit derivative of $\check{\Delta}t$ with (4.4), we notice that the signs will alternate in each subsequent derivative. Theorem 4.3 is obtained after taking derivatives of (4.4). ∎

From Theorems 4.2 and 4.3, the difference between the solution of Problem 2.6 or Problem 2.7 and that of Problem 2.8 is locally of order ϵ^2. Similarly, the difference between the solutions of Problems 2.6 and 2.7 is also of order ϵ^2. Moreover, as shown in Figure 3, given that the first derivative is zero and $\hat{\Delta}t''(0) = -\check{\Delta}t''(0)$ (Theorem 4.3), the natural projection will typically fall *between* the I-projection and the reverse I-projection.

Besides the fact that it is a "middle point" between I-projection and reverse I-projection, there are other reasons to prefer natural projection. First, as noted earlier, it does not require assigning a probability distribution to the data. Second, the natural projection inherits the invariance properties of the natural distance, the most important of which are symmetry of the arguments and invariance to inversion; the

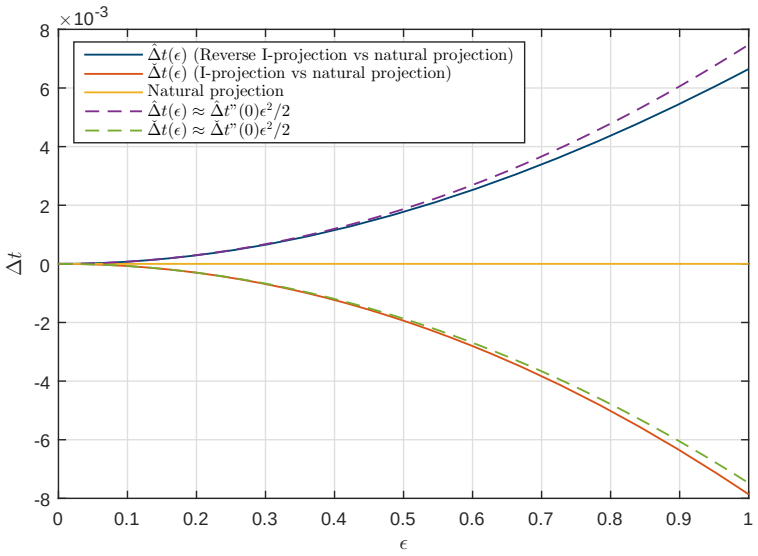

FIG. 3. *We draw three matrix realizations from a Wishart distribution of size* $n = 10$ *(*A*,* A_2*, and* C*).* A_1 *is then constructed as the minimizer of the natural distance from* C *to the geodesic* $\varphi_{A\to A_2}(t)$ *(cf. construction used in Theorem* 4.2*). We show* $\hat{\Delta}t$ *and* $\check{\Delta}t$ *as a function of the distance* ϵ *from* A_1 *to* C*. We control* ϵ *by defining* $\widehat{C} = \varphi_{A_1\to C}\left(\frac{1}{d(A_1,C)}\right)$*; thus,* $d(A_1,\widehat{C}) = 1$ *and* $\epsilon = d(A_1,\varphi_{A_1\to\widehat{C}}(\epsilon))$ *(cf.* (2.5)*). By construction,* Δt *for the natural projection is always zero.*

latter property is particularly useful when working with precision matrices, e.g., to take advantage of sparsity. (Reverse) I-projection does not enjoy these properties. Third, as we showed in section 3, the optimality conditions of the I-projections are first-order Taylor expansions of the natural projection; therefore, by maximizing the likelihood we are only solving a "flat" version of the geodesic problem. Finally, the natural distance is equivalent (up to a constant) to the Fisher information metric/Rao distance between two normal distributions with common mean, and thus the natural projection also minimizes these loss functions. If one uses geodesics (lines that minimize the natural distance) to build a covariance family, it is consistent to use the natural projection to select the most representative member of the family. In section 6, we will show that these advantages of natural projection translate into modeling benefits in practical applications, e.g., robustness to noise-corrupted data.

Up to now, we have not been concerned with asymptotics in the sample size q, but it is worth recalling that as $q/n \to \infty$, the sample covariance matrix converges to the true (population) covariance. If the former is close to the family, we have seen that minimizing the natural distance, maximizing the likelihood, and performing I-projection coincide up to second order. But it is additionally true that if the true covariance matrix is a member of the geodesic covariance family, natural projection of the sample covariance yields a consistent estimator of the population covariance, i.e., an estimate that converges in probability to the true covariance as $q \to \infty$. For a precise statement and proof of this result, see Proposition A.5.

5. Generalization to *p*-parameter covariance families. Thus far, we have only presented results for one-parameter covariance functions and families. In this section, we present a generalization to the multiparameter case. We employ geodesics

to construct a function of p parameters using $p+1$ matrices in $\mathbf{S}_+(n)$.

DEFINITION 5.1 (the *unbalanced* p-parameter covariance family). *By combining two one-parameter covariance functions, we obtain*

$$\varphi_{A_1 \to A_2 \to A_3}(t_1, t_2) \coloneqq \varphi_{\left(\varphi_{A_1 \to A_2}(t_1)\right) \to A_3}(t_2).$$

Analogously, by combining three one-parameter covariance functions, we obtain

$$\varphi_{A_1 \to A_2 \to A_3 \to A_4}(t_1, t_2, t_3) \coloneqq \varphi_{\left(\varphi_{\left(\varphi_{A_1 \to A_2}(t_1)\right) \to A_3}(t_2)\right) \to A_4}(t_3).$$

Recursively, we can construct the unbalanced p-parameter covariance function, which we denote as

$$\varphi_{A_1 \to \cdots \to A_{p+1}}(t_1, \ldots, t_p).$$

The image of the resulting function is the unbalanced *p-parameter covariance family.*

DEFINITION 5.2 (the *balanced* p-parameter covariance family). *By combining two one-parameter covariance functions, we obtain*

$$\varphi_{(A_1 \to A_2) \to (A_3 \to A_4)}(t_1, t_2, t_3) \coloneqq \varphi_{\left(\varphi_{A_1 \to A_2}(t_1)\right) \to \left(\varphi_{A_3 \to A_4}(t_2)\right)}(t_3).$$

Recursively, we can construct the balanced p-parameter covariance function. The image of the resulting function is the balanced *p-parameter covariance family.*

Figure 4 illustrates the construction of the two pure p-parameter covariance functions. The structure can be understood as a tree, where the anchor matrices are represented by leaf nodes and every other node has two parents. Each child node is associated with a parameter t_i. A *mixed* p-parameter covariance function can be obtained by combining balanced and unbalanced covariance functions. In the balanced tree structure, we require the number of anchor matrices to be a power of two.

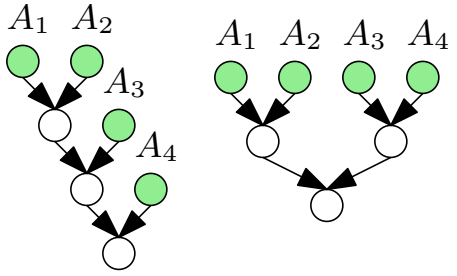

FIG. 4. *Unbalanced (left) versus balanced (right) trees. Green nodes represents anchor matrices and white circles are combinations of one-parameter covariance families. Color is available online only.*

Using the tree representation in Figure 4, we see that a covariance family is invariant to swapping the order of the two parents of any node; this follows from the second property listed after Definition 2.1. Any other permutation of the anchor matrices will change the image of the multiparameter covariance function and hence the family. As a specific example, consider the unbalanced two-parameter covariance function $\varphi_{A_1 \to A_2 \to A_3}(t_1, t_2)$ from Definition 5.1. There are six possible orderings of the anchors. Swapping the first two matrices of any ordering does not change the image

of the covariance function, but any other permutation does; thus, we can describe three different covariance families.

