Calculating Optimistic Likelihoods Using (Geodesically) Convex Optimization

Viet Anh Nguyen **Soroosh Shafieezadeh-Abadeh**
École Polytechnique Fédérale de Lausanne, Switzerland
{viet-anh.nguyen, soroosh.shafiee}@epfl.ch

Man-Chung Yue
The Hong Kong Polytechnic University, Hong Kong
manchung.yue@polyu.edu.hk

Daniel Kuhn
École Polytechnique Fédérale de Lausanne, Switzerland
daniel.kuhn@epfl.ch

Wolfram Wiesemann
Imperial College Business School, United Kingdom
ww@imperial.ac.uk

Abstract

A fundamental problem arising in many areas of machine learning is the evaluation of the likelihood of a given observation under different nominal distributions. Frequently, these nominal distributions are themselves estimated from data, which makes them susceptible to estimation errors. We thus propose to replace each nominal distribution with an ambiguity set containing all distributions in its vicinity and to evaluate an *optimistic likelihood*, that is, the maximum of the likelihood over all distributions in the ambiguity set. When the proximity of distributions is quantified by the Fisher-Rao distance or the Kullback-Leibler divergence, the emerging optimistic likelihoods can be computed efficiently using either geodesic or standard convex optimization techniques. We showcase the advantages of working with optimistic likelihoods on a classification problem using synthetic as well as empirical data.

1 Introduction

Assume that a set of i.i.d. data points $x_1^M \triangleq x_1, \ldots, x_M \in \mathbb{R}^n$ is generated from one of several Gaussian distributions $\mathbb{P}_c$, $c \in \mathcal{C}$ with $|\mathcal{C}| < \infty$. To determine the distribution $\mathbb{P}_{c^\star}$, $c^\star \in \mathcal{C}$, under which x_1^M has the highest likelihood $\ell(x_1^M, \mathbb{P}_c)$ across all $\mathbb{P}_c$, $c \in \mathcal{C}$, we can solve the problem

$$c^\star \in \operatorname*{arg\,max}_{c \in \mathcal{C}} \left\{ \ell(x_1^M, \mathbb{P}_c) \triangleq -\frac{1}{M} \sum_{m=1}^{M} (x_m - \mu_c)^\top \Sigma_c^{-1} (x_m - \mu_c) - \log\det \Sigma_c \right\}, \qquad (1)$$

where μ_c and Σ_c denote the means and covariance matrices that unambiguously characterize the distributions $\mathbb{P}_c$, $c \in \mathcal{C}$, and the log-likelihood function $\ell(x_1^M, \mathbb{P}_c)$ quantifies the (logarithm of the) relative probability of observing x_1^M under the Gaussian distribution $\mathbb{P}_c$. Problem (1) naturally arises in various machine learning applications. In quadratic discriminant analysis, for example,

Preprint. Under review.

x_1^M denotes the input values of data samples whose categorical outputs $y_1, \ldots, y_M \in \mathcal{C}$ are to be predicted based on the class-conditional distributions $\mathbb{P}_c$, $c \in \mathcal{C}$ [32]. Likewise, in Bayesian inference with synthetic likelihoods, a Bayesian belief about the models $\mathbb{P}_c$, $c \in \mathcal{C}$, assumed to be Gaussian for computational tractability, is performed based on an observation x_1^M [38, 50]. Problem (1) also arises in likelihood-ratio tests where the null hypothesis 'x_1^M is generated by a distribution $\mathbb{P}_c$, $c \in \mathcal{C}_0$' is compared with the alternative hypothesis 'x_1^M is generated by a distribution $\mathbb{P}_c$, $c \in \mathcal{C}_1$' [20, 21].

In practice, the parameters (μ_c, Σ_c) of the candidate distributions $\mathbb{P}_c$, $c \in \mathcal{C}$, are typically not known and need to be estimated from data. In quadratic discriminant analysis, for example, it is common to replace the means μ_c and covariance matrices Σ_c with their empirical counterparts $\hat{\mu}_c$ and $\hat{\Sigma}_c$ that are estimated from data. Similarly, the rival model distributions $\mathbb{P}_c$, $c \in \mathcal{C}$, in Bayesian inference with synthetic likelihoods are Gaussian estimates derived from (costly) sampling processes. Unfortunately, problem (1) is highly sensitive to misspecification of the candidate distributions $\mathbb{P}_c$. To combat this problem, we propose to replace the likelihood function in (1) with the *optimistic likelihood*

$$\max_{\mathbb{P} \in \mathcal{P}_c} \ell(x_1^M, \mathbb{P}) \;\text{ with }\; \mathcal{P}_c = \left\{ \mathbb{P} \in \mathcal{M} : \varphi(\hat{\mathbb{P}}_c, \mathbb{P}) \leq \rho_c \right\}, \tag{2}$$

where $\mathcal{M}$ is the set of all non-degenerate Gaussian distributions on $\mathbb{R}^n$, φ is a dissimilarity measure satisfying $\varphi(\mathbb{P}, \mathbb{P}) = 0$ for all $\mathbb{P} \in \mathcal{M}$, and $\rho_c \in \mathbb{R}_+$ are the radii of the ambiguity sets $\mathcal{P}_c$. Problem (2) assumes that the true candidate distributions $\mathbb{P}_c$ are unknown but close to the nominal distributions $\hat{\mathbb{P}}_c$ that are estimated from the training data. In contrast to the log-likelihood $\ell(x_1^M, \mathbb{P}_c)$ that is maximized in problem (1), the optimistic likelihood (2) is of interest in its own right. A common problem in constrained likelihood estimation, for example, is to determine a Gaussian distribution $\mathbb{P}^\star \sim (\mu^\star, \Sigma^\star)$ that is close to a nominal distribution $\mathbb{P}^0 \sim (\mu^0, \Sigma^0)$ reflecting the available prior information such that x_1^M has high likelihood under $\mathbb{P}^\star$ [41]. This task is an instance of the optimistic likelihood evaluation problem (2) with a suitably chosen dissimilarity measure φ.

Of crucial importance in the generalized likelihood problem (2) is the choice of the dissimilarity measure φ as it impacts both the statistical properties as well as the computational complexity of the estimation procedure. A natural choice appears to be the Wasserstein distance, which has recently been popularized in the field of optimal transport [46, 49]. The Wasserstein distance on the space of Gaussian distributions is a Riemannian distance, that is, the distance corresponding the curvilinear geometry on the set of Gaussian distributions induced by the Wasserstein distance, as opposed to the usual distance obtained by treating it as a subset of the space of symmetric matrices. However, since the Wasserstein manifold has a non-negative sectional curvature [46], calculating the associated optimistic likelihood (2) appears to be computationally intractable. Instead, we study the optimistic likelihood under the Fisher-Rao (FR) distance, which is commonly used in signal and image processing [3, 37] as well as computer vision [25, 48]. The FR distance is also a Riemannian metric, and it enjoys many attractive statistical properties that we review in Section 2 of this paper. Most importantly, the FR distance has a non-positive sectional curvature, which implies that the optimistic likelihood (2) reduces to the solution of a geodesically convex optimization problem that is amenable to an efficient solution [7, 43, 47, 51, 52, 53]. We also study problem (2) under the Kullback–Leibler (KL) divergence (or relative entropy), which is intimately related to the FR metric. While the KL divergence lacks some of the desirable statistical features of the FR metric, we will show that it gives rise to optimistic likelihoods that can be evaluated in quasi-closed form by reduction to a one dimensional problem.

While this paper focuses on the parametric approximation of the likelihood where $\mathbb{P}$ belongs to the family of Gaussian distributions, we emphasize that the optimistic likelihood approach can also be utilized in a *non*-parametric setting [35].

The contributions of this paper may be summarized as follows.

1. We show that for Fisher-Rao ambiguity sets, the optimistic likelihood (2) reduces to a geodesically convex problem and is hence amenable to an efficient solution via a Riemannian gradient descent algorithm. We analyze the optimality as well as the convergence of the resulting algorithm.
2. We show that for Kullback-Leibler ambiguity sets, the optimistic likelihood (2) can be evaluated in quasi-closed form by reduction to a one dimensional convex optimization problem.
3. We evaluate the numerical performance of our optimistic likelihoods on a classification problem with artificially generated as well as standard benchmark instances.

Our optimistic likelihoods follow a broader optimization paradigm that exercises optimism in the face of ambiguity. This strategy has been shown to perform well, among others, in multi-armed bandit problems and Bayesian optimization, where the Upper Confidence Bound algorithm takes decisions based on optimistic estimates of the reward [14, 15, 33, 45]. Optimistic optimization has also been successfully applied in support vector machines [10], and it closely relates to sparsity inducing non-convex regularization schemes [36].

The remainder of the paper proceeds as follows. We study the optimistic likelihood (2) under FR and KL ambiguity sets in Sections 2 and 3, respectively. We test our theoretical findings in the context of a classification problem, and we report on numerical experiments in Section 4. Supplementary material and all proofs are provided in the online companion.

Notation. Throughout this paper, $\mathbb{S}^n$, $\mathbb{S}_+^n$ and $\mathbb{S}_{++}^n$ denote the spaces of n-dimensional symmetric, symmetric positive semi-definite and symmetric positive definite matrices, respectively. For any $A \in \mathbb{R}^{n\times n}$, the trace of A is defined as $\mathrm{Tr}\left(A\right) = \sum_{i=1}^n A_{ii}$. For any $A \in \mathbb{S}^n$, $\lambda_{\min}(A)$ and $\lambda_{\max}(A)$ denote the minimum and maximum eigenvalues of A, respectively. The base of $\log(\cdot)$ is e.

2 Optimistic Likelihood Problems under the FR Distance

Consider a family of distributions with density functions $p_\theta(x)$, where the parameter θ ranges over a finite-dimensional smooth manifold Θ. At each point $\theta \in \Theta$, the Fisher information matrix $I_\theta = \mathbb{E}_x[\nabla_\theta \log(p_\theta(x))\nabla_\theta \log(p_\theta(x))^\top|\theta]$ defines an inner product $\langle\cdot,\cdot\rangle_\theta$ on the tangent space $T_\theta\Theta$ by $\langle\zeta_1,\zeta_2\rangle_\theta = \zeta_1^T I_\theta \zeta_2$ for $\zeta_1, \zeta_2 \in T_\theta\Theta$. The family of inner products $\{\langle\cdot,\cdot\rangle_\theta\}_{\theta\in\Theta}$ on the tangent spaces then defines a Riemannian metric, called the FR metric. The FR distance on Θ is the geodesic distance associated with the FR metric, *i.e.*, the FR distance between the two points $\theta_0, \theta_1 \in \Theta$ is

$$d(\theta_0,\theta_1) = \inf_\gamma \int_0^1 \sqrt{\langle\gamma'(t),\gamma'(t)\rangle_{\gamma(t)}}dt,$$

where the infimum is taken over all smooth curves $\gamma : [0,1] \to \Theta$ with $\gamma(0) = \theta_0$ and $\gamma(1) = \theta_1$. Any curve γ attaining the infimum is said to be a geodesic from θ_0 to θ_1. The FR metric represents a natural distance measure for parametric families of probability distributions as it is invariant under transformations on the data space (the x space) by a class of statistically important mappings, and it is the unique (up to a scaling) Riemannian metric enjoying such a property, see [6, 16, 18].

Since the covariance matrix is more difficult to estimate than the mean (see Appendix Appendix A), we focus here on the family of all Gaussian distributions with a fixed mean vector $\hat{\mu} \in \mathbb{R}^n$. These distributions are parameterized by $\theta = \Sigma$, that is, the covariance matrix. The parameter manifold is thus given by $\Theta = \mathbb{S}_{++}^n$. On this manifold, the FR distance is available in closed form.[1]

Proposition 2.1 (FR distance for Gaussian distributions [4])**.** If $\mathcal{N}(\hat{\mu},\Sigma_0)$ and $\mathcal{N}(\hat{\mu},\Sigma_1)$ are Gaussian distributions with identical mean $\hat{\mu} \in \mathbb{R}^n$ and covariance matrices $\Sigma_0, \Sigma_1 \in \mathbb{S}_{++}^n$, we have

$$d(\Sigma_0,\Sigma_1) = \frac{1}{\sqrt{2}}\left\|\log(\Sigma_1^{-\frac{1}{2}}\Sigma_0\Sigma_1^{-\frac{1}{2}})\right\|_F, \tag{3}$$

where $\log(\cdot)$ represents the matrix logarithm, and $\|\cdot\|_F$ stands for the Frobenius norm.

