Consistency Of Restricted Maximum Likelihood Estimators Of Principal Components

Debashis Paul and Jie Peng

Technical Report #2008-2
March 26, 2008

This material was based upon work supported by the National Science Foundation under Agreement No. DMS-0112069. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science

Statistical and Applied Mathematical Sciences Institute
PO Box 14006
Research Triangle Park, NC 27709-4006
www.samsi.info

Submitted to the Annals of Statistics

CONSISTENCY OF RESTRICTED MAXIMUM LIKELIHOOD ESTIMATORS OF PRINCIPAL COMPONENTS

By Debashis Paul* and Jie Peng

University of California, Davis

In this paper we consider two closely related problems : estimation of eigenvalues and eigenfunctions of the covariance kernel of functional data based on (possibly) irregular measurements, and the problem of estimating the eigenvalues and eigenvectors of the covariance matrix for high-dimensional Gaussian vectors. In [23], a restricted maximum likelihood (REML) approach has been developed to deal with the first problem. In this paper, we establish consistency and derive rate of convergence of the REML estimator for the functional data case, under appropriate smoothness conditions. Moreover, we prove that when the number of measurements per sample curve is bounded, under squared-error loss, the rate of convergence of the REML estimators of eigenfunctions is near-optimal. In the case of Gaussian vectors, asymptotic consistency and an efficient score representation of the estimators are obtained under the assumption that the effective dimension grows at a rate slower than the sample size. These results are derived through an explicit utilization of the intrinsic geometry of the parameter space, which is non-Euclidean. Moreover, the results derived in this paper suggest an asymptotic equivalence between the inference on functional data with dense measurements and that of the high dimensional Gaussian vectors.

1. Introduction. Analysis of functional data, where the measurements per subject, or replicate, are taken on a finite interval, has been one of the growing branches of statistics in recent times. In fields such as longitudinal data analysis, chemometrics, econometrics, the functional data analysis viewpoint has been successfully used to summarize data and gain better understanding of the problems at hand. The monographs of [24] and [13] give detailed accounts of the applications of functional data approach to various problems in these fields. Depending on how the individual curves

*Paul's research was partially supported by the National Science Foundation under Agreement No. DMS-0112069. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

AMS 2000 subject classifications: Primary 62G20, 60K35; secondary 62H25

Keywords and phrases: Functional data analysis, principal component analysis, high-dimensional data, Stiefel manifold, intrinsic geometry, consistency

are measured, one can think of two different scenarios - (i) when the curves are measured on a dense grid; and (ii) when the measurements are observed on an irregular, and typically sparse set of points on an interval. The first situation usually arises when the data are recorded by some automated instrument, e.g. in chemometrics, where the curves represent the spectra of certain chemical substances. The second scenario is more typical in longitudinal studies where the individual curves could represent the level of concentration of some substance, and the measurements on the subjects may be taken only at irregular time points. In the first scenario, i.e., data on a regular grid, as long as the individual curves are smooth, the measurement noise level is low, and the grid is dense enough, one can essentially treat the data to be on a continuum, and employ techniques similar to the ones used in classical multivariate analysis. For example, [14] derive stochastic expansions of sample PCA when the sample curves are noise-free and measured on a continuum. However, in the second scenario, the irregular nature of the data, and the presence of measurement noise pose challenges and require a different treatment. Under such a scenario, data corresponding to individual subjects can be viewed as partially observed, and noise-corrupted, independent realizations of an underlying stochastic process. The estimation of the eigenvalues and eigenfunctions of a smooth covariance kernel, from sparse, irregular measurements, has been studied by various authors including [17], [26] and [23], among others.

In [23], a *restricted maximum likelihood (REML)* approach is taken to obtain the estimators. REML estimators are widely used and studied in statistics. For example, the usefulness of REML and profile REML estimation has been recently demonstrated in the context of functional linear mixed effects model by [1]. In [23], it is assumed that the covariance kernel can be well-approximated by a positive-semidefinite kernel of finite rank r whose eigenfunctions can be represented by $M(\geq r)$ known orthonormal basis functions. Thus the basis coefficient matrix B of the approximant belongs to the *Stiefel manifold* of $M \times r$ matrices with orthonormal columns. The working assumption of Gaussianity allows the authors to derive the log-likelihood of the observed data given the measurement times. Then a Newton-Raphson procedure, that respects the geometry of the parameter space, is employed to obtain the estimates by maximizing the log-likelihood. This procedure is based on the formulation of a general Newton-Raphson scheme on Stiefel manifold developed in [11]. [23] also derive a computationally efficient approximate cross-validation score for selecting M and r. Through extensive simulation studies, it is demonstrated that the REML estimator is much more efficient than an alternative procedure ([26]) based on local linear smoothing of empirical covariances. The latter estimator does not naturally reside in the parameter space, even though it has been proved to achieve the optimal non-parametric convergence rate in the minimax sense under l^2 loss, under the optimal choice of the bandwidth and when the number of measurements per curve is bounded ([15]). Also, in most situations, our method outperforms the EM approach of [17]. Although the latter estimator also aims to maximize the log-likelihood, it does not naturally reside in the parameter space either, and thus it does not utilize its geometry efficiently.

The superior numerical performance of the REML estimator motivates us to conduct a detailed study of its asymptotic properties. In this paper, we establish consistency and derive the rate of convergence (under l^2 loss) of the REML estimator when the eigenfunctions have a certain degree of smoothness, and when a stable and smooth basis, e.g., the cubic B-spline basis with a pre-determined set of knots, is used for approximating them. The techniques used to prove consistency differ from the standard asymptotic analysis tools when the parameter space is Euclidean. Specifically, we restrict our attention to small ellipsoids around zero in the tangent space to establish a mathematically manageable neighborhood around an "optimal parameter" (a good approximation of the "true parameter" within the *model space*). We derive asymptotic results when the number of measurements per curve grows sufficiently *slowly* with the sample size (referred as the *sparse case*). We also show that for a special scenario of the *sparse case*, when there is a bounded number of measurements per curve, the risk of the REML estimator (measured in squared-error loss) of the eigenfunctions has asymptotically near-optimal rate (i.e., within a factor of $\log n$ of the optimal rate) under an appropriate choice of the number of basis functions.

Besides the *sparse case*, we consider two other closely related problems: (i) the estimation of the eigenvalues and eigenfunctions of a smooth covariance kernel, from dense, possibly irregular, measurements (referred as the *dense case*); and (ii) the estimation of the eigenvalues and eigenvectors of a high-dimensional covariance matrix (referred as the *matrix case*). In the *matrix case*, we assume that there is preliminary information so that the data can be efficiently approximated in a lower dimensional *known* linear space whose effective dimension grows at a rate slower than the sample size n. The proofs of the results in all three cases utilize the intrinsic geometry of the parameter space through a decomposition of the Kullback-Leibler divergence. However, the *matrix case* and the *dense case* are more closely related, and the techniques for proving the results in these cases are different in certain aspects from the treatment of the *sparse case*, as described in Sections 2 and 3.

Moreover, in the *matrix case*, we also derive a *semiparametric efficient score representation* of the REML estimator (Theorem 3.2), that is given in terms of the *intrinsic Fisher information operator* (note that the residual term is not necessarily $o_P(n^{-1/2})$). This result is new, and explicitly quantifies the role of the intrinsic geometry of the parameter space on the asymptotic behavior of the estimators. Subsequently, it points to an *asymptotic optimality* of the REML estimators. Here, asymptotic optimality means achieving the asymptotic minimax risk under l^2 loss within a suitable class of covariance matrices (kernels). We want to point out that, in the *matrix case*, the REML estimators coincide with the usual PCA estimates, i.e., the eigenvalues and eigenvectors of the sample covariance matrix ([19]). In [22], a first order approximation of the PCA estimators is obtained by matrix perturbation analysis. Our current results show that the efficient score representation coincides with this approximation, and thereby gives a geometric interpretation to this. The asymptotically optimal rate of the l^2-risk of the REML estimator in the *matrix case* follows from this representation and the lower bound on the minimax rate obtained in [22]. Asymptotic properties of high-dimensional PCA under a similar context have also been studied by [12]. Recently several approaches have been proposed for estimating large dimensional covariance matrices and their eigenvalues and eigenvectors under suitable sparsity assumptions on the population covariance, e.g. [2], [3] and [9].

At this point, we would like to highlight the main contributions of this paper. First, we have established the consistency and derived the rate of convergence of REML estimators for functional principal components in two different regimes - the *sparse case* and the *dense case*. In [15], it is shown that an estimator of functional principal component based on a local polynomial approach achieves the optimal nonparametric rate when the number of measurements per curve is bounded. However, to the best of our knowledge, no results exist regarding the consistency, or rate of convergence, of the REML estimators in the functional data context. Secondly, we have derived an efficient score representation for sample principal components of high-dimensional, i.i.d. Gaussian vectors. This involves calculation of the intrinsic Fisher information operator and its inverse, and along the line we also provide an independent verification that the REML estimates under a rank-restricted covariance model are indeed the PCA estimates. Thirdly, we expect that the current framework can be refined to establish efficient score representation of the REML estimators of the functional principal components, and therefore the results obtained in this paper serve as first steps towards studying the asymptotic optimality of these estimators. Moreover, results obtained in this paper suggest an *asymptotic equivalence* between the inference on functional data with dense measurements and that of the high dimensional Gaussian vectors. Finally, our work provides useful techniques for dealing with the analysis of estimation procedures based on minimization of a loss function (e.g. MLE, or more generally M-estimators) over a non-Euclidean parameter space for semiparametric problems. There has been some work on analysis of maximum likelihood estimators for parametric problems when the parameter space is non-Euclidean (see e.g. [20]). However, there has been very limited work for non/semi-parametric problems with non-Euclidean parameter space. Recently, [5] establish semiparametric efficiency of estimators in ICA (Independent Component Analysis) problems using a sieve maximum likelihood approach.

