An Empirical Bayes Approach to Shrinkage Estimation on the Manifold of Symmetric Positive-Definite Matrices*

Chun-Hao Yang[1], Hani Doss[2] and Baba C. Vemuri[3]
[1]Institute of Applied Mathematical Sciences, National Taiwan University
[2]Department of Statistics, University of Florida
[3]Department of CISE, University of Florida

Abstract

The James-Stein estimator is an estimator of the multivariate normal mean and dominates the maximum likelihood estimator (MLE) under squared error loss. The original work inspired great interest in developing shrinkage estimators for a variety of problems. Nonetheless, research on shrinkage estimation for manifold-valued data is scarce. In this paper, we propose shrinkage estimators for the parameters of the Log-Normal distribution defined on the manifold of $N \times N$ symmetric positive-definite matrices. For this manifold, we choose the Log-Euclidean metric as its Riemannian metric since it is easy to compute and has been widely used in a variety of applications. By using the Log-Euclidean distance in the loss function, we derive a shrinkage estimator in an analytic form and show that it is asymptotically optimal within a large class of estimators that includes the MLE, which is the sample Fréchet mean of the data. We demonstrate the performance of the proposed shrinkage estimator via several simulated data experiments. Additionally, we apply the shrinkage estimator to perform statistical inference in both diffusion and functional magnetic resonance imaging problems.

Keywords: Stein's unbiased risk estimate, Fréchet mean, Tweedie's estimator

*This research was funded in part by NSF grant IIS-1724174, and NIH NINDS and NIA grant R01NS121099 to Vemuri, NSF grant DMS-1854476 to Doss, and MOST grant 110-2118-M-002-005-MY3 and NTU grant 111L7431 to Yang.

1 Introduction

Symmetric positive-definite (SPD) matrices are common in applications of science and engineering. In computer vision problems, they are encountered in the form of covariance matrices, e.g., region covariance descriptors (Tuzel et al. 2006), and in diffusion magnetic resonance imaging, SPD matrices manifest themselves as diffusion tensors which are used to model the diffusion of water molecules (Basser et al. 1994), and as Cauchy deformation tensors in morphometry to model the deformations (see Frackowiak et al. (2004, Ch. 36)). Many other applications can be found in Cherian & Sra (2016). In such applications, the statistical analysis of data must perform geometry-aware computations, i.e., employ methods that take into account the nonlinear geometry of the data space. In most data analysis applications, it is useful to describe the entire dataset with a few summary statistics. For data residing in Euclidean space, this may be simply the sample mean, and for data residing in non-Euclidean spaces, e.g. Riemannian manifolds, the corresponding statistic is the sample Fréchet mean (FM) (Fréchet 1948). The sample FM also plays an important role in different statistical inference methods, e.g. principal geodesic analysis (Fletcher et al. 2003), clustering algorithms, etc. If M is a metric space with metric d, and $x_1, \ldots, x_n \in M$, the sample FM is defined by $\bar{x} = \arg\min_m \sum_{i=1}^n d^2(x_i, m)$. For Riemannian manifolds, the distance is usually chosen to be the intrinsic distance induced by the Riemannian metric. Then, the above optimization problem can be solved by Riemannian gradient descent algorithms (Pennec 2006, Groisser 2004, Afsari 2011, Moakher 2005, Gabay 1982, Udriste 2013). However, Riemannian gradient descent algorithms are usually computationally expensive, and efficient recursive algorithms for computing the sample FM have been presented in the literature for various Riemannian manifolds by Sturm (2003), Ho et al. (2013), Salehian et al. (2015), Chakraborty & Vemuri (2015), Lim & Pálfia (2014) and Chakraborty & Vemuri (2019).

In $\mathbb{R}^p$ with the Euclidean metric, the sample FM is just the ordinary sample mean. Suppose that $X_1, \ldots, X_n$ are a random sample from the multivariate normal distribution on $\mathbb{R}^p$. The sample mean $\bar{X} = n^{-1} \sum_{i=1}^n X_i$ is the maximum likelihood estimator (MLE) for the mean of the underlying normal distribution, and the James-Stein (shrinkage) estimator (James & Stein 1961) was shown to be better (under squared error loss) than the MLE when

$p > 2$ and the covariance matrix of the underlying normal distribution is assumed to be known. Inspired by this result, the goal of this paper is to develop shrinkage estimators for data residing in P_N, the space of $N \times N$ SPD matrices.

For the model $X_i \overset{\text{ind}}{\sim} N(\mu_i, \sigma^2)$, $i = 1, \ldots, p$, where $p > 2$ and σ^2 is known, the MLE of $\mu = [\mu_1, \ldots, \mu_p]^T$ is $\hat{\mu}^{\text{MLE}} = [X_1, \ldots, X_p]^T$ and it is natural to ask whether it is admissible. Stein (1956) gave a negative answer to this question and provided a class of estimators for μ that dominate the MLE. Subsequently, James & Stein (1961) proposed the estimator

$$\left(1 - \frac{(p-2)\sigma^2}{\|X\|^2}\right) X \tag{1}$$

where $X = [X_1, \ldots, X_p]^T$, which is now referred to as the James-Stein (shrinkage) estimator.

Ever since the work reported in James & Stein (1961), shrinkage estimators have been developed for the parameters of other distributions, such as the Poisson and Gamma (Clevenson & Zidek (1975) and Berger (1980)). In order to understand shrinkage estimation fully, one must understand why the process of shrinkage improves estimation. In this context, Efron and Morris presented a series of works to provide an empirical Bayes interpretation by modifying the original James-Stein estimator to suit different problems (Efron & Morris 1973*a*,*b*). The empirical Bayes approach to designing a shrinkage estimator can be described as follows. First, reformulate the model as a Bayesian model, i.e., place a prior on the parameters. Then, the hyperparameters of the prior are estimated from the data. Efron & Morris (1973*b*) presented several examples of different shrinkage estimators developed within this empirical Bayes framework.

In all the works cited above, the domain of the data has invariably been a vector space and, as mentioned earlier, many applications naturally encounter data residing in non-Euclidean spaces. Hence, generalizing shrinkage estimation to non-Euclidean spaces is a worthwhile pursuit. In this paper, we focus on shrinkage estimation for the Riemannian manifold P_N. We assume that the observed SPD matrices are drawn from a Log-Normal distribution defined on P_N (Schwartzman 2016) and we are interested in estimating the mean and the covariance matrix of this distribution. We point out that a simple method to derive a shrinkage estimator in this case is to apply James-Stein shrinkage to the log-transformed SPD matrices. However, this does not lead to an optimal shrinkage estimator.

Hence, we derive shrinkage estimators for the parameters of the Log-Normal distribution using an empirical Bayes framework, which is described in detail subsequently, and show that the proposed estimator is asymptotically optimal within a class of estimators including the MLE. We discuss this issue in more detail in Section 3.2.

We present simulated data experiments which demonstrate that the proposed shrinkage estimator of the mean of the Log-Normal distribution is better (in terms of risk) than the sample FM, which is the MLE, and the shrinkage estimator proposed by Yang & Vemuri (2019). Further, we also apply the shrinkage estimator to find group differences between patients with Parkinson's disease and controls (normal subjects) from their respective brain scans acquired using diffusion magnetic resonance images (dMRIs). Additionally, we empirically demonstrate the advantage of shrinkage estimation applied to simultaneous estimation of the parameters of Log-Normal distributions via an experiment involving brain connectivity networks derived from resting state functional MRI (rs-fMRI) human brain scans. This experiment is presented to show that the advantages of the proposed shrinkage estimator persist as we vary the size N of the SPD matrices.

Besides estimation of the mean of different distributions, estimation of the covariance matrix (or the precision matrix) of a multivariate normal distribution is an important problem in statistics, finance, engineering and many other fields. The usual estimator, namely the sample covariance matrix, performs poorly in high-dimensional problems and many researchers have endeavored to improve covariance estimation by applying the concept of shrinkage in this context (Stein 1975, Daniels & Kass 2001, Ledoit & Wolf 2003, Donoho et al. 2018). In this literature, it is assumed that for each $i = 1, \ldots, p$, we observe iid vectors $X_{i,1}, \ldots, X_{i,n_i} \in \mathbb{R}^N$ where for each $j = 1, \ldots, n_i$, the covariance matrix of X_{ij} is $\Sigma_i \in P_N$. This framework is distinct from our setup, where we assume that our observations are estimates $\widehat{\Sigma}_1, \ldots, \widehat{\Sigma}_p \in P_N$.

The rest of this paper is organized as follows. In Section 2, we present relevant material on the Riemannian geometry of P_N and shrinkage estimation. The main theoretical results are stated in Section 3, with the proofs of the theorems relegated to the supplement. In Section 4, we demonstrate how the proposed shrinkage estimators perform via several synthetic data examples and present applications to (real data) diffusion tensor imaging (DTI), a clin-

ically popular version of dMRI, and rs-fMRI. Specifically, we apply the proposed shrinkage estimator to (i) estimation of the brain atlases (templates) of patients with Parkinson's disease and a control group, (ii) identification of the regions of the brain where the two groups differ significantly, and (iii) estimation of connectivity networks from rs-fMRI. Finally, in Section 5 we discuss our contributions and present some future research directions.

2 Preliminaries

In this section, we briefly review the commonly used Log-Euclidean metric for P_N proposed by Arsigny et al. (2007) and the concept of Stein's unbiased risk estimate, which will form the framework for deriving the shrinkage estimators.

2.1 Riemannian Geometry of P_N

In this work, we endow the manifold P_N with the Log-Euclidean metric. We note that there is another commonly used Riemannian metric on P_N, called the affine-invariant metric (see Terras (2016, Ch. 1) for its introduction and Lenglet, Rousson, Deriche & Faugeras (2006) and Moakher (2005) for its applications). The affine-invariant metric is computationally more expensive; however, because in some applications it provides results that are indistinguishable from those obtained under the Log-Euclidean metric, as demonstrated in Arsigny et al. (2007) and Schwartzman (2016), we choose to work with the Log-Euclidean metric. For other metrics on P_N used in a variety of applications, we refer the reader to the recent survey by Feragen & Fuster (2017).

The Log-Euclidean metric is a bi-invariant Riemannian metric on the abelian Lie group $(P_N, \odot)$ where $X \odot Y = \exp(\log X + \log Y)$. The intrinsic distance $d_{\text{LE}} : P_N \times P_N \to \mathbb{R}$ induced by the Log-Euclidean metric has a very simple form, namely $d_{\text{LE}}(X, Y) = \|\log X - \log Y\|$, where $\|\cdot\|$ is the Frobenius norm. Let $\mathsf{Sym}(N)$ be the vector space of $N \times N$ symmetric matrices. Consider the map $\text{vec} : \mathsf{Sym}(N) \to \mathbb{R}^{\frac{N(N+1)}{2}}$ given by $\text{vec}(Y) = \left[y_{11}, \ldots, y_{nn}, \sqrt{2}(y_{ij})_{i<j}\right]^T$ (Schwartzman 2016). This map is actually an isomorphism between $\mathsf{Sym}(N)$ and $\mathbb{R}^{\frac{N(N+1)}{2}}$. To make the notation more concise, for $X \in P_N$, we denote $\widetilde{X} = \text{vec}(\log X) \in \mathbb{R}^{\frac{N(N+1)}{2}}$. From the definition of vec, we see that $d_{\text{LE}}(X, Y) = \|\widetilde{X} - \widetilde{Y}\|$.