The distance $d(\cdot,\cdot)$ is invariant under *inversions* and *congruent transformations* of the input parameters [39, Proposition 1], *i.e.*, for any $\hat{\Sigma}, \Sigma \in \mathbb{S}_{++}^n$ and invertible matrix $A \in \mathbb{R}^{n\times n}$, we have

$$d(\hat{\Sigma}^{-1},\Sigma^{-1}) = d(\hat{\Sigma},\Sigma) \tag{4}$$

$$\text{and} \quad d(A\hat{\Sigma}A^\top, A\Sigma A^\top) = d(\hat{\Sigma},\Sigma). \tag{5}$$

By the inversion invariance (4), the distance $d(\cdot,\cdot)$ is independent of whether we use the covariance matrix Σ or the precision matrix Σ^{-1} to parametrize normal distributions. Note that if $x_1 \sim \mathcal{N}(\mu,\Sigma_1)$ and $x_2 \sim \mathcal{N}(\mu,\Sigma_2)$, then $Ax_1+b \sim \mathcal{N}(A\mu+b, A\Sigma_1A^\top)$ and $Ax_2+b \sim \mathcal{N}(A\mu+b, A\Sigma_2A^\top)$. By

[1] We can also handle the case where the covariance matrix is fixed but the mean is subject to ambiguity, see Appendix Appendix B. However, as there is no closed-form expression for the FR distance between two generic Gaussian distributions, we cannot handle the case where both the mean and the covariance matrix are subject to ambiguity.

the congruence invariance (5), the distance $d(\cdot,\cdot)$ thus remains unchanged under affine transformations $x \to Ax + b$. Remarkably, the invariance property (5) uniquely characterizes the distance $d(\cdot,\cdot)$. More precisely, any Riemannian distance satisfying the invariance property (5) coincides (up to a scaling) with the distance $d(\cdot,\cdot)$, see, for example, [40, Section 3] and [11, Section 2].

We now study the optimistic likelihood problem (2), where the FR distance is used as the dissimilarity measure. Given a data batch x_1^M and a radius $\rho > 0$, the optimistic likelihood problem reduces to

$$\min_{\Sigma \in \mathcal{B}^{\mathrm{FR}}} L(\Sigma), \quad \text{where} \quad \left\{ \begin{array}{l} L(\Sigma) \triangleq \left\langle S, \Sigma^{-1} \right\rangle + \log\det \Sigma, \\ \mathcal{B}^{\mathrm{FR}} \triangleq \{\Sigma \in \mathbb{S}_{++}^n : d(\Sigma, \hat{\Sigma}) \leq \rho\}, \end{array} \right. \tag{6}$$

and $S = M^{-1} \sum_{m=1}^M (x_m - \hat{\mu})(x_m - \hat{\mu})^\top$ stands for the sample covariance matrix.

We next prove that problem (6) is solvable, which justifies the use of the minimization operator.

Lemma 2.2. The optimal value of problem (6) is finite and is attained by some $\Sigma^\star \in \mathcal{B}^{\mathrm{FR}}$.

Even though the objective function of (6) involves a concave log-det term, it can be shown to be convex over the region $0 \prec \Sigma \preceq 2S$ [12, Exercise 7.4]. However, in practice S may be singular, in which case this region becomes empty. Maximum likelihood estimation problems akin to (6) are often reparameterized in terms of the precision matrix $X = \Sigma^{-1}$. In this case, (6) becomes

$$\min \left\{ \left\langle S, X \right\rangle - \log\det X : \; X \in \mathbb{S}_{++}^n, \; \| \log(X^{\frac{1}{2}} \hat{\Sigma} X^{\frac{1}{2}}) \|_F \leq \sqrt{2}\rho \right\}.$$

Even though this reparameterization convexifies the objective, it renders the feasible set non-convex.

2.1 Geodesic Convexity of the Optimistic Likelihood Problem

As problem (6) cannot be addressed with standard methods from convex optimization, we re-interpret it as a constrained minimization problem on the Riemannian manifold $\Theta = \mathbb{S}_{++}^n$ endowed with the FR metric. The key advantage of this approach is that we can show problem (6) to be *geodesically convex*. Geodesic convexity generalizes the usual notion of convexity in Euclidean spaces to Riemannian manifolds. We can thus solve problem (6) via algorithms from geodesically convex optimization, which inherit many benefits of the standard algorithms of convex optimization in Euclidean spaces.

The Riemannian manifold $\Theta = \mathbb{S}_{++}^n$ endowed with the FR metric is in fact a Hadamard manifold, that is, a complete simply connected Riemannian manifold with non-positive sectional curvature, see [27, Theorem XII 1.2]. Thus, any two points are connected by a *unique* geodesic [13]. By [9, Theorem 6.1.6], for $\Sigma_0, \Sigma_1 \in \mathbb{S}_{++}^n$, the unique geodesic $\gamma : [0,1] \to \mathbb{S}_{++}^n$ from Σ_0 to Σ_1 is given by

$$\gamma(t) = \Sigma_0^{\frac{1}{2}} \left(\Sigma_0^{-\frac{1}{2}} \Sigma_1 \Sigma_0^{-\frac{1}{2}} \right)^t \Sigma_0^{\frac{1}{2}}. \tag{7}$$

We are now ready to give precise definitions of geodesically convex sets and functions on Hadamard manifolds. We emphasize that these definitions would be more subtle for general Riemannian manifolds, which can have several geodesics between two points.

Definition 2.3 (Geodesically convex set)**.** A set $\mathcal{U} \subseteq \mathbb{S}_{++}^n$ is said to be geodesically convex if for all $\Sigma_0, \Sigma_1 \in \mathcal{U}$, the image of the unique geodesic from Σ_0 to Σ_1 is contained in $\mathcal{U}$, *i.e.*, $\gamma([0,1]) \subseteq \mathcal{U}$.

Definition 2.4 (Geodesically convex function)**.** A function $f : \mathbb{S}_{++}^n \to \mathbb{R}$ is said to be geodesically convex if for all $\Sigma_0, \Sigma_1 \in \mathbb{S}_{++}^n$, the unique geodesic γ from Σ_0 to Σ_1 satisfies $f(\gamma(t)) \leq (1-t) f(\Sigma_0) + t f(\Sigma_1)\ \forall t \in [0,1]$.

In order to prove that (6) is a geodesically convex optimization problem, we need to establish the geodesic convexity of the feasible region $\mathcal{B}^{\mathrm{FR}}$ and the loss function $L(\cdot)$. Note that, in stark contrast to Euclidean geometry, a geodesic ball on a general manifold may not be geodesically convex.[2]

Theorem 2.5 (Geodesic convexity of problem (6))**.** For any $\hat{\Sigma} \in \mathbb{S}_{++}^n$, $S \in \mathbb{S}_{+}^n$ and $\rho \in \mathbb{R}_+$, $\mathcal{B}^{\mathrm{FR}}$ is a geodesically convex set, and $L(\cdot)$ is a geodesically convex function over $\mathbb{S}_{++}^n$.

Theorem 2.5 establishes that the optimistic likelihood problem (6), which is non-convex with respect to the usual Euclidean geometry on the embedding space $\mathbb{R}^{n \times n}$, is actually convex with respect to the Riemannian geometry on $\mathbb{S}_{++}^n$ induced by the FR metric.

[2]For example, consider the circle $S^1 \triangleq \{x \in \mathbb{R}^2 : \|x\|_2 = 1\}$ which is a 1-dimensional manifold. Any major arc is a geodesic ball but *not* a geodesically convex subset of S^1.

Algorithm 1 Projected Geodesic Gradient Descent Algorithm

Input: $\hat{\Sigma} \succ 0$, $\rho > 0$, $S \succeq 0$, $K \in \mathbb{N}$, $\{\alpha_k\}_{k=1}^K \subseteq \mathbb{R}_{++}$
Initialization: Set $\Sigma_1 \leftarrow \hat{\Sigma}$, $\bar{\Sigma}_1 \leftarrow \hat{\Sigma}$
for $k = 1, 2, \ldots, K-1$ **do**
  Compute the Riemannian gradient at Σ_k: $G_k \leftarrow 2(\Sigma_k - S)$
  Perform a gradient descent step using the exponential map:

$$\Sigma_{k+\frac{1}{2}} \leftarrow \mathrm{Exp}_{\Sigma_k}(-\alpha_k G_k) = \Sigma_k^{\frac{1}{2}} \exp\big(-\alpha_k \Sigma_k^{-\frac{1}{2}} G_k \Sigma_k^{-\frac{1}{2}}\big)\Sigma_k^{\frac{1}{2}}$$

  Project $\Sigma_{k+\frac{1}{2}}$ onto $\mathcal{B}^{\mathrm{FR}}$: $\Sigma_{k+1} \leftarrow \mathrm{Proj}_{\mathcal{B}^{\mathrm{FR}}}(\Sigma_{k+\frac{1}{2}})$
  Compute the new iterate by interpolation: $\bar{\Sigma}_{k+1} \leftarrow \mathrm{Exp}_{\bar{\Sigma}_k}\big(\frac{1}{k+1}\mathrm{Exp}_{\bar{\Sigma}_k}^{-1}(\Sigma_{k+1})\big)$
end for
Output: Report the last iterate $\bar{\Sigma}_K$ as an approximate solution

2.2 Projected Geodesic Gradient Descent Algorithm

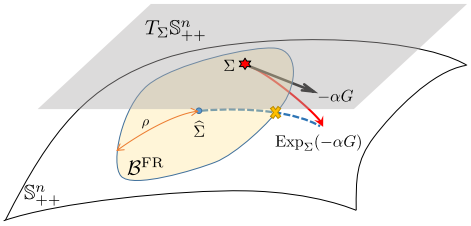


Figure 1: Visualization of the FR ball $\mathcal{B}^{\mathrm{FR}}$ (yellow set) within the manifold $\mathbb{S}^n_{++}$ (white set).

In the same way as the convexity of a standard constrained optimization problem can be exploited to find a global minimizer via a projected gradient descent algorithm, the geodesic convexity of problem (6) can be exploited to find a global minimizer by using a *projected geodesic gradient descent* algorithm of the type described in [52]. The mechanics of a generic iteration are visualized in Figure 1. As in any gradient descent method, given the current iterate Σ, we first need to compute the direction along which the objective function L decreases fastest. In the context of optimization on manifolds, this direction corresponds to the negative Riemannian gradient $-G$ at point Σ, which belongs to the tangent space $T_\Sigma \mathbb{S}^n_{++} \simeq \mathbb{S}^n$. Unfortunately, the curve $\gamma(\alpha) = \Sigma - \alpha G$ fails to be a geodesic and will eventually leave the manifold for sufficiently large step sizes α. This prompts us to construct the (unique) geodesic that emanates from point Σ with initial velocity $-G$. Formally, this geodesic can be represented as $\gamma(\alpha) = \mathrm{Exp}_\Sigma(-\alpha G)$, where $\mathrm{Exp}_\Sigma(\cdot)$ denotes the *exponential map* at Σ. As we will see below, this geodesic (visualized by the red curve) remains within the manifold for any $\alpha > 0$ but may eventually leave the feasible region $\mathcal{B}^{\mathrm{FR}}$. If this happens for the chosen step size α, we project $\mathrm{Exp}_\Sigma(-\alpha G)$ back onto the feasible region, that is, we map it to its closest point in $\mathcal{B}^{\mathrm{FR}}$ with respect to the FR distance (visualized by the yellow cross). Denoting this FR projection by $\mathrm{Proj}_{\mathcal{B}^{\mathrm{FR}}}(\cdot)$, the next iterate of the projected geodesic gradient descent algorithm can thus be expressed as $\Sigma^+ = \mathrm{Proj}_{\mathcal{B}^{\mathrm{FR}}}(\mathrm{Exp}_\Sigma(-\alpha G))$.

Starting from $\Sigma_1 = \hat{\Sigma}$, the proposed algorithm constructs K iterates $\{\Sigma_k\}_{k=1}^K$ via the above recursion. As in [52], the algorithm also constructs a second sequence $\{\bar{\Sigma}_k\}_{k=1}^K$ of feasible covariance matrices with $\bar{\Sigma}_1 = \hat{\Sigma}$ and $\bar{\Sigma}_{k+1} = \bar{\gamma}(1/(k+1))$ for $k = 1, \ldots, K-1$, where $\bar{\gamma}(t)$ represents the geodesic (7) connecting $\bar{\Sigma}_k$ with Σ_{k+1}. Thus, $\bar{\Sigma}_{k+1}$ is defined as a *geodesic convex combination* of $\bar{\Sigma}_k$ and Σ_{k+1}. A precise description of the proposed algorithm in pseudocode is provided in Algorithm 1.

In the following we show that the Riemannian gradient, the exponential map $\mathrm{Exp}_\Sigma(\cdot)$ as well as the projection $\mathrm{Proj}_{\mathcal{B}^{\mathrm{FR}}}(\cdot)$ can all be evaluated in closed form in $\mathcal{O}(n^3)$.