As in subsection 2.4, here we present the natural projection for a generic p-parameter covariance function. I-projection and reverse I-projection can be defined analogously. Let $\widehat{C}$ be the sample covariance matrix of the data $\{y_k\}_{k=1}^q$, and assume that $\widehat{C}$ is full rank.

PROBLEM 5.3 (natural projection to a p-parameter covariance function).

$$\arg\min_{t_1,\dots,t_p\in(-\infty,\infty)} d\big(\varphi_{A_1\to\cdots\to A_{p+1}}(t_1,\dots,t_p),\widehat{C}\big).$$

Algorithm 5.1 Coordinate descent for an unbalanced p-variate covariance function

Input: $A_1,\dots,A_j,\dots,A_{p+1}\in\mathbf{S}_+(n)$.

1. Use $\mathbf{t}^{(0)}=[t_1^{(0)},\dots,t_p^{(0)}]=\mathbf{0}$ as initial guess.
2. For $j=1:p$, find
   $t_j^{(1)}=\arg\min_{t_j\in(-\infty,\infty)} d(\varphi_{A_1\to\cdots\to A_{p+1}}(t_1^{(1)},\dots,t_{j-1}^{(1)},t_j,t_{j+1}^{(0)},\dots,t_p^{(0)}),\widehat{C})$.
3. Repeat step 2 for N iterations until convergence.

Return: The approximate minimizer is then $\varphi_{A_1\to\cdots\to A_{p+1}}(t_1^{(N)},\dots,t_p^{(N)})$.

To solve Problem 5.3, we propose Algorithm 5.1 based on coordinate descent. The objective function of Problem 5.3 is convex with respect to the first variable (Lemma A.3). However, it is not necessarily convex in other directions. Therefore, Algorithm 5.1 is not guaranteed to converge to a global minimum. By construction, however, the distance obtained via the algorithm is nonincreasing as we increase the size of the family p or the number of iterations N. In addition to providing a matrix within the family, the algorithm outputs the corresponding parameter values $t_1^*,\dots,t_p^*$. In practice, as with many coordinate descent algorithms, we find that this simple approach performs well.

Algorithm 5.1 can also be extended to the balanced p-variate covariance function. To do so, it suffices to define an order for step 2. The simplest strategy is first to minimize with respect to each parameter connecting each pair of parents (cf. Figure 4) and subsequently each pair of children, descending through the hierarchy. The same process can be applied for the mixed covariance function.

6. Case study: Hydraulic head in an aquifer. In this section, we use an example from groundwater hydrology to understand the capabilities of the covariance families and estimation methods developed above. We consider a simple model of the hydraulic head in an aquifer, illustrated in Figure 5, where the spatially heterogeneous permeability is modeled as a random field. We are interested in estimating the covariance of the resulting hydraulic head, across multiple points in the spatial domain. The stochastic model for the permeability field, which reflects various scenarios of geostatistical knowledge, directly impacts the covariance of the hydraulic head.

6.1. Analytical model. The hydraulic head h in the aquifer can be modeled by a one-dimensional Poisson equation with a stochastic permeability coefficient κ,

$$\frac{\partial}{\partial x}\left(\kappa(x,\omega)\frac{\partial h(x,\omega)}{\partial x}\right)+Q(x)=0,\quad x\in\Omega=[0,L], \tag{6.1}$$

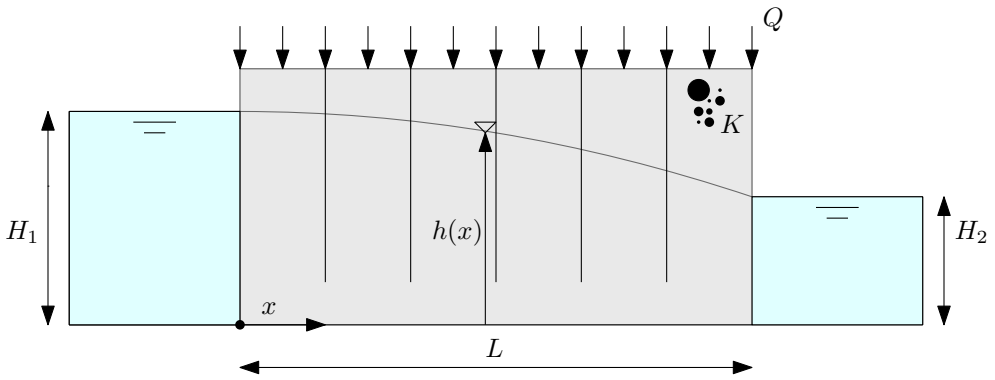

FIG. 5. *Illustration of the considered aquifer.*

where the source term is uniform $Q(x) = Q = 0.02$ and the boundary conditions are Dirichlet:

$$h(0) = H_1 = 50, \quad h(L) = H_2 = 20.$$

The permeability field $\kappa(x, \omega)$ is here defined as the exponential of a Gaussian process on $[0, L]$ with constant mean $\mu(x) = 1$ and covariance kernel

$$C(x, x') = \sigma^2 \exp\left(-\frac{1}{p}\left(\frac{|x - x'|}{l}\right)^p\right).$$

In the examples below, we will use $p = 2$ and $L = 100$, with various values of the correlation length l and variance σ^2 as indicated.

6.2. Construction of the covariance family. For any realization of the permeability field κ, we solve the equation above using a second-order finite difference scheme. By drawing many realizations of the log-permeability from the Gaussian process defined above, we can use Monte Carlo simulation to construct a sample estimate of the covariance of $\{h(x_i, \omega)\}_{i=1}^n$, taken at $n = 20$ equally spaced points $\{x_i\}$ on the domain. We do so for two different values of the correlation length l, called l_1 and l_2, fixing $\sigma^2 = 0.3$, and build a one-parameter covariance family using the corresponding covariance matrices of h (called A_1 and A_2) as anchors.

In Figure 6, we show an initial comparison of the one-parameter geodesic covariance family $\varphi_{A_1 \to A_2}(t)$ with the "flat" covariance family given by $t \mapsto (1-t)A_1 + tA_2$. We plot the distance from each point in the family to another given matrix (A_3, obtained with a log-permeability correlation length of l_3) for two cases: one where A_1 is closer to A_2 (left) and the other where A_1 is farther from A_2 (right). We see that if the two anchors are far apart, the natural distance from A_3 to the one-parameter *flat* covariance family loses convexity; moreover, it is not well defined for the entire real line, as the covariance matrices in the family lose rank for certain values of t. In contrast, the distance to the geodesic covariance family is convex and well defined for all $t \in \mathbb{R}$. In all of these cases, we use a very large number of Monte Carlo samples ($q = 10^6$) to construct A_1, A_2, and A_3 so that sampling error is negligible.