Given $X_1, \ldots, X_n \in P_N$, we denote the sample FM with respect to the Log-Euclidean distance given above by

$$\bar{X} = \underset{M \in P_N}{\arg\min}\, n^{-1} \sum_{i=1}^{n} d_{LE}^2(X_i, M) = \exp\left(n^{-1} \sum_{i=1}^{n} \log X_i \right).$$

2.2 The Log-Normal Distribution on P_N

In this work, we assume that the observed SPD matrices follow the Log-Normal distribution introduced by Schwartzman (2016), which can be viewed as a generalization of the Log-Normal distribution on $\mathbb{R}^+$ to P_N. The definition is as follows.

Definition 1 Let X be a P_N-valued random variable. We say X follows a Log-Normal distribution with mean $M \in P_N$ and covariance matrix $\Sigma \in P_{N(N+1)/2}$, or $X \sim \mathrm{LN}(M, \Sigma)$, if $\widetilde{X} \sim N(\widetilde{M}, \Sigma)$.

From the definition, it is easy to see that $E \log X = \log M$ and $E\|\log X - \log M\|^2 = E\|\widetilde{X} - \widetilde{M}\|^2 = \mathrm{tr}(\Sigma)$. Some important results regarding this distribution were obtained in Schwartzman (2016). The following proposition, proved in Schwartzman (2016), for the MLEs of the parameters will be useful subsequently.

Proposition 1 *Let $X_1, \ldots, X_n \overset{iid}{\sim} LN(M, \Sigma)$. Then, the MLEs of M and Σ are $\widehat{M}^{MLE} = \bar{X}$ and $\widehat{\Sigma}^{MLE} = n^{-1} \sum_{i=1}^{n} \left(\widetilde{X}_i - \widetilde{\widehat{M}^{MLE}} \right) \left(\widetilde{X}_i - \widetilde{\widehat{M}^{MLE}} \right)^T$. The MLE of M is the sample FM under the Log-Euclidean metric.*

2.3 Bayesian Formulation of Shrinkage Estimation in $\mathbb{R}^p$

As discussed earlier, the James-Stein estimator originated from the problem of simultaneous estimation of multiple means of (univariate) normal distributions. The derivation relied heavily on properties of the univariate normal distribution. Later on, Efron & Morris (1973*b*) gave an empirical Bayes interpretation for the James-Stein estimator, which is presented by considering the hierarchical model

$$X_i | \theta_i \overset{\text{ind}}{\sim} N(\theta_i, A), \quad i = 1, \ldots, p,$$
$$\theta_i \overset{\text{iid}}{\sim} N(\mu, \lambda),$$

where A is known and μ and λ are unknown. The posterior mean for θ_i is

$$\hat{\theta}_i^{\lambda,\mu} = \frac{\lambda}{\lambda + A} X_i + \frac{A}{\lambda + A} \mu. \tag{2}$$

The parametric empirical Bayes method for estimating the θ_i's consists of first estimating the prior parameters λ and μ and then substituting them into (2). The prior parameters λ and μ can be estimated by the MLE. For the special case of $\mu = 0$, this method produces an estimator similar to the James-Stein estimator (1). Although this estimator is derived in an (empirical) Bayesian framework, it is of interest to determine whether it has good frequentist properties. For example, if we specify a loss function L and consider the induced risk function R, one would like to determine whether the estimator has uniformly smallest risk within a reasonable class of estimators. For (2), the optimal choice of λ and μ is

$$\left(\hat{\lambda}^{\text{opt}}, \hat{\mu}^{\text{opt}}\right) = \arg\min_{\lambda,\mu} R\left(\hat{\boldsymbol{\theta}}^{\lambda,\mu}, \boldsymbol{\theta}\right),$$

where $\boldsymbol{\theta} = [\theta_1, \ldots, \theta_p]^T$, $\hat{\boldsymbol{\theta}}^{\lambda,\mu} = [\hat{\theta}_1^{\lambda,\mu}, \ldots, \hat{\theta}_p^{\lambda,\mu}]^T$, and $\hat{\lambda}^{\text{opt}}$ and $\hat{\mu}^{\text{opt}}$ depend on $\boldsymbol{\theta}$, which is unknown. Instead of minimizing the risk function directly, we minimize Stein's unbiased risk estimate (SURE) (Stein 1981), denoted by $\text{SURE}(\lambda, \mu)$, which satisfies $E_{\boldsymbol{\theta}}\left[\text{SURE}(\lambda, \mu)\right] = R(\hat{\boldsymbol{\theta}}^{\lambda,\mu}, \boldsymbol{\theta})$. Thus, we use

$$\left(\hat{\lambda}^{\text{SURE}}, \hat{\mu}^{\text{SURE}}\right) = \arg\min_{\lambda,\mu} \text{SURE}(\lambda, \mu).$$

The challenging part of this endeavor is to derive SURE, which depends heavily on the risk function and the underlying distribution of the data. This approach has been used to derive estimators for many models. For example, Xie et al. (2012) derived the (asymptotically) optimal shrinkage estimator for a heteroscedastic hierarchical model, and their result is further generalized in Jing et al. (2016) and Kong et al. (2017).

3 An Empirical Bayes Shrinkage Estimator for Log-Normal Distributions

In this section, we consider the model

$$X_{ij} \overset{\text{ind}}{\sim} \text{LN}(M_i, \Sigma_i), \quad i = 1, \ldots, p, \; j = 1, \ldots, n,$$

and develop shrinkage estimators for the vector of means $\boldsymbol{M} = [M_1, \ldots, M_p]$ and the array of covariance matrices $\boldsymbol{\Sigma} = [\Sigma_1, \ldots, \Sigma_p]$. The motivation for this model is that for the applications of DTI, we have n DT images and each DT image contains p voxels. The diffusive behavior of water molecules in each voxel is characterized by a 3×3 SPD matrix. The X_{ij}'s are P_N-valued random matrices. For completeness, we first briefly review the shrinkage estimator of $\boldsymbol{M}$ proposed by Yang & Vemuri (2019), who assumed $\Sigma_i = A_i I$, where the A_i's are known positive numbers and I is the identity matrix. The assumption on $\boldsymbol{\Sigma}$ is useful when n is small since for small sample sizes the MLE for $\boldsymbol{\Sigma}$ is very unstable. Next, we present estimators for both $\boldsymbol{M}$ and $\boldsymbol{\Sigma}$. Besides presenting these estimators, we establish asymptotic optimality results for the proposed estimators. To be more precise, we show that the proposed estimators are asymptotically optimal within a large class of estimators that contains the MLE.

Another related interesting problem often encountered in practice involves group testing and estimating the "difference" between two given groups. Consider the model

$$X_{ij} \overset{\text{ind}}{\sim} \text{LN}(M_i^{(1)}, \Sigma_i^{(1)}), \quad i = 1, \ldots, p,\ j = 1, \ldots, n_x,$$
$$Y_{ij} \overset{\text{ind}}{\sim} \text{LN}(M_i^{(2)}, \Sigma_i^{(2)}), \quad i = 1, \ldots, p,\ j = 1, \ldots, n_y,$$

where the X_{ij}'s and Y_{ij}'s are independent. We want to estimate the differences between $M_i^{(1)}$ and $M_i^{(2)}$ for $i = 1, \ldots, p$ and select the i's for which the differences are large. However, the selected estimates tend to overestimate the corresponding true differences. The bias introduced by the selection process is termed *selection bias* (Dawid 1994). Selection bias originates from the fact that there are two possible reasons for the selected differences to be large: (i) the true differences are large and (ii) the random errors contained in the estimates are large. Tweedie's formula (Efron 2011), which we discuss and briefly review in Section 3.3, deals with precisely this selection bias, in the context of the normal means problem. In this work, we apply an analogue of Tweedie's formula designed for the context of SPD matrices.

3.1 An Estimator of $\boldsymbol{M}$ When $\boldsymbol{\Sigma}$ is Known

For completeness, we briefly review the work of Yang & Vemuri (2019) where the authors presented the estimator for $\boldsymbol{M}$ assuming that $\Sigma_i = A_i I$ and the A_i's are known positive

numbers. Under this assumption, they considered the class of estimators given by

$$\widehat{M}_i^{\lambda,\mu} = \exp\left(\frac{n\lambda}{n\lambda + A_i}\log \bar{X}_i + \frac{A_i}{n\lambda + A_i}\log \mu\right), \tag{3}$$

where $\mu \in P_N$, $\lambda > 0$, and $\bar{X}_i$ is the sample FM of $X_{i1}, \ldots, X_{in}$. Using the Log-Euclidean distance as the loss function $L(\widehat{\boldsymbol{M}}, \boldsymbol{M}) = p^{-1}\sum_{i=1}^{p} d_{\mathrm{LE}}^2(\widehat{M}_i, M_i)$, they showed that the SURE for the corresponding risk function $R(\widehat{\boldsymbol{M}}, \boldsymbol{M}) = EL(\widehat{\boldsymbol{M}}, \boldsymbol{M})$ is given by

$$\mathrm{SURE}(\lambda, \mu) = \frac{1}{p}\sum_{i=1}^{p} \frac{A_i}{(n\lambda + A_i)^2}\left(A_i\|\log \bar{X}_i - \log \mu\|^2 + \frac{q(n^2\lambda^2 - A_i^2)}{n}\right),$$

where $q = N(N+1)/2$. Hence, λ and μ can be estimated by

$$\left(\hat{\lambda}^{\mathrm{SURE}}, \hat{\mu}^{\mathrm{SURE}}\right) = \arg\min_{\lambda,\mu} \mathrm{SURE}(\lambda, \mu).$$

Their shrinkage estimator for M_i is given by

$$\widehat{M}_i^{\mathrm{SURE}} = \exp\left(\frac{n\hat{\lambda}^{\mathrm{SURE}}}{n\hat{\lambda}^{\mathrm{SURE}} + A_i}\log \bar{X}_i + \frac{A_i}{n\hat{\lambda}^{\mathrm{SURE}} + A_i}\log \hat{\mu}^{\mathrm{SURE}}\right). \tag{4}$$

They also presented the following two theorems showing the asymptotic optimality of the shrinkage estimator.

Theorem 1 *Assume the following conditions:*

(i) $\limsup_{p\to\infty} p^{-1}\sum_{i=1}^{p} A_i^2 < \infty$,

(ii) $\limsup_{p\to\infty} p^{-1}\sum_{i=1}^{p} A_i\|\log M_i\|^2 < \infty$,

(iii) $\limsup_{p\to\infty} p^{-1}\sum_{i=1}^{p} \|\log M_i\|^{2+\delta} < \infty$ *for some* $\delta > 0$.