By [4, Page 362], the FR metric on the tangent space $T_\Sigma \mathbb{S}^n_{++}$ at $\Sigma \in \mathbb{S}^n_{++}$ can be re-expressed as

$$\big\langle \Omega_1, \Omega_2 \big\rangle_\Sigma \triangleq \frac{1}{2}\,\mathrm{Tr}\left(\Omega_1 \Sigma^{-1} \Omega_2 \Sigma^{-1}\right) \quad \forall \Omega_1, \Omega_2 \in T_\Sigma \mathbb{S}^n_{++}. \tag{8}$$

Using (8) and [1, Equation 3.32], the Riemannian gradient $G = \mathrm{grad}\, L$ of the objective function $L(\cdot)$ at Σ can be computed from the Euclidean gradient $\nabla L(\Sigma)$ as

$$\mathrm{grad}\ L(\Sigma) = 2\Sigma(\nabla L(\Sigma))\Sigma = 2\Sigma(\Sigma^{-1} - \Sigma^{-1} S \Sigma^{-1})\Sigma = 2(\Sigma - S). \tag{9}$$

Moreover, by [43, Equation (3.2)], the exponential map $\mathrm{Exp}_\Sigma : T_\Sigma \mathbb{S}^n_{++} \to \mathbb{S}^n_{++}$ at Σ is given by

$$\mathrm{Exp}_\Sigma(G) = \Sigma^{\frac{1}{2}} \exp(\Sigma^{-\frac{1}{2}} G \Sigma^{-\frac{1}{2}}) \Sigma^{\frac{1}{2}}, \quad G \in T_\Sigma \mathbb{S}^n_{++} \simeq \mathbb{S}^n,$$

where $\exp(\cdot)$ denotes the matrix exponential. The inverse map $\mathrm{Exp}_\Sigma^{-1} : \mathbb{S}^n_{++} \to T_\Sigma \mathbb{S}^n_{++}$ satisfies

$$\mathrm{Exp}_\Sigma^{-1}(A) = \Sigma^{\frac{1}{2}} \big(\log \Sigma^{-\frac{1}{2}} A \Sigma^{-\frac{1}{2}} \big) \Sigma^{\frac{1}{2}}, \quad A \in \mathbb{S}^n_{++}.$$

Finally, the projection $\mathrm{Proj}_{\mathcal{B}}(\cdot)$ onto $\mathcal{B}^{\mathrm{FR}}$ with respect to the FR distance is defined through

$$\mathrm{Proj}_{\mathcal{B}^{\mathrm{FR}}}(\Sigma') \triangleq \operatorname*{arg\,min}_{\Sigma \in \mathcal{B}^{\mathrm{FR}}} \; d(\Sigma, \Sigma'), \quad \Sigma' \in \mathbb{S}^n_{++}. \tag{10}$$

The following lemma ensures that this projection is well-defined and admits a closed-form expression.

Lemma 2.6 (Projection onto $\mathcal{B}^{\mathrm{FR}}$)**.** For any $\Sigma' \in \mathbb{S}^n_{++}$ with $d(\hat{\Sigma}, \Sigma') = \rho'$ the following hold.

(i) There $\arg\min$-mapping in (10) is a singleton, and thus $\mathrm{Proj}_{\mathcal{B}^{\mathrm{FR}}}(\Sigma')$ is well-defined.

(ii) The projection of Σ' onto $\mathcal{B}^{\mathrm{FR}}$ is given by

$$\mathrm{Proj}_{\mathcal{B}^{\mathrm{FR}}}(\Sigma') = \begin{cases} \hat{\Sigma}^{\frac{1}{2}} \big(\hat{\Sigma}^{-\frac{1}{2}} \Sigma' \hat{\Sigma}^{-\frac{1}{2}} \big)^{\frac{\rho}{\rho'}} \hat{\Sigma}^{\frac{1}{2}} & \text{if } \rho' > \rho, \\ \Sigma' & \text{otherwise.} \end{cases} \tag{11}$$

By comparison with (7), one easily verifies that $\mathrm{Proj}_{\mathcal{B}^{\mathrm{FR}}}(\Sigma')$ constitutes a geodesic convex combination between Σ' and $\hat{\Sigma}$. Figure 1 visualizes the geodesic from $\hat{\Sigma}$ to Σ' by the blue dashed line. Therefore, the projection $\mathrm{Proj}_{\mathcal{B}^{\mathrm{FR}}}$ onto the FR ball $\mathcal{B}^{\mathrm{FR}}$ within $\mathbb{S}^n_{++}$ endowed with the FR metric is constructed in a similar manner as the projection onto a Euclidean ball within a Euclidean space.

The following theorem asserts that Algorithm 1 enjoys a sublinear convergence rate.

Theorem 2.7 (Sublinear convergence rate)**.** With a constant stepsize

$$\alpha_k \equiv 2^{1/4} \sqrt{\rho \tanh(2\sqrt{2}\rho)} / (\Gamma \sqrt{K}),$$

where $\Gamma \triangleq 2^{-1/2} \sqrt{n} \cdot e^{2\sqrt{2}\rho} \cdot \lambda_{\min}^{-2}(\hat{\Sigma}) \cdot \max\{|1 - e^{\sqrt{2}\rho} \lambda_{\min}^{-1}(\hat{\Sigma}) \lambda_{\max}(S)|, 1\}$, Algorithm 1 satisfies

$$L(\bar{\Sigma}_K) - L(\Sigma^\star) \leq \frac{2^{\frac{7}{4}} \rho^{\frac{3}{2}} \Gamma}{\sqrt{K \tanh(2\sqrt{2}\rho)}}.$$

The proof of Theorem 2.7 closely follows that of [52, Theorem 9]. The difference is that [52, Theorem 9] requires the objective function to be Lipschitz continuous on $\mathbb{S}^n_{++}$. Unfortunately, such an assumption is not satisfied by $L(\cdot)$. We circumvent this by proving that the Riemannian gradient of $L(\cdot)$ is bounded uniformly on $\mathcal{B}^{\mathrm{FR}}$.

Endeavors are currently underway to devise algorithms for minimizing a geodesically strongly convex objective function over a geodesically convex feasible set that offer a linear convergence guarantee, see, e.g., [52, Theorem 15]. The next lemma shows that the objective function of problem (6) is indeed geodesically smooth and geodesically strongly convex[3] whenever $S \succ 0$. This suggests that the empirical performance of Algorithm 1 could be significantly better than the theoretical guarantee of Theorem 2.7. Indeed, our numerical results in Section 4.1 confirm that if $S \succ 0$, then Algorithm 1 displays a linear convergence rate.

Lemma 2.8 (Strong convexity and smoothness of $L(\cdot)$)**.** The objective function $L(\cdot)$ of problem (6) is geodesically β-smooth on $\mathcal{B}^{\mathrm{FR}}$ with

$$\beta = \frac{2\lambda_{\max}(S)}{\lambda_{\min}(\widehat{\Sigma}) \exp(-\sqrt{2}\rho)}.$$

If $S \succ 0$, then $L(\cdot)$ is also geodesically σ-strongly convex on $\mathcal{B}^{\mathrm{FR}}$ with

$$\sigma = \frac{2\lambda_{\min}(S)}{\lambda_{\max}(\widehat{\Sigma}) \exp(\sqrt{2}\rho)}.$$

Remark 2.9. Problem (6) could also be addressed with the algorithmic framework developed in [31]. Due to space limitations, we leave this for future research.

[3] The strong convexity and smoothness properties are defined in Definitions C.4 and C.5, respectively.

3 Generalized Likelihood Estimation under the KL Divergence

The KL divergence, which is widely used in information theory [19, § 2], can be employed as an alternative dissimilarity measure in the optimistic likelihood problem (2). If both $\hat{\mathbb{P}}$ and $\mathbb{P}$ are Gaussian distributions, then the KL divergence from $\hat{\mathbb{P}}$ to $\mathbb{P}$ admits an analytical expression.

Proposition 3.1 (KL divergence for Gaussian distributions)**.** For any $\hat{\mu} \in \mathbb{R}^n$ and $\Sigma_0, \Sigma_1 \in \mathbb{S}_{++}^n$, the KL divergence from $\mathbb{P}_0 = \mathcal{N}(\hat{\mu}, \Sigma_0)$ to $\mathbb{P}_1 = \mathcal{N}(\hat{\mu}, \Sigma_1)$ amounts to

$$\mathrm{KL}(\mathbb{P}_0 \,\|\, \mathbb{P}_1) = \frac{1}{2}\big(\mathrm{Tr}\left(\Sigma_1^{-1}\Sigma_0\right) + \log\det(\Sigma_1\Sigma_0^{-1}) - n\big).$$

Unlike the FR distance, the KL divergence is not symmetric. Proposition 3.1 implies that if the KL divergence is used as the dissimilarity measure, then the optimistic likelihood problem (2) reduces to

$$\min_{\Sigma \succ 0} \quad \left\{ \mathrm{Tr}\left(S\Sigma^{-1}\right) + \log\det\Sigma : \mathrm{Tr}\left(\Sigma^{-1}\hat{\Sigma}\right) + \log\det(\Sigma\hat{\Sigma}^{-1}) - n \leq 2\rho \right\}, \tag{12}$$

where $S = M^{-1}\sum_{m=1}^M (x_m - \hat{\mu})(x_m - \hat{\mu})^\top$ denotes again the sample covariance matrix. Because of the concave log-det terms in the objective and the constraints, problem (12) is non-convex. By using the variable substitution $X \leftarrow \Sigma^{-1}$, however, problem (12) can be reduced to a univariate convex optimization problem and thereby solved in quasi-closed form.

Theorem 3.2. For any $\hat{\Sigma} \succ 0$ and $\rho > 0$, the optimal value of problem (12) amounts to

$$(1+\gamma^\star)\,\mathrm{Tr}\left(S(S+\gamma^\star\hat{\Sigma})^{-1}\right) + \log\det(S+\gamma^\star\hat{\Sigma}) - n\log(1+\gamma^\star),$$

where $\gamma^\star$ is the unique optimal solution of the univariate convex optimization problem

$$\min_{\gamma>0} \left\{ \gamma(2\rho + \log\det\hat{\Sigma}) + n(1+\gamma)\log(1+\gamma) - (1+\gamma)\log\det(S+\gamma\hat{\Sigma}) \right\}. \tag{13}$$

Problem (13) can be solved efficiently using state-of-the-art first- or second-order methods, see Appendix Appendix E. However, in each iteration we still need to evaluate the determinant of a positive definite n-by-n matrix, which requires $\mathcal{O}(n^3)$ arithmetic operations. The following corollary shows that this computational burden can be alleviated when the sample covariance matrix S has low rank.

Corollary 3.3 (Singular sample covariance matrices)**.** If $S = \Lambda\Lambda^\top$ for some $\Lambda \in \mathbb{R}^{n\times k}$ and $k \in \mathbb{N}$ with $k < n$, then problem (13) simplifies to

$$\min_{\gamma>0} \left\{ 2\gamma\rho + n(1+\gamma)\log(1+\gamma) - (n-k)(1+\gamma)\log\gamma - (1+\gamma)\log\det(\gamma I_k + \Lambda^\top\hat{\Sigma}^{-1}\Lambda) \right\}.$$

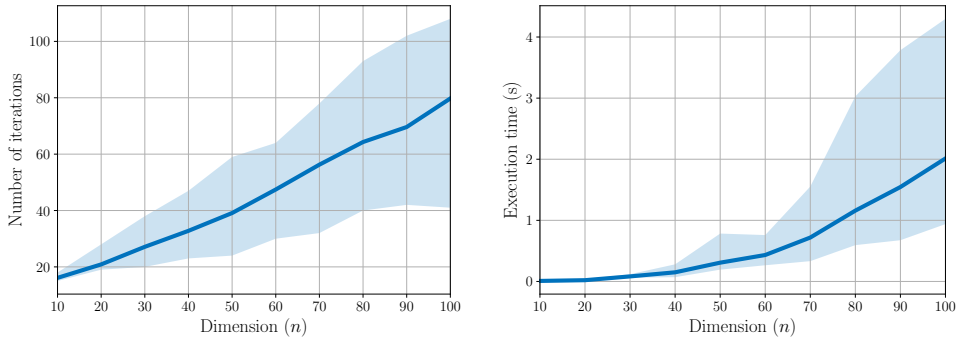
We will see that for classification problems the matrix S has rank 1, in which case the log-det term in the above univariate convex program reduces to the scalar logarithm. In Appendix Appendix E we provide explicit first- and second-order derivatives of the objective of problem (13) and its simplification.

4 Numerical Results

We investigate the empirical behavior of our projected geodesic gradient descent algorithm (Section 4.1) and the predictive power of our flexible discriminant rules (Section 4.2). Our algorithm and all tests are implemented in Python, and the source code is available from `https://github.com/sorooshafiee/Optimistic_Likelihoods`.