6.3. Regularization of a solution obtained with a reduced number of Monte Carlo instances. As described above, a standard method for solving (6.1)—

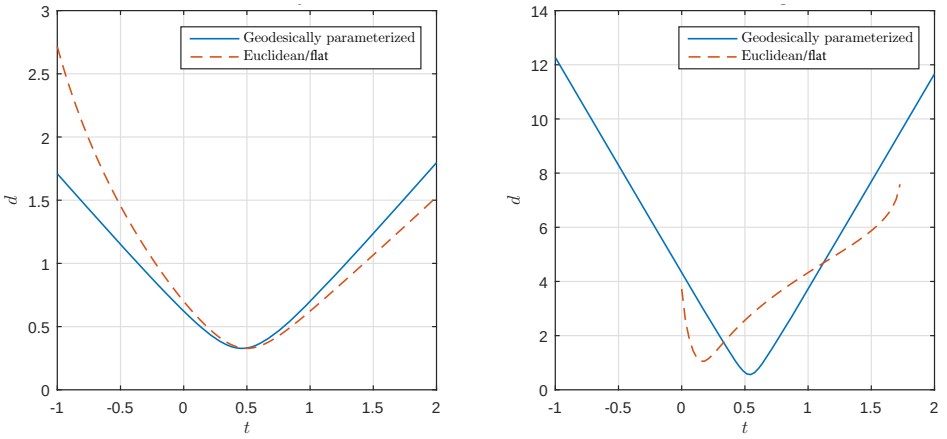

FIG. 6. *Contrasting the geodesic and "flat" one-parameter covariance families, by evaluating the natural distance to a third matrix. Left: when the anchor matrices are close ($l_1 = 20$, $l_2 = 30$, and $l_3 = 25$), the two families are similar. Right: when the anchors are far apart ($l_1 = 20$, $l_2 = 100$, and $l_3 = 60$), the natural distance to the flat family is not convex; outside of a certain range, where the red dashed line disappears, it is not even well defined. Color is available online only.*

e.g., computing the covariance of the solution field h—is Monte Carlo simulation. However, this approach converges slowly and can require a significant number of samples to produce an accurate estimate, thus incurring significant computational cost. Indeed, a central concern of forward uncertainty quantification (UQ) is the development of methods to characterize $h(x, \omega)$ with a cost that is much smaller than that of direct Monte Carlo simulation. Yet most UQ approaches focus on solving (6.1) and similar stochastic PDEs only for a *single* specification of the stochastic process $\kappa(x, \omega)$ [31]; if the parameters describing the stochastic inputs change, the problem typically must be resolved entirely.

Here we explore how to use tailored covariance estimation to solve this "outer" problem accurately using a rather small number of Monte Carlo samples. We propose to compute the covariance matrix of the solution (here called A_3) for new values of the input correlation length as follows:

1. Construct a covariance family using related problem instances. In the current example, we build a one-parameter covariance family using the anchor matrices A_1 and A_2 described in the previous subsection, corresponding to permeability field correlation lengths $l_1 = 20$ and $l_2 = 30$. These anchors are obtained with a large number of Monte Carlo samples ($q = 10^6$, $q/n = 5 \times 10^4$).
2. Compute the sample covariance matrix $\hat{A}_3$ at the desired new value of the permeability correlation length (here $l_3 = 25$) using a reduced number of Monte Carlo samples ($q = 10^3$, $q/n = 50$).
3. Project $\hat{A}_3$ to the family, via natural projection, to obtain a covariance matrix estimate A_3^* that is ideally closer to the actual solution A_3.

Figure 7 illustrates the proposed method. In Figure 8, we show the impact of this regularization scheme by performing 1,000 instances of the numerical experiment. In each instance, we repeat steps 2 to 3 above; i.e., we keep the anchors A_1 and A_2 fixed and only recalculate the noisier sample covariance $\hat{A}_3$. On average, the natural distance b' from the initial sample covariance estimate $\hat{A}_3$ to the true covariance matrix A_3 is 0.77 units. The average distance b from the projected matrix A_3^* to

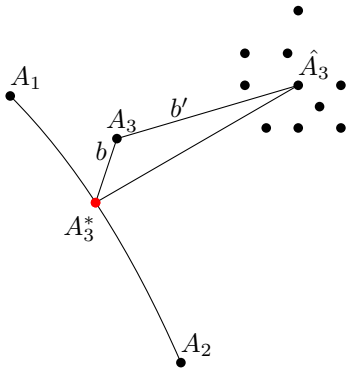

FIG. 7. *Illustration of regularizing by projection.*

the true covariance is 0.07. The average reduction in error (i.e., b'/b averaged over problem instances), which can be understood as a regularization ratio, is 11.8. The blue histogram of distances has a hard minimum at 0.04, which reflects limitations of the covariance family: the true solution A_3 is close to the geodesic but does not actually belong to it.

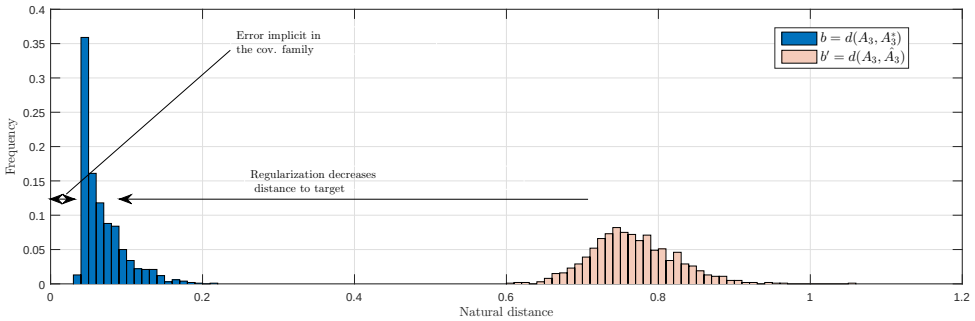

FIG. 8. *Using* 1,000 *solutions obtained via the process outlined in subsection* 6.3*, we plot a histogram of the initial distance* b' *to the true covariance matrix and a histogram of the distance* b *once the sample covariance is regularized via natural projection (cf. Figure* 7*). The averages of* b' *and* b *are* 0.77 *and* 0.07*, respectively. On average, we reduce the error by a regularization ratio of* 11.8.

Setting aside the offline cost of computing the anchors, the computational cost of performing natural projection of $\hat{A}_3$ onto the geodesic family is the cost of solving Problem 2.8. This problem is univariate and convex, and thus easily tackled by a variety of methods; here we simply use direct search, which usually requires fewer than 10 iterations to find the optimal t^* with an absolute precision of 10^{-4}. Evaluating the cost function requires solving a symmetric definite generalized eigenvalue problem at each iteration. Generically, these problems have $O(n^3)$ cost for covariance matrices of size n, though the complexity may be lower with iterative solvers and problems with particular structure (since the natural distance depends most strongly on the extreme eigenvalues, i.e., those that differ most from one). In this example, the actual cost of solving the eigenproblem is negligible, particularly compared to the cost of drawing Monte Carlo samples to form $\hat{A}_3$. In general, we note that the cost of drawing Monte Carlo samples typically *also* scales with n. For instance, each Monte Carlo draw

might involve a linear PDE solve (say of $O(n^2)$ complexity), and to maintain a fixed q/n, the sample size q must itself increase linearly with n.

6.4. Regularization of a noisy solution. Projection onto the geodesic family, as illustrated in subsection 6.3, can also be performed using maximum likelihood or I-projection; we find that these approaches yield similar regularization performance for the previous problem. Now we consider a more difficult regularization task, where Monte Carlo samples are perturbed with independent and identically distributed realizations of zero-mean Gaussian noise before they are used to construct the sample covariance matrix. This problem mimics a situation where noisy observational data are used to estimate the covariance of the hydraulic head h. The anchor matrices A_1, A_2 and true covariance matrix A_3 are the same as in the previous problem. Figure 9 shows the regularization ratios b'/b obtained for both natural projection (left) and maximum likelihood estimation (right), at 10 different noise magnitudes, each using 500 instances. On each box, the central mark indicates the median value of b'/b, and the bottom and top edges of the box indicate the 25th and 75th percentiles, respectively. The whiskers extend to the most extreme data points not considered outliers, and the outliers are plotted individually using the "+" symbol. The horizontal axis corresponds to the standard deviation of the Gaussian noise, scaled by $\alpha 0.05\sqrt{\mathrm{tr}(A_3)/n}$.