The rest of the paper is organized as follows. In Section 2, we present the data model for the functional principal components, and state the consistency results of the REML estimators. In Section 3, we describe the model for high-dimensional Gaussian vectors and derive asymptotic consistency and an efficient score representation of the corresponding REML estimators. Section 4 is devoted to giving an overview of the proof of the consistency result for the functional data case (Theorems 2.1 and 2.2). Section 5 gives an outline of the proof of consistency in the *matrix case* (Theorem 3.1), in particular emphasizing the major differences with the proof of Theorem 2.1. Section 6 is concerned with the proof of the score representation in the *matrix case* (Theorem 3.2). Section 7 has a summary of the results and a discussion on some future works. Technical details are given in the appendices.

2. Functional data.

In this section, we start with a description of the functional principal components analysis, and then make a distinction between the *sparse case* and the *dense case*. We then present the asymptotic results and relevant conditions for consistency under these two settings.

2.1. *Model.*

Suppose that we observe data $Y_i = (Y_{ij})_{j=1}^{m_i}$, at the design points $T_i = (T_{ij})_{j=1}^{m_i}$, $i = 1, \ldots, n$, with

$$Y_{ij} = X_i(T_{ij}) + \sigma\varepsilon_{ij}, \tag{2.1}$$

where $\{\varepsilon_{ij}\}$ are i.i.d. $N(0,1)$, $X_i(\cdot)$ are i.i.d. Gaussian processes on the interval $[0,1]$ (or, more generally, $[a,b]$ for some $a < b$) with mean 0 and covariance kernel $\overline{\Sigma}_0(u,v) = \mathbb{E}[X_i(u)X_i(v)]$. $\overline{\Sigma}_0$ has the spectral decomposition

$$\overline{\Sigma}_0(u,v) = \sum_{k=1}^{\infty} \overline{\lambda}_k \overline{\psi}_k(u)\overline{\psi}_k(v)$$

where $\{\overline{\psi}_k\}_{k=1}^{\infty}$ are orthonormal eigenfunctions and $\overline{\lambda}_1 > \cdots > \overline{\lambda}_r > \overline{\lambda}_{r+1} \geq \cdots \geq 0$ are the eigenvalues. The assumption that the stochastic process has mean zero is simply to focus only on the asymptotics of the estimates of eigenvalues and eigenfunctions of the covariance kernel (i.e., the functional principal components).

Throughout this paper we assume Gaussianity of the observations. We want to emphasize that, Gaussianity is more of a working assumption in deriving the REML estimators. But it plays a less significant role in asymptotic analysis. For the functional data case, only place where Gaussianity is used is in the proof of Proposition 4.2, and even this can be relaxed by assuming appropriate tail behavior of the observations. Gaussianity is more crucial in the analysis for the matrix case. The proofs of Proposition 5.2 and Theorem 3.2 depend on an exponential inequality on the extreme eigenvalues of a Wishart matrix (based on a result of [10]), even though we expect the non-asymptotic bound to hold more generally.

In this paper, we are primarily interested in the situation where the design points are i.i.d. from a distribution with density g (random design). We shall consider two scenarios, to be referred as the *sparse case* and the *dense case*, respectively. The *sparse case* refers to the situation when the number of measurements, m_i, are comparatively small (see **B1**). The *dense case* refers to the situation where the m_i's are large so that the design matrix (i.e., the matrix of basis functions evaluated at the time points) has a concentration property (see **B1'** and **D**). In the latter case, we also allow for the possibility that the design is non-random.

Next, we describe the *model space*, to be denoted by $\mathcal{M}_{M,r} := \mathcal{M}_{M,r}(\phi)$, (for $1 \leq r \leq M$) for the REML estimation procedure. The model space $\mathcal{M}_{M,r}$ consists of the class of covariance kernels $C(\cdot,\cdot)$, which have rank r, and whose eigenfunctions are represented in a known orthonormal basis $\{\phi_k\}_{k=1}^{M}$ of smooth functions. Furthermore, the nonzero eigenvalues are all distinct. For example, in [23], $\{\phi_k\}_{k=1}^{M}$ is taken to be an orthonormalized cubic B-spline basis with equally spaced knots. Thus, the model space consists of the elements $C(\cdot,\cdot) = \sum_{k=1}^{r} \lambda_k \psi_k(\cdot)\psi_k(\cdot)$, where $\lambda_1 > \cdots > \lambda_r > 0$, and $(\psi_1(\cdot), \ldots, \psi_r(\cdot)) = (\boldsymbol{\phi}(\cdot))^T B$, where B is an $M \times r$ matrix satisfying $B^T B = I_r$, and $\boldsymbol{\phi}(\cdot) = (\phi_1(\cdot), \ldots, \phi_M(\cdot))^T$. Note that we do not assume that $\overline{\Sigma}_0$ belongs to the *model space*. For the asymptotic analysis, we only assume that it can be well-approximated by a member of the *model space* (see condition **C** and Lemma 2.1). We define the *best approximation error* of the model as $\inf_{\widetilde{C} \in \mathcal{M}_{M,r}(\phi)} \parallel \overline{\Sigma}_0 - \widetilde{C} \parallel_F$, where $\parallel \cdot \parallel_F$ denotes the Hilbert-Schmidt norm. A rank r approximation to $\overline{\Sigma}_0$ in $\mathcal{M}_{M,r}(\phi)$ can be defined as

$$\Sigma_{*0}(u,v) = \sum_{k=1}^{r} \lambda_{*k}\psi_{*k}(u)\psi_{*k}(v),$$

with $\lambda_{*1} > \cdots > \lambda_{*r} > 0$, and

$$(\psi_{*1}(t), \cdots, \psi_{*r}(t)) = (\boldsymbol{\phi}(t))^T B_*,$$

where B_* is an $M \times r$ matrix satisfying $B_*^T B_* = I_r$. We refer to $\{(\psi_{*k}, \lambda_{*k})\}_{k=1}^r$, or equivalently, the pair (B_*, Λ_*), as an *optimal parameter*, if the corresponding Σ_{*0} is a close approximation to $\overline{\Sigma}_0$ in the sense that, the approximation error $\| \overline{\Sigma}_0 - \Sigma_{*0} \|_F$ has the same rate (as a function of M) as the best approximation error. Henceforth, (B_*, Λ_*) is used to denote an optimal parameter.

Observe that, under model (2.1), Y_i are independent, and conditionally on T_i they are distributed as $N_{m_i}(0, \overline{\Sigma}_i)$. Here, the $m_i \times m_i$ matrix $\overline{\Sigma}_i$ is of the form $\overline{\Sigma}_i = ((\overline{\Sigma}_0(T_{ij}, T_{ij'})))_{j,j'=1}^{m_i} + \sigma^2 I_{m_i}$. Then the matrix $\Sigma_{*i} = \Phi_i^T B_* \Lambda_* B_*^T \Phi_i + \sigma^2 I_{m_i}$ is an approximation to $\overline{\Sigma}_i$, where $\Phi_i := [\boldsymbol{\phi}(T_{i1}) : \cdots : \boldsymbol{\phi}(T_{im_i})]$ is an $M \times m_i$ matrix. We shall use Λ to denote interchangeably the $r \times r$ diagonal matrix $\text{diag}(\lambda_1, \ldots, \lambda_r)$ and the $r \times 1$ vector $(\lambda_1, \ldots, \lambda_r)^T$. Note that, the parameter (B, Λ) belongs to the parameter space $\Omega := \mathcal{S}_{M,r} \otimes \mathbb{R}_+^r$, where $\mathcal{S}_{M,r} = \{A \in \mathbb{R}^{M \times r} : A^T A = I_r\}$ is the *Stiefel manifold* of $M \times r$ matrices with orthonormal columns. For fixed r and M, the *REML estimator* of $\{(\overline{\psi}_k, \overline{\lambda}_k)\}_{k=1}^r$ is defined as a minimizer over Ω of the negative log-likelihood (up to an additive constant and the scale factor n):

$$L_n(B, \Lambda) = \frac{1}{2n}\sum_{i=1}^{n} \text{tr}\,(\Sigma_i^{-1} Y_i Y_i^T) + \frac{1}{2n}\sum_{i=1}^{n} \log|\Sigma_i|, \tag{2.2}$$

where $\Sigma_i = \Phi_i^T B \Lambda B^T \Phi_i + \sigma^2 I_{m_i}$.

2.2. *Consistency.* We shall present results on consistency of the REML estimators of functional principal components in the two different regimes considered above, namely, the *sparse case* (i.e., when the number of measurements per curve is "small") and the *dense case* (i.e., when the number of measurements per curve is "large"). Throughout this paper, we assume that σ^2 is known, even though [23] provide estimate of σ^2 as well. This assumption is primarily to simplify the exposition. It can be verified that all the consistency results derived in this paper hold even when σ^2 is estimated. We make the following assumptions about the covariance kernel $\overline{\Sigma}_0$.