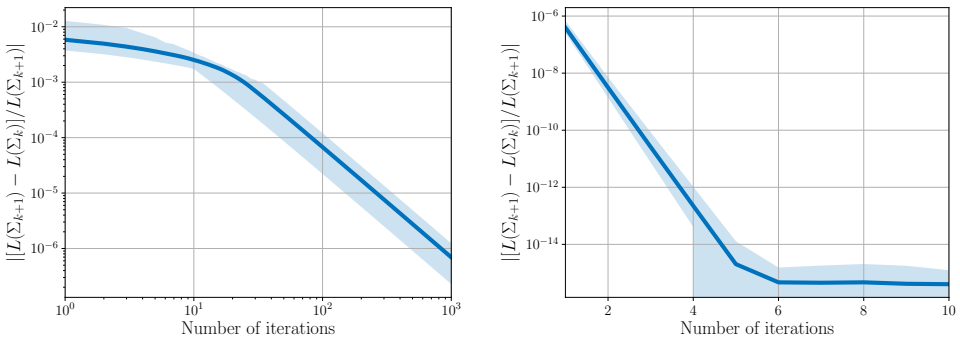
4.1 Convergence Behavior of the Projected Geodesic Descent Algorithm

To study the empirical convergence behavior of Algorithm 1, for $n \in \{10, 20, \ldots, 100\}$ we generate 100 covariance matrices $\hat{\Sigma}$ according to the following procedure. We *(i)* draw a standard normal random matrix $B \in \mathbb{R}^{n\times n}$ and compute $A = B + B^\top$; we *(ii)* conduct an eigenvalue decomposition $A = RDR^T$; we *(iii)* replace D with a random diagonal matrix $\hat{D}$ whose diagonal elements are



(a) Scaling of iteration count for $S \succeq 0$

(b) Scaling of execution time for $S \succeq 0$



(c) Convergence for $n = 100$ for $S \succeq 0$

(d) Convergence for $n = 100$ for $S \succ 0$

Figure 2: Convergence behavior of the projected geodesic gradient descent algorithm. Solid lines (shaded regions) represent averages (ranges) across 100 independent simulations.

sampled uniformly from $[1, 10]^n$; and we *(iv)* set $\hat{\Sigma} = R\hat{D}R^\top$. For each of these covariance matrices, we set $\hat{\mu} = 0$, $M = 1$, $x_1^M \triangleq x$ for a standard normal random vector $x \in \mathbb{R}^n$ and calculate the optimistic likelihood (6) for $\rho = \sqrt{n}/100$. This choice of ρ ensures that the radius of the ambiguity set scales with n in the same way as the Frobenius norm. Figures 2(a) and 2(b) report the number of iterations as well as the overall execution time of Algorithm 1 when we terminate the algorithm as soon as the relative improvement $|[L(\Sigma_{k+1}) - L(\Sigma_k)]/L(\Sigma_{k+1})|$ drops below 0.01%. Notice that the number of required iterations scales linearly with n while the overall runtime grows polynomially with n. Figure 2(c) shows the relative improvement as a function of the iteration count. Empirically, the number of iterations scales with $\mathcal{O}(1/k^2)$, which is faster than the theoretical rate established in Theorem 2.7. We also study the empirical convergence behavior of Algorithm 1 when the input matrix S is positive definite. We repeat the first experiment with $M = 100$, and we set $S = \delta I + \sum_{i=1}^M x_i x_i^\top / M$ for $\delta = 10^{-6}$ to ensure that S is positive definite. Figure 2(d) indicates that, in this case, the empirical convergence rate of Algorithm 1 is linear.

4.2 Application: Flexible Discriminant Rules

Consider a classification problem where a categorical response $Y \in \mathcal{C}, \mathcal{C} = \{1, \ldots, C\}$, should be predicted from continuous inputs $X \in \mathbb{R}^n$. In this context, Bayes' Theorem implies that $\mathbb{P}(Y = c|X = x) \propto \pi_c \cdot f_c(x)$, $c \in \mathcal{C}$, where $\pi_c = \mathbb{P}(Y = c)$ denotes the prior probability of the response belonging to class c, and f_c is the density function of X for an observation of class c. In practice, π_c and f_c are unknown and need to be estimated from a training data set $(\hat{x}_1, \hat{y}_1), \ldots, (\hat{x}_N, \hat{y}_N)$. Assuming that the densities f_c, $c \in \mathcal{C}$, correspond to Gaussian distributions with (unknown) class-specific means μ_c and covariance matrices Σ_c, the quadratic discriminant analysis (QDA) replaces π_c with $\hat{\pi}_c = N_c/N$, where $N_c = |\{i : \hat{y}_i = c\}|$, and f_c with the density of the Gaussian distribution $\hat{\mathbb{P}}_c \sim \mathcal{N}(\hat{\mu}_c, \hat{\Sigma}_c)$, whose mean and covariance matrix are estimated from the training data, to classify

Table 1: Average correct classification rates on the benchmark instances

	FQDA	KQDA	QDA	RQDA	SQDA	WQDA
Australian	80.68	**83.68**	80.03	79.76	80.73	79.94
Banknote authentication	99.07	**99.47**	98.56	98.54	98.53	98.54
Climate model	94.46	**94.55**	91.78	92.72	94.42	92.78
Cylinder	70.69	70.67	67.10	70.33	**70.99**	70.34
Diabetic	**75.97**	74.53	74.19	74.60	74.70	75.04
Fourclass	**80.13**	79.97	79.32	79.32	79.32	79.33
German credit	74.50	74.60	71.41	**76.18**	74.99	76.31
Haberman	74.87	**75.41**	74.92	74.96	75.04	74.97
Heart	**84.23**	83.31	81.42	82.62	84.17	82.42
Housing	88.89	**92.90**	88.54	87.01	81.69	88.31
Ilpd	57.42	**57.83**	55.18	54.97	55.45	55.15
Mammographic mass	80.66	80.85	80.37	80.88	**81.05**	80.65
Pima	**75.97**	74.53	74.19	74.60	74.70	75.04
Ringnorm	**98.69**	98.65	98.56	98.56	98.65	98.56

a new observation x using the discriminant rule

$$\mathcal{C}_{\text{QDA}}(x) \in \arg\max_{c\in\mathcal{C}} \left\{ \frac{1}{2}\ell(x, \hat{\mathbb{P}}_c) + \log(\hat{\pi}_c) \right\}.$$

Here, the likelihood $\ell(x, \hat{\mathbb{P}}_c)$ is defined as in (1) for $M = 1$. If $\hat{\pi}_1 = \ldots = \hat{\pi}_C$, this classification rule reduces to the maximum likelihood discrimant rule [24, § 14].

QDA can be sensitive to misspecifications of the empirical moments. To reduce this sensitivity, we replace the nominal Gaussian distributions $\hat{\mathbb{P}}_c$ with the Gaussian distributions $\mathbb{P}_c^\star$ that would have generated the sample x with highest likelihood, among all Gaussian distributions in the vicinity of the nominal distributions $\hat{\mathbb{P}}_c$. This results in a *flexible discriminant rule* of the form

$$\mathcal{C}_{\text{flex}}(x) \in \arg\max_{c\in\mathcal{C}} \max_{\mathbb{P}\in\mathcal{P}_c} \left\{ \frac{1}{2}\ell(x, \mathbb{P}) + \log(\hat{\pi}_c) \right\},$$

which makes use of the optimistic likelihoods (2). Here, $\mathcal{P}_c$ is the FR or KL ball centered at the nominal distribution $\hat{\mathbb{P}}_c$. To ensure that $\hat{\Sigma}_c \succ 0$ for all $c \in \mathcal{C}$, we use the Ledoit-Wolf covariance estimator [28], which is parameter-free and returns a well-conditioned matrix by minimizing the mean squared error between the estimated and the real covariance matrix.

We compare the performance of our flexible discriminant rules with standard QDA implementations from the literature on datasets from the UCI repository [5]. Specifically, we compare the following methods.

- **FQDA** and **KQDA**: our flexible discriminant rules based on FR (FQDA) and KL (KQDA) ambiguity sets with radii ρ_c;
- **QDA**: regular QDA with empirical means and covariance matrices estimated from data;
- **RQDA**: regularized QDA based on the linear shrinkage covariance estimator $\hat{\Sigma}_c + \rho_c I_n$;
- **SQDA**: sparse QDA based on the graphical lasso covariance estimator [23] with parameter ρ_c;
- **WQDA**: Wasserstein QDA based on the nonlinear shrinkage approach [34] with parameter ρ_c.

All results are averaged across 100 independent trials for $\rho_c \in \{a\sqrt{n} \cdot 10^b : a \in \{1, \ldots, 9\},\ b \in \{-3, -2, -1\}\}$. In each trial, we randomly select 75% of the data for training and the remaining 25% for testing. The size of the ambiguity set and the regularization parameter are selected using stratified 5-fold cross validation. The performance of the classifiers is measured by the *correct classification rate* (CCR). The average CCR scores over the 100 trials are reported in Table 1.

Appendix A Justification for using Ambiguity Sets with Fixed Mean Vector

We provide some empirical evidence to justify the ambiguity sets with a fixed mean vector used in Sections 2 and 3. Towards this end, we first fix a matrix $A \in \mathbb{R}^{n \times n}$, where each element is drawn independently from a standard Gaussian distribution, and set $\Sigma \triangleq AA^\top$. We then generate $N \in \{20, \ldots, 100\}$ i.i.d. samples $\hat{x}_1, \ldots, \hat{x}_N$ from the Gaussian distribution $\mathbb{Q} = \mathcal{N}(0, \Sigma)$, and compute the empirical mean $\hat{\mu}_N$ and the empirical covariance matrix $\hat{\Sigma}_N$ as

$$\hat{\mu}_N = \frac{1}{N} \sum_{i=1}^{N} \hat{x}_i \quad \text{and} \quad \hat{\Sigma}_N = \frac{1}{N-1} \sum_{i=1}^{N} (\hat{x}_i - \hat{\mu}_N)(\hat{x}_i - \hat{\mu}_N)^\top.$$

We now construct two probability distributionss based on $\hat{\mu}_N$ and $\hat{\Sigma}_N$, that is, we set

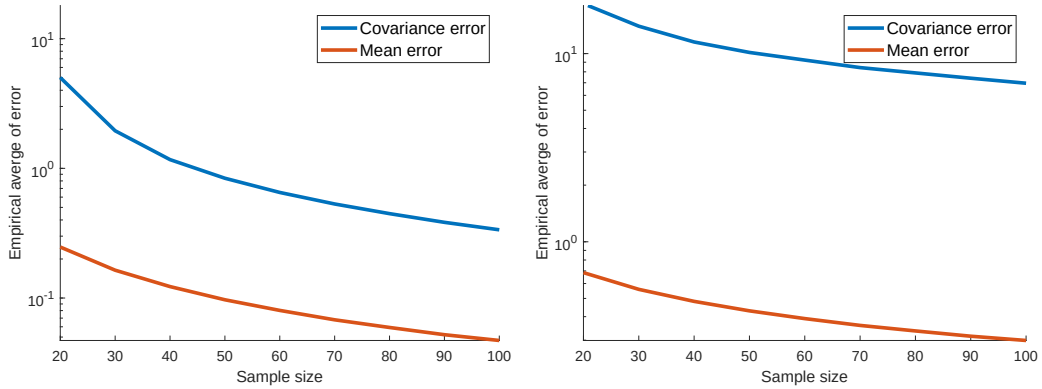
$$\hat{\mathbb{P}}_{N,0} = \mathcal{N}(0, \hat{\Sigma}_N) \quad \text{and} \quad \hat{\mathbb{P}}_{N,\Sigma} = \mathcal{N}(\hat{\mu}_N, \Sigma).$$

Notice that $\hat{\mathbb{P}}_{N,0}$ has the same mean as the unknown probability distribution $\mathbb{Q}$ that generates the data, while $\hat{\mathbb{P}}_{N,\Sigma}$ has the same covariance matrix as $\mathbb{Q}$. In the following, we define the mean vector estimation error $\delta_{N,\text{mean}}$ and the covariance matrix estimation error $\delta_{N,\text{covariance}}$ as

$$\delta_{N,\text{mean}} = \varphi(\hat{\mathbb{P}}_{N,\Sigma}, \mathbb{Q}) \quad \text{and} \quad \delta_{N,\text{cov}} = \varphi(\hat{\mathbb{P}}_{N,0}, \mathbb{Q}),$$

respectively, where $\varphi(\cdot, \cdot)$ is a dissimilarity measure for distributions that can be set either to the Fisher-Rao metric (see Section 2) or to the Kullback-Leibler divergence (see Section 3).

Figure A.3 shows the average estimation error for different sample sizes N, where the average is taken over 500 independent simulation runs. We observe that the error in estimating the covariance matrix is one order of magnitude higher than the error in estimating the mean vector under both the KL divergence and the FR metric.



(a) KL divergence (b) FR metric

Figure A.3: Average estimation error for different sample sizes N using the KL divergence or the FR metric as a dissimilarity measure.