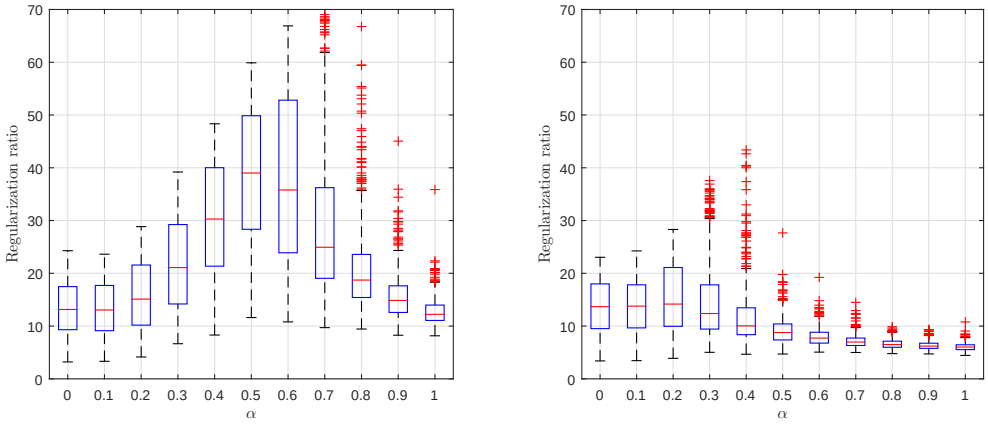

FIG. 9. *Boxplot of regularization ratios obtained with natural projection (left) and maximum likelihood (right) for increasing magnitudes of noise. Larger values are better. Each box is the result of* 500 *instances. Standard deviation of the noise perturbations is* $\alpha 0.05\sqrt{\mathrm{tr}(A_3)/n}$.

Figure 9 suggests that maximum likelihood and natural projection perform quite differently in the presence of noise. Both projections yield similar results for $\alpha \leq 0.2$, but natural projection is more robust to noise-corrupted samples for larger noise perturbations. The covariance families for both cases are exactly the same, but maximum likelihood appears less able to identify the closest covariance matrix in the family as the noise magnitude increases. Both regularization methods display an up-down trend with α. As α first increases, b' increases while b stays relatively constant; in other words, the sample covariance moves further from the true A_3 but the quality of the regularized matrix A_3^* does not deteriorate, and thus we observe larger regularization ratios (much more so for natural projection). As α grows even larger, eventually the distance b starts increasing as well, and thus b'/b begins to fall. The variance of the

regularization ratios is somewhat larger for natural projection, but the median ratio obtained with maximum likelihood is generally even smaller than the minimum ratio achieved with natural projection.

As discussed in section 3, the optimality equation for maximum likelihood estimation in the covariance family is a linearized version of natural projection. It may be that maximum likelihood is more sensitive to noise-corrupted data because it is missing certain higher-order terms. Further exploration of natural projection's ability to overcome noise might employ a perturbation analysis of the generalized eigenvalues of the spectral functions in Remark 2.4; we leave this to future work.

6.5. Performance of proposed algorithm for multiparametric families. With more than two anchor matrices, one can build a multiparametric covariance function as described in section 5. To illustrate, suppose that we have three anchor covariance matrices A_1, A_2, and A_3, corresponding to solutions of (6.1) for ($l_1 = 20$, $\sigma_1^2 = 0.3$), ($l_2 = 30$, $\sigma_2^2 = 0.3$), and ($l_3 = 25$, $\sigma_3^2 = 0.4$), respectively. We now solve a problem similar to that in subsection 6.3, but with two parameters. The objective is to approximate the covariance matrix A_4, obtained with ($l_4 = 25$, $\sigma_4^2 = 0.35$). First, we build an unbalanced two-parameter covariance family $\varphi_{A_1 \to A_2 \to A_3}(t_1, t_2)$ (see Definition 5.1). Then we project an approximation $\hat{A}_4$ of A_4, obtained with 1,000 Monte Carlo simulations of (6.1), onto the family. Repeating this experiment with 1,000 independent realizations of $\hat{A}_4$, we obtain an average regularization ratio of 6.5. These results are illustrated in Figure 10.

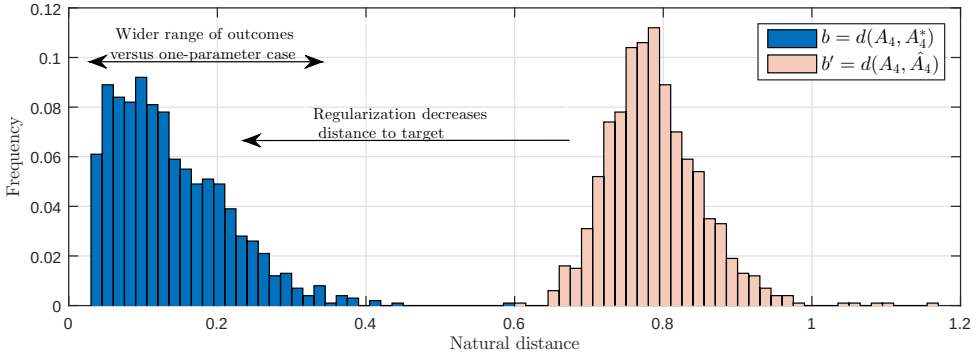

FIG. 10. *Multiparametric case: Using* 1,000 *solutions obtained via the process outlined in subsection* 6.5*, we plot a histogram of the initial distance* b' *to the true covariance matrix and a histogram of the distance* b *obtained after natural projection. The average distances are* 0.79 *and* 0.14*, respectively. On average, we reduce the error by a regularization ratio of* 6.5*.*

To perform natural projection onto this multiparametric family, we used Algorithm 5.1. To illustrate the performance of this algorithm, the colored contours in Figure 11(left) show the distance to $\hat{A}_4$ as a function of the covariance family's parameters t_1, t_2. Iterations of the algorithm $1, \ldots, N$ are marked with green triangles: intermediate iterations are unfilled triangles, and the filled triangle denotes the final point. We say that convergence is achieved when consecutive values of t_1 and t_2 each do not differ by more than 10^{-4}. In this particular example, convergence requires four iterations, three of which fall inside the perimeter of Figure 11(left).

In Figure 11(right), color contours illustrate the distance from the family to the true covariance matrix A_4, with iterations of the coordinate descent algorithm again overlaid. As the algorithm progresses, the distance to A_4 does not necessarily decrease;

as expected, the minimization is blind to A_4. Indeed, the goal of Algorithm 5.1 is to obtain the minimizer of $d\big(\hat{A}_4, \varphi_{A_1\to A_2\to A_3}(t_1,t_2)\big)$ (green triangle), which is generally not the same as the minimizer of $d\big(A_4, \varphi_{A_1\to A_2\to A_3}(t_1,t_2)\big)$ (red square). The natural distance between the covariance matrices corresponding to these two minimizers is 0.08 units. This distance reflects the fact that $\hat{A}_4$ is noisy, and thus its best approximation in the family is not the best approximation of A_4. Separately, the limit of the two-parameter covariance family's ability to represent A_4 is captured by the value of the contour in Figure 11(right) at the red square, which is 0.03 units.