Then,

$$\sup_{\lambda>0, \|\log\mu\| < \max_i \|\log \bar{X}_i\|} \left|SURE(\lambda, \mu) - L(\widehat{\boldsymbol{M}}^{\lambda,\mu}, \boldsymbol{M})\right| \xrightarrow{prob} 0 \quad as\ p \to \infty.$$

Theorem 2 *If assumptions (i)–(iii) in Theorem 1 hold, then for every* $\lambda > 0$ *and* $\mu \in P_N$,

$$\limsup_{p\to\infty}\left[R(\widehat{\boldsymbol{M}}^{SURE}, \boldsymbol{M}) - R(\widehat{\boldsymbol{M}}^{\lambda,\mu}, \boldsymbol{M})\right] \le 0.$$

3.2 Estimators for M and Σ

In Yang & Vemuri (2019), the covariance matrices of the underlying distributions were assumed to be known, to simplify the derivation. In real applications however, the covariance matrices are rarely known, and in practice they must be estimated. In this paper, we consider the general case of unknown covariance matrices, which is more challenging and pertinent in real applications. Let

$$\begin{aligned} X_{ij}|(M_i, \Sigma_i) &\overset{\text{ind}}{\sim} \text{LN}(M_i, \Sigma_i) \\ M_i|\Sigma_i &\overset{\text{ind}}{\sim} \text{LN}(\mu, \lambda^{-1}\Sigma_i) \\ \Sigma_i &\overset{\text{iid}}{\sim} \text{Inv-Wishart}(\Psi, \nu), \end{aligned} \tag{5}$$

for $i = 1, \ldots, p$ and $j = 1, \ldots, n$. The prior for (M_i, Σ_i) is called the Log-Normal-Inverse-Wishart (LNIW) prior, and it is motivated by the normal-inverse-Wishart prior in the Euclidean space setting. We emphasize that the main reason for choosing the LNIW prior over others is the property of conjugacy which leads to a closed-form expression for our estimators. Let

$$\bar{X}_i = \exp\left(n^{-1}\sum_{j=1}^{n}\log X_{ij}\right) \quad \text{and} \quad S_i = \sum_{j=1}^{n}\big(\widetilde{X}_{ij} - \widetilde{\bar{X}}_i\big)\big(\widetilde{X}_{ij} - \widetilde{\bar{X}}_i\big)^T. \tag{6}$$

Then the posterior distributions of M_i and Σ_i are given by

$$\begin{aligned} M_i|\big(\{X_{ij}\}_{i,j}, \{\Sigma_i\}_{i=1}^p\big) &\sim \text{LN}\Big(\exp\Big(\frac{n\log\bar{X}_i + \lambda\log\mu}{\lambda+n}\Big), (\lambda+n)^{-1}\Sigma_i\Big), \\ \Sigma_i|S_i &\sim \text{Inv-Wishart}(\Psi + S_i, \nu + n - 1), \end{aligned}$$

and the posterior means for M_i and Σ_i are given by

$$\widehat{M}_i = \exp\Big(\frac{n\log\bar{X}_i + \lambda\log\mu}{\lambda+n}\Big) \quad \text{and} \quad \widehat{\Sigma}_i = \frac{\Psi + S_i}{\nu + n - q - 2}. \tag{7}$$

Consider the loss function

$$L\big((\widehat{\boldsymbol{M}}, \widehat{\boldsymbol{\Sigma}}), (\boldsymbol{M}, \boldsymbol{\Sigma})\big) = p^{-1}\sum_{i=1}^{p} d^2_{\text{LE}}(\widehat{M}_i, M_i) + p^{-1}\sum_{i=1}^{p}\|\widehat{\Sigma}_i - \Sigma_i\|^2 = L_1(\widehat{\boldsymbol{M}}, \boldsymbol{M}) + L_2(\widehat{\boldsymbol{\Sigma}}, \boldsymbol{\Sigma}).$$

Its induced risk function is

$$
\begin{aligned}
R\big((\widehat{\boldsymbol{M}}, \widehat{\boldsymbol{\Sigma}}), (\boldsymbol{M}, \boldsymbol{\Sigma})\big) &= p^{-1} \sum_{i=1}^{p} \big[E d_{\mathrm{LE}}^2(\widehat{M}_i, M_i) + E\|\widehat{\Sigma}_i - \Sigma_i\|^2\big] \\
&= p^{-1}(\lambda+n)^{-2} \sum_{i=1}^{p} \big[n \mathrm{tr}\Sigma_i + \lambda^2 d_{\mathrm{LE}}^2(\mu, M_i)\big] \\
&\quad + p^{-1} \sum_{i=1}^{p} (\nu+n-q-2)^{-2} \Big[\big(n-1+(\nu-q-1)^2\big)\mathrm{tr}(\Sigma_i^2) \\
&\quad - 2(\nu-q-1)\mathrm{tr}(\Psi\Sigma_i) + (n-1)(\mathrm{tr}\Sigma_i)^2 + \mathrm{tr}(\Psi^2)\Big],
\end{aligned}
$$

with the detailed derivation given in the supplement. The SURE for this risk function is

$$
\begin{aligned}
\mathrm{SURE}(\lambda, \Psi, \nu, \mu) = p^{-1} \Bigg\{ & \sum_{i=1}^{p} (\lambda+n)^{-2} \Big[\frac{n-\lambda^2/n}{n-1} \mathrm{tr}\, S_i + \lambda^2 d_{\mathrm{LE}}^2(\bar{X}_i, \mu)\Big] \\
& + (\nu+n-q-2)^{-2} \Big[\frac{n-3+(\nu-q-1)^2}{(n+1)(n-2)} \mathrm{tr}(S_i^2) \\
& + \frac{(n-1)^2-(\nu-q-1)^2}{(n-1)(n+1)(n-2)} \big(\mathrm{tr}\, S_i\big)^2 - 2\frac{\nu-q-1}{n-1} \mathrm{tr}(\Psi S_i) + \mathrm{tr}(\Psi^2)\Big] \Bigg\},
\end{aligned} \tag{8}
$$

with the detailed derivation given in the supplement.

Remark Note that instead of the LNIW prior, one may also consider a prior that captures the correlation structure (if there is any) among both the M_i's and the Σ_i's, which may be a more appropriate prior for some applications encountered in image analysis. However, such a prior will make the ensuing mathematical analysis much more complicated than it already is. Hence we stay with the LNIW prior, and in Section 4 we take a more practical and effective approach (involving smoothing) to deal with the aforementioned correlation structure.

The hyperparameter vector $(\lambda, \Psi, \nu, \mu)$ is estimated by minimizing the risk estimate $\mathrm{SURE}(\lambda, \Psi, \nu, \mu)$, and the resulting shrinkage estimators of M_i and Σ_i are obtained by plugging in the minimizing vector into (7). This optimization step allows us to determine the shrinkage from the data. So, unlike the original James-Stein estimator which shrinks the estimate towards a fixed target, we are able to obtain nearly optimal estimates. Note that this is a non-convex optimization problem, and for such problems convergence relies heavily

on the choice of the initialization. We suggest the following initialization, which is discussed in the supplemental material:

$$\begin{aligned}
\mu_0 &= \exp\left(p^{-1}\sum_{i=1}^{p}\log \bar{X}_i\right),\\
\lambda_0 &= \frac{np^{-1}\sum_{i=1}^{p} d^2_{\mathrm{LE}}(\bar{X}_i,\mu_0)}{\frac{n}{p(n-1)}\sum_{i=1}^{p}\operatorname{tr} S_i - p^{-1}\sum_{i=1}^{p} d^2_{\mathrm{LE}}(\bar{X}_i,\mu_0)},\\
\nu_0 &= \frac{q+1}{\frac{n-q-2}{p^2q(n-1)}\operatorname{tr}\left[\left(\sum_{i=1}^{p} S_i\right)\left(\sum_{i=1}^{p} S_i^{-1}\right)\right]-1} + q + 1,\\
\Psi_0 &= \frac{\nu_0 - q - 1}{p(n-1)}\sum_{i=1}^{p} S_i.
\end{aligned}$$

In all our experiments, the algorithm converged in fewer than 20 iterations with this initialization. This concludes the description of our estimators of the unknown means and covariance matrices. Theorem 3 below states that $\mathrm{SURE}(\lambda,\Psi,\nu,\mu)$ approximates the true loss $L\big(\big(\widehat{\boldsymbol{M}}^{\lambda,\mu},\widehat{\boldsymbol{\Sigma}}^{\Psi,\nu}\big),(\boldsymbol{M},\boldsymbol{\Sigma})\big)$ well in the sense that the difference between the two random variables converges to 0 in probability as $p\to\infty$. Additionally, Theorem 4 below shows that the estimators of $\boldsymbol{M}$ and $\boldsymbol{\Sigma}$ obtained by minimizing $\mathrm{SURE}(\lambda,\Psi,\nu,\mu)$ are asymptotically optimal in the class of estimators of the form (7).

Theorem 3 *Assume the following conditions:*

(i) $\limsup_{p\to\infty} p^{-1}\sum_{i=1}^{p}\left(\operatorname{tr}\Sigma_i\right)^4 < \infty$,

(ii) $\limsup_{p\to\infty} p^{-1}\sum_{i=1}^{p}\widetilde{M}_i^T\Sigma_i\widetilde{M}_i < \infty$,

(iii) $\limsup_{p\to\infty} p^{-1}\sum_{i=1}^{p}\|\log M_i\|^{2+\delta} < \infty$ *for some* $\delta>0$.

Then

$$\sup_{\substack{\lambda>0,\nu>q+1,\|\Psi\|\le\max_{1\le i\le p}\|S_i\|,\\ \|\log\mu\|\le\max_{1\le i\le p}\|\log\bar{X}_i\|}}\left|SURE(\lambda,\Psi,\nu,\mu) - L\Big(\big(\widehat{\boldsymbol{M}}^{\lambda,\mu},\widehat{\boldsymbol{\Sigma}}^{\Psi,\nu}\big),(\boldsymbol{M},\boldsymbol{\Sigma})\Big)\right| \xrightarrow{prob} 0 \quad as\ p\to\infty.$$

Note that the optimization has some constraints. However, in practice, with proper initialization as suggested earlier, the constraints on Ψ and μ can be safely ignored. The reason is that, for Ψ and μ far from S_i's and $\bar{X}_i$ respectively, the value of SURE will be large. The constraints on λ and ν can easily be handled by standard constrained optimization algorithms, e.g. L-BFGS-B (Byrd et al. 1995).

Theorem 4 *If assumptions (i)–(iii) in Theorem 3 hold, then*

$$\limsup_{p\to\infty}\left[R\Big((\widehat{\boldsymbol{M}}^{SURE},\widehat{\boldsymbol{\Sigma}}^{SURE}),(\boldsymbol{M},\boldsymbol{\Sigma})\Big)-R\Big((\widehat{\boldsymbol{M}}^{\lambda,\mu},\widehat{\boldsymbol{\Sigma}}^{\Psi,\nu}),(\boldsymbol{M},\boldsymbol{\Sigma})\Big)\right]\leq 0.$$

Note that in all the theorems above, we consider the asymptotic regime $p\to\infty$ while the size of the SPD matrix N is held fixed. The main reason for fixing the size of the SPD matrix is that in our first application, namely the DTI analysis, the size of the diffusion tensors is always 3×3, because diffusion magnetic resonance images are 3-dimensional images. However, the number of voxels p can increase, as they are determined by the resolution of the acquired image, which can increase due to advances in medical imaging technology. This is different from the usual high-dimensional covariance matrix estimation problem in which the size of the covariance matrix is allowed to grow.

Remark The proofs of Theorems 1 and 2 in Yang & Vemuri (2019) use arguments similar to those that already exist in the literature, and in that sense they are not very difficult. In contrast, the proofs of our Theorems 3 and 4 do not proceed along familiar lines. Indeed, they are rather complicated, the difficulty being that bounding the moments of Wishart matrices or the moments of the trace of Wishart matrices is nontrivial when the orders of the required moments are higher than two. We present these proofs in the supplement.