Appendix B Optimistic Likelihood with Ambiguous Mean Vector

We now consider the FR and KL ambiguity sets for the family of Gaussian distributions with a fixed covariance matrix $\hat{\Sigma} \in \mathbb{S}_{++}^n$. We thus consider the manifold $\Theta = \mathbb{R}^n$ of the mean vector $\theta = \mu$. The FR distance induced by the FR metric on this manifold is denoted by $\bar{d}(\cdot, \cdot)$ and is again available in closed form.

Proposition B.1 (FR distance between Gaussian distributions [4])**.** If $\mathcal{N}(\mu_0, \hat{\Sigma})$ and $\mathcal{N}(\mu_1, \hat{\Sigma})$ are Gaussian distributions with identical covariance matrix $\hat{\Sigma} \in \mathbb{S}_{++}^n$ and mean vectors $\mu_0, \mu_1 \in \mathbb{R}^n$, we have

$$\bar{d}(\mu_0, \mu_1) = \sqrt{(\mu_0 - \mu_1)^\top \hat{\Sigma}^{-1} (\mu_0 - \mu_1)}.$$

Similarly, the KL divergence between two distributions with the same covariance matrix admits a simple analytical expression.

Proposition B.2 (KL divergence between Gaussian distributions)**.** For any $\hat{\Sigma} \in \mathbb{S}_{++}^n$ and $\mu_0, \mu_1 \in \mathbb{R}^n$, the KL divergence from $\mathbb{P}_0 = \mathcal{N}(\mu_0, \hat{\Sigma})$ to $\mathbb{P}_1 = \mathcal{N}(\mu_1, \hat{\Sigma})$ amounts to

$$\mathrm{KL}(\mathbb{P}_0 \parallel \mathbb{P}_1) = \frac{1}{2}(\mu_0 - \mu_1)^\top \hat{\Sigma}^{-1}(\mu_0 - \mu_1).$$

Throughout this section, we denote by $\hat{\mathbb{P}} = \mathcal{N}(\hat{\mu}, \hat{\Sigma})$ and $\mathbb{P} = (\mu, \hat{\Sigma})$ two Gaussian distributions with the same covariance matrix $\hat{\Sigma} \in \mathbb{S}_{++}^n$ but different mean vectors $\hat{\mu}, \mu \in \mathbb{R}^n$, respectively. Propositions B.1 and B.2 imply that the FR distance and the KL divergence of $\hat{\mathbb{P}}$ and $\mathbb{P}$ satisfy the relation[4] $2\,\mathrm{KL}(\hat{\mathbb{P}} \parallel \mathbb{P}) = \bar{d}^2(\hat{\mu}, \mu)$.

With the Fisher-Rao distance as our dissimilarity measure φ and given the observations x_1^M, the optimistic likelihood problem (2) becomes

$$\min_{\mu} \left\{ \frac{1}{M} \sum_{m=1}^{M} (x_m - \mu)^\top \hat{\Sigma}^{-1}(x_m - \mu) + \log\det \hat{\Sigma} : \ (\mu - \hat{\mu})^\top \hat{\Sigma}^{-1}(\mu - \hat{\mu}) \leq \rho^2 \right\}. \tag{B.1}$$

Problem (B.1) is already a finite convex program but can be further simplified to a univariate convex optimization problem and therefore solved in quasi-closed form.

Theorem B.3 (Optimistic likelihood with mean ambiguity set)**.** For any $\hat{\mu} \in \mathbb{R}^n, \hat{\Sigma} \in \mathbb{S}_{++}^n$ and $\rho > 0$, the optimal value of problem (B.1) is given by

$$\frac{1}{M} \sum_{m=1}^{M} (x_m - \mu^\star)^\top \hat{\Sigma}^{-1}(x_m - \mu^\star) + \log\det \hat{\Sigma},$$

where $\mu^\star = (1 + \gamma^\star)^{-1} (\bar{x} + \gamma^\star \hat{\mu})$, and $\gamma^\star$ solves the univariate convex optimization problem

$$\min_{\gamma \geq 0} \ \gamma \left(\rho^2 - \hat{\mu}^\top \hat{\Sigma}^{-1} \hat{\mu} \right) + \frac{(\bar{x} + \gamma\hat{\mu})^\top \hat{\Sigma}^{-1} (\bar{x} + \gamma\hat{\mu})}{1 + \gamma} \tag{B.2}$$

with $\bar{x} = M^{-1} \sum_{m=1}^{M} x_m$.

Proof. As $\hat{\Sigma}$ is constant, the minimizers of (B.1) also solve

$$\min_{\mu} \left\{ M^{-1} \sum_{m=1}^{M} (x_m - \mu)^\top \hat{\Sigma}^{-1}(x_m - \mu) : \ (\mu - \hat{\mu})^\top \hat{\Sigma}^{-1}(\mu - \hat{\mu}) \leq \rho^2 \right\}. \tag{B.3}$$

Problem (B.3) is equivalent to

$$\begin{aligned} &\min_{\mu} \max_{\gamma \geq 0} \left\{ \ \left\langle \hat{\Sigma}^{-1}, M^{-1} \sum_{m=1}^{M} (\mu - x_m)(\mu - x_m)^\top \right\rangle + \gamma \left(\left\langle \hat{\Sigma}^{-1}, (\mu - \hat{\mu})(\mu - \hat{\mu})^\top \right\rangle - \rho^2 \right) \ \right\} \\ &= \max_{\gamma \geq 0} \min_{\mu} \left\{ \ -\gamma\rho^2 + \left\langle \hat{\Sigma}^{-1}, M^{-1} \sum_{m=1}^{M} (\mu - x_m)(\mu - x_m)^\top + \gamma(\mu - \hat{\mu})(\mu - \hat{\mu})^\top \right\rangle \ \right\}, \end{aligned}$$

where the equality follows from strong duality, which holds because $\rho > 0$ and because $\mu = \hat{\mu}$ constitutes a Slater point for the primal problem (B.3). For any fixed $\gamma \geq 0$, the inner minimization problem over μ admits the optimal solution

$$\mu^\star(\gamma) = (1 + \gamma)^{-1} (\bar{x} + \gamma\hat{\mu})$$

[4]More generally, for arbitrary parametric families of distributions, the second-order Taylor expansion of the KL divergence is given by the FR distance because $\mathrm{KL}(\hat{\mathbb{P}} \parallel \mathbb{P}) = \frac{1}{2}\,\mathrm{FR}^2(\hat{\mathbb{P}}, \mathbb{P}) + \mathcal{O}(\mathrm{FR}^3(\hat{\mathbb{P}}, \mathbb{P}))$. See [26, § 7.2.2] for further details.

with $\bar{x} = M^{-1} \sum_{m=1}^{M} x_m$. Thus, the optimal value of (B.3) equals

$$\max_{\gamma \geq 0} \ \gamma \left(\hat{\mu}^\top \hat{\Sigma}^{-1} \hat{\mu} - \rho^2 \right) - \frac{(\bar{x} + \gamma \hat{\mu})^\top \hat{\Sigma}^{-1} (\bar{x} + \gamma \hat{\mu})}{1 + \gamma},$$

which is equivalent to the minimization problem (B.2). By strong duality, given any minimizer $\gamma^\star$ of problem (B.2), an optimal solution for (B.3) and also for (B.1) can be constructed as

$$\mu^\star = (1 + \gamma^\star)^{-1} (\bar{x} + \gamma^\star \hat{\mu}).$$

Substituting $\mu^\star$ into the objective function of (B.1) yields the postulated optimal value. □

In the following, we provide the first- and second-order derivatives of the objective function of (B.2), which can be used for implementing the optimization algorithm to solve for $\gamma^\star$. To this end, we denote by $g(\gamma)$ the objective function of (B.2). A direct calculation shows that

$$g'(\gamma) = \left(\rho^2 - \hat{\mu}^\top \hat{\Sigma}^{-1} \hat{\mu} \right) + \frac{((2 + \gamma)\hat{\mu} - \bar{x})^\top \hat{\Sigma}^{-1} (\bar{x} + \gamma \hat{\mu})}{(1 + \gamma)^2}.$$

Moreover, the second-order derivative of $g(\gamma)$ is given by

$$g''(\gamma) = \frac{2 \hat{\mu}^\top \hat{\Sigma}^{-1} \hat{\mu}}{(1 + \gamma)} - \frac{2[(2 + \gamma)\hat{\mu} - \bar{x}]^\top \hat{\Sigma}^{-1} (\bar{x} + \gamma \hat{\mu})}{(1 + \gamma)^3}.$$

Appendix C Proofs of Section 2

To prove Lemma 2.2, we require the following preparatory lemma.

Lemma C.1 (Properties of $\mathcal{B}^{\mathrm{FR}}$)**.** The FR ball has the following properties:

(i) $\mathcal{B}^{\mathrm{FR}}$ is compact and complete on $\mathbb{S}^n_{++}$.

(ii) For any $\Sigma \in \mathcal{B}^{\mathrm{FR}}$, we have $\lambda_{\min}(\hat{\Sigma}) e^{-\sqrt{2}\rho} \cdot I_n \preceq \Sigma \preceq \lambda_{\max}(\hat{\Sigma}) e^{\sqrt{2}\rho} \cdot I_n$.

Proof. To prove assertion (i), we first show that $\mathcal{B}^{\mathrm{FR}}$ is compact and complete with respect to the topology induced by the Riemannian distance $d(\cdot, \cdot)$. Recall that $\mathbb{S}^n_{++}$ is a Hadamard manifold and thus constitutes a complete metric space. By the Hopf-Rinow theorem [17, § 8, Theorem 2.8(b)], $\mathcal{B}^{\mathrm{FR}}$ is compact in the usual topology because $\mathcal{B}^{\mathrm{FR}}$ is a metric ball and therefore closed and bounded. Moreover, $\mathcal{B}^{\mathrm{FR}}$ is complete in the usual topology because any closed subset of a complete metric space is complete as well. By [30, Theorem 13.29], the metric topology with respect to $d(\cdot, \cdot)$ on $\mathbb{S}^n_{++}$ coincides with the subspace topology of $\mathbb{S}^n_{++}$ with respect to the usual topology on $\mathbb{S}^n$. This completes the proof of assertion (i).

To prove assertion (ii), pick any $\Sigma \in \mathcal{B}^{\mathrm{FR}}$ and let $0 \leq \lambda_1(A) \leq \ldots \leq \lambda_n(A)$ denote the eigenvalues of any symmetric positive definite n-by-n matrix A in increasing order. Then, we have

$$\sqrt{\log^2(\lambda_i(\hat{\Sigma}^{-\frac{1}{2}} \Sigma \hat{\Sigma}^{-\frac{1}{2}}))} \leq \sqrt{\sum_{j=1}^{n} \log^2(\lambda_j(\hat{\Sigma}^{-\frac{1}{2}} \Sigma \hat{\Sigma}^{-\frac{1}{2}}))} = \| \log(\hat{\Sigma}^{-\frac{1}{2}} \Sigma \hat{\Sigma}^{-\frac{1}{2}}) \|_F \leq \sqrt{2}\rho$$

for any $i = 1, \ldots, n$, where the equality follows from the definition of the Frobenius norm, and the last inequality follows from the definition of $\mathcal{B}^{\mathrm{FR}}$. Note that $\lambda_i(\hat{\Sigma}^{-\frac{1}{2}} \Sigma \hat{\Sigma}^{-\frac{1}{2}}) = 1/\lambda_{n-i+1}(\Sigma^{-\frac{1}{2}} \hat{\Sigma} \Sigma^{-\frac{1}{2}})$, and hence any eigenvalue $\lambda_i(\hat{\Sigma}^{-\frac{1}{2}} \Sigma \hat{\Sigma}^{-\frac{1}{2}})$ obeys the bounds

$$e^{-\sqrt{2}\rho} \leq \lambda_i(\hat{\Sigma}^{-\frac{1}{2}} \Sigma \hat{\Sigma}^{-\frac{1}{2}}) \leq e^{\sqrt{2}\rho}.$$

This implies that

$$\lambda^{-1}_{\max}(\hat{\Sigma}) \lambda_{\max}(\Sigma) \leq e^{\sqrt{2}\rho} \quad \text{and} \quad \lambda^{-1}_{\min}(\hat{\Sigma}) \lambda_{\min}(\Sigma) \geq e^{-\sqrt{2}\rho},$$

which completes the proof of assertion (ii). □

We are now ready to prove Lemma 2.2.