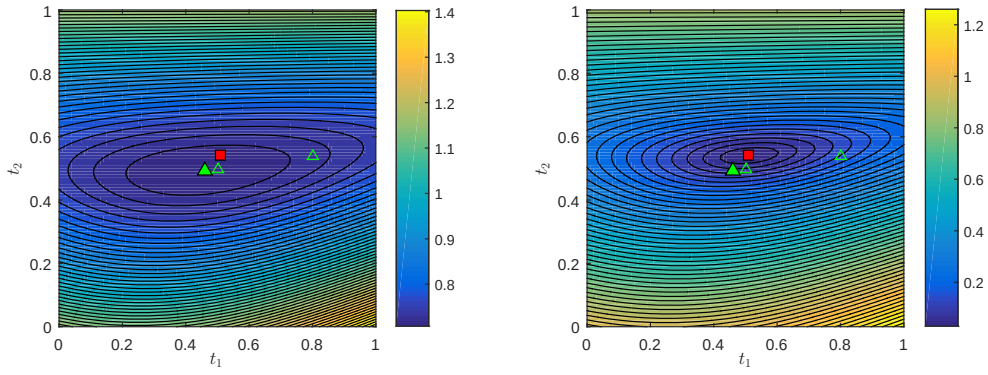

FIG. 11. *Multiparametric regularization for one instance of $\hat{A}_4$, described in subsection 6.5. Left panel: contours/colors represent the value of $d\big(\hat{A}_4, \varphi_{A_1\to A_2\to A_3}(t_1,t_2)\big)$, and triangles show iterations of Algorithm 5.1 until convergence. The red square is the minimizer of $d\big(A_4, \varphi_{A_1\to A_2\to A_3}(t_1,t_2)\big)$. Right panel: contours/colors represent $d\big(A_4, \varphi_{A_1\to A_2\to A_3}(t_1,t_2)\big)$. Color is available online only.*

7. Conclusions. We have proposed a framework for building expressive and problem-tailored parametric covariance families by connecting representative "anchor" covariance matrices through geodesics. The building block of the framework is the one-parameter covariance family, corresponding to a single geodesic. These geodesics may be combined to yield multiparameter covariance families. Given some new data, one can then choose the most appropriate member of such a family by minimizing the natural distance (on the manifold of symmetric positive-definite matrices) to the sample covariance matrix of the data. We call this notion natural projection. Unlike maximum likelihood estimation (reverse I-projection) or I-projection within the family, natural projection is consistent with the notion of distance employed to build the covariance family. We elucidate the differences among these estimation techniques and show that I-projection and reverse I-projection can be seen as linearizations of the natural projection.

We also illustrate the advantages of geodesic covariance families and natural projection in several numerical experiments. Analogous covariance families that do not employ the geodesic structure may lose rank, and the distance from such a "flat" family to another matrix is in general not convex—especially if the anchors are far apart. The geodesic families avoid these difficulties. When performing parameter estimation within the geodesic family, maximum likelihood and natural projection provide similar results in the absence of noise. If the data are corrupted by noise, however, then natural projection is better able to regularize the solution.

Given a covariance family, the choice between natural projection and maximum

likelihood estimation within the family also depends on the number of data points q and the size of the matrices n. When $q < n$, the natural distance cannot be used because the sample covariance matrix will be rank deficient and thus will not belong to the manifold of symmetric positive-definite matrices. On the other hand, when $q > n$, we suggest that it is preferable to use natural projection because it is consistent with the geodesic construction of the family (with a cost function that is a proper notion of distance) and because it does not require assigning a distribution to the data. Moreover, it has superior noise rejection properties. If the sample covariance matrix is close to the family, however, we have also shown that minimizing the natural distance, maximizing the likelihood, and performing I-projection within the family coincide up to second order. Moreover, as $q/n \to \infty$, the sample covariance matrix itself becomes a good representation of the true (population) covariance, and in this limit, the practical need to project to any parametric family diminishes. Nonetheless, we have demonstrated numerically that even for $q/n \approx 50$, regularization of the sample covariance matrix via projection can yield significant reductions in error. And if the population covariance matrix is contained within the geodesic family, then the natural projection of the sample covariance yields a consistent estimator of the true covariance, as one would certainly desire.

A natural extension of this work is to weaken the role of the anchor matrices: not to project directly to a parametric family defined by the anchors, but rather to seek only "closeness" to one or more anchors, where the degree of closeness might depend on the quality of the sample covariance matrix. A popular strategy along these lines is linear shrinkage, which consists in selecting a linear combination of the sample covariance matrix and a single reference covariance matrix (often chosen to be the identity). In particular, linear shrinkage seeks the combination that is closest (in expected Frobenius distance) to the true covariance; thus, both the loss and the effective covariance family are flat. The obvious challenge is that the true covariance matrix is not known. In the spirit of the present paper, future work could develop a geodesic version of shrinkage. We expect that such a nonlinear shrinkage (cf. [28]) could extend some of the favorable properties of the geodesic framework to the nonparametric case.

Another promising application of the covariance families developed here may lie in hierarchical Bayesian modeling, e.g., for inverse problems. Specifically, we suggest that geodesically parameterized covariance matrices could be used to describe particularly flexible classes of prior models, where the covariance family's parameters $(t_1, \ldots, t_p)$ may serve as hyperparameters—rather than the correlation/scale/smoothness parameters of standard covariance kernels. The parameters of the covariance family could then be inferred either in a fully Bayesian formulation or via an empirical Bayesian approach. A different possibility is to use geodesic families to continuously interpolate among the posterior covariance matrices that follow from particular prior choices.

Appendix A. Technical results.

LEMMA A.1 (spectral function minimization). *Let $F = f \circ \lambda$ be a spectral function, and let $X(t) := \widehat{C}^{-\frac{1}{2}} \varphi_{A_1 \to A_2}(t) \widehat{C}^{-\frac{1}{2}} = M \Lambda^t M^\top$, where $\varphi_{A_1 \to A_2}(t)$ is the geodesic defined in* (2.3)*, $M = \widehat{C}^{-\frac{1}{2}} A_1^{\frac{1}{2}} U$, and $\widehat{C}$ is full rank. Minimizing $F\big(X(t)\big)$ over*

t is equivalent to finding t^+ such that[1]

$$\operatorname{tr}\left(V(t^+)\left(\left.\frac{df}{d\lambda}\right|_{\lambda\left(X(t^+)\right)}\right)V(t^+)^\top M\Lambda^{t^+}\log_m(\Lambda)M^\top\right)=0,$$

where $V(t)\Sigma(t)V(t)^\top$ is an orthonormal eigendecomposition of $X(t)$.

Proof. Notice that the generalized eigenvalues of the pencil $\left(\varphi_{A_1\to A_2}(t),\widehat{C}\right)$ are the eigenvalues of $X(t)$. Since this is an unconstrained minimization problem, the idea is to impose the following condition:

$$\left.\frac{dF\big(X(t)\big)}{dt}\right|_{X(t^+)}=0.$$

The difficulty here is that $F\big(X(t)\big)=f\circ\lambda\circ X(t)$, where $X(t)$ is defined as in the present lemma, λ is the function that extracts the eigenvalues of a given matrix, and f is a mapping from these eigenvalues to $\mathbb{R}$. Applying the chain rule and [30, Theorem 1.1], we obtain

$$\frac{dF\big(X(t)\big)}{dt}=\operatorname{tr}\left(V(t)\left.\frac{df}{d\lambda}\right|_{\lambda\left(X(t)\right)}V^\top(t)\frac{dX(t)}{dt}\right),$$

where $\frac{dX(t)}{dt}=M\Lambda^t\log_m(\Lambda)M^\top$. ∎

The next three proofs of results from section 3 follow similar strategies. We first use Lemma A.1 for the corresponding spectral function in Remark 2.4, and then we derive the orthogonality condition.

Proof of Proposition 3.3. We start with (cf. (2.6))

$$f(\lambda_1,\dots,\lambda_n)=\sqrt{\sum_{k=1}^n\log^2\lambda_k}.$$

After omitting the square root, the derivative is

$$\frac{df(\lambda)}{d\lambda_k}=\frac{2\log_m\left(\lambda_k\big(X(t)\big)\right)}{\lambda_k\big(X(t)\big)}=\left(2\Sigma^{-1}(t)\log_m\big(\Sigma(t)\big)\right)_{(k,k)},$$

where $\Sigma(t)$ is defined in Lemma A.1. We have to find a t^* such that

(A.1)
$$\left.\frac{dF\big(X(t)\big)}{dt}\right|_{X(t^*)}=2\operatorname{tr}\left(V(t^*)\Sigma^{-1}(t^*)\log_m\big(\Sigma(t^*)\big)V^\top(t^*)M\Lambda^{t^*}\log_m(\Lambda)M^\top\right)=0.$$

Now, notice that by construction $M^\top V(t)\Sigma^{-1}(t)=\Lambda^{-t}M^{-1}V(t)$ and applying the cyclical property of the trace, we obtain

$$\operatorname{tr}\left(M^{-1}V(t^*)\log_m\big(\Sigma(t^*)\big)V^\top(t^*)M\Lambda^{t^*}\log_m(\Lambda)\Lambda^{-t^*}\right)=0.$$

[1] $\frac{df}{d\lambda}$ is our shorthand notation for $\operatorname{diag}\left(\frac{df}{d\lambda_1},\dots,\frac{df}{d\lambda_n}\right)$.