Remark SURE.Full-FM estimates the covariance matrices Σ_i's, and this results in performance that is worse than if the covariance matrices were known. On the other hand, FM.LE and SURE-FM assume that the covariance matrices are known, and since this is not the case, we need to estimate them. So we're not really comparing SURE.Full-FM with FM.LE and SURE-FM, but rather with versions of FM.LE and SURE-FM in which the covariance matrices are estimated. In SURE.Full-FM they are estimated jointly via shrinkage, whereas in FM.LE/SURE-FM they are estimated separately.

3.3 Tweedie's Formula for F Statistics

One of the motivations for the development of our approach for estimating $\boldsymbol{M}$ and $\boldsymbol{\Sigma}$ is a problem in neuroimaging involving detection of differences between a patient group and a control group. The problem can be stated as follows. There are n_x patients in a disease

group and n_y normal subjects in a control group. We consider a region of the brain image consisting of p voxels. As explained in Section 4.2, the local diffusional property of water molecules in the human brain is of clinical importance, and it is common to capture this diffusional property at each voxel in diffusion magnetic resonance imaging (dMRI) via a zero-mean Gaussian with a 3×3 covariance matrix. Using any of the existing state-of-the-art dMRI analysis techniques, it is possible to estimate, from each patient image, the diffusion tensor M_i corresponding to voxel i, for $i = 1, \ldots, p$. Let $M_i^{(1)}$ and $M_i^{(2)}$ denote the diffusion tensors corresponding to voxel i for the disease and control groups respectively. The goal is to identify the indices i for which the difference between $M_i^{(1)}$ and $M_i^{(2)}$ is large. The model we consider is

$$X_{ij} \overset{\text{ind}}{\sim} \text{LN}(M_i^{(1)}, \Sigma_i), \quad i = 1, \ldots, p, \ j = 1, \ldots, n_x,$$
$$Y_{ij} \overset{\text{ind}}{\sim} \text{LN}(M_i^{(2)}, \Sigma_i), \quad i = 1, \ldots, p, \ j = 1, \ldots, n_y.$$

In this work, we use the Hotelling T^2 statistic for each $i = 1, \ldots, p$ as a measure of the difference between $M_i^{(1)}$ and $M_i^{(2)}$. The Hotelling T^2 statistic for SPD matrices has been proposed by Schwartzman et al. (2010), and is given by

$$t_i^2 = \big(\widetilde{\bar{X}}_i - \widetilde{\bar{Y}}_i\big)^T \Big[\Big(\frac{1}{n_x} + \frac{1}{n_y}\Big) S_i\Big]^{-1} \big(\widetilde{\bar{X}}_i - \widetilde{\bar{Y}}_i\big), \tag{9}$$

where $\bar{X}_i$ and $\bar{Y}_i$ are the FMs of $\{X_{ij}\}_j$ and $\{Y_{ij}\}_j$, and $S_i = (n_x + n_y - 2)^{-1}\big(S_i^{(1)} + S_i^{(2)}\big)$ is the pooled estimate of Σ_i where $S_i^{(1)}$ and $S_i^{(2)}$ are computed using (6). Since the X_{ij}'s and Y_{ij}'s are Log-normally distributed, one can easily verify that the distribution of t_i^2 is given by

$$\frac{\nu - q - 1}{\nu q} t_i^2 \overset{\text{ind}}{\sim} F_{q, \nu - q - 1, \lambda_i}, \tag{10}$$

where $\nu = n_x + n_y - 2$ is a degrees of freedom parameter (recall that $q = N(N+1)/2$). Note that we make the assumption that the covariance matrices for the two groups are equal, i.e. $\Sigma_i^{(1)} = \Sigma_i^{(2)} = \Sigma_i$. Similar results can be obtained for the unequal covariance case, but with more complicated expressions for the T^2 statistics and the degrees of freedom parameters. The λ_i's are the non-centrality parameters for the non-central F distribution and are given by

$$\lambda_i = \Big(\frac{1}{n_x} + \frac{1}{n_y}\Big)^{-1} \big(\widetilde{M}_i^{(1)} - \widetilde{M}_i^{(2)}\big)^T \Sigma_i^{-1} \big(\widetilde{M}_i^{(1)} - \widetilde{M}_i^{(2)}\big).$$

These non-centrality parameters can be interpreted as the (squared) differences between $M_i^{(1)}$ and $M_i^{(2)}$, and they are the parameters we would like to estimate using the statistics (9) computed from the data. Then, based on estimates $\hat{\lambda}_i$, we select those i's with large estimates, say the largest 1% of all $\hat{\lambda}_i$. However, the process of selection from the computed estimates introduces selection bias. The bias comes from the fact that it is possible to select some indices i's for which the actual λ_i's are not large but the random errors are large, so that the estimates $\hat{\lambda}_i$ are pushed away from the true parameters λ_i. There are several ways to correct for this bias, and Efron (2011) proposed to use *Tweedie's formula* for such a purpose.

Tweedie's formula was first proposed by Robbins (1956), and we review this formula here in the context of the classical normal means problem, which is stated as follows. We observe $Z_i \overset{\text{ind}}{\sim} N(\mu_i, \sigma^2)$, $i = 1, \ldots, p$, where the μ_i's are unknown and σ^2 is known, and the goal is to estimate the μ_i's. In the empirical Bayes approach to this problem we assume that the μ_i's are iid according to some distribution G. The marginal density of the Z_i's is then $f(z) = \int \phi_\sigma(z - \mu)\, dG(\mu)$, where ϕ_σ is the density of the $N(0, \sigma^2)$ distribution. With this notation, if G is known (so that f is known), the best estimator of μ_i (under squared error loss) is the so-called Tweedie estimator given by $\hat{\mu}_i = Z_i + \sigma^2[f'(Z_i)/f(Z_i)]$. A feature of this estimator is that it depends on G only through f, and this is desirable because it is fairly easy to estimate f from the Z_i's (so we don't need to specify G). Another interesting observation about this estimator is that $\hat{\mu}_i$ is shrinking the MLE $\hat{\mu}_i^{\text{MLE}} = Z_i$ and can be viewed as a generalization of the James-Stein estimator which assumes $\mu_i \overset{\text{iid}}{\sim} N(0, \lambda)$ with unknown λ. The Tweedie estimator can be generalized to exponential families. Suppose that $Z_i|\eta_i \overset{\text{ind}}{\sim} f_{\eta_i}(z) = \exp(\eta_i z - \phi(\eta_i)) f_0(z)$, and the prior for ϕ is G. Then the Tweedie estimator for η_i is $\hat{\eta}_i = l'(Z_i) - l_0'(Z_i)$, where $l(z) = \log \int f_\eta(z)\, dG(\eta)$ is the log of the marginal likelihood of the Z_i's and $l_0(z) = \log f_0(z)$.

Although this formula is elegant and useful, it applies only to exponential families. Recently, Du & Hu (2020) derived a Tweedie-type formula for non-central χ^2 statistics, for situations where one is interested in estimating the non-centrality parameters. Suppose $Z_i|\lambda_i \overset{\text{ind}}{\sim} \chi^2_{\nu,\lambda_i}$ and $\lambda_i \overset{\text{iid}}{\sim} G$. Then,

$$E(\lambda_i|Z_i) = \Big[(Z_i - \nu + 4) + 2Z_i\Big(\frac{2l_\nu''(Z_i)}{1 + 2l_\nu'(Z_i)} + l_\nu'(Z_i)\Big)\Big]\big(1 + 2l_\nu'(Z_i)\big),$$

where $l_\nu(\cdot)$ is the marginal log-likelihood of the Z_i's (see Theorem 1 in Du & Hu (2020)).

For our situation, if we define $Z_i = [(\nu - q - 1)/\nu q]t_i^2$, then $Z_i|\lambda_i \overset{\text{ind}}{\sim} F_{q,\nu-q-1,\lambda_i}$ (recall that $\nu = n_x + n_y - 2$, see (9) and (10)). Assume that the λ_i's are iid according to some distribution G. We now address the problem of how to obtain empirical Bayes estimates of the λ_i's. Let $\Phi_{\nu_1,\nu_2,\lambda}$ be the cumulative distribution function (cdf) of the non-central F distribution, $F_{\nu_1,\nu_2,\lambda}$, and let $\tilde{\Phi}_{\nu,\lambda}$ be the cdf of the non-central χ^2 distribution, $\chi^2_{\nu,\lambda}$. Then the transformed variable $Y_i = \tilde{\Phi}^{-1}_{\nu_1,\lambda_i}(\Phi_{\nu_1,\nu_2,\lambda_i}(Z_i))$ follows a non-central χ^2 distribution with degrees of freedom parameter ν_1 and non-centrality parameter λ_i, and we note that when ν_2 is large, $\Phi_{\nu_1,\nu_2,\lambda_i}$ and $\tilde{\Phi}_{\nu_1,\lambda_i}$ are nearly equal, so that this quantile transformation is nearly the identity. However, the transformation depends on λ_i, which is the parameter to be estimated, so we propose the following iterative algorithm for estimating $E(\lambda_i|Z_i)$. Let $\lambda_i^{(t)}$ be the estimate of λ_i at the t-th iteration. Then our iterative update of λ_i is given by

$$\lambda_i^{(t+1)} = \Big[(Y_i^{(t)} - \nu_1 + 4) + 2Y_i^{(t)}\Big(\frac{2l''_{\nu_1}(Y_i^{(t)})}{1 + 2l'_{\nu_1}(Y_i^{(t)})} + l'_{\nu_1}(Y_i^{(t)})\Big)\Big]\big(1 + 2l'_{\nu_1}(Y_i^{(t)})\big),$$

where $Y_i^{(t)} = \tilde{\Phi}^{-1}_{\nu_1,\lambda_i^{(t)}}\big(\Phi_{\nu_1,\nu_2,\lambda_i^{(t)}}(Z_i)\big)$, $\nu_1 = q$, and $\nu_2 = \nu - q - 1$. Now the marginal log-likelihood $l_{\nu_1}(y)$ is not available since the prior G for λ_i is unknown. There are several ways to estimate the marginal density of the $Y_i^{(t)}$'s. One of these is through kernel density estimation. However, the iterative formula involves the first and second derivatives of the marginal log-likelihood, and estimates of the derivatives of a density produced through kernel methods are notoriously unstable. There exist different approaches for dealing with this problem (see Sasaki et al. (2016) and Shen & Ghosal (2017)). Here we follow Efron (2011) and postulate that l_{ν_1} is well approximated by a polynomial of degree K, and write $l_{\nu_1}(y) = \sum_{k=0}^K \beta_k y^k$. The coefficients β_k, $k = 1, \ldots, K$, can be estimated via *Lindsey's method* (Efron & Tibshirani 1996), which is a Poisson regression technique for (parametric) density estimation; the coefficient β_0 is determined by the requirement that $f_{\nu_1}(y) = \exp(l_{\nu_1}(y))$ integrates to 1. The advantage of Lindsey's method over methods that use kernel density estimation is that it does not require us to estimate the derivatives separately, since $l'_{\nu_1}(y) = \sum_{k=1}^K k\beta_k y^{k-1}$ and $l''_{\nu_1}(y) = \sum_{k=2}^K k(k-1)\beta_k y^{k-2}$. In our experience, with l'_{ν_1} and l''_{ν_1} estimated in this way, if we initialize the scheme by setting $\lambda_i^{(0)}$ to be the estimate of λ_i given by the Du & Hu (2020) procedure, then the algorithm converges in fewer than 10 iterations.