A1 The r largest eigenvalues of $\overline{\Sigma}_0$ satisfy, (i) $c_1 \geq \overline{\lambda}_1 > \cdots > \overline{\lambda}_r > \overline{\lambda}_{r+1}$ for some $c_1 < \infty$; (ii) $\max_{1 \leq j \leq r}(\overline{\lambda}_j - \overline{\lambda}_{j+1})^{-1} \leq c_2 < \infty$.

A2 The eigenfunctions $\{\overline{\psi}_k\}_{k=1}^r$ are four times continuously differentiable and satisfy

$$\max_{1\leq k\leq r} \| \overline{\psi}_k^{(4)} \|_\infty \leq C_0 \quad \text{for some} \quad 0 < C_0 < \infty.$$

SPARSE case. In this case, we only consider the situation when σ^2 is fixed (i.e., it does not vary with n). We shall first deal with the case when m_i's are bounded. Then we extend our results to the situation when m_i's increase slowly with sample size, and are of the same order of magnitude for all i (condition **B1**). We also assume a boundedness condition for the random design (condition **B2**).

B1 The number of measurements m_i satisfy $\underline{m} \leq m_i \leq \overline{m}$ with $4 \leq \underline{m}$ and $\overline{m}/\underline{m}$ is bounded by some constant $d_2 > 0$. Also, $\overline{m} = O(n^\kappa)$ for some $\kappa \geq 0$.

B2 For each i, $\{T_{ij} : j = 1, \ldots, m_i\}$ are i.i.d. from a distribution with density g, where g satisfies

$$c_{g,0} \leq g(x) \leq c_{g,1} \ \text{ for all } \ x \in [0,1], \ \text{ where } \ 0 < c_{g,0} \leq c_{g,1} < \infty. \tag{2.3}$$

Finally, we have a condition on the l^2 error for approximating the covariance kernel in the model space $\mathcal{M}_{M,r}$. Define the *maximal approximation error* for an optimal parameter (B_*, Λ_*) as:

$$\overline{\beta}_n := \max_{1\leq i\leq n} \frac{1}{m_i} \| \overline{\Sigma}_i - \Sigma_{*i} \|_F . \tag{2.4}$$

C $\overline{m}\overline{\beta}_n = O(\sqrt{\frac{M \log n}{n}})$.

If we use orthonormalized cubic B-spline basis for representing the eigenfunctions, then **C** follows from **A1**-**A2** and **B1**-**B2**, if the covariance kernel is indeed of rank r:

LEMMA 2.1. *If* **A1**-**A2** *and* **B1**-**B2** *hold,* $\overline{\Sigma}_0$ *is of rank* r*, and we use the orthonormalized cubic B-spline basis with equally spaced knots to represent the eigenfunctions, then* **C** *holds, if* $M^{-1}(n\overline{m}^2/\log n)^{1/9} = O(1)$.

Proof of Lemma 2.1 follows from the fact that for a cubic B-spline basis, for sufficiently large M, we can choose (B_*, Λ_*) such that (i) $\max_{1\leq k\leq r} \| \overline{\psi}_k - \psi_{*k} \|_\infty = O(M^{-4})$ (by **A1** and **A2**), and (ii) $\overline{\beta}_n = O(M^{-4})$ (see Part I of Appendix). This implies that $\| \overline{\Sigma}_0 - \Sigma_{*0} \|_F = O(M^{-4})$. The assumption that the covariance kernel is of finite rank can be relaxed somewhat by considering the true parameter as a sequence of covariance kernels $\overline{\Sigma}_{0,n}$ such that the $(r+1)$-th largest eigenvalue $\overline{\lambda}_{r+1,n}$ decays to zero sufficiently fast. Note that in Lemma 2.1, the use of B-spline basis is not essential. The result holds under the choice of any *stable basis* (i.e., the Gram matrix has a bounded condition number) with sufficient smoothness.

We now state the main result in the following theorem.

THEOREM 2.1. *(SPARSE case) Suppose that* **A1**-**A2**, **B1**-**B2** *and* **C** *hold, and $\overline{m}$ is bounded. Suppose further that M satisfies*

$$M^{-1}(n/\log n)^{1/9} = O(1) \quad \text{and} \quad M = o(\sqrt{n/\log n}) \tag{2.5}$$

as $n \to \infty$. Then, given $\eta > 0$, there exists $c_{0,\eta} > 0$ such that for $\alpha_n = c_{0,\eta}\sigma\sqrt{\frac{\overline{m}^2 M \log n}{n}}$, with probability at least $1 - O(n^{-\eta})$, there is a minimizer $(\widehat{B}, \widehat{\Lambda})$ of (2.2) satisfying

$$\begin{aligned} \| \widehat{B} - B_* \|_F &\leq \alpha_n, \\ \| \widehat{\Lambda} - \Lambda_* \|_F &\leq \alpha_n. \end{aligned}$$

Moreover, the corresponding estimate of the covariance kernel, viz., $\widehat{\Sigma}_0(u,v) = \sum_{k=1}^r \widehat{\lambda}_k \widehat{\psi}_k(u)\widehat{\psi}_k(v)$, satisfies, with probability at least $1 - O(n^{-\eta})$,

$$\| \widehat{\Sigma}_0 - \overline{\Sigma}_0 \|_F = O(\alpha_n).$$

COROLLARY 2.1. *Suppose that the conditions of Theorem 2.1 hold. Then the best rate of convergence holds if $M \asymp (n/\log n)^{1/9}$, and the corresponding rate is given by $\alpha_n \asymp (\log n/n)^{4/9}$. For estimating the eigenfunctions, this is within a factor of $\log n$ of the optimal rate. The optimal rate over a class $\mathcal{C}$ of covariance kernels of rank r satisfying conditions* **A1**-**A2**, *and the random design points satisfying conditions* **B1**-**B2** *(with $\overline{m}$ bounded), is $n^{-4/9}$.*

Notice that, the rate obtained here for the estimated eigenvalues is not optimal. We expect a parametric rate of convergence for the latter, which can be achieved by establishing an efficient score representation of the estimators along the line of Theorem 3.2. The following result generalizes Theorem 2.1 by allowing for m_i's to slowly increase with n, and its proof is encapsulated in the proof of Theorem 2.1.

COROLLARY 2.2. *Suppose that,* **A1**-**A2**, **B1**-**B2** *and* **C** *hold. Suppose further that,* $\overline{m}$ *and* M *satisfy*

$$
\begin{aligned}
&(i)\ \overline{m}^4 M \log n = o(n),\\
&(ii)\ \max\{\overline{m}^3 M^{5/2}(\log n)^2, \overline{m}^{7/2} M^2 (\log n)^{3/2}\} = o(n),\\
&(iii)\ M^{-1}(n\overline{m}^2/\log n)^{1/9} = O(1), \quad (iv)\ \overline{m}^2 M^2 \log n = o(n).
\end{aligned} \tag{2.6}
$$

Then the conclusion of Theorem 2.1 holds. Also, the best rate is obtained when $M \asymp (n\overline{m}^2/\log n)^{1/9}$*, and the corresponding* $\alpha_n \asymp \overline{m}^{10/9}(\log n/n)^{4/9}$.

Condition (i) is required to ensure that $\overline{m}^2\alpha_n^2 = o(1)$; condition (ii) is needed to ensure that the upper bound in (A.12) in Lemma A.2 is $o(1)$; condition (iii) ensures that **C** holds; and finally, condition (iv) is used in proving Lemmas A.2, A.3 and A.4 in the Appendix (Part II). A sufficient condition for (2.6) to hold is that $\overline{m} = O(n^{1/5})$ and $M \asymp (n\overline{m}^2/\log n)^{1/9}$. Notice that the best rate obtained in Corollary 2.2 is not optimal in general. It is near-optimal (up to a factor of $\log n$ of the optimal rate) only when $\overline{m}$ is bounded above (Theorem 2.1).

DENSE case. This case refers to the scenario where the number of time points per curve is large, such that $\min_{1\le i\le n} m_i \to \infty$ sufficiently fast (see condition **D** and the corresponding discussion). For simplicity, we assume further that the number of design points is the same for all the sample curves, which is not essential for the validity of the results. Denote this common value by m. In terms of the asymptotic analysis, there is an important distinction between the *sparse case* and *dense case*. For the purpose of further exposition and the proof of the result on consistency of REML estimator in the *dense case*, it is more convenient to work with the transformed data $\widetilde{Y}_i = \Phi_i Y_i$. Let $\Gamma_i = \frac{1}{m}\Phi_i\Sigma_i\Phi_i^T$ and $R_i = \frac{1}{m}\Phi_i\Phi_i^T$. Then $\Gamma_i = mR_iB\Lambda B^T R_i + \sigma^2 R_i$. Then, a way of estimating $\{(\overline{\lambda}_k, \overline{\psi}_k)\}_{k=1}^r$ is by minimizing the negative log-likelihood of the transformed data:

$$
\widetilde{L}_n(B,\Lambda) = \frac{1}{2n}\sum_{i=1}^n \mathrm{tr}\,(\Gamma_i^{-1}\frac{1}{m}\widetilde{Y}_i\widetilde{Y}_i^T) + \frac{1}{2n}\sum_{i=1}^n \log|\Gamma_i|. \tag{2.7}
$$

Notice that, if R_i's are non-singular for all i, then by direct computation, we have that the negative log-likelihoods for the raw data: (2.2) and that of the transformed data: (2.7) differ only by a constant independent of the parameters B and Λ. Hence, on the set $\{R_i$ are non-singular for all $i\}$, the estimators obtained by minimizing (2.2) and (2.7) are the same. Assumptions **B1** and **B2** are now replaced by:

B1' $m = O(n^\kappa)$ for some $\kappa > 0$.