Since diagonal matrices commute,

$$\operatorname{tr}\Big(\log_m\big(M^{-1}V(t^*)\Sigma(t^*)V^\top(t^*)M\big)\log_m(\Lambda)\Big)=0.$$

After that, we obtain

$$\operatorname{tr}\big(\log_m(\Lambda^{t^*}M^\top M)\log_m(\Lambda)\big)=0.$$

The first part of the proof (cf. (A.2)) is concluded after realizing that by construction $M^\top M = U^\top A_1^{\frac{1}{2}}\widehat{C}^{-1}A_1^{\frac{1}{2}}U$ and that

$$\operatorname{tr}\big(\log_m(Z\Lambda^{-t^*})\log_m(\Lambda)\big)=0. \tag{A.2}$$

Then note that (3.1) can be rewritten as

$$\operatorname{tr}\left(\log_m\left(R\left(-\frac{t^*}{2}\right)A_1^{-\frac{1}{2}}\widehat{C}A_1^{-\frac{1}{2}}R\left(-\frac{t^*}{2}\right)\right)\log_m\left(R\left(-\frac{t^*}{2}\right)R(1+t^*)R\left(-\frac{t^*}{2}\right)\right)\right)=0,$$

which is equivalent to (3.2). ∎

Proof of Proposition 3.4. For reverse I-projection, we start with (cf. (2.7))

$$f(\lambda_1,\dots,\lambda_n)=\sum_{k=1}^{n}\frac{\lambda_k^{-1}+\log\lambda_k-1}{2}.$$

Then

$$\frac{df(\lambda)}{d\lambda_k}=-\frac{1}{2\lambda_k\big(X(t)\big)^2}+\frac{1}{2\lambda_k\big(X(t)\big)}=\frac{1}{2}\big(-\Sigma^{-2}(t)+\Sigma^{-1}(t)\big)_{(k,k)}.$$

Similarly to the previous case, now we have to find a $\hat{t}$ such that

$$\operatorname{tr}\Big(V(\hat{t})\big(-\Sigma^{-2}(\hat{t})+\Sigma^{-1}(\hat{t})\big)V^\top(\hat{t})M\Lambda^{\hat{t}}\log_m(\Lambda)M^\top\Big)=0.$$

Applying $M\Lambda^tM^\top = V(t)\Sigma(t)V(t)^\top$ and the cyclical property of the trace, we obtain

$$\operatorname{tr}\big(-M^{-T}\log_m(\Lambda)\Lambda^{-\hat{t}}M^{-1}+\log_m(\Lambda)\big)=0.$$

The result follows after applying $M=\widehat{C}^{-\frac{1}{2}}A_1^{\frac{1}{2}}U$, that is,

$$\operatorname{tr}\big((Z\Lambda^{-\hat{t}}-\mathrm{Id})\log_m(\Lambda)\big)=0. \tag{A.3}$$

The orthogonality condition is obtained as in Proposition 3.3. ∎

Proof of Proposition 3.5. For I-projection, we start with (cf. (2.8))

$$f(\lambda_1,\dots,\lambda_n)=\sum_{k=1}^{n}\frac{\lambda_k-\log\lambda_k-1}{2}.$$

Then

$$\frac{df(\lambda)}{d\lambda_k}=\frac{1}{2}-\frac{1}{2\lambda_k\big(X(t)\big)}=\frac{1}{2}\big(\mathrm{Id}-\Sigma^{-1}(t)\big)_{(k,k)}.$$

Similarly to the previous case, now we have to find a $\check{t}$ such that

$$\operatorname{tr}\Big(V(\check{t})\big(\operatorname{Id}-\Sigma^{-1}(\check{t})\big)V^\top(\check{t})M\Lambda^{\check{t}}\log_m(\Lambda)M^\top\Big)=0.$$

Applying $M\Lambda^t M^\top = V(t)\Sigma(t)V(t)^\top$ and the cyclical property of the trace, we obtain

$$\operatorname{tr}\big(M^\top M\Lambda^{\check{t}}\log_m(\Lambda)-\log_m(\Lambda)\big)=0.$$

The result follows after applying $M=\widehat{C}^{-\frac{1}{2}}A_1^{\frac{1}{2}}U$, that is,

$$\operatorname{tr}\big((\Lambda^{\check{t}}Z^{-1}-\operatorname{Id})\log_m(\Lambda)\big)=0. \tag{A.4}$$

The orthogonality condition is obtained as in Proposition 3.3. ∎

LEMMA A.2 (equivalence of likelihood maximization and reverse I-projection). *Let $p_Y(\cdot\,;t)$ be a Gaussian density on $\mathbb{R}^n$ centered at zero, with covariance $\varphi_{A_1\to A_2}(t)$. Let $y_1,\dots,y_q \overset{iid}{\sim} p_Y(\cdot\,;\bar{t})$ for some $\bar{t}\in\mathbb{R}$, such that the sample covariance matrix $\widehat{C}=\frac{1}{q}\sum_{i=1}^q y_i y_i^\top$ is full rank. Maximizing the log-likelihood $\sum_{i=1}^q \log p_Y(y_i;t)$ with respect to t is equivalent to minimizing the KL divergence $D_{KL}\big(N(0,\widehat{C})\;\|\;N(0,\varphi_{A_1\to A_2}(t))\big)$.*

Proof. Each observation y_i has the following density:

$$p_Y(y_i;t)=\frac{1}{(2\pi)^{n/2}\sqrt{|\varphi_{A_1\to A_2}(t)|}}\exp\left(-\frac{1}{2}y_i^\top\varphi^{-1}_{A_1\to A_2}(t)y_i\right).$$

The joint log-likelihood can be expressed as

$$\log\prod_{i=1}^q p_Y(y_i;t)=\sum_{i=1}^q\left(-\log^{n/2}2\pi-\frac{1}{2}\log|\varphi_{A_1\to A_2}(t)|-\frac{1}{2}y_i^\top\varphi^{-1}_{A_1\to A_2}(t)y_i\right).$$

Now notice that $\varphi_{A_1\to A_2}(t)=A_1^{\frac{1}{2}}U\Lambda^tU^\top A_1^{\frac{1}{2}}$, $|\varphi_{A_1\to A_2}(t)|=|A_1||\Lambda^t|$, and ignore the constant terms. Since this is an unconstrained and concave problem, the extremum t^+ can be found by setting the derivative of the function to zero, that is,

$$q\operatorname{tr}\big(\log_m(\Lambda)\big)-\sum_{i=1}^q y_i^\top A_1^{-1/2}U\Lambda^{-t^+}\log_m(\Lambda)U^\top A_1^{-1/2}y_i=0. \tag{A.5}$$

In matrix form, the above expression reads as

$$\operatorname{tr}\big(\widehat{C}A_1^{-\frac{1}{2}}U\Lambda^{-t^+}\log_m(\Lambda)U^\top A_1^{-\frac{1}{2}}-\log_m(\Lambda)\big)=0. \tag{A.6}$$

The proof is concluded after realizing that the last expression is precisely (A.3). ∎