4 Experimental Results

Here we describe the performance of our methods on three synthetic data sets and three sets of real data from diffusion and functional magnetic resonance imaging. Details on these data sets will be given subsequently. For the synthetic data experiments, we show the following.

(i) The proposed shrinkage estimator for the FM (SURE.Full-FM, *with* simultaneous estimation of the covariance matrices) outperforms the sample FM (FM.LE) and the shrinkage estimator proposed by Yang & Vemuri (2019) (SURE-FM, *with fixed* covariance matrices); see Section 4.1.1.

(ii) Our estimator outperforms its competitors for different (increasing) values of N, the size of the SPD matrices; see Section 4.1.2.

(iii) The shrinkage estimates of the group differences capture the regions that are significantly distinct between two groups of SPD matrix-valued images; see Section 4.1.3.

For the real data experiments, we demonstrate the following.

(iv) The SURE.Full-FM provides improvement over the two competing estimators (FM.LE and SURE-FM) for (a) computing an atlas (template) of diffusion tensor images acquired from human brains (Section 4.2.1) and (b) computing the mean connectivity networks from resting state functional MRI (fMRI) measurements (Section 4.2.2). The former experiment tests the framework for the accuracy by varying the spatial dimension, p, of the images (i.e. the number of voxels), and the latter tests the accuracy under varying size, N, of the SPD matrices.

(v) The proposed shrinkage estimator for detecting group differences is able to identify the regions that are significantly distinct between patients with Parkinson's disease and control subjects; see Section 4.2.3.

Details of these experiments are presented in the following paragraphs.

4.1 Synthetic Data Experiments

We present three synthetic data experiments here to show that the proposed shrinkage estimator, SURE.Full-FM, outperforms the sample FM and SURE-FM under varying sample

sizes and sizes of SPD matrices and that the shrinkage estimates of the group differences can accurately localize the regions that are significantly different between the two groups.

4.1.1 Comparison Between SURE.Full-FM and Competing Estimators

Using generated noisy SPD fields (P_3) as data, we now present performance comparisons of three estimators of $\boldsymbol{M}$: (i) SURE.Full-FM, which is the proposed shrinkage estimator, (ii) SURE-FM, proposed by Yang & Vemuri (2019) which assumes that the covariance matrices are known and (iii) the MLE, which is denoted by FM.LE (since by proposition 1 it is the FM based on the Log-Euclidean metric). The synthetic data are generated according to (5). Specifically, we set $\mu = I_3$, $\Psi = I_6$, and $n = 10$, and we vary the variance λ and the degrees of freedom parameter ν of the prior distribution as follows: $\lambda = 10, 50$, and $\nu = 15, 30$. Figure 1 shows the relationship between the average loss (averaged over $m = 1000$ repetitions) and the spatial dimension p under varying conditions for the three estimators. Note that, since the covariance matrices Σ_i's are unknown in our synthetic data experiment and $(n-1)^{-1}S_i$ is an unbiased estimate for Σ_i, the A_i's in (4) can be unbiasedly estimated by $[(n-1)q]^{-1}\,\mathrm{tr}\,S_i$. As is evident from Figure 1, for large λ the gains from using SURE.Full-FM are greater.

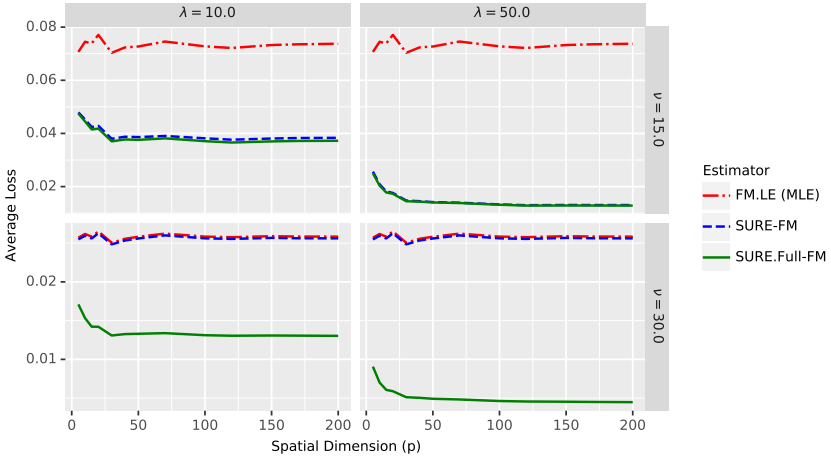


Figure 1: Average loss for each of the three estimators. Results for varying λ and degrees of freedom parameter ν are shown across the columns and rows respectively. Note that in the bottom two panels, the curve corresponding to FM.LE is essentially the same as the one for SURE.FM, but is barely visible.

This observation is in accordance with our intuition, which is that for large λ, the M_i's are clustered, and it is beneficial to shrink the MLEs of the M_i's towards a common value. The main difference between SURE-FM and SURE.Full-FM is that the former requires knowledge of the Σ_i's and in general such information is not available, and estimates for the Σ_i's are needed to compute the SURE-FM. Hence, the performance of SURE-FM depends heavily on how good the estimates for the Σ_i's are. In our synthetic data experiment, we consider the unbiased estimate $\hat{A}_i = [(n-1)q]^{-1} \operatorname{tr} S_i$ for SURE-FM. In this case, the prior mean for Σ_i is $E(\Sigma_i) = (\nu - q - 1)^{-1} I_q$, for which the assumption $\Sigma_i = A_i I$ seems reasonable. For large ν, the generated Σ_i's are far from being identity matrices, which violates the assumption (this can be observed in Figure 1, where we see that SURE-FM is almost identical to FM.LE for $\nu = 30$).

On the other hand, we can fix λ and ν to see how different choices of μ and Ψ affect the performance of our shrinkage estimator SURE.Full-FM. To do this, we fix $n = 10$, $\lambda = 10$, and $\nu = 15$ (so that we can compare with the top-left panel of Figure 1). We consider $\mu = \operatorname{diag}(2, 0.5, 0.5)$ and $\Psi_{ij} = 0.5^{|i-j|}$. The results are shown in Figure 2. The top-left panel of Figure 1 shows that when $\mu = I$ and $\Psi = I$, there is no difference between SURE-FM and SURE.Full-FM, but Figure 2 shows that when one of μ and Ψ is not the identity, our shrinkage estimator outperforms SURE-FM. For different choices of λ and ν, the improvement is more significant, following the trend we observed in Figure 1. Note that the chosen hyperparameters here, $\lambda = 10, 50$ and $\nu = 15, 30$, are extreme. What we aim to show with this particular choice in the simulation is that the SURE-FM can perform either as well as the SURE.Full-FM or as poorly as the MLE. That is, the performance of SURE.Full-FM is the best possible performance that SURE-FM can achieve.

4.1.2 Performance Comparisons as Matrix Size Varies

In this subsection, we present some experiments to assess the improvement when we vary the size of the matrices. In particular, we consider the size N of the SPD matrices increasing from 10 to 100, and we fix $p = 10$, $n = 5$, $\mu = I_3$, and $\Psi = I_q = I_{N(N+1)/2}$ throughout the experiments. The results are shown in Figure 3. As we can see from the figure, although the average loss increases with increasing size (of the SPD matrices), the improvement of our

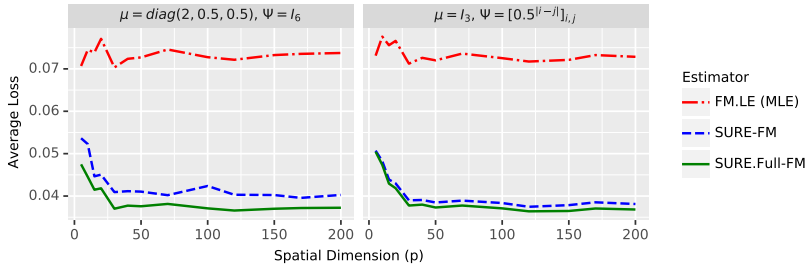


Figure 2: Average loss for each of the three estimators. The left panel assumes $\mu = \text{diag}(2, 0.5, 0.5)$ and $\Psi = I$ and the right panel assumes $\mu = I$ and $\Psi_{ij} = 0.5^{|i-j|}$.

estimator over the competitors is more significant for large N.

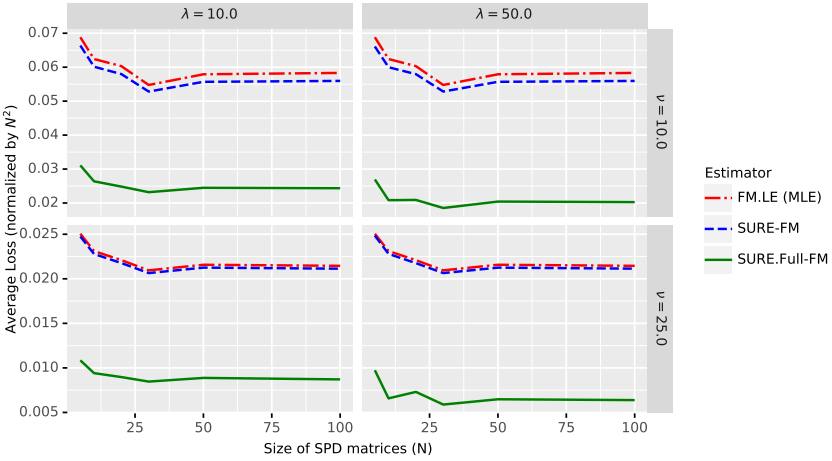


Figure 3: Average loss (normalized by N^2) for each of the three estimators. Results for varying λ and degrees of freedom parameter ν are shown across the columns and rows, respectively.

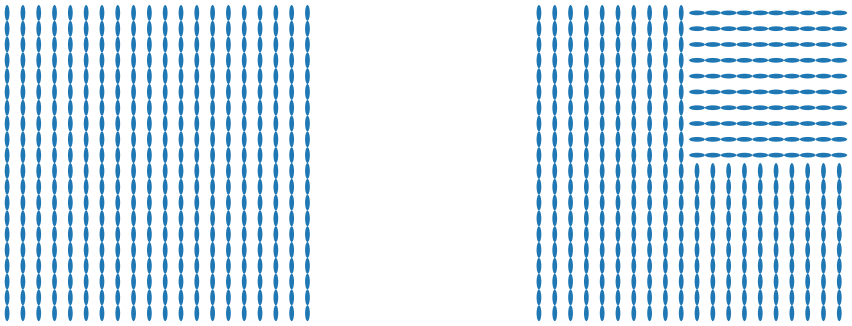
4.1.3 Differences Between Two Groups of SPD-Valued Images

In this subsection, we demonstrate the method proposed in Section 3.3 for evaluating the difference between two groups of SPD-valued images. For this synthetic data experiment, we use P_2, the manifold of 2×2 SPD matrices, since it is easy to visualize these matrices. For the visualization, we represent each 2×2 SPD matrix of the SPD-valued image by an ellipse with the two eigenvectors as the axes of the ellipse and the two eigenvalues as the

width and height along the corresponding axes. The data are created as follows. Given n_k, $M_i^{(k)}$, σ_i^2, $k = 1, 2$, $i = 1, \ldots, p$, generate

$$X_{ij} \overset{\text{ind}}{\sim} \text{LN}(M_i^{(1)}, \sigma_i^2 I), \quad j = 1, \ldots, n_1,$$
$$Y_{ij} \overset{\text{ind}}{\sim} \text{LN}(M_i^{(2)}, \sigma_i^2 I), \quad j = 1, \ldots, n_2.$$

We generate $n_1 = n_2 = 30$ P_2-valued images for the two groups, and the size of each P_2-valued image is 20×20, which gives $p = 20 \times 20 = 400$. For the variances σ_i^2, we consider a low-variance scenario $\sigma_i^2 \overset{\text{iid}}{\sim} U(0.1, 0.3)$ and a high-variance scenario $\sigma_i^2 \overset{\text{iid}}{\sim} U(0.3, 0.8)$. The means $M_i^{(k)}$ are depicted visually in Figure 4 (in the form of images with ellipses instead of gray values at each pixel), and the region in which the means are different is the top-right corner, containing a quarter of the pixels; this is the 'ground truth' data.