D Given $\eta > 0$, there exist constants $c_{1,\eta}, c_{2,\eta} > 0$ such that the event $A_{1,\eta}$ defined by

$$A_{1,\eta} = \left\{ \max_{1\leq i\leq n} \parallel R_i - I_M \parallel \leq c_{1,\eta}\sqrt{\frac{\sigma^2}{m\log n}}, \right. \\ \left. \text{and} \quad \max_{1\leq i\leq n} \parallel B_*^T R_i B_* - I_r \parallel \leq c_{2,\eta}\frac{\sigma^2}{m\log n} \right\}. \tag{2.8}$$

has probability at least $1 - O(n^{-\eta})$. Note that $A_{1,\eta}$ is defined in terms of $\mathbf{T} := \{T_{ij} : j = 1, \dots, m; i = 1, \dots, n\}$ alone. We assume throughout that $\sigma^2 \leq m$ (note that σ^2/m can be viewed as the signal-to-noise ratio. Therefore, for n large enough, on $A_{1,\eta}$, R_i is invertible for all i. The condition **D** gives concentration of individual R_i's around the identity matrix and is discussed in more detail at the end of this section. Finally, we make an assumption about the maximal approximation error $\overline{\beta}_n$ defined through (2.4) which differs slightly from the condition **C** in the sparse case.

C' Given $\eta > 0$, there is a constant $c_\eta > 0$ such that $\overline{\beta}_n \leq c_\eta \frac{\sigma^2}{m}\sqrt{\frac{M\log n}{n}}$ with probability at least $1 - O(n^{-\eta})$.

A result similar to Lemma 2.1 can be proved to ensure condition **C'** when a stable basis is used.

THEOREM 2.2. *(DENSE case) Suppose that* **A1**-**A2**, **B1'**, **C'** *and* **D** *hold, and $m \geq \sigma^2 > 0$. Then, given $\eta > 0$, there exists $c_{0,\eta} > 0$ such that for $\alpha_n = c_{0,\eta}\sigma\sqrt{\frac{M\log n}{nm}}$, with probability at least $1-O(n^{-\eta})$, there is a minimizer $(\widehat{B}, \widehat{\Lambda})$ of (2.7) satisfying*

$$\begin{aligned} \parallel (I_M - B_*B_*^T)(\widehat{B} - B_*) \parallel_F &\leq \alpha_n, \\ \parallel B_*^T(\widehat{B} - B_*) \parallel_F &\leq \sqrt{\frac{m}{\sigma^2}}\alpha_n, \\ \parallel \widehat{\Lambda} - \Lambda_* \parallel_F &\leq \sqrt{\frac{m}{\sigma^2}}\alpha_n. \end{aligned}$$

Further, the corresponding estimated covariance kernel $\widehat{\Sigma}_0$ defined as $\widehat{\Sigma}_0(u,v) = \sum_{k=1}^r \widehat{\lambda}_k\widehat{\psi}_k(u)\widehat{\psi}_k(v)$ satisfies, with probability at least $1 - O(n^{-\eta})$,

$$\parallel \widehat{\Sigma}_0 - \overline{\Sigma}_0 \parallel_F = O\left(\sqrt{\frac{M\log n}{n}}\right).$$

The proof of Theorem 2.2 requires a slight refinement of the techniques used in proving Theorem 3.1 stated in Section 3.2, making heavy use of condition **D**. To save space, we omit the proof. Note that the best rate in Theorem 2.2 implicitly depends on conditions **C'** and **D** in a complicated way, which is not the optimal rate for estimating the principal components. The optimal rate for l^2 risk of the eigenfunctions in this context is conjectured to be of the order $\max\{(\sigma^2/nm)^{8/9}, 1/n\}$, with the second term within brackets appearing only when $r > 1$. This can be verified for the case $r = 1$ using a refinement of the arguments used in proving Corollary 2.1.

Discussion on condition **D**. We shall only consider the setting of an uniform design - either fixed, or random. The condition (2.8) clearly requires m to be sufficiently large, since it gives concentration of individual R_i's around the identity matrix. To fulfil **D**, we also need some conditions on the basis functions used. Specifically, we concentrate on the following classes of basis functions. We assume that the basis functions are at least 3 times continuously differentiable.

- **E1** (*Sinusoidal basis*) $\max_{1\leq k\leq M} \parallel \phi_k \parallel_\infty = O(1)$.
- **E2** (*Spline-type basis*) (i) For any $k \in \{1,\ldots,M\}$, at most for a bounded number of basis functions ϕ_l, $\text{supp}(\phi_k)\cap$ $\text{supp}(\phi_l)$ is nonempty; (ii) $\max_{1\leq k\leq M} \parallel \phi_k \parallel_\infty = O(\sqrt{M})$.

One of the key observations in the case of functional data is that, the eigenfunctions $\{\psi_{*k}\}_{k=1}^r$ of the kernel Σ_{0*} (belonging to the model space) have the same degree of smoothness as the basis $\{\phi_k\}_{k=1}^M$, and the functions $\{\psi_{*k}\}_{k=1}^r$ and their derivatives are bounded. Also, notice that, $B_*^T R_i B_* = ((\frac{1}{m}\sum_{j=1}^m \psi_{*k}(T_{ij})\psi_{*l}(T_{ij})))_{k,l=1}^r$. Based on these observations, we present some sufficient conditions for (2.8) to hold under the uniform design and bases of type **E1** or **E2**. We omit the proof, which uses Bernstein's inequality (in the random design case) and the *Trapezoidal rule* (in the fixed design case).

Proposition 2.1. *Suppose that the basis is of type* **E1** *or* **E2**. *In the case of random, uniform design, (2.8) is satisfied if* $(M\log n)^2/\sigma^2 = O(1)$, *and* $\sqrt{m}\log n/\sigma^2 = O(1)$. *In the case of fixed, uniform design, (2.8) holds (with probability 1) if* $\frac{M^2\log n}{m}(1 + \frac{M^{5/2}}{m})^2/\sigma^2 = O(1)$, *and* $\log n/\sigma^2 = O(1)$. *Moreover, in this setting, if the eigenfunctions* $\{\overline{\psi}_k\}_{k=1}^r$ *vanish at the boundaries, and if the basis functions are chosen so that they also vanish at the boundaries, it is sufficient that* $\frac{M^7\log n}{m^{7/2}}/\sigma^2 = O(1)$ *and* $\frac{\log n}{m}/\sigma^2 = O(1)$.

Note that, two obvious implications of Proposition 2.1 are that (i) m needs to be rather large; and (ii) σ^2 may need to grow with n, in order that **D** holds.

Remark 1. It is to be noted that even though the consistency results for the functional data problem are proved under a specific choice of the basis for representing the eigenfunctions, viz., the (orthonormalized) cubic B-spline basis with equally spaced knots, this is by no means essential. The main features of this basis are given in terms of the various properties described in the Appendix (Part I). The crucial aspects are: (a) the basis is stable; (b) the basis functions have a certain order of smoothness; and (c) the basis functions have fast decay away from an interval of length $O(M^{-1})$ where M is the number of basis functions used. Same consistency results can be proved as long as those properties are satisfied.

Remark 2. When $\overline{m}$, the number of measurements is bounded, we can relax condition **A2** to that the eigenfunctions are twice continuously differentiable and with bounded second derivative, and under this assumption we can prove a result analogous to Theorem 2.1 and Corollary 2.1, with the corresponding optimal rate of convergence being $n^{-2/5}$ instead of $n^{-4/9}$.

3. High-dimensional vector. In this section, we describe a scenario where the observations are i.i.d. Gaussian vectors, which can be approximately represented in a known lower dimensional space (see **C"**), where the effective dimensionality of the observations grows at a rate slower than the sample size. For convenience, we refer this setting as the *matrix case*. It can be seen that besides the proofs of the results derived in this section sharing a lot of common features with those in Section 2, these results also suggest an asymptotic equivalence between the *dense case* for functional data, and the *matrix case*. This means that understanding one problem helps in understanding the other problem. In particular, we conjecture that the results derived for the Gaussian vectors, such as the efficient score representation (Theorem 3.2), can be carried over to the functional data case with dense measurements.