Proof of Lemma 2.2. First, assertion (i) of Lemma C.1 ensures that the feasible region $\mathcal{B}^{\mathrm{FR}}$ is compact. Second, we note that the objective function $L(\cdot)$ is continuous at any positive definite matrix. By assertion (ii) of Lemma C.1, there is a uniform positive lower bound on the eigenvalues of all matrices in $\mathcal{B}^{\mathrm{FR}}$. Therefore $L(\cdot)$ is continuous on $\mathcal{B}^{\mathrm{FR}}$. The solvability of problem (6) then follows from Weierstrass' extreme value theorem [2, Corollary 2.35]. □

Proof of Theorem 2.5. We first show that $\mathcal{B}^{\mathrm{FR}}$ is a geodesically convex set. By [13, Proposition II.1.4], balls in CAT(κ) spaces[5] of radius less than $D_\kappa/2$ are geodesically convex, where D_κ is the diameter of the model space of constant curvature κ (see [13, Definition I.2.10]). It is known that the smooth manifold $\mathbb{S}^n_{++}$ is a CAT(0) space [13, Theorem II.10.39], which implies via [13, Point I.2.12] that $D_0 = \infty$. The claim thus follows.

The proof that $L(\cdot)$ is a geodesically convex function over $\mathbb{S}^n_{++}$ closely follows from [54, Lemma III.2] and [43, Corollary 5.3] and is thus omitted. □

Proof of Lemma 2.6. The claim trivially holds if $\rho' \leq \rho$. We thus prove the two statements under the assumption that $\rho' > \rho$. By [13, Theorem II.10.39], $\mathbb{S}^n_{++}$ is a CAT(0) space. Furthermore, by Lemma C.1, the geodesic ball $\mathcal{B}^{\mathrm{FR}}$ is both complete and compact. Assertion (i) in Lemma 2.6 then follows from [13, Proposition II.2.4].

To prove assertion (ii), we define

$$\Sigma_p \triangleq \hat{\Sigma}^{\frac{1}{2}}(\hat{\Sigma}^{-\frac{1}{2}}\Sigma'\hat{\Sigma}^{-\frac{1}{2}})^{\frac{\rho}{\rho'}}\hat{\Sigma}^{\frac{1}{2}}.$$

One readily verifies that $d(\hat{\Sigma}, \Sigma_p) = \rho$, and hence $\Sigma_p \in \mathcal{B}^{\mathrm{FR}}$. Recall that $\Sigma' \in \mathbb{S}^n_{++}$ and $d(\hat{\Sigma}, \Sigma') = \rho'$. Given any $\Sigma \in \mathcal{B}^{\mathrm{FR}}$, by the triangle inequality, we thus have

$$d(\Sigma, \Sigma') \geq d(\hat{\Sigma}, \Sigma') - d(\hat{\Sigma}, \Sigma) \geq d(\hat{\Sigma}, \Sigma') - \max_{\Sigma'' \in \mathcal{B}^{\mathrm{FR}}} d(\hat{\Sigma}, \Sigma'') = \rho' - \rho.$$

This reasoning implies that

$$\min_{\Sigma \in \mathcal{B}^{\mathrm{FR}}} d(\Sigma, \Sigma') \geq \rho' - \rho. \tag{C.1}$$

By definition, the geodesic $\gamma(t) = \hat{\Sigma}^{\frac{1}{2}}(\hat{\Sigma}^{-\frac{1}{2}}\Sigma'\hat{\Sigma}^{-\frac{1}{2}})^t\hat{\Sigma}^{\frac{1}{2}}$ connecting $\hat{\Sigma}$ and Σ' has constant-speed, that is, $d(\gamma(t), \gamma(s)) = d(\gamma(0), \gamma(1)) \cdot |t - s|$ for any $t, s \in [0, 1]$ (see [9, Theorem 6.1.6]). Therefore, we have

$$d(\Sigma_p, \Sigma') = d(\gamma(\tfrac{\rho}{\rho'}), \gamma(1)) = d(\hat{\Sigma}, \Sigma') \cdot \left| \frac{\rho}{\rho'} - 1 \right| = \rho' - \rho,$$

which implies that the lower bound (C.1) is attained by Σ_p. The uniqueness result of assertion (i) thus allows us to conclude that Σ_p is the projection of Σ' onto $\mathcal{B}^{\mathrm{FR}}$. □

The proof of Theorem 2.7 is based on the following two technical lemmas.

Lemma C.2 (Bounded gradient)**.** For any $X \in T_\Sigma\mathbb{S}^n_{++}$, denote by $\|X\|_\Sigma \triangleq \sqrt{\langle X, X\rangle_\Sigma}$ the norm induced by the inner product $\langle \cdot, \cdot \rangle_\Sigma$ defined in (8). The Riemannian gradient of the objective function $L(\cdot)$ of problem (6) satisfies

$$\|\operatorname{grad} L(\Sigma)\|_\Sigma \leq \sqrt{n} \cdot e^{2\sqrt{2}\rho} \cdot \lambda^{-2}_{\min}(\hat{\Sigma}) \cdot \max\{|1 - e^{\sqrt{2}\rho}\lambda^{-1}_{\min}(\hat{\Sigma})\lambda_{\max}(S)|, 1\} \quad \forall \Sigma \in \mathcal{B}^{\mathrm{FR}}.$$

Proof. By (9) and the definition of $\|\cdot\|_\Sigma$, we have

$$\|\mathrm{grad} L(\Sigma)\|^2_\Sigma = \frac{1}{2}\operatorname{Tr}\left(\mathrm{grad} L(\Sigma) \cdot \Sigma^{-1} \cdot \mathrm{grad} L(\Sigma) \cdot \Sigma^{-1}\right) = \frac{1}{2}\operatorname{Tr}\left(A\Sigma^{-2}A\Sigma^{-2}\right),$$

where $A \triangleq (I_n - \Sigma^{-\frac{1}{2}}S\Sigma^{-\frac{1}{2}})$. Lemma C.1(ii) thus implies that

$$(1 - e^{\sqrt{2}\rho}\lambda^{-1}_{\min}(\hat{\Sigma})\lambda_{\max}(S))I_n \preceq (1 - \lambda^{-1}_{\min}(\Sigma)\lambda_{\max}(S))I_n \preceq A \preceq I_n,$$

[5] A formal definition of CAT spaces can be found in [13, Definition II.1.1], and the upper bound κ of the curvature of a metric space is defined in [13, Definition II.1.2]

and therefore we have

$$\begin{aligned}\|\text{grad}L(\Sigma)\|_\Sigma &\leq \sqrt{\frac{n}{2}\cdot\frac{\lambda_{\max}^2(A)}{\lambda_{\min}^4(\Sigma)}}\\ &\leq \sqrt{\frac{n}{2}\cdot\frac{\max\{1,(1-e^{\sqrt{2}\rho}\lambda_{\min}^{-1}(\hat{\Sigma})\lambda_{\max}(S))^2\}}{\lambda_{\min}^4(\hat{\Sigma})e^{-4\sqrt{2}\rho}}}\\ &= \frac{\sqrt{n}\cdot\max\left\{1,\left|1-e^{\sqrt{2}\rho}\lambda_{\min}^{-1}(\hat{\Sigma})\lambda_{\max}(S)\right|\right\}}{\sqrt{2}\lambda_{\min}^2(\hat{\Sigma})e^{-2\sqrt{2}\rho}},\end{aligned}$$

where the last inequality follows from Lemma C.1(ii). This observation completes the proof. □

Lemma C.3 (Lower bounded sectional curvature)**.** The sectional curvature of the Riemannian manifold $\mathbb{S}_{++}^n$ equipped with the FR metric (8) is lower bounded by -2.

Proof. Select $\Sigma \in \mathbb{S}_{++}^n$, and let $X, Y \in T_\Sigma \mathbb{S}_{++}^n$ be two orthonormal tangent vectors at Σ, that is,

$$\|X\|_\Sigma = 1 = \|Y\|_\Sigma \quad \text{and} \quad \langle X, Y\rangle_\Sigma = 0.$$

Using the formula for the Riemannian curvature tensor $R(\cdot,\cdot,\cdot,\cdot)$ from [42, Theorem 2.1 (ii)], we have

$$R(X,Y,Y,X) = -\frac{1}{4}\operatorname{Tr}\left(Y\Sigma^{-1}X\Sigma^{-1}X\Sigma^{-1}Y\Sigma^{-1}\right) + \frac{1}{4}\operatorname{Tr}\left(X\Sigma^{-1}Y\Sigma^{-1}X\Sigma^{-1}Y\Sigma^{-1}\right). \quad \text{(C.2)}$$

Then, the sectional curvature $\kappa(X,Y)$ associated with the 2-plane spanned by $\{X,Y\}$ satisfies

$$\begin{aligned}\kappa(X,Y) &= -R(X,Y,X,Y)\\ &= -\frac{1}{4}\operatorname{Tr}\left(Y\Sigma^{-1}X\Sigma^{-1}X\Sigma^{-1}Y\Sigma^{-1}\right) + \frac{1}{4}\operatorname{Tr}\left(X\Sigma^{-1}Y\Sigma^{-1}X\Sigma^{-1}Y\Sigma^{-1}\right)\\ &\geq -\left(\|X\|_\Sigma^2\|Y\|_\Sigma^2 + \|X\|_\Sigma^2\|Y\|_\Sigma^2\right) = -2,\end{aligned}$$

where the first equality follows from [29, Proposition 8.8], the second equality exploits (C.2), and the inequality holds due to the Cauchy-Schwarz inequality. □

We are now equipped with all the necessary ingredients to prove Theorem 2.7.

Proof of Theorem 2.7. The proof closely follows that of [52, Theorem 9]. The main difference is that we replace the assumption of Lipschitz continuity of the objective function with the assumption of a bounded Riemannian gradient. Due to Theorem 2.5, the function $L(\cdot)$ is geodesically convex. So we have (see the sentence following Definition 2 in [52])

$$L(\Sigma') \geq L(\Sigma) + \langle \text{grad}L(\Sigma), \text{Exp}_\Sigma^{-1}(\Sigma')\rangle_\Sigma, \quad \forall \Sigma, \Sigma' \in \mathbb{S}_{++}^n.$$

Therefore, for any $k \geq 1$,

$$L(\Sigma_k) - L(\Sigma^\star) \leq -\langle \text{grad}L(\Sigma_k), \text{Exp}_{\Sigma_k}^{-1}(\Sigma^\star)\rangle_{\Sigma_k}. \quad \text{(C.3)}$$

By [52, Corollary 8], Lemma C.3 and because the diameter of the feasible region is 2ρ, the right hand side of (C.3) is upper bounded by

$$\frac{1}{2\alpha}\left(d^2(\Sigma_k,\Sigma^\star) - d^2(\Sigma_{k+1},\Sigma^\star)\right) + \frac{\alpha\cdot(2\rho)\cdot\sqrt{2}\cdot\|\text{grad}L(\Sigma_k)\|_{\Sigma_k}^2}{2\tanh((2\rho)\cdot\sqrt{2})}, \quad \text{(C.4)}$$

where the norm $\|\cdot\|_{\Sigma_k}$ is defined as in Lemma C.2. Therefore, substituting the upper bound (C.4) into (C.3) and using C.2, we find

$$L(\Sigma_k) - L(\Sigma^\star) \leq \frac{1}{2\alpha}\left(d^2(\Sigma_k,\Sigma^\star) - d^2(\Sigma_{k+1},\Sigma^\star)\right) + \frac{\sqrt{2}\alpha\rho\Gamma^2}{\tanh(2\sqrt{2}\rho)}, \quad \text{(C.5)}$$

where $\Gamma \triangleq 2^{-1/2}\sqrt{n} \cdot e^{2\sqrt{2}\rho} \cdot \lambda_{\min}^{-2}(\hat{\Sigma}) \cdot \max\{|1 - e^{\sqrt{2}\rho}\lambda_{\min}^{-1}(\hat{\Sigma})\lambda_{\max}(S)|, 1\}$. By telescoping, we then obtain

$$\begin{aligned}\frac{1}{K}\sum_{k=1}^{K} L(\Sigma_k) - L(\Sigma^\star) &\leq \frac{1}{2\alpha K}\left(d^2(\Sigma_1, \Sigma^\star) - d^2(\Sigma_{K+1}, \Sigma^\star)\right) + \frac{\sqrt{2}\alpha\rho\Gamma^2}{\tanh(2\sqrt{2}\rho)} \\ &\leq \frac{2\rho^2}{\alpha K} + \frac{\sqrt{2}\alpha\rho\Gamma^2}{\tanh(2\sqrt{2}\rho)} = \frac{2^{\frac{7}{4}}\rho^{\frac{3}{2}}\Gamma}{\sqrt{K\tanh(2\sqrt{2}\rho)}},\end{aligned}$$

where the second inequality follows from the bounds $d^2(\Sigma_{K+1}, \Sigma^\star) \geq 0$ and $d(\Sigma_1, \Sigma^\star) \leq 2\rho$, and the equality holds because $\alpha = 2^{1/4}\sqrt{\rho\tanh(2\sqrt{2}\rho)}/(\Gamma\sqrt{K})$. Note that although the matrix Σ_{K+1} is not actually computed because Algorithm 1 is terminated at $k = K - 1$, it is well-defined, and inequality (C.5) is valid for $k = K$. If we can show that

$$L(\bar{\Sigma}_K) \leq \frac{1}{K}\sum_{k=1}^{K} L(\Sigma_k),$$

the desired result follows. Towards that end, we prove by induction that

$$L(\bar{\Sigma}_T) \leq \frac{1}{T}\sum_{k=1}^{T} L(\Sigma_k) \quad \forall T \in \mathbb{N}t. \tag{C.6}$$

The inequality trivially holds for $T = 1$. Suppose now that the inequality in (C.6) holds for some $T \geq 1$. Then, we have

$$\begin{aligned}L(\bar{\Sigma}_{T+1}) &= L\left(\bar{\Sigma}_T^{\frac{1}{2}}\left(\bar{\Sigma}_T^{-\frac{1}{2}}\Sigma_{T+1}\bar{\Sigma}_T^{-\frac{1}{2}}\right)^{\frac{1}{T+1}}\bar{\Sigma}_T^{\frac{1}{2}}\right) \\ &= L\left(\gamma_T\left(\frac{1}{T+1}\right)\right) \\ &\leq \frac{T}{T+1}L\left(\bar{\Sigma}_T\right) + \frac{1}{T+1}L\left(\Sigma_{T+1}\right) \\ &\leq \frac{1}{T+1}\sum_{k=1}^{T} L(\Sigma_k) + \frac{1}{T+1}L\left(\Sigma_{T+1}\right) \\ &= \frac{1}{T+1}\sum_{k=1}^{T+1} L(\Sigma_k),\end{aligned}$$

where γ_T denotes the geodesic from $\bar{\Sigma}_T$ to Σ_{T+1}. The first inequality follows from the geodesic convexity of $L(\cdot)$ (see Theorem 2.5 and Definition 2.4), and the second inequality holds due to the induction hypothesis (C.6). The claim now follows because T was chosen arbitrarily. □

Next, we formally define the notions of geodesic strong convexity and geodesic smoothness for functions on $\mathbb{S}_{++}^n$.