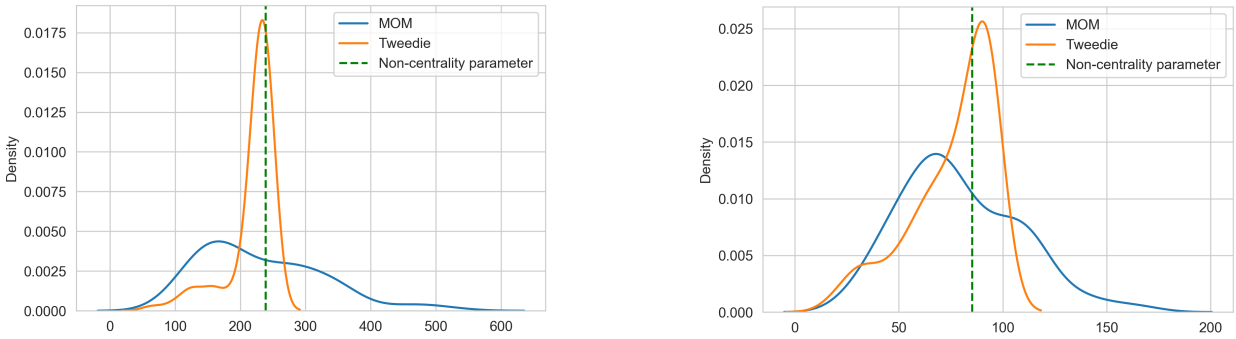


(a) Group 1 (b) Group 2

Figure 4: The mean P_2-valued images $M_i^{(k)}$, $k = 1, 2$, used to generate random P_2-valued images for the two groups. The vertical ellipse represents the matrix $\text{diag}(0.3, 1)$ and the horizontal ellipse represents the matrix $\text{diag}(1, 0.3)$.

As described in Section 3.3, we first compute the Hotelling T^2 statistic from $\{X_{ij}\}_{j=1}^{n_1}$ and $\{Y_{ij}\}_{j=1}^{n_2}$ for each i and transform each of them to an F statistic. We then have p non-central F statistics, $f_i \overset{\text{ind}}{\sim} F_{\nu_1, \nu_2, \lambda_i}$, $i = 1, \ldots, p$, where $\nu_1 = q = 3$, and $\nu_2 = n_1 + n_2 - 2 - q - 1 = n_1 + n_2 - 6$. With the resulting F statistics, we can apply the algorithm described in Section 3.3 to estimate the non-centrality parameters (at each location), and for the estimation of the marginal log likelihood, we adopt Lindsey's method and fit a polynomial of degree $K = 5$ to the log-likelihood l_{ν_1}. We have experimented using different values of

K, and we found that the results are robust to changes in K, at least for relatively small K. In our experiments, we set $n_1 = n_2 = 30$. As we can see from Figure 4, we expect the method to yield large values on the top-right corner of the image and small values for the rest of the matrix-valued image (field). We compare the proposed estimator $\hat{\lambda}_i^{\text{Tweedie}}$ to the estimator $\hat{\lambda}_i^{\text{MOM}} = \max\left(\frac{\nu_1(\nu_2-2)}{\nu_2} f_i - \nu_1, 0\right)$, which is obtained by the method of moments (MOM) and truncated at 0, and also compare them for different σ_i^2's. We compare with the MOM estimator instead of the MLE for two reasons: (i) the MLE for the non-centrality parameter of the non-central F distribution is expensive to compute, and (ii) the MOM is commonly used as a standard for comparison, see for example Kubokawa et al. (1993). Figure 5 gives the results. As we can see, the density of the Tweedie-adjusted estimates concentrates in a smaller region when compared to that of the MOM estimates. This shows that the Tweedie-adjusted estimator allows us to capture the true region of difference better than does the MOM estimator, especially for small σ_i^2's. This is due to the shrinkage effect in Tweedie's formula.



(a) $\sigma_i^2 \overset{\text{iid}}{\sim} U(0.1, 0.3)$ (b) $\sigma_i^2 \overset{\text{iid}}{\sim} U(0.3, 0.8)$

Figure 5: Comparison of densities of the proposed shrinkage estimates of the non-centrality parameters and those of the MOM estimates. The green dashed lines indicate the true non-centrality parameters.

4.2 Real Data Experiments

In this section, we present three real data experiments involving dMRI and rs-fMRI datasets. The dMRI data we use are publicly available via https://pdbp.ninds.nih.gov/our-data.

dMRI is a diagnostic imaging technique that allows one to non-invasively probe the axonal fiber connectivity in the body by making the magnetic resonance signal sensitive to water diffusion through the tissue being imaged. In dMRI, the water diffusion is fully characterized by the probability density function of the displacement of water molecules, called the ensemble average propagator (EAP) (Callaghan 1993). A simple model that has been widely used to describe the displacement of water molecules is a zero mean Gaussian; its covariance matrix defines the diffusion tensor and characterizes the diffusivity function locally. The diffusion tensors are 3×3 SPD matrices and hence have 6 unique entries that need to be determined. Thus, the diffusion imaging technique employed in this case involves the application of at least 6 diffusion sensitizing magnetic gradients for acquisition of full 3D MR images (Basser et al. 1994). This dMRI technique is called diffusion tensor imaging (DTI). Some practical techniques for estimating the diffusion tensors and the population mean of diffusion tensors have been reported in Wang & Vemuri (2004), Chefd'Hotel et al. (2004), Fletcher & Joshi (2004), Alexander (2005), Zhou et al. (2008), Lenglet, Rousson & Deriche (2006), and Dryden et al. (2009).

DTI has been the de facto non-invasive dMRI diagnostic imaging technique of choice in the clinic for a variety of neurological ailments. After fitting/estimating the diffusion tensors at each voxel, scalar-valued or vector-valued measures are derived from the diffusion tensors for further analysis. For instance, fractional anisotropy (FA) is a scalar-valued function of the eigenvalues of the diffusion tensor and it was found that FA was reduced in the neuro-anatomical structure called the Substantia Nigra in patients with Parkinson's disease compared to control subjects (Vaillancourt et al. 2009). In Schwartzman et al. (2010) the authors used the full tensor information, and we adopt the same strategy here since the full tensor captures both the eigenvalues and eigenvectors, which can prove to be much more useful (compared to FA or other scalar measures) in order to assess the changes caused by pathologies to the underlying tissue micro-architecture revealed by dMRI.

4.2.1 Estimation of the Motor Sensory Tracts of Patients with Parkinson's Disease

In this section, we demonstrate the performance of SURE.Full-FM on the dMRI scans of human brain data acquired from 50 patients with Parkinson's disease and 44 control (normal) subjects. The dMRI acquisition parameters were as follows: repetition time = 7748ms, echo time = 86ms, flip angle = 90°, number of diffusion gradients = 64, field of view = 224×224 mm, in-plane resolution = 2mm isotropic, slice-thickness = 2mm, and SENSE factor = 2. All the dMRI data were pre-registered into a common coordinate frame prior to any further data processing.

The motor sensory area fiber tracts (M1 fiber tracts) are extracted from each patient of the two groups using the template described in Archer et al. (2017), which is freely available from `http://lrnlab.org`. The size (length) of each tract is 33 voxels for the left hemisphere tract and 34 voxels for the right hemisphere tract. Diffusion tensors are then fit to each of the voxels along each of the tracts to obtain $p = 33$ ($p = 34$) 3×3 SPD matrices. We then compute the Log-Euclidean FM tract for each group over the patients, i.e., the FM tract here also has 33 (34) diffusion tensors along the tract. We will use these FMs computed from the full population for each group as the 'ground truth'; thus, the underlying distribution in this experiment is the empirical distribution formed by the observed data, i.e. the 33 (34) SPD matrices. Then, we randomly draw a subsample of size $n = 10, 20, 50, 100$, with replacement, from each group and compute the SURE.Full-FM (our proposed estimator) and the two competing estimators (FM.LE and SURE-FM respectively) for each group and for each subsample size n. An explanation of why sampling is done with replacement is given in Section 5 of the supplementary document. We compare the performance of the different estimators by the Log-Euclidean distance between the estimator and the 'ground truth' FMs. The entire procedure is repeated for $m = 100$ random draws of subsamples and the average distances are reported in Table 1. Since our proposed shrinkage estimator jointly estimates the FM and the covariance matrices, we also compare our covariance estimates, denoted SURE.Full-Cov, with the MLE of the covariance matrices, i.e., the sample covariance matrices. The results are shown in Table 2.

Table 1: Average loss for the three estimators in estimating the population FM for varying n (with the standard errors in parentheses).

n	10	20	50	100
FM.LE	0.774 (0.03)	0.405 (0.01)	0.159 (0.005)	0.079 (0.002)
SURE-FM	0.772 (0.03)	0.404 (0.01)	0.159 (0.005)	0.080 (0.002)
SURE.Full-FM	**0.388 (0.02)**	**0.169 (0.003)**	**0.094 (0.002)**	**0.057 (0.001)**

Table 2: Average loss for the two estimators, MLE and SURE.Full-Cov, in estimating the population covariance matrices for varying n (with the standard errors in parentheses).

n	10	20	50	100
MLE	123.69 (5.71)	66.80 (2.69)	25.54 (0.91)	12.91 (0.41)
SURE.Full-Cov	**111.13 (5.01)**	**63.27 (2.53)**	**24.99 (0.88)**	**12.80 (0.40)**

As is evident from Table 1, the SURE.Full-FM outperforms the competing estimators under varying size of subsamples. Also note that, as the sample size increases, the improvement is less significant, which is consistent with the observations on the synthetic data experiments in Section 4.1. Recall that in Section 4.1.1, the SURE.FM and the SURE.Full-FM perform equally well when the assumption $\Sigma_i = A_i I$ is not violated severely. For real data, it is impossible to check this assumption and it is unlikely to be true. Hence, in this real data experiment, SURE.Full-FM outperforms SURE-FM by a large margin. The improvement of the proposed shrinkage estimator for the covariance matrices over the MLEs is evident from Table 2. Another important issue is that for real data the independence assumption is unrealistic. For this reason we have described a simple simulation study to see how the dependence affects the performance of our estimator; see Section 6.2 of the supplement.