3.1. *Model.* Suppose that we have i.i.d. observations $Y_1, \cdots, Y_n$ from $N_m(0, \overline{\Sigma})$. Assume the covariance matrix $\overline{\Sigma}$ has the following structure

$$\overline{\Sigma} = \overline{\Sigma}_0 + \sigma^2 I_m.$$

This may be regarded as a "signal-plus-noise" model, with σ^2 representing the variance of the isotropic noise component. We further assume that $\overline{\Sigma}_0$

has at least r positive eigenvalues, for some $r \geq 1$. The eigenvalues of $\overline{\Sigma}_0$ are given by $s\overline{\lambda}_1 > \cdots > s\overline{\lambda}_r > s\overline{\lambda}_{r+1} \geq \cdots \geq 0$, where $s > 0$ is a parameter representing the "signal strength" (so that s/σ^2 represents the signal-to-noise ratio). We assume that the observations can be well represented in a known M dimensional basis Φ with $M \leq m$ (condition **C"**). Then the model space $\mathcal{M}_{M,r}(\Phi)$ (with $r \leq M \leq m$) is defined as the set of all $m \times m$ matrices Σ of the form $\Sigma = s\Phi^T B\Lambda B^T\Phi + \sigma^2 I_m$, where Φ is an $M \times m$ matrix satisfying $\Phi\Phi^T = I_M$, $B \in \mathcal{S}_{M,r}$ and Λ is $r \times r$, diagonal with positive diagonal elements. Note that, in order to prove consistency of the REML estimator, we require that the intrinsic dimension M grows with n sufficiently slowly. In fact, it has been shown (e.g. in [21]) that, when $s/\sigma^2 = O(1)$, M must be $o(n)$ to achieve consistency.

Throughout we assume that σ^2 and s are known. Of course, we can estimate the eigenvalues of $\overline{\Sigma}$ without any knowledge of s. The unknown parameters of the model are B and Λ. The parameter space is therefore $\Omega = \mathcal{S}_{M,r} \otimes \mathbb{R}_+^r$. The estimate $(\widehat{B}, \widehat{\Lambda})$ of (B, Λ) is obtained by minimizing over Ω the negative log-likelihood (up to an additive constant and the multiplicative factor n),

$$L_n(B,\Lambda) = \frac{1}{2n}\mathrm{tr}\,(\Sigma^{-1}\sum_{i=1}^{n} Y_i Y_i^T) + \frac{1}{2}\log|\Sigma|. \tag{3.1}$$

We then set the estimator of the first r eigenvectors of $\overline{\Sigma}$ as $\widehat{\Psi} = \Phi^T\widehat{B}$.

Similar to the *dense case*, for asymptotic analysis, it is more convenient to work with the transformed data $\widetilde{Y}_i = \Phi Y_i$. Let $\Gamma = \Phi\Sigma\Phi^T = sB\Lambda B^T + \sigma^2 I_M$. Then one can obtain estimates of (B, Λ) by minimizing over Ω the negative log-likelihood of the transformed data:

$$\widetilde{L}_n(B,\Lambda) = \frac{1}{2n}\mathrm{tr}\,(\Gamma^{-1}\sum_{i=1}^{n} \widetilde{Y}_i \widetilde{Y}_i^T) + \frac{1}{2}\log|\Gamma|, \tag{3.2}$$

which results in the same estimate obtained by minimizing (3.1).

Remark 3. It is known that ([19]), in the setting described above, the REML estimators of (B, Λ) coincide with the first r principal components of the sample covariance matrix of $\widetilde{Y}_i = \Phi Y_i$, $i = 1, \ldots, n$. On the other hand, based on the calculations carried out in Part IV of the Appendix, it is easy to see that the PCA estimators $(\widehat{B}^{PC}, \widehat{\Lambda}^{PC})$ satisfy the likelihood equations $\nabla_B \widetilde{L}_n(\widehat{B}^{PC}, \widehat{\Lambda}^{PC}) = 0$ and $\nabla_\zeta \widetilde{L}_n(\widehat{B}^{PC}, \widehat{\Lambda}^{PC}) = 0$. Thus, our approach provides an independent verification of the known result that the PCA estimates are REML estimators under the rank-restricted covariance model studied here.

3.2. *Consistency.* We make the following assumptions about the covariance matrix.

- **A1'** The eigenvalues of $\overline{\Sigma}_0$ are given by $s\overline{\lambda}_1 \geq \cdots \geq s\overline{\lambda}_m \geq 0$ and satisfy, for some $r \geq 1$ (fixed), (i) $\overline{c}_1 \geq \overline{\lambda}_1 > \cdots > \overline{\lambda}_r > \overline{\lambda}_{r+1}$ for some $\overline{c}_1 < \infty$; (ii) $\max_{1\leq j\leq r}(\overline{\lambda}_j - \overline{\lambda}_{j+1})^{-1} \leq \overline{c}_2 < \infty$.
- **C"** Assume that there exists $(B_*, \Lambda_*) \in \Omega$ (referred as "optimal parameter") such that, the matrix $\Sigma_{*0} = s\Phi^T B_* \Lambda_* B_*^T \Phi$ is a close approximation to $\overline{\Sigma}_0$ in the sense that $\beta_n := \parallel \overline{\Sigma}_0 - \Sigma_{*0} \parallel_F = O(\sigma^2\sqrt{\frac{M\log n}{n}})$.

Note that **C"** implies that the observation vectors can be closely approximated in the basis Φ.

THEOREM 3.1. *(MATRIX case) Suppose that* **A1'** *and* **C"** *hold, and* $s \geq \sigma^2 > 0$*. Then given* $\eta > 0$*, there exists* $c_{0,\eta} > 0$ *such that for* $\alpha_n = c_{0,\eta}\sigma\sqrt{\frac{M\log n}{ns}}$*, with probability at least* $1 - O(n^{-\eta})$*, there is a minimizer* $(\widehat{B}, \widehat{\Lambda})$ *of (3.2) satisfying*

$$\begin{aligned} \parallel (I_M - B_* B_*^T)(\widehat{B} - B_*) \parallel_F &\leq \alpha_n, \\ \parallel B_*^T(\widehat{B} - B_*) \parallel_F &\leq \sqrt{\frac{s}{\sigma^2}}\alpha_n, \\ \parallel \widehat{\Lambda} - \Lambda_* \parallel_F &\leq \sqrt{\frac{s}{\sigma^2}}\alpha_n. \end{aligned}$$

Observe that the rates obtained in Theorem 2.2 and Theorem 3.1 are identical once we replace m in Theorem 2.2 by s. Thus the number of measurements m in the *dense case* is an analog of the signal strength s in the *matrix case*. This important observation suggests an *asymptotic equivalence* between these two problems. This is a result of the concentration of the matrices $\{R_i\}_{i=1}^n$ around I_M for the *dense case* (condition **D**). Under the matrix case, the analogs of R_i exactly equal the identity matrix. Moreover, Theorem 3.1 establishes the closeness of the REML estimator to the optimal parameter, which serves as an important step towards proving Theorem 3.2.

3.3. *Efficient score representation.* When the observations are i.i.d. Gaussian vectors, we can get a more refined result than the one stated in Theorem 3.1. In this section, we show that by using the intrinsic geometry, we can get an *efficient score representation* of the REML estimator (and hence PCA estimator). In [22], a first order approximation to the sample eigenvectors (i.e. PCA estimates) is obtained using matrix perturbation theory ([18]). Subsequently, it has also been shown there that the rate of convergence of

l^2-risk of PCA estimators is optimal. Here, we show that the efficient score representation of the REML estimator coincides with this first order approximation when the signal-to-noise ratio s/σ^2 is bounded (Corollary 3.1). Our approach is different from the perturbation analysis. It also quantifies the role of intrinsic geometry of the parameter space explicitly. Our result gives an alternative interpretation of this approximation, and consequently, the score representation points to an asymptotic optimality of the REML (and hence PCA) estimator.

We first introduce some notations. More details can be found in the Appendix (Part IV). Let $\zeta = \log \Lambda$ (treated interchangeably as an $r \times 1$ vector and an $r \times r$ diagonal matrix). The the parameter space for (B, ζ) is $\tilde{\Omega} := \mathcal{S}_{M,r} \otimes \mathbb{R}^r$. Let $\mathcal{T}_B := \{U \in \mathbb{R}^{M \times r} : B^T U = -U^T B\}$ denote the tangent space of the *Stiefel manifold* $\mathcal{S}_{M,r}$ at B. Then the tangent space for the product manifold $\tilde{\Omega}$ at (B, ζ) is $\mathcal{T}_B \oplus \mathbb{R}^r$ (see Part V of the Appendix for the definition of the product manifold and its tangent space).

For notational simplicity, we use θ_* to denote (B_*, ζ_*) and θ_0 to denote (B_0, ζ_0). Define $L(\theta_0; \theta_*) = \mathbb{E}_{\theta_*} \widetilde{L}_n(\theta_0)$. Let $\nabla \widetilde{L}_n(\cdot)$ and $\nabla L(\cdot; \theta_*)$ denote the *intrinsic gradient* of the functions $\widetilde{L}_n(\cdot)$ and $L(\cdot; \theta_*)$ with respect to (B, ζ), respectively. Also, let $H_n(\cdot)$ and $H(\cdot; \theta_*)$ denote the *intrinsic Hessian operator* of the functions $\widetilde{L}_n(\cdot)$ and $L(\cdot; \theta_*)$ with respect to (B, ζ), respectively. Let $H^{-1}(\cdot; \theta_*)$ denote the inverse Hessian operator of $L(\cdot; \theta_*)$. Also we use $H_B(\cdot; \theta_*)$ to denote the Hessian of $L(\cdot; \theta_*)$ w.r.t. B. Notations for Hessian w.r.t. ζ and gradients w.r.t B and ζ are defined similarly.