Definition C.4 (Strong convexity)**.** Let $\mathcal{B} \subseteq \mathbb{S}_{++}^n$ be a subset and $\sigma > 0$. A differentiable function $F : \mathcal{B} \to \mathbb{R}$ is said to be (geodesically) σ-strongly convex on $\mathcal{B}$ if

$$F(Y) \geq F(X) + \langle \operatorname{grad} F(X), \operatorname{Exp}_X^{-1}(Y)\rangle_X + \frac{\sigma}{2}d^2(X, Y). \tag{C.7}$$

Definition C.5 (Smoothness)**.** Let $\mathcal{B} \subseteq \mathbb{S}_{++}^n$ be a subset and $\beta > 0$. A differentiable function $F : \mathcal{B} \to \mathbb{R}$ is said to be (geodesically) β-smooth on $\mathcal{B}$ if

$$F(Y) \leq F(X) + \langle \operatorname{grad} F(X), \operatorname{Exp}_X^{-1}(Y)\rangle_X + \frac{\beta}{2}d^2(X, Y). \tag{C.8}$$

The proof of Lemma 2.8 is based on the following preparatory results.

Lemma C.6. Let $F : \mathbb{S}^n_{++} \to \mathbb{R}$ be a twice continuously differentiable function and $\mathcal{B} \subseteq \mathbb{S}^n_{++}$ be a geodesically convex subset. The following implications hold.

(i) If the smallest eigenvalue of the Riemannian Hessian $\operatorname{hess} F(X)$ (interpreted as an operator on $T_X\mathbb{S}^N_{++}$) of F at X is lower bounded uniformly on $\mathcal{B}$ by $\sigma > 0$, *i.e.*,

$$\min\left\{\langle \operatorname{hess} F(X)[V], V\rangle : V \in T_X\mathbb{S}^N_{++}, \|V\|_X = 1\right\} \geq \sigma \quad \forall X \in \mathcal{B}, \tag{C.9}$$

then F is σ-strongly convex on $\mathcal{B}$.

(ii) If the largest eigenvalue of the Riemannian Hessian $\operatorname{hess} F(X)$ (interpreted as an operator on $T_X\mathbb{S}^N_{++}$) of F at X is upper bounded uniformly on $\mathcal{B}$ by $\beta > 0$, *i.e.*,

$$\max\left\{\langle \operatorname{hess} F(X)[V], V\rangle : V \in T_X\mathbb{S}^N_{++}, \|V\|_X = 1\right\} \leq \beta \quad \forall X \in \mathcal{B}, \tag{C.10}$$

then F is β-smooth on $\mathcal{B}$.

The proof of Lemma C.6 closely follows that of its Euclidean counterpart and is omitted here.

Proof of Lemma 2.8. Define $f(\Sigma) = \operatorname{Tr}\left(\Sigma^{-1}S\right)$. Because $\log\det\Sigma$ is a geodesically linear function [44, Proposition 12], it suffices to study the smoothness and convexity properties of $f(\cdot)$. By [22, Equations (28)], the Riemannian Hessian $\operatorname{hess} f(\Sigma)$ at Σ is given by

$$\operatorname{hess} f(\Sigma)[V] = \Sigma\left(\nabla^2 f(\Sigma)[V]\right)\Sigma + \frac{1}{2}\left(V\nabla f(\Sigma)\Sigma + \Sigma\nabla f(\Sigma)V\right) \quad \forall V \in T_\Sigma\mathbb{S}^n_{++}. \tag{C.11}$$

By elementary matrix calculus, we know that

$$\nabla f(\Sigma) = -\Sigma^{-1}S\Sigma^{-1} \tag{C.12}$$

and

$$\nabla^2 f(\Sigma)[V] = \Sigma^{-1}V\Sigma^{-1}S\Sigma^{-1} + \Sigma^{-1}S\Sigma^{-1}V\Sigma^{-1} \quad \forall V \in \mathbb{S}^n, \tag{C.13}$$

where the Hessian $\nabla^2 f(\Sigma)$ is interpreted as a linear operator on $\mathbb{S}^n$. Noting that $T_\Sigma\mathbb{S}^n_{++} = \mathbb{S}^n$ and combining (C.11), (C.12) and (C.13), we obtain

$$\langle \operatorname{hess} f(\Sigma)[V], V\rangle_\Sigma = \operatorname{Tr}\left(\Sigma^{-1}S\Sigma^{-1}V\Sigma^{-1}V\right) \quad \forall V \in \mathbb{S}^n. \tag{C.14}$$

Using these preparatory results, we now demonstrate that $f(\cdot)$ is β-smooth and σ-strongly convex in the geodesic sense.

Smoothness. In order to establish the smoothness properties of $f(\cdot)$, we consider the maximization problem

$$\max\left\{\langle \operatorname{hess} f(\Sigma)[V], V\rangle_\Sigma : V \in \mathbb{S}^n,\ \|V\|_\Sigma = 1\right\},$$

which, by (C.14) and the definition of $\|\cdot\|_\Sigma$, is equivalent to

$$\max\left\{\operatorname{Tr}\left(\Sigma^{-1}S\Sigma^{-1}V\Sigma^{-1}V\right) : V \in \mathbb{S}^n,\ \tfrac{1}{2}\operatorname{Tr}\left(\Sigma^{-1}V\Sigma^{-1}V\right) = 1\right\}.$$

The optimal value of this problem is upper bounded by $2\,\lambda_{\max}(S)/\lambda_{\min}(\Sigma)$. Using the bound from Lemma C.1(ii), we have

$$\frac{2\,\lambda_{\max}(S)}{\lambda_{\min}(\Sigma)} \leq \frac{2\lambda_{\max}(S)}{\lambda_{\min}(\widehat{\Sigma})\exp(-\sqrt{2}\rho)} = \beta.$$

By Lemma C.6(ii), $f(\cdot)$ is β-smooth.

Strong convexity. In order to establish the convexity properties of $f(\cdot)$, we consider the minimization problem

$$\min\left\{\langle \operatorname{hess} f(\Sigma)[V], V\rangle_\Sigma : V \in \mathbb{S}^n,\ \|V\|_\Sigma = 1\right\},$$

which, by (C.14) and the definition of $\|\cdot\|_\Sigma$, is equivalent to

$$\min\left\{\operatorname{Tr}\left(\Sigma^{-1}S\Sigma^{-1}V\Sigma^{-1}V\right) : V \in \mathbb{S}^n,\ \tfrac{1}{2}\operatorname{Tr}\left(\Sigma^{-1}V\Sigma^{-1}V\right) = 1\right\}.$$

The optimal value of this problem is lower bounded by $2\,\lambda_{\min}(S)/\lambda_{\max}(\Sigma)$. Using the bound in Lemma C.1(ii), we have

$$\frac{2\,\lambda_{\min}(S)}{\lambda_{\max}(\Sigma)} = \frac{2\lambda_{\min}(S)}{\lambda_{\max}(\widehat{\Sigma})\exp(\sqrt{2}\rho)} = \sigma.$$

Since $S \succ 0$, $\sigma > 0$. By Lemma C.6(i), $f(\cdot)$ is thus σ-strongly convex. This completes the proof. $\square$

Appendix D Proofs of Section 3

Proof of Theorem 3.2. By applying the change of variables $Z \leftarrow \Sigma^{-1}$, problem (12) can be reformulated as

$$\inf_{Z} \left\{ \mathrm{Tr}\left(SZ\right) - \log\det Z : Z \succ 0,\ \mathrm{Tr}\left(\hat{\Sigma}Z\right) - \log\det Z \leq \bar{\rho} \right\}, \tag{D.1}$$

where $\bar{\rho} \triangleq 2\rho + n + \log\det\hat{\Sigma}$. Note that (D.1) is equivalent to

$$\begin{aligned}
&\inf_{Z\succ 0} \sup_{\gamma\geq 0}\ \mathrm{Tr}\left(SZ\right) - \log\det Z + \gamma\left(\mathrm{Tr}\left(\hat{\Sigma}Z\right) - \log\det Z - \bar{\rho}\right) \\
=&\sup_{\gamma\geq 0} \inf_{Z\succ 0}\ -\gamma\bar{\rho} + \mathrm{Tr}\left((S+\gamma\hat{\Sigma})Z\right) - (1+\gamma)\log\det Z \\
=&\sup_{\gamma\geq 0}\left\{ -\gamma\bar{\rho} + \inf_{Z\succ 0}\left\{ \mathrm{Tr}\left((S+\gamma\hat{\Sigma})Z\right) - (1+\gamma)\log\det Z \right\}\right\},
\end{aligned} \tag{D.2}$$

where the first equality follows from strong duality, which holds because $\rho > 0$ and because $\hat{\Sigma}^{-1}$ is a Slater point for the primal problem (D.1).

To analyze problem (D.2), assume first that S is singular. If $\gamma = 0$, then the inner minimization problem over Z is unbounded, and thus $\gamma = 0$ is never optimal for the outer maximization problem. For any $\gamma > 0$, the inner minimization problem over Z admits the optimal solution $Z^\star(\gamma) = (1+\gamma)(S+\gamma\hat{\Sigma})^{-1}$. Problem (D.1) is thus equivalent to

$$\begin{aligned}
&\sup_{\gamma>0}\Big\{ -\gamma\bar{\rho} + n(1+\gamma) - (1+\gamma)\log\det[(1+\gamma)(S+\gamma\hat{\Sigma})^{-1}]\Big\} \\
=&\sup_{\gamma>0}\Big\{ -\gamma\bar{\rho} + n(1+\gamma) - (1+\gamma)\log[(1+\gamma)^n\det(S+\gamma\hat{\Sigma})^{-1}]\Big\} \\
=&\sup_{\gamma>0}\Big\{ -\gamma\bar{\rho} + n(1+\gamma) - n(1+\gamma)\log(1+\gamma) - (1+\gamma)\log\det(S+\gamma\hat{\Sigma})^{-1}\Big\}.
\end{aligned} \tag{D.3}$$

By strong duality, any minimizer $\gamma^\star$ of (13) can be used to construct a minimizer

$$\Sigma^\star = (1+\gamma^\star)^{-1}(S+\gamma^\star\hat{\Sigma})$$

for problem (12). This observation establishes the claim if S is singular.