4.2.2 Simultaneous Estimation of Resting State Functional MRI Connectivity Networks

In this section, we present an experiment on simultaneous estimation of connectivity networks from resting state functional MRI (rs-fMRI). Briefly, rs-fMRI is an MRI technique to measure human brain activity in the resting state, and a (functional) connectivity network computed from rs-fMRI measurements describes how different regions of the human brain

are correlated functionally. Van Den Heuvel & Pol (2010) give a nice review on connectivity network analysis for rs-fMRI and its applications. There is a large literature on the relationship between disruption in functional connectivity (dis-connectivity) and neurologic and psychiatric brain disorders, including Alzheimer's disease, depression, and attention deficit hyperactivity disorder (ADHD) (Van Den Heuvel & Pol 2010, pp. 529). Hence, functional connectivity networks have served as an important tool in such studies.

A functional connectivity matrix is constructed as follows. At each region of the brain, the blood oxygenation level dependent (BOLD) signal, which is a scalar measure of the neuronal activity in the region, can be detected by an MR scanner. Over the course of time, we obtain a time sequence of BOLD signals $b = [b_1, \ldots, b_n]$, where n is the number of time points, at each region of the brain. To describe the connectivity between two regions, we compute the correlation between the two time sequences of BOLD signals from the two regions. In other words, the connectivity between the two regions measures how in-sync or out-of-sync the two regions are in terms of the BOLD signals. For N regions, the connectivity matrix stores the pairwise connectivity (correlation) of each pair of regions. Hence, a connectivity network is essentially a correlation matrix. To apply the results discussed in Section 3, we consider the problem of simultaneous estimation of the connectivity networks from different rs-fMRI studies. We use the pre-processed networks from the USC Multimodal Connectivity Database (http://umcd.humanconnectomeproject.org/) (Brown et al. 2012). The datasets we used are ADHD200_CC200, PRURIM, and UCSF_MAC_PSP. There are in total 7 groups emanating from these three datasets (ADHD200_CC200: Typically Developing, ADHD-Combined, and ADHD-Inattentive; PRURIM: Healthy and Psoriasis; UCSF_MAC_PSP: Control and Progressive Supranuclear Palsy), so here $p = 7$. (Because 7 is not large, our asymptotic optimality results do not apply; however, it is reasonable to expect that shrinkage gives an improvement here in the same way that shrinkage gives an improvement in the normal means problem when $p > 2$.)

Since the networks are from different studies targeting different disorders, the sizes of the connectivity networks and the sizes of the studies are all different. The sizes of the connectivity networks in these three datasets are 190×190 (ADHD200_CC200), 116×116 (PRURIM), and 27×27 (UCSF_MAC_PSP). Thus, we extract an $N \times N$ (where N is fixed

Table 3: Average loss for the three estimators in estimating the population FM for varying sub-network size N (with the standard errors in parentheses).

N	3	5	7	10
FM.LE	0.409 (0.01)	0.738 (0.01)	1.387 (0.012)	1.969 (0.016)
SURE-FM	0.296 (0.005)	0.641 (0.007)	1.238 (0.009)	1.854 (0.015)
SURE.Full-FM	**0.246 (0.006)**	**0.534 (0.006)**	**0.977 (0.006)**	**1.581 (0.012)**

across all networks from different datasets) highly correlated sub-network from each network using a hierarchical clustering algorithm (Rokach & Maimon 2005) and the experiments are based on these sub-networks rather than on the original networks. As in the procedure in the previous section, we treat these (sub-)networks as the population and randomly draw subsamples (with replacement) from each group. Then we compute the average loss for the three estimators. The entire procedure is repeated $m = 1000$ times and the results are presented in Table 3. As was seen in Section 4.1.2, the improvement of our estimator over the competitors is more significant for large sub-network sizes. Note that in contrast to our experiment with synthetic data in Section 4.1.2, here we cannot keep increasing N: the maximum value is 27 because the connectivity network for the UCSF_MAC_PSP dataset is 27×27. This experiment demonstrates versatility of our shrinkage estimator to application domains beyond the analysis of dMRI datasets.

4.2.3 Tweedie-Adjusted Estimator as an Imaging Biomarker

Finally, we apply the shrinkage estimator proposed in Section 3.3 to identify the regions that are significantly distinct in diffusional properties (as captured via diffusion tensors) between patients with Parkinson's disease and control subjects. In this experiment, the dataset consists of DTI scans of 46 patients with Parkinson's disease and 24 control subjects. To identify the differences between the two groups, we use the DTI of the whole brain, which contains $p = 112 \times 112 \times 60$ voxels, without pre-selecting any region of interest. The diffusion tensors are fit at each voxel across the whole brain volume. The goal of this experiment is to see if we are able to automatically identify the regions capturing the large differences between the Parkinson's disease group and control groups and qualitatively validate our

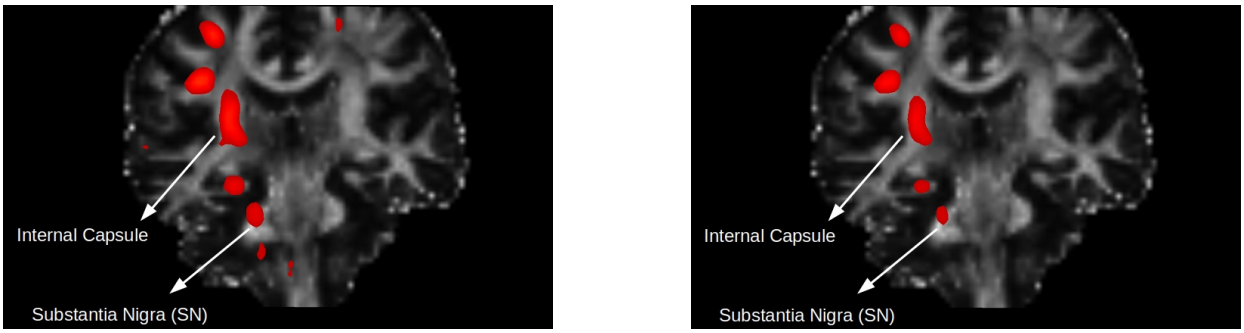
findings against what is expected by expert neurologists. In this context, Prodoehl et al. (2013) observed that the region most affected by Parkinson's disease is the Substantia Nigra, which is contained in the Basal Ganglia region of the human brain.

After computing both the Tweedie-adjusted estimates and the MOM estimates of the non-centrality parameters, we select the voxels with the largest 1% estimates of the non-centrality parameters and mark those voxels in bright red. (There are other ways to determine the threshold for the selection, for example by using the false discovery rate (FDR) in hypothesis testing problems. However, this is beyond the scope of this paper and we refer the reader to Schwartzman (2008) and Schwartzman et al. (2010) for interesting work on FDR analysis for DTI datasets.) These voxels are where the large differences between Parkinson's disease and control groups are observed. The results are shown in Figure 6. For better visualization, we threshold the estimates by the top 1%. To take into account the spatial structure, we apply a $4 \times 4 \times 4$ average mask to smooth the results. This smoothing may also be achieved by incorporating a spatial regularization term in the expression for SURE (8); however, the ensuing analysis becomes much more complicated and will be addressed in future work.

From the results, we can see that the shrinkage effect of our Tweedie-adjusted estimate successfully corrects selection bias and produces a more accurate identification of the affected regions. Our method is able to capture the Substantia Nigra, which is the region known to be affected by Parkinson's disease. *Notably, our method did not point to the apparently spurious and isolated regions selected by the MOM estimator (the tiny red spots in Figure 6(a)).* We also mention that past research using FA-based analysis did not report the Internal Capsule as a region affected by Parkinson's disease. We suspect that this discrepancy is due to the fact that FA discards the directional information of the diffusion tensors while we use the full diffusion tensor which contains the directional information. We plan to conduct a large-scale experiment in our future work to see if this observation continues to hold.

5 Discussion and Conclusions

In this work, we have presented shrinkage estimators for the mean and covariance of the Log-Normal distribution defined on the manifold P_N of $N \times N$ SPD matrices. We also



(a) MOM estimates

(b) Tweedie-adjusted estimates

Figure 6: Differences between scans of Parkinson's disease group and those of the control group are superimposed on a dMRI scan of a randomly-chosen Parkinson's disease patient and indicated in red.

showed that the proposed shrinkage estimators are asymptotically optimal in a large class of estimators including the MLE. The proposed shrinkage estimators are in closed form and resemble (in form) the James-Stein estimator in Euclidean space $\mathbb{R}^p$. We demonstrated that the proposed shrinkage estimators outperform the MLE via several synthetic data examples and real data experiments using diffusion tensor MRI and rs-fMRI datasets. The improvements of the proposed shrinkage estimators are significant especially in the small sample size scenarios, which is very pertinent to medical imaging applications. Further, we also empirically demonstrated that the improvement in the distribution parameter estimates is achieved with increasing size of the SPD matrices as well.

Our work reported here is however based on the Log-Euclidean metric, and one of the drawbacks of this metric is that it is not invariant under affine transformations, which may be a desirable property in some applications. Unfortunately, the derivation of the shrinkage estimators under the affine-invariant metric is challenging due to the fact that there is no closed-form expression for some elementary quantities such as the sample FM, which makes it almost impossible to derive the corresponding closed form for the SURE. Our future research efforts will focus on developing a general framework for designing shrinkage estimators that are applicable to general Riemannian manifolds.

For applications in localizing the regions of the brain where two groups differ, our approach already works well, but it can potentially be improved if we take into account the fact

that some features of neighboring voxels within a region are close. For instance, $M_i^{(k)}$ and $M_j^{(k)}$ should be close if voxels i and j are close. Currently, our approach is to apply a spatial smoother to the Tweedie-adjusted estimates. Instead, the improvement can be achieved by imposing regularization constraints, e.g. a spatial process prior, in the proposed framework. However, the ensuing analysis becomes rather complicated and will be the focus of our future efforts.

Acknowledgments We thank the reviewers for constructive criticism that greatly improved the paper and to Dr. David Vaillancourt for providing us with the real dMRI data used in the experiments.

References

Afsari, B. (2011), 'Riemannian L^p center of mass: Existence, uniqueness, and convexity', *Proceedings of the American Mathematical Society* **139**(02), 655–655.
URL: *http://www.ams.org/jourcgi/jour-getitem?pii=S0002-9939-2010-10541-5*

Alexander, D. C. (2005), 'Multiple-fiber reconstruction algorithms for diffusion MRI', *White Matter in Cognitive Neuroscience: Advances in Diffusion Tensor Imaging and Its Applications* **1064**, 113–133.

Archer, D. B., Vaillancourt, D. E. & Coombes, S. A. (2017), 'A template and probabilistic atlas of the human sensorimotor tracts using diffusion MRI', *Cerebral Cortex* **28**(5), 1685–1699.

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

Basser, P. J., Mattiello, J. & LeBihan, D. (1994), 'MR diffusion tensor spectroscopy and imaging', *Biophysical Journal* **66**(1), 259–267.

Berger, J. (1980), 'Improving on inadmissible estimators in continuous exponential families with applications to simultaneous estimation of gamma scale parameters', *The Annals of Statistics* pp. 545–571.

Brown, J. A., Rudie, J. D., Bandrowski, A., Van Horn, J. D. & Bookheimer, S. Y. (2012), 'The UCLA multimodal connectivity database: a web-based platform for brain connectivity matrix sharing and analysis', *Frontiers in neuroinformatics* **6**, 28.