LEMMA A.3 (convexity of distance function). *Distance function $d(\varphi_{A_1\to A_2}(t),\widehat{C})$ is convex in t. Therefore, Problem 2.8 is an unconstrained convex minimization problem.*

Proof. The strategy is to take the derivative of (A.1) and realize that it is nonnegative:

$$\frac{dF\big(X(t)\big)}{dt}=2\operatorname{tr}\big(\log_m(M\Lambda^tM^\top)\log_m(M\Lambda M^{-1})\big).$$

Recall that

$$\frac{d\log_m X(t)}{dt} = \int_0^1 \big((X(t)-\mathrm{Id})s+\mathrm{Id}\big)^{-1}\left(\frac{dX(t)}{dt}\right)\big((X(t)-\mathrm{Id})s+\mathrm{Id}\big)^{-1}ds.$$

Notice that in our case, $X(t) = M\Lambda^t M^\top$ and

$$\frac{dX(t)}{dt} = M\Lambda^t \log_m(\Lambda) M^\top.$$

Performing an orthonormal eigendecomposition of the form $X(t) = V(t)\Sigma V(t)^\top$, the other main part of the integrand reads as

$$\big((X(t)-\mathrm{Id})s+\mathrm{Id}\big)^{-1} = \Big(\big(V\Sigma V^\top-\mathrm{Id}\big)s+\mathrm{Id}\Big)^{-1} = V\big(s\Sigma+\mathrm{Id}(1-s)\big)^{-1}V^\top \coloneqq VJV^\top,$$

where for easy presentation, we omit the explicit dependence of V, Σ, and J on t. J is clearly positive since Σ is and $s \in [0,1]$.

The second derivative can be expressed as

$$\frac{d^2F\big(X(t)\big)}{dt^2} = 2\int_0^1 \mathrm{tr}\left(VJV^\top M\Lambda^t \log_m(\Lambda)M^\top VJV^\top M\log_m(\Lambda)M^{-1}\right)ds.$$

It suffices to show that the trace is positive for $s \in [0,1]$. Using the equality $V^\top M\Lambda^t = \Sigma V^\top M^{-T}$ from the eigendecomposition and the cyclical property of the trace, we obtain

$$\begin{aligned}\frac{d^2F\big(X(t)\big)}{2dt^2} &= \int_0^1 \mathrm{tr}\left(VJ\Sigma V^\top M^{-T}\log_m(\Lambda)M^\top VJV^\top M\log_m(\Lambda)M^{-1}\right)ds\\ &= 2\int_0^1 \mathrm{tr}\left((J\Sigma)^{\frac12}V^\top M^{-T}\log_m(\Lambda)M^\top VJ^{\frac12}J^{\frac12}V^\top M\log_m(\Lambda)M^{-1}V(J\Sigma)^{\frac12}\right)ds\\ &= 2\int_0^1 \mathrm{tr}(KK^\top)ds > 0,\end{aligned}$$

where $K = (J\Sigma)^{\frac12}V^\top M^{-T}\log_m(\Lambda)M^\top VJ^{\frac12}$. ∎

LEMMA A.4 (convexity of the KL divergence function). *The following KL divergences $D_{KL}\big(N(0,\widehat{C}) \,\|\, N\big(0,\varphi_{A_1\to A_2}(t)\big)\big)$ and $D_{KL}\big(N\big(0,\varphi_{A_1\to A_2}(t)\big) \,\|\, N(0,\widehat{C})\big)$ are convex functions of t. Therefore, Problems* 2.6 *and* 2.7 *are unconstrained convex minimization problems.*

Proof. It suffices to evaluate the second derivative and conclude that it is always positive. Taking derivatives of (A.3), we obtain

$$\mathrm{tr}\left(Z\Lambda^{-t}\log_m^2(\Lambda)\right) > 0 \quad \forall t.$$

Similarly, taking derivatives of (A.4), we obtain

$$\mathrm{tr}\left(\Lambda^t Z^{-1}\log_m^2(\Lambda)\right) > 0 \quad \forall t.$$ ∎

PROPOSITION A.5 (consistency with a perfect family). *If the true covariance matrix A is a member of the geodesic covariance family $\varphi_{A_1\to A_2}(t)$, natural projection of the sample covariance matrix yields a consistent estimator of A.*

Proof. Without loss of generality, let the observations $(y_i)_{i=1}^q$ be zero-mean identically distributed random vectors drawn independently from a distribution with covariance matrix A, define the sample covariance matrix as $\widehat{A}_q \coloneqq \frac{1}{q}\sum_{i=1}^q y_i y_i^\top$, and let $A_1 = A$. The natural projection of $\widehat{A}_q$ into $\varphi_{A_1\to A_2}$ is denoted as $A_q^* = \varphi_{A_1\to A_2}(t_q^*)$, where

$$t_q^* = \operatorname*{argmin}_t d(\varphi_{A_1\to A_2}(t), \widehat{A}_q).$$

Define $\widehat{Q}_q(t) \coloneqq d\big(\varphi_{A_1\to A_2}(t), \widehat{A}_q\big)$ and $Q(t) \coloneqq d\big(\varphi_{A_1\to A_2}(t), A\big)$. By Lemmas A.3 and 3.2, we have that $Q(t)$ is uniquely minimized at $t = 0$, which is clearly in the interior of $(-\infty, \infty)$, and that $\widehat{Q}_q(t)$ is convex for any q. Now we show that $\widehat{Q}_q(t)$ converges in probability to $Q(t)$ for any $t \in \mathbb{R}$:

$$\lim_{q\to\infty} \mathbb{P}\big(|Q_q(t) - Q(t)| > \epsilon\big) \le \lim_{q\to\infty} \mathbb{P}\big(d(A, \widehat{A}_q) > \epsilon\big) = \lim_{q\to\infty} \mathbb{P}\left(\sum_{k=1}^n \log^2 \lambda_k^{(A,\widehat{A}_q)} > \epsilon^2\right) \xrightarrow{p} 0$$

for any $\epsilon > 0$, where the first step follows from the triangle inequality and the last follows from the Marčenko–Pastur law [32], which gives the distribution of the eigenvalues of $A^{-\frac{1}{2}}\widehat{A}_q A^{-\frac{1}{2}}$ and hence the generalized eigenvalues $\lambda_k^{(A,\widehat{A}_q)}$, for sufficiently large n. Alternatively, for Gaussian y_i, the final limit holds at any n via [43, Theorem 7]. Invoking [36, Theorem 2.7], we then obtain $t_q^* \xrightarrow{p} 0$. Since convergence in probability is preserved under continuous mappings, we also have $A_q^* \xrightarrow{p} A$. ∎

REFERENCES

[1] P.-A. Absil, P.-Y. Gousenbourger, P. Striewski, and B. Wirth, *Differentiable piecewise-Bézier surfaces on Riemannian manifolds*, SIAM J. Imaging Sci., 9 (2016), pp. 1788–1828, https://doi.org/10.1137/16M1057978.

[2] H. Akaike, *Determination of the Number of Factors by an Extended Maximum Likelihood Principle*, Institute of Statistical Mathematics, Tokyo, 1971.

[3] H. Akaike, *Information theory and an extension of the maximum likelihood principle*, in Selected Papers of Hirotugu Akaike, Springer, New York, 1998, pp. 199–213.

[4] S.-i. Amari, *Information Geometry and Its Applications*, Appl. Math. Sci. 194, Springer, Tokyo, 2016.

[5] D. Amsallem, J. Cortial, K. Carlberg, and C. Farhat, *A method for interpolating on manifolds structural dynamics reduced-order models*, Internat. J. Numer. Methods Engrg., 80 (2009), pp. 1241–1258.

[6] D. Amsallem and C. Farhat, *Interpolation method for adapting reduced-order models and application to aeroelasticity*, AIAA J., 46 (2008), pp. 1803–1813.