The following result gives the efficient score representation of the REML estimator in the situation when $\sigma^2 = 1$ and $s = 1$. The result can be extended via rescaling to the case for arbitrary σ^2 and s with $s \geq \sigma^2 > 0$, and s/σ^2 being bounded.

THEOREM 3.2. *(Score representation) Suppose that* **A1'** *and* **C"** *hold with* $\sigma^2 = 1$, $s = 1$, *and* $M = o(n^a)$ *for some* $a \in (0, 1)$. *Let* $\gamma_n = \max\{\sqrt{\frac{M \vee \log n}{n}}, \beta_n\}$. *Then there is a minimizer* $(\widehat{B}, \widehat{\Lambda})$ *of the negative log-likelihood (3.2) such that, with probability tending towards 1,*

$$\widehat{B} - B_* = -H_B^{-1}(\theta_*; \theta_*)(\nabla_B \widetilde{L}_n(\theta_*)) + O(\gamma_n^2) \tag{3.3}$$

$$\widehat{\Lambda} - \Lambda_* = -\Lambda_* H_\zeta^{-1}(\theta_*; \theta_*)(\nabla_\zeta \widetilde{L}_n(\theta_*)) + O(\gamma_n^2). \tag{3.4}$$

In particular, from this representation, we have, with probability tending *towards 1,*

$$\| \widehat{B} - B_* \|_F = O(\gamma_n); \tag{3.5}$$

$$\| \widehat{\Lambda} - \Lambda_* \|_F = O(\sqrt{\frac{\log n}{n}} + \gamma_n^2). \tag{3.6}$$

Note that Theorem 3.2 gives the optimal rate of convergence for l^2-risk of the estimated eigenvectors when $\beta_n = 0$ (i.e., no model bias). This result follows from the minimax lower bound on the risk obtained by [22]. Note that, this lower bound under the current setting follows essentially from the proof of Corollary 2.1. Also, when $a \leq 1/2$, this result shows that $\widehat{\Lambda}$ converges at a parametric rate. Indeed, the representation (3.4) implies asymptotic normality of $\widehat{\Lambda}$ when $a \leq 1/2$. In the derivation of Theorem 3.2 we need to compute the Hessian and its inverse, which leads to the following representation.

COROLLARY 3.1. *Under the assumptions of Theorem 3.2, we have the following representation:*

$$H_B^{-1}(\theta_*;\theta_*)(\nabla_B \widetilde{L}_n(\theta_*)) = [\mathbf{R}_1 \widetilde{S} B_{*1} : \cdots : \mathbf{R}_r \widetilde{S} B_{*r}],$$

where B_{*j} is the j-th column of B_*, and

$$\mathbf{R}_j = \sum_{1 \leq i \neq j \leq r} \frac{1}{(\lambda_{*i} - \lambda_{*j})} B_{*i} B_{*i}^T - \frac{1}{\lambda_{*j}} (I_M - B_* B_*^T),$$

is the resolvent operator corresponding to Γ_* "evaluated at" $(1 + \lambda_{*j})$.

Combining Corollary 3.1 with (3.3), we get a first order approximation to $\widehat{B}$ which coincides with the approximation for sample eigenvectors obtained in [22]. However, Theorem 3.2 has deeper implications. Since it gives an efficient score representation, it suggests an asymptotic optimality of the REML estimators in the minimax sense.

4. Proof of Theorem 2.1. Since σ^2 is fixed and assumed known, without loss of generality, we take $\sigma^2 = 1$. In this section, we give an outline of the main ideas/steps. The details of the proofs are given in the Appendix (Part II). The strategy of the proof is as follows. We restrict our attention to a subset $\Theta(\alpha_n)$ of the parameter space (referred as the *restricted parameter space*), which is the image under exponential map of the boundary of an ellipsoid centered at 0, in the tangent space of an "optimal parameter". We then show that with probability tending towards 1, for every parameter value in this restricted parameter space, the value of the negative log-likelihood is greater than the value of the negative log-likelihood at the optimal parameter. Due to the Euclidean geometry of the tangent space, this implies that with probability tending towards 1, there is a local maximum of the log-likelihood within the image (under exponential map) of the closed ellipsoid. The key steps of the proof are:

(i) Decompose the difference between the negative log-likelihood at the optimal parameter and an arbitrary parameter in the restricted space as a sum of three terms - a term representing the average Kullback-Leibler divergence between the distributions, a term representing random fluctuation in the log-likelihood, and a term representing the *model bias* (equation (4.5)).
(ii) For every fixed parameter in the restricted parameter space: (a) provide upper and lower bounds (dependent on α_n) for the average Kullback-Leibler divergence; (b) provide upper bounds for the random term and the model bias term. In both cases, the bounds are probabilistic with exponentially small tails.
(iii) Use a covering argument combined with a union bound to extend the above probabilistic bounds on difference between log-likelihoods corresponding to a single parameter in $\Theta(\alpha_n)$ to the infimum of the difference over the entire $\Theta(\alpha_n)$.

The strategy of this proof is standard. However, in order to carry it out we need to perform detailed computations involving the geometry of the parameter space such as the structure of the tangent space and the exponential map. Note that, in the current case the geometry of the parameter space is well-understood, so that there exist an explicit form of the exponential map and a precise description of the tangent space. This helps in obtaining the precise form of the local Euclidean approximations around an optimal parameter in the derivations.

4.1. *Parameter space and exponential map.* We use the following characterization of the tangent space $\mathcal{T}_B$ of the Stiefel manifold $\mathcal{S}_{M,r}$ at a point B. Any element $U \in \mathcal{T}_B$ can be expressed as $U = BA_U + C_U$, where $A_U = -A_U^T$ and $B^T C_U = O$. We then define the restricted parameter space centered at an optimal parameter (B_*, Λ_*) by

$$\begin{aligned}(4.1)\quad \Theta(\alpha_n) \;:=\; &\{(\mathbf{exp}(1, B_* A_U + C_U), \Lambda_* \exp(D)) : \\ &A_U = -A_U^T, B_*^T C_U = O, D \in \mathbb{R}^r, \\ &\text{such that} \quad \| A_U \|_F^2 + \| C_U \|_F^2 + \| D \|_F^2 = \alpha_n^2\}.\end{aligned}$$

In the definition of $\Theta(\alpha_n)$ and henceforth, we shall treat Λ and D interchangeably as an $r \times 1$ vector, and an $r \times r$ diagonal matrix. The function $\mathbf{exp}(t, U)$ is the *exponential map* on $\mathcal{S}_{M,r}$ at B_*, mapping a tangent vector in $\mathcal{T}_{B_*}$ to a point on the manifold. For $U \in \mathcal{T}_{B_*}$ and $t \geq 0$, it is defined as

$$\mathbf{exp}(t, U) = B_*\mathbf{M}(t, U) + Q\mathbf{N}(t, U),$$

where

$$\begin{bmatrix} \mathbf{M}(t, U) \\ \mathbf{N}(t, U) \end{bmatrix} = \exp\left(t \begin{bmatrix} B_*^T U & -R^T \\ R & O \end{bmatrix} \right) \begin{bmatrix} I_r \\ O \end{bmatrix},$$

where $\exp(\cdot)$ is the usual matrix exponential, and $QR = (I_M - B_* B_*^T)U$ is the QR-decomposition. The properties of the map $\mathbf{exp}(1, \cdot)$ that we shall heavily use in the subsequent analysis (see Parts II and III of the Appendix) are : for $U \in \mathcal{T}_{B_*}$,

$$\begin{aligned} (4.2) \qquad & B_*^T(\mathbf{exp}(1, U) - B_*) \\ = \quad & B_*^T U + O((\| B_*^T U \|_F + \| (I_M - B_* B_*^T)U \|_F) \| U \|_F), \\ (4.3) \qquad & (I_M - B_* B_*^T)\mathbf{exp}(1, U) \\ = \quad & (I_M - B_* B_*^T)U + O\left(\| (I_M - B_* B_*^T)U \|_F \| U \|_F \right) \end{aligned}$$

as $\| U \|_F \to 0$. These properties are easily verified by using the definition of the matrix exponential $\exp(\cdot)$, and the Taylor series expansion.

4.2. *Loss decomposition.* We shall show that, given $\eta > 0$, for an appropriate choice of the constant $c_{0,\eta}$ in the definition of α_n (in Theorem 2.1), for large enough n, we have

$$\mathbb{P}\left(\inf_{(B,\Lambda) \in \Theta(\alpha_n)} L_n(B, \Lambda) > L_n(B_*, \Lambda_*) \right) \geq 1 - O(n^{-\eta}). \tag{4.4}$$

From this, it follows immediately that with probability tending towards 1, there is a local minimum $(\widehat{B}, \widehat{\Lambda})$ of $L_n(B, \Lambda)$ in the set $\overline{\Theta}(\alpha_n)$ defined as

$$\begin{aligned} \overline{\Theta}(\alpha_n) \quad = \quad & \{(\mathbf{exp}(1, B_* A_U + C_U), \Lambda_* \exp(D)) : A_U = -A_U^T, B_*^T C_U = O, \\ & D \in \mathbb{R}^r \ \text{ such that } \ \| A_U \|_F^2 + \| C_U \|_F^2 + \| D \|_F^2 \leq \alpha_n^2\}, \end{aligned}$$

which concludes the proof of Theorem 2.1.