Assume next that S has full rank. In this case, the inner minimization problem in (D.2) admits the optimal solution $Z^\star(\gamma) = (1+\gamma)(S+\gamma\hat{\Sigma})^{-1}$ for any fixed $\gamma \geq 0$, and thus problem (D.2) is equivalent to

$$\sup_{\gamma\geq 0}\Big\{ -\gamma\bar{\rho} + n(1+\gamma) - n(1+\gamma)\log(1+\gamma) - (1+\gamma)\log\det(S+\gamma\hat{\Sigma})^{-1}\Big\},$$

which differs from (D.3) only in that it has a closed feasible set, that is, $\gamma = 0$ is feasible. Because the objective function of the above optimization problem is continuous in γ, we can in fact optimize over $\gamma > 0$ without reducing the supremum. The claim now follows by replacing $\bar{\rho}$ with its definition and eliminating the constant term from the objective function. □

Proof of Corollary 3.3. For any $\hat{\Sigma} \in \mathbb{S}^n_{++}$ and $\gamma > 0$, the Woodbury formula [8, Corollary 2.8.8] implies that

$$(S+\gamma\hat{\Sigma})^{-1} = \gamma^{-1}\hat{\Sigma}^{-\frac{1}{2}}(\gamma^{-1}\hat{\Sigma}^{-\frac{1}{2}}S\hat{\Sigma}^{-\frac{1}{2}} + I_n)^{-1}\hat{\Sigma}^{-\frac{1}{2}},$$

and thus we have

$$\begin{aligned}
\log\det(\gamma\hat{\Sigma}+S)^{-1} + n\log\gamma + \log\det\hat{\Sigma} &= -\log\det\left(I_n + \gamma^{-1}\hat{\Sigma}^{-\frac{1}{2}}\Lambda\Lambda^\top\hat{\Sigma}^{-\frac{1}{2}}\right) \\
&= -\log\det\left(I_k + \gamma^{-1}\Lambda^\top\hat{\Sigma}^{-1}\Lambda\right) \\
&= k\log\gamma - \log\det\left(\gamma I_k + \Lambda^\top\hat{\Sigma}^{-1}\Lambda\right),
\end{aligned}$$

where the second equality follows from [8, Equation 2.8.14]. Substituting the expression for $\log\det(\gamma\hat{\Sigma}+S)^{-1}$ into (13) and removing the irrelevant constant term $(n+\log\det\hat{\Sigma})$ yields the equivalent minimization problem

$$\inf_{\gamma>0}\Big\{2\gamma\rho+n(1+\gamma)\log(1+\gamma)-(n-k)(1+\gamma)\log\gamma-(1+\gamma)\log\det(\gamma I_k+\Lambda^\top\hat{\Sigma}^{-1}\Lambda)\Big\}.$$

This observation completes the proof. □

Appendix E Derivatives of Problem (13)

Use $g_1(\gamma)$ as a shorthand for the objective function of problem (13). In the following, we provide the first- and second-order derivatives of $g_1(\cdot)$, which are needed by the optimization algorithm that solves (13). In particular, the first-order derivative is given by

$$g_1'(\gamma)=2\rho+n\left(\log(1+\gamma)+1\right)-\log\det\Big(\gamma I_n+S\hat{\Sigma}^{-1}\Big)-(1+\gamma)\operatorname{Tr}\big((\gamma I_n+S\hat{\Sigma}^{-1})^{-1}\big),$$

and the second-order derivative can be expressed as

$$g_1''(\gamma)=\frac{n}{1+\gamma}-\operatorname{Tr}\big((\gamma I_n+S\hat{\Sigma}^{-1})^{-1}\Big(2I_n+(1+\gamma)(\gamma I_n+S\hat{\Sigma}^{-1})^{-1}\Big)\big).$$

Next, denote by g_2 the objective function of the singular reduction problem of Corollary 3.3, that is,

$$g_2(\gamma)=2\gamma\rho+n(1+\gamma)\log(1+\gamma)-(n-k)(1+\gamma)\log\gamma-(1+\gamma)\log\det(\gamma I_k+\Lambda^\top\hat{\Sigma}^{-1}\Lambda).$$

The first- and second-order derivative of g_2 are given by

$$\begin{aligned}g_2'(\gamma)=&\,2\rho+n\left(\log(1+\gamma)+1\right)-(n-k)\left(\log\gamma+\gamma^{-1}+1\right)\\&-\log\det\Big(\gamma I_k+\Lambda^\top\hat{\Sigma}^{-1}\Lambda\Big)-(1+\gamma)\operatorname{Tr}\big((\gamma I_k+\Lambda^\top\hat{\Sigma}^{-1}\Lambda)^{-1}\big),\end{aligned}$$

and

$$\begin{aligned}g_2''(\gamma)=&\frac{n}{1+\gamma}-(n-k)(\gamma^{-1}-\gamma^{-2})\\&-\operatorname{Tr}\big(\gamma I_k+\Lambda^\top\hat{\Sigma}^{-1}\Lambda)^{-1}\Big(2I_k+(1+\gamma)(\gamma I_k+\Lambda^\top\hat{\Sigma}^{-1}\Lambda)^{-1}\Big)\big),\end{aligned}$$

respectively.

Acknowledgments We gratefully acknowledge financial support from the Swiss National Science Foundation under grant BSCGI0_157733 as well as the EPSRC grants EP/M028240/1, EP/M027856/1 and EP/N020030/1.

References

[1] P.-A. Absil, R. Mahony, and R. Sepulchre. *Optimization Algorithms on Matrix Manifolds*. Princeton University Press, 2009.

[2] C. D. Aliprantis and K. C. Border. *Infinite Dimensional Analysis: A Hitchhiker's Guide*. Springer, 2006.

[3] M. Arnaudon, F. Barbaresco, and L. Yang. Riemannian medians and means with applications to radar signal processing. *IEEE Journal of Selected Topics in Signal Processing*, 7(4):595–604, 2013.

[4] C. Atkinson and A. F. Mitchell. Rao's distance measure. *Sankhyā: The Indian Journal of Statistics, Series A*, 43(3):345–365, 1981.

[5] K. Bache and M. Lichman. UCI machine learning repository, 2013. Available from `http://archive.ics.uci.edu/ml`.

[6] M. Bauer, M. Bruveris, and P. W. Michor. Uniqueness of the Fisher–Rao metric on the space of smooth densities. *Bulletin of the London Mathematical Society*, 48(3):499–506, 2016.

[7] G. C. Bento, O. P. Ferreira, and J. G. Melo. Iteration-complexity of gradient, subgradient and proximal point methods on Riemannian manifolds. *Journal of Optimization Theory and Applications*, 173(2):548–562, 2017.

[8] D. S. Bernstein. *Matrix Mathematics: Theory, Facts, and Formulas*. Princeton University Press, 2009.

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

[10] J. Bi and T. Zhang. Support vector classification with input data uncertainty. In *Advances in Neural Information Processing Systems*, pages 161–168, 2005.

[11] S. Bonnabel and R. Sepulchre. Riemannian metric and geometric mean for positive semidefinite matrices of fixed rank. *SIAM Journal on Matrix Analysis and Applications*, 31(3):1055–1070, 2009.

[12] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

[13] M. R. Bridson and A. Haefliger. *Metric Spaces of Non-Positive Curvature*. Springer, 2013.

[14] E. Brochu, V. M. Cora, and N. de Freitas. A tutorial on Bayesian optimization of expensive cost functions, with application to active user modeling and hierarchical reinforcement learning. Technical Report UBC TR-2009-023 and arXiv:1012.2599, University of British Columbia, Department of Computer Science, 2009.

[15] S. Bubeck and N. Cesa-Bianchi. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends® in Machine Learning*, 5(1):1–122, 2012.

[16] L. L. Campbell. An extended Čencov characterization of the information metric. *Proceedings of the American Mathematical Society*, 98(1):135–141, 1986.

[17] M. P. d. Carmo. *Riemannian Geometry*. Birkhäuser, 1992.

[18] N. N. Čencov. *Statistical Decision Rules and Optimal Inference*. American Mathematical Society, 2000.

[19] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. Wiley-Interscience, 2006.

[20] D. R. Cox. Tests of separate families of hypotheses. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, pages 105–123, 1961.

[21] D. R. Cox. A return to an old paper: 'tests of separate families of hypotheses'. *Journal of the Royal Statistical Society: Series B*, 75(2):207–215, 2013.

[22] O. Ferreira, M. Louzeiro, and L. Prudente. Gradient method for optimization on Riemannian manifolds with lower bounded curvature. *arXiv preprint arXiv:1806.02694*, 2018.

[23] J. Friedman, T. Hastie, and R. Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441, 2008.

[24] W. K. Härdle and L. Simar. *Applied Multivariate Statistical Analysis*. Springer, 2015.

[25] S. Jayasumana, R. Hartley, M. Salzmann, H. Li, and M. Harandi. Kernel methods on the Riemannian manifold of symmetric positive definite matrices. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 73–80, 2013.

[26] R. E. Kass and P. W. Vos. *Geometrical Foundations of Asymptotic Inference*. John Wiley & Sons, 2011.

[27] S. Lang. *Fundamentals of Differential Geometry*. Springer, 2012.

[28] O. Ledoit and M. Wolf. A well-conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis*, 88(2):365–411, 2004.

[29] J. M. Lee. *Riemannian Manifolds: An Introduction to Curvature*. Springer, 1997.

[30] J. M. Lee. *Introduction to Smooth Manifolds*. Springer, 2013.

[31] C. Liu and N. Boumal. Simple Algorithms for Optimization on Riemannian Manifolds with Constraints. *Applied Mathematics and Optimization*, 2019.

[32] G. McLachlan. *Discriminant Analysis and Statistical Pattern Recognition*. John Wiley & Sons, 2004.

[33] R. Munos. From bandits to Monte-Carlo tree search: The optimistic principle applied to optimization and planning. *Foundations and Trends® in Machine Learning*, 7(1):1–129, 2014.

[34] V. A. Nguyen, D. Kuhn, and P. M. Esfahani. Distributionally robust inverse covariance estimation: The Wasserstein shrinkage estimator. *arXiv preprint arXiv:1805.07194*, 2018.

[35] V. A. Nguyen, S. Shafieezadeh-Abadeh, M.-C. Yue, D. Kuhn, and W. Wiesemann. Optimistic distributionally robust optimization for nonparametric likelihood approximation. In *Advances in Neural Information Processing Systems*, 2019.

[36] M. Norton, A. Takeda, and A. Mafusalov. Optimistic robust optimization with applications to machine learning. *arXiv preprint arXiv:1711.07511*, 2017.

[37] X. Pennec, P. Fillard, and N. Ayache. A Riemannian framework for tensor computing. *International Journal of Computer Vision*, 66(1):41–66, 2006.

[38] L. F. Price, C. C. Drovandi, A. Lee, and D. J. Nott. Bayesian synthetic likelihood. *Journal of Computational and Graphical Statistics*, 27(1):1–11, 2018.

[39] S. Said, L. Bombrun, Y. Berthoumieu, and J. H. Manton. Riemannian Gaussian distributions on the space of symmetric positive definite matrices. *IEEE Transactions on Information Theory*, 63(4):2153–2170, 2017.

[40] R. P. Savage. The space of positive definite matrices and Gromov's invariant. *Transactions of the American Mathematical Society*, 274(1):239–263, 1982.

[41] R. Schoenberg. Constrained maximum likelihood. *Computational Economics*, 10(3):251–266, 1997.

[42] L. T. Skovgaard. A Riemannian geometry of the multivariate normal model. *Scandinavian Journal of Statistics*, 11(4):211–223, 1984.

[43] S. Sra and R. Hosseini. Conic geometric optimization on the manifold of positive definite matrices. *SIAM Journal on Optimization*, 25(1):713–739, 2015.

[44] S. Sra, N. K. Vishnoi, and O. Yildiz. On geodesically convex formulations for the Brascamp-Lieb constant. In *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques (APPROX/RANDOM 2018)*. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2018.

[45] N. Srinivas, A. Krause, S. Kakade, and M. Seeger. Gaussian process optimization in the bandit setting: no regret and experimental design. In *International Conference on Machine Learning*, pages 1015–1022, 2010.

[46] A. Takatsu. Wasserstein geometry of Gaussian measures. *Osaka Journal of Mathematics*, 48(4):1005–1026, 2011.

[47] N. Tripuraneni, N. Flammarion, F. Bach, and M. Jordan. Averaging stochastic gradient descent on Riemannian manifolds. In *Conference On Learning Theory*, volume 75 of *Proceedings of Machine Learning Research*, pages 650–687, 2018.

[48] O. Tuzel, F. Porikli, and P. Meer. Pedestrian detection via classification on Riemannian manifolds. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(10):1713–1727, 2008.

[49] C. Villani. *Optimal Transport: Old and New*. Springer, 2008.

[50] S. N. Wood. Statistical inference for noisy nonlinear ecological dynamic systems. *Nature*, 466(7310):1102–1104, 2010.

[51] H. Zhang, S. J. Reddi, and S. Sra. Riemannian SVRG: Fast stochastic optimization on Riemannian manifolds. In *Advances in Neural Information Processing Systems*, pages 4592–4600, 2016.

[52] H. Zhang and S. Sra. First-order methods for geodesically convex optimization. In *Conference on Learning Theory*, volume 49 of *Proceedings of Machine Learning Research*, pages 1617–1638, 2016.

[53] H. Zhang and S. Sra. An estimate sequence for geodesically convex optimization. In *Conference On Learning Theory*, volume 75 of *Proceedings of Machine Learning Research*, pages 1703–1723, 2018.

[54] T. Zhang. Robust subspace recovery by geodesically convex optimization. *arXiv preprint arXiv:1206.1386*, 2012.