Byrd, R. H., Lu, P., Nocedal, J. & Zhu, C. (1995), 'A limited memory algorithm for bound constrained optimization', *SIAM Journal on Scientific Computing* **16**(5), 1190–1208.

Callaghan, P. T. (1993), *Principles of Nuclear Magnetic Resonance Microscopy*, Oxford University Press.

Chakraborty, R. & Vemuri, B. C. (2015), Recursive Fréchet mean computation on the Grassmannian and its applications to computer vision, *in* 'Proceedings of the IEEE International Conference on Computer Vision', pp. 4229–4237.

Chakraborty, R. & Vemuri, B. C. (2019), 'Statistics on the Stiefel manifold: Theory and applications', *The Annals of Statistics* **47**(1), 415–438.

Chefd'Hotel, C., Tschumperlé, D., Deriche, R. & Faugeras, O. (2004), 'Regularizing flows for constrained matrix-valued images', *Journal of Mathematical Imaging and Vision* **20**(1-2), 147–162.

Cherian, A. & Sra, S. (2016), Positive definite matrices: data representation and applications to computer vision, *in* 'Algorithmic Advances in Riemannian Geometry and Applications', Springer, pp. 93–114.

Clevenson, M. L. & Zidek, J. V. (1975), 'Simultaneous estimation of the means of independent Poisson laws', *Journal of the American Statistical Association* **70**(351a), 698–705.

Daniels, M. J. & Kass, R. E. (2001), 'Shrinkage estimators for covariance matrices', *Biometrics* **57**(4), 1173–1184.

Dawid, A. P. (1994), 'Selection paradoxes of Bayesian inference', *Multivariate Analysis and Its Applications (Hong Kong, 1992). IMS Lecture Notes Monograph Series.* **24**, 211–220.

Donoho, D. L., Gavish, M. & Johnstone, I. M. (2018), 'Optimal shrinkage of eigenvalues in the spiked covariance model', *The Annals of Statistics* **46**(4), 1742–1778.

Dryden, I. L., Koloydenko, A. & Zhou, D. (2009), 'Non-Euclidean statistics for covariance matrices, with applications to diffusion tensor imaging', *The Annals of Applied Statistics* **3**(3), 1102–1123.

Du, L. & Hu, I. (2020), 'An empirical Bayes method for chi-squared data', *Journal of the American Statistical Association* **0**(0), 1–14.
URL: *https://doi.org/10.1080/01621459.2020.1777137*

Efron, B. (2011), 'Tweedie's formula and selection bias', *Journal of the American Statistical Association* **106**(496), 1602–1614.

Efron, B. & Morris, C. (1973*a*), 'Combining possibly related estimation problems', *Journal of the Royal Statistical Society: Series B* **35**(3), 379–402.

Efron, B. & Morris, C. (1973*b*), 'Stein's estimation rule and its competitors: An empirical Bayes approach', *Journal of the American Statistical Association* **68**(341), 117–130.

Efron, B. & Tibshirani, R. (1996), 'Using specially designed exponential families for density estimation', *The Annals of Statistics* **24**(6), 2431–2461.

Feragen, A. & Fuster, A. (2017), Geometries and interpolations for symmetric positive definite matrices, *in* 'Modeling, Analysis, and Visualization of Anisotropy', Springer, pp. 85–113.

Fletcher, P. T. & Joshi, S. (2004), Principal geodesic analysis on symmetric spaces: Statistics of diffusion tensors, *in* 'Computer Vision and Mathematical Methods in Medical and Biomedical Image Analysis', Springer, pp. 87–98.

Fletcher, P. T., Joshi, S., Lu, C. & Pizer, S. M. (2003), Gaussian distributions on Lie groups and their application to statistical shape analysis, *in* 'Biennial International Conference on Information Processing in Medical Imaging', Springer, pp. 450–462.

Frackowiak, R. S. J., Friston, K. J., Frith, C. D., Dolan, R. J., Price, C. J., Zeki, S., Ashburner, J. T. & Penn, W. D. (2004), *Human Brain Function*, Elsevier.

Fréchet, M. (1948), 'Les éléments aléatoires de nature quelconque dans un espace distancié', *Annales de l'Institut Henri Poincaré* **10**(4), 215–310.

Gabay, D. (1982), 'Minimizing a differentiable function over a differential manifold', *Journal of Optimization Theory and Applications* **37**(2), 177–219.

Groisser, D. (2004), 'Newton's method, zeroes of vector fields, and the Riemannian center of mass', *Advances in Applied Mathematics* **33**(1), 95–135.

Ho, J., Cheng, G., Salehian, H. & Vemuri, B. (2013), Recursive Karcher expectation estimators and geometric law of large numbers, *in* 'Artificial Intelligence and Statistics', PMLR, pp. 325–332.

James, W. & Stein, C. (1961), Estimation with quadratic loss, *in* 'Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability', Vol. 1, University of California Press, pp. 361–379.

Jing, B.-Y., Li, Z., Pan, G. & Zhou, W. (2016), 'On SURE-type double shrinkage estimation', *Journal of the American Statistical Association* **111**(516), 1696–1704.

Kong, X., Liu, Z., Zhao, P. & Zhou, W. (2017), 'SURE estimates under dependence and heteroscedasticity', *Journal of Multivariate Analysis* **161**, 1–11.

Kubokawa, T., Robert, C. P. & Saleh, A. K. M. E. (1993), 'Estimation of noncentrality parameters', *The Canadian Journal of Statistics/La Revue Canadienne de Statistique* **21**(1), 45–57.

Ledoit, O. & Wolf, M. (2003), 'Improved estimation of the covariance matrix of stock returns with an application to portfolio selection', *Journal of Empirical Finance* **10**(5), 603–621.

Lenglet, C., Rousson, M. & Deriche, R. (2006), 'DTI segmentation by statistical surface evolution', *IEEE Transactions on Medical Imaging* **25**(6), 685–700.

Lenglet, C., Rousson, M., Deriche, R. & Faugeras, O. (2006), 'Statistics on the manifold of multivariate normal distributions: Theory and application to diffusion tensor MRI processing', *Journal of Mathematical Imaging and Vision* **25**(3), 423–444.

Lim, Y. & Pálfia, M. (2014), 'Weighted inductive means', *Linear Algebra and its Applications* **453**, 59–83.

Moakher, M. (2005), 'A differential geometric approach to the geometric mean of symmetric positive-definite matrices', *SIAM Journal on Matrix Analysis and Applications* **26**(3), 735–747.

Pennec, X. (2006), 'Intrinsic statistics on Riemannian manifolds: Basic tools for geometric measurements', *Journal of Mathematical Imaging and Vision* **25**(1), 127.

Prodoehl, J., Li, H., Planetta, P. J., Goetz, C. G., Shannon, K. M., Tangonan, R., Comella, C. L., Simuni, T., Zhou, X. J., Leurgans, S., Corcos, D. M. & Vaillancourt, D. E. (2013),

'Diffusion tensor imaging of Parkinson's disease, atypical Parkinsonism, and essential tremor', *Movement Disorders* **28**(13), 1816–1822.

Robbins, H. (1956), An empirical Bayes approach to statistics, *in* 'Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics', University of California Press, Berkeley, Calif., pp. 157–163. **URL:** *https://projecteuclid.org/euclid.bsmsp/1200501653*

Rokach, L. & Maimon, O. (2005), Clustering methods, *in* 'Data mining and knowledge discovery handbook', Springer, pp. 321–352.

Salehian, H., Chakraborty, R., Ofori, E., Vaillancourt, D. & Vemuri, B. C. (2015), 'An efficient recursive estimator of the Fréchet mean on a hypersphere with applications to medical image analysis', *Mathematical Foundations of Computational Anatomy* **3**, 143–154.

Sasaki, H., Noh, Y.-K., Niu, G. & Sugiyama, M. (2016), 'Direct density derivative estimation', *Neural Computation* **28**(6), 1101–1140.

Schwartzman, A. (2008), 'Empirical null and false discovery rate inference for exponential families', *The Annals of Applied Statistics* **2**(4), 1332–1359.

Schwartzman, A. (2016), 'Lognormal distributions and geometric averages of symmetric positive definite matrices', *International Statistical Review* **84**(3), 456–486.

Schwartzman, A., Dougherty, R. F. & Taylor, J. E. (2010), 'Group comparison of eigenvalues and eigenvectors of diffusion tensors', *Journal of the American Statistical Association* **105**(490), 588–599.

Shen, W. & Ghosal, S. (2017), 'Posterior contraction rates of density derivative estimation', *Sankhya A* **79**(2), 336–354.

Stein, C. (1956), Inadmissibility of the usual estimator for the mean of a multivariate normal distribution, *in* 'Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, 1954–1955, vol. I', University of California Press, Berkeley and Los Angeles, pp. 197–206.

Stein, C. (1975), Estimation of a covariance matrix, *in* 'Reitz Lecture, 39th Annual Meeting IMS. Atlanta, Georgia'.

Stein, C. (1981), 'Estimation of the mean of a multivariate normal distribution', *The Annals of Statistics* **9**(6), 1135–1151.

Sturm, K.-T. (2003), 'Probability measures on metric spaces of nonpositive curvature', *Heat Kernels and Analysis on Manifolds, Graphs, and Metric Spaces: Lecture Notes from a Quarter Program on Heat Kernels, Random Walks, and Analysis on Manifolds and Graphs: April 16-July 13, 2002, Emile Borel Centre of the Henri Poincaré Institute, Paris, France* **338**, 357–390.

Terras, A. (2016), *Harmonic analysis on symmetric spaces—higher rank spaces, positive definite matrix space and generalizations*, Springer.

Tuzel, O., Porikli, F. & Meer, P. (2006), Region covariance: A fast descriptor for detection and classification, *in* 'European Conference on Computer Vision', Springer, pp. 589–600.

Udriste, C. (2013), *Convex functions and optimization methods on Riemannian manifolds*, Vol. 297, Springer Science & Business Media.

Vaillancourt, D., Spraker, M., Prodoehl, J., Abraham, I., Corcos, D., Zhou, X., Comella, C. & Little, D. (2009), 'High-resolution diffusion tensor imaging in the substantia nigra of de novo Parkinson disease', *Neurology* **72**(16), 1378–1384.

Van Den Heuvel, M. P. & Pol, H. E. H. (2010), 'Exploring the brain network: a review on resting-state fMRI functional connectivity', *European Neuropsychopharmacology* **20**(8), 519–534.

Wang, Z. & Vemuri, B. C. (2004), Tensor field segmentation using region based active contour model, *in* 'European Conference on Computer Vision', Springer, pp. 304–315.

Xie, X., Kou, S. C. & Brown, L. D. (2012), 'SURE estimates for a heteroscedastic hierarchical model', *Journal of the American Statistical Association* **107**(500), 1465–1479.

Yang, C.-H. & Vemuri, B. C. (2019), Shrinkage estimation on the manifold of symmetric positive-definite matrices with applications to neuroimaging, *in* 'International Conference on Information Processing in Medical Imaging', Springer, pp. 566–578.

Zhou, D., Dryden, I. L., Koloydenko, A. & Li, B. (2008), A Bayesian method with reparameterization for diffusion tensor imaging, *in* 'Medical Imaging 2008: Image Processing', Vol. 6914, International Society for Optics and Photonics, p. 69142J.