We start with the basic decomposition:

$$
\begin{aligned}
(4.5) \qquad & L_n(B,\Lambda) - L_n(B_*,\Lambda_*) \\
= \quad & [\mathbb{E}L_n(B,\Lambda) - \mathbb{E}L_n(B_*,\Lambda_*)] \\
& + [(L_n(B,\Lambda) - \mathbb{E}L_n(B,\Lambda)) - (L_n(B_*,\Lambda_*) - \mathbb{E}L_n(B_*,\Lambda_*))] \\
= \quad & \frac{1}{n}\sum_{i=1}^n K(\Sigma_i,\Sigma_{*i}) + \frac{1}{2n}\sum_{i=1}^n \mathrm{tr}\left((\Sigma_i^{-1} - \Sigma_{*i}^{-1})(S_i - \overline{\Sigma}_i)\right) \\
& + \frac{1}{2n}\sum_{i=1}^n \mathrm{tr}\left((\Sigma_i^{-1} - \Sigma_{*i}^{-1})(\overline{\Sigma}_i - \Sigma_{*i})\right),
\end{aligned}
$$

where $S_i = Y_i Y_i^T$ and $K(\Sigma_i, \Sigma_{*i})$ equals

$$
\frac{1}{2}\mathrm{tr}\left(\Sigma_i^{-1/2}(\Sigma_{*i} - \Sigma_i)\Sigma_i^{-1/2}\right) - \frac{1}{2}\log|I_{m_i} + \Sigma_i^{-1/2}(\Sigma_{*i} - \Sigma_i)\Sigma_i^{-1/2}|,
$$

which is the Kullback-Leibler divergence corresponding to observation i. Note that the proofs of Theorems 2.2 and 3.1 share a lot of commonality with the *sparse case* discussed here, in that these proofs depend on the same basic decomposition of the loss function.

4.3. *Probabilistic bounds for a fixed parameter in* $\Theta(\alpha_n)$. In order to derive the results in the following three propositions, we need to restrict our attention to an appropriate subset of the space of the design points $\mathbf{T}$ which has high probability. Accordingly, given $\eta > 0$, we define such a set A_η through (A.6) in Proposition A.1 (in Part I of the Appendix). The following proposition gives probabilistic bounds for the average Kullback-Leibler divergence in terms of α_n.

PROPOSITION 4.1. *Given $\eta > 0$, for every $(B,\Lambda) \in \Theta(\alpha_n)$, there is a set $A_{1,\eta}^{B,\Lambda}$ (depending on (B,Λ)), defined as*

$$
(4.6) \qquad A_{1,\eta}^{B,\Lambda} := \left\{ d'_\eta \alpha_n^2 \le \frac{1}{n}\sum_{i=1} K(\Sigma_i,\Sigma_{*i}) \le d''_\eta \overline{m}^2 \alpha_n^2 \right\},
$$

for appropriate positive constants d'_η and d''_η (depending on $\overline{\lambda}_1$ and r), such that for n large enough, $\mathbb{P}(A_\eta \cap (A_{1,\eta}^{B,\Lambda})^c) = O(n^{-(2+2\kappa)Mr-\eta})$.

Note that the bound in (4.6) is not sharp when $\overline{m} \to \infty$, which leads the suboptimal rates in Corollary 2.2. The following propositions bound the random term and the bias term in (4.5), respectively.

PROPOSITION 4.2. *Given $\eta > 0$, for each $(B, \Lambda) \in \Theta(\alpha_n)$, there is a set $A_{2,\eta}^{B,\Lambda}$, defined as,*

$$A_{2,\eta}^{B,\Lambda} = \left\{ \left| \frac{1}{2n} \sum_{i=1}^{n} tr((\Sigma_i^{-1} - \Sigma_{*i}^{-1})(S_i - \overline{\Sigma}_i)) \right| \leq d_\eta \overline{m} \alpha_n \sqrt{\frac{M \log n}{n}} \right\},$$

for some $d_\eta > 0$, such that, $\mathbb{P}(A_\eta \cap (A_{2,\eta}^{B,\Lambda})^c) = O(n^{-(2+2\kappa)Mr-\eta})$.

PROPOSITION 4.3. *Given $\eta > 0$, for each $(B, \Lambda) \in \Theta(\alpha_n)$, there is a set $A_{3,\eta}^{B,\Lambda}$, defined as,*

$$A_{3,\eta}^{B,\Lambda} = \Big\{ \left| \frac{1}{2n} \sum_{i=1}^{n} tr[(\Sigma_i^{-1} - \Sigma_{*i}^{-1})(\overline{\Sigma}_i - \Sigma_{*i})] \right| \quad \leq \quad d_\eta \overline{m} \alpha_n \sqrt{\frac{M \log n}{n}} \Big\},$$

for some constant $d_\eta > 0$, such that for large enough n, $\mathbb{P}(A_\eta \cap (A_{3,\eta}^{B,\Lambda})^c) = O(n^{-(2+2\kappa)Mr-\eta})$.

Combining Propositions 4.1-4.3, we obtain that, given $\eta > 0$, there is a constant $c_{0,\eta}$, such that, for every $(B, \Lambda) \in \Theta(\alpha_n)$,

$$\mathbb{P}(\{L_n(B, \Lambda) - L_n(B_*, \Lambda_*) \leq \frac{1}{2}\alpha_n^2\} \cap A_\eta) = O(n^{-(2+2\kappa)Mr-\eta}). \tag{4.7}$$

4.4. *Covering of the space $\Theta(\alpha_n)$.* To complete the proof of Theorem 2.1, we construct a δ_n-*net* in the set $\Theta(\alpha_n)$, for some $\delta_n > 0$ sufficiently small. This means that, for any $(B_1, \Lambda_1) \in \Theta(\alpha_n)$ there exists an element (B_2, Λ_2) of the net (with $B_k = \mathbf{exp}(1, B_* A_{U_k} + C_{U_k})$ and $\Lambda_k = \Lambda_* \exp(D_k)$, $k = 1, 2$), such that we have $\| B_1 - B_2 \|_F^2 + \| \Lambda_1 - \Lambda_2 \|_F^2 \leq \delta_n^2$. The spaces $\{A \in \mathbb{R}^{r \times r} : A = -A^T\}$ and $\{C \in \mathbb{R}^{M \times r} : B_*^T C = O\}$ are Euclidean subspaces of dimension $r(r-1)/2$ and $Mr - r^2$, respectively. Therefore, $\Theta(\alpha_n)$ is the image under $(\mathbf{exp}(1, \cdot), \exp(\cdot))$ of a hyper-ellipse of dimension $p = Mr - r(r+1)/2$. Thus, using standard construction of nets on spheres in $\mathbb{R}^p$, we can find such a δ_n-net $\mathcal{C}[\delta_n]$, with at most $d_1 \max\{1, (\alpha_n \delta_n^{-1})^p\}$ elements, for some $d_1 < \infty$.

If we take $\delta_n = (\overline{m}^2 n)^{-1}$, then from (4.7) using union bound it follows that, for n large enough,

$$\mathbb{P}\left(\left\{ \inf_{(B,\Lambda) \in \mathcal{C}[\delta_n]} L_n(B, \Lambda) - L_n(B_*, \Lambda_*) > \frac{1}{2}\alpha_n^2 \right\} \ \cap \ A_\eta \right) \geq 1 - O(n^{-\eta}).$$

This result, together with the following lemma and the fact that $\mathbb{P}(A_\eta) \geq 1 - O(n^{-\eta})$ (Proposition A.1), as well as the definition of $\mathcal{C}[\delta_n]$, proves (4.4). The proof of Lemma 4.1 is given in the Appendix (Part II).

LEMMA 4.1. *Let (B_k, Λ_k), $k = 1, 2$, be any two elements of $\Theta(\alpha_n)$ satisfying $\| B_1 - B_2 \|_F^2 + \| \Lambda_1 - \Lambda_2 \|_F^2 \leq \delta_n^2$, with $\delta_n = (\overline{m}^2 n)^{-1}$. Then, given $\eta > 0$, there are constants $d_{3,\eta}, d_{4,\eta} > 0$, such that, the set $A_{4,\eta} := \left\{\max_{1\leq i\leq n} \| \overline{\Sigma}_i^{-1/2} S_i \overline{\Sigma}_i^{-1/2} - I_{m_i} \|_F \leq d_{3,\eta}\overline{m}\log n\right\}$ satisfies $\mathbb{P}(A_{4,\eta}|\mathbf{T}) \geq 1 - O(n^{-\eta-1})$, for $\mathbf{T} \in A_\eta$; and on $A_{4,\eta}$, we have $|L_n(B_1, \Lambda_1) - L_n(B_2, \Lambda_2)| = o(\alpha_n^2)$.*

5. Proof of Theorem 3.1. There is essentially only one step where the proof of Theorem 3.1 differs from that of Theorem 2.1. It involves providing sharper bounds for the Kullback-Leibler divergence between an "optimal parameter", and an arbitrary parameter in the restricted parameter space $\widetilde{\Theta}(\alpha_n)$, an ellipsoid in the tangent space at the "optimal parameter":

$$\begin{aligned}\widetilde{\Theta}(\alpha_n) &= \{(\mathbf{exp}(1, B_* A_U + C_U), \Lambda_* \exp(D)) : A_U = -A_U^T, B_*^T C_U = O,\\ &\qquad D \in \mathbb{R}^r \text{ such that } \frac{\sigma^2}{s} \| A_U \|_F^2 + \| C_U \|_F^2 + \frac{\sigma^2}{s} \| D \|_F^2 = \alpha_n^2\}.\end{aligned}$$

Note that now the restricted parameter space is the image (under exponential maps) of an ellipse, whose principal axes can differ substantially depending on the signal-to-noise ratio s/σ^2. This is crucial for obtaining the sharper bounds for the Kullback-Leibler divergence (see equation (A.30)). As in Section 4, our strategy is to show that, given $\eta > 0$, for an appropriate choice of $c_{0,\eta}$, for large enough n, we have

$$\mathbb{P}\left(\inf_{(B,\Lambda)\in\widetilde{\Theta}(\alpha_n)} \widetilde{L}_n(B, \Lambda) > \widetilde{L}_n(B_*, \Lambda_*)\right) \geq 1 - O(n^{-\eta}).$$

From this, we conclude the proof of Theorem 3.1 using similar arguments as in the proof of Theorem 2.1.