[7] D. Amsallem and C. Farhat, *An online method for interpolating linear parametric reduced-order models*, SIAM J. Sci. Comput., 33 (2011), pp. 2169–2198, https://doi.org/10.1137/100813051.

[8] C. Atkinson and A. F. Mitchell, *Rao's distance measure*, Sankhyā Ser. A, 43 (1981), pp. 345–365.

[9] R. Bhatia, *Positive-Definite Matrices*, Princeton University Press, Princeton, NJ, 2009.

[10] S. Bonnabel and R. Sepulchre, *Riemannian metric and geometric mean for positive semidefinite matrices of fixed rank*, SIAM J. Matrix Anal. Appl., 31 (2009), pp. 1055–1070, https://doi.org/10.1137/080731347.

[11] T. Cai and W. Liu, *Adaptive thresholding for sparse covariance matrix estimation*, J. Amer. Statist. Assoc., 106 (2011), pp. 672–684.

[12] T. Cai, W. Liu, and X. Luo, *A constrained L1 minimization approach to sparse precision matrix estimation*, J. Amer. Statist. Assoc., 106 (2011), pp. 594–607.

[13] N. Cressie, *Statistics for Spatial Data*, Vol. 4, Wiley Online Library, 1992.

[14] I. Csiszár and F. Matus, *Information projections revisited*, IEEE Trans. Inform. Theory, 49 (2003), pp. 1474–1490.

[15] I. S. Dhillon and J. A. Tropp, *Matrix nearness problems with Bregman divergences*, SIAM J. Matrix Anal. Appl., 29 (2007), pp. 1120–1146, https://doi.org/10.1137/060649021.

[16] D. L. Donoho, M. Gavish, and I. M. Johnstone, *Optimal shrinkage of eigenvalues in the spiked covariance model*, Ann. Statist., 46 (2018), pp. 1742–1778.
[17] A. Edelman, T. A. Arias, and S. T. Smith, *The geometry of algorithms with orthogonality constraints*, SIAM J. Matrix Anal. Appl., 20 (1998), pp. 303–353, https://doi.org/10.1137/S0895479895290954.
[18] J. Faraut and A. Korányi, *Analysis on Symmetric Cones*, Oxford University Press, New York, 1994.
[19] W. Förstner and B. Moonen, *A metric for covariance matrices*, in Geodesy-The Challenge of the 3rd Millennium, Springer, Berlin, Heidelberg, 2003, pp. 299–309.
[20] J. Friedman, T. Hastie, and R. Tibshirani, *Sparse inverse covariance estimation with the graphical lasso*, Biostatistics, 9 (2008), pp. 432–441.
[21] R. Furrer, M. G. Genton, and D. Nychka, *Covariance tapering for interpolation of large spatial datasets*, J. Comput. Graph. Statist., 15 (2006), pp. 502–523.
[22] P.-Y. Gousenbourger, E. M. Massart, A. Musolas, P.-A. Absil, J. M. Hendrickx, L. Jacques, and Y. Marzouk, *Piecewise-Bézier C1 smoothing on manifolds with application to wind field estimation*, in Proceedings of the 25th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN), 2017, pp. 305–3010.
[23] P.-Y. Gousenbourger, C. Samir, and P.-A. Absil, *Piecewise-Bézier C1 interpolation on Riemannian manifolds with application to 2D shape morphing*, in Proceedings of the 22nd International IEEE Conference on Pattern Recognition (ICPR), 2014, pp. 4086–4091.
[24] J. R. Guerci, *Theory and application of covariance matrix tapers for robust adaptive beamforming*, IEEE Trans. Signal Process., 47 (1999), pp. 977–985.
[25] O. Ledoit and M. Wolf, *Honey, i shrunk the sample covariance matrix*, J. Portf. Manag., 30 (2004), pp. 110–119.
[26] O. Ledoit and M. Wolf, *A well-conditioned estimator for large-dimensional covariance matrices*, J. Multivariate Anal., 88 (2004), pp. 365–411.
[27] O. Ledoit and M. Wolf, *Non-linear shrinkage estimation of large-dimensional covariance matrices*, Ann. Statist., 40 (2012), pp. 1024–1060.
[28] O. Ledoit and M. Wolf, *Optimal estimation of a large-dimensional covariance matrix under Stein's loss*, Bernoulli, 24 (2018), pp. 3791–3832.
[29] C. Lenglet, M. Rousson, R. Deriche, O. Faugeras, S. Lehericy, and K. Ugurbil, *A Riemannian approach to diffusion tensor images segmentation*, in Proceedings of the Biennial International Conference on Information Processing in Medical Imaging (IPMI), Springer, Berlin, Heidelberg, 2005, pp. 591–602.
[30] A. S. Lewis, *Derivatives of spectral functions*, Math. Oper. Res., 21 (1996), pp. 576–588.
[31] G. J. Lord, C. E. Powell, and T. Shardlow, *An Introduction to Computational Stochastic PDEs*, Cambridge Texts Appl. Math. 50, Cambridge University Press, New York, 2014.
[32] V. A. Marčenko and L. A. Pastur, *Distribution of eigenvalues for some sets of random matrices*, Math. USSR-Sb., 1 (1967), pp. 457–483.
[33] H. Markowitz, *Portfolio selection*, J. Finance, 7 (1952), pp. 77–91.
[34] M. Moakher, *A differential geometric approach to the geometric mean of symmetric positive-definite matrices*, SIAM J. Matrix Anal. Appl., 26 (2005), pp. 735–747, https://doi.org/10.1137/S0895479803436937.
[35] M. Moakher and P. G. Batchelor, *Symmetric positive-definite matrices: From geometry to applications and visualization*, in Visualization and Processing of Tensor Fields, Springer, Berlin, 2006, pp. 285–298.
[36] K. Newey and D. McFadden, *Large sample estimation and hypothesis testing*, in Handbook of Econometrics, Vol. 4., R.F. Engle and D.L. McFadden, eds., North-Holland, Amsterdam, 1994, pp. 2111–2245.
[37] X. Pennec, P. Fillard, and N. Ayache, *A Riemannian framework for tensor computing*, Int. J. Comput. Vis., 66 (2006), pp. 41–66.
[38] C. R. Rao, *On the distance between two populations*, Sankhyā , 9 (1949), pp. 246–248.
[39] C. R. Rao, *Differential metrics in probability spaces*, Diff. Geom. Stat. Inference, 10 (1987), pp. 217–240.
[40] C. E. Rasmussen, *Gaussian Processes for Machine Learning*, MIT Press, Cambridge, MA, 2006.
[41] J. Schafer and K. Strimmer, *A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics*, Stat. Appl. Genet. Mol. Biol., 4 (2005), pp. 1175–1189.
[42] S. T. Smith, *Adaptive Radar*, Wiley Encyclopedia of Electrical and Electronics Engineering, 2001.

[43] S. T. Smith, *Covariance, subspace, and intrinsic Cramér-Rao bounds*, IEEE Trans. Signal Process., 53 (2005), pp. 1610–1630.
[44] C. Stein, *Estimation of a covariance matrix*, in 39th Annual Meeting IMS, Atlanta, GA, 1975.
[45] C. Stein, *Lectures on the theory of estimation of many parameters*, J. Sov. Math., 34 (1986), pp. 1373–1403.
[46] B. Vandereycken, P.-A. Absil, and S. Vandewalle, *A Riemannian geometry with complete geodesics for the set of positive-semidefinite matrices of fixed rank*, IMA J. Numer. Anal., 33 (2013), pp. 481–514.
[47] Y. Yuan, H. Zhu, W. Lin, and J. Marron, *Local polynomial regression for symmetric positive-definite matrices*, J. R. Stat. Soc. Ser. B. Stat. Methodol., 74 (2012), pp. 697–719.