Define $\widetilde{S} = \frac{1}{n}\sum_{i=1}^n \widetilde{Y}_i \widetilde{Y}_i^T$, where $\widetilde{Y}_i = \Phi Y_i$. Then, for an arbitrary $(B, \Lambda) \in \widetilde{\Theta}(\alpha_n)$, we have the following decomposition:

$$\begin{aligned}(5.1)\quad & \widetilde{L}_n(B, \Lambda) - \widetilde{L}_n(B_*, \Lambda_*)\\ &= \left[\left(\widetilde{L}_n(B, \Lambda) - \mathbb{E}\widetilde{L}_n(B, \Lambda)\right) - \left(\widetilde{L}_n(B_*, \Lambda_*) - \mathbb{E}\widetilde{L}_n(B_*, \Lambda_*)\right)\right]\\ &= K(\Gamma, \Gamma_*) + \frac{1}{2}\mathrm{tr}\left((\Gamma^{-1} - \Gamma_*^{-1})(\widetilde{S} - \overline{\Gamma})\right) + \frac{1}{2}\mathrm{tr}\left((\Gamma^{-1} - \Gamma_*^{-1})(\overline{\Gamma} - \Gamma_*)\right)\end{aligned}$$

with

$$\begin{aligned}K(\Gamma, \Gamma_*) &= \frac{1}{2}\mathrm{tr}\,(\Gamma^{-1}(\Gamma_* - \Gamma)) - \frac{1}{2}\log|I_M + \Gamma^{-1}(\Gamma_* - \Gamma)|\\ &= \frac{1}{2}\mathrm{tr}\,(\Gamma^{-1/2}(\Gamma_* - \Gamma)\Gamma^{-1/2}) - \frac{1}{2}\log|I_M + \Gamma^{-1/2}(\Gamma_* - \Gamma)\Gamma^{-1/2}|,\end{aligned}$$

being the Kullback-Leibler divergence between the probability distributions $N_M(0,\Gamma)$ and $N_M(0,\Gamma_*)$, where $\Gamma^{-1/2} = (\Gamma^{1/2})^{-1}$, and $\Gamma^{1/2}$ is a symmetric, positive definite, square root of Γ. The following is an analogue of Proposition 4.1.

PROPOSITION 5.1. *Under the assumptions of Theorem 3.1, there exist constants $c', c'' > 0$ such that, for sufficiently large n,*

$$c'\alpha_n^2\left(\frac{s}{\sigma^2}\right) \leq K(\Gamma,\Gamma_*) \leq c''\alpha_n^2\left(\frac{s}{\sigma^2}\right), \tag{5.2}$$

for all $(B,\Lambda) \in \widetilde{\Theta}(\alpha_n)$, where $\Gamma = sB\Lambda B^T + \sigma^2 I_M$ and $\Gamma_* = sB_*\Lambda_* B_*^T + \sigma^2 I_M$.

The following are analogues of the Propositions 4.2 and 4.3, respectively.

PROPOSITION 5.2. *Given $\eta > 0$, there exists a constant $c_\eta > 0$, such that for each $(B,\Lambda) \in \widetilde{\Theta}(\alpha_n)$,*

$$\begin{aligned} &\mathbb{P}\left(\left|tr((\Gamma^{-1} - \Gamma_*^{-1})(\widetilde{S} - \overline{\Gamma}))\right| \leq c_\eta\sqrt{\frac{M\log n}{n}}\sqrt{\frac{s}{\sigma^2}}\alpha_n\right) \\ \geq\ & 1 - O(n^{-(2+2\kappa)Mr-\eta}). \end{aligned}$$

This proposition can be easily proved using an exponential inequality by [10] on the fluctuations of the extreme eigenvalues of a Wishart matrix.

PROPOSITION 5.3. *There is a constant $c > 0$ such that, uniformly over $(B,\Lambda) \in \widetilde{\Theta}(\alpha_n)$,*

$$\begin{aligned} \left|tr((\Gamma^{-1} - \Gamma_*^{-1})(\overline{\Gamma} - \Gamma_*))\right| &\leq \parallel \Gamma^{-1} - \Gamma_*^{-1} \parallel_F \parallel \overline{\Gamma} - \Gamma_* \parallel_F \\ &\leq c\frac{1}{\sigma^2}\sqrt{\frac{s}{\sigma^2}}\alpha_n\beta_n. \end{aligned}$$

Propositions 5.1-5.3 (together with conditions **A1'** and **C"**) show that, for an appropriate choice of $c_{0,\eta}$, $\widetilde{L}_n(B,\Lambda) - \widetilde{L}_n(B_*,\Lambda_*) \geq c'\alpha_n^2$, for some $c' > 0$ with very high probability, for every fixed $(B,\Lambda) \in \widetilde{\Theta}(\alpha_n)$. The proof of Theorem 3.1 is finished by constructing a δ_n-*net* similarly as in Section 4.4 for the *sparse case*.

6. Proof of Theorem 3.2. The basic strategy of the proof is similar to that in classical inference with Euclidean parameter space. The main difficulty in present context lies in dealing with the Hessian operator of the log-likelihood (intrinsic Fisher information operator) and its inverse. Details of these calculations are given in the Appendix (Part IV).

Rewrite the negative log-likelihood (3.2) (up to a multiplicative constant) as

$$\widetilde{L}_n(B,\Lambda) = \mathrm{tr}\,(\Gamma^{-1}\widetilde{S}) + \log|\Gamma|, \quad \text{where } \widetilde{S} = \frac{1}{n}\sum_{i=1}^{n}\widetilde{Y}_i\widetilde{Y}_i^T. \tag{6.1}$$

By Theorem 3.1, given $\eta > 0$, there is a constant $c_{3,\eta} > 0$ such that the set

$$\widetilde{A}_{3,\eta} := \left\{\| \widehat{U} \|_F^2 + \| \widehat{D} \|_F^2 \leq c_{3,\eta}\alpha_n^2\right\}$$

has probability at least $1 - O(n^{-\eta})$, where $\alpha_n = c_{0,\eta}\sqrt{\frac{M\log n}{n}}$, and $(\widehat{U},\widehat{D}) \in \mathcal{T}_{B_*} \oplus \mathbb{R}^r$ is such that $(\widehat{B},\widehat{\Lambda}) := (\mathbf{exp}(1,\widehat{U}), \Lambda_* \exp(\widehat{D}))$ is a minimizer of (6.1).

First, by the same concentration bound for singular values of random matrices with i.i.d. Gaussian entries ([10]) used in the proof of Proposition 5.2, there exists $c_{4,\eta} > 0$, such that the set

$$\widetilde{A}_{4,\eta} := \left\{\| \widetilde{S} - \overline{\Gamma} \| \leq c_{4,\eta}\sqrt{\frac{M \vee \log n}{n}}\right\}$$

has probability at least $1 - O(n^{-\eta})$. It then follows that, we can choose an appropriate constant $c_{5,\eta} > 0$ such that, on $\widetilde{A}_{3,\eta} \cap \widetilde{A}_{4,\eta}$, $\| \nabla\widetilde{L}_n(\theta_*) \| \leq c_{5,\eta}\gamma_n$, where $\gamma_n = \max\{\sqrt{\frac{M\vee\log n}{n}}, \beta_n\}$ and $\theta_* = (B_*,\Lambda_*)$. Next, for any $X = (X_B, X_\zeta) \in \mathcal{T}_{B_*} \oplus \mathbb{R}^r$, define

$$\| X \| := \left[\| X_B \|_F^2 + \| X_\zeta \|_F^2\right]^{1/2}.$$

Also, let $\langle\cdot,\cdot\rangle_g$ denote the canonical metric on $\mathcal{T}_{B_*} \oplus \mathbb{R}^r$ (see Part IV of the Appendix). Using the fact that $\nabla\widetilde{L}_n(\widehat{\theta}) = 0$, where $\widehat{\theta} = (\widehat{B},\widehat{\Lambda})$, and defining $\widehat{\Delta} := (\widehat{U},\widehat{D})$, then on $\widetilde{A}_{3,\eta} \cap \widetilde{A}_{4,\eta}$, for any $X \in \mathcal{T}_{B_*} \oplus \mathbb{R}^r$ with $\| X \| \leq 1$,

$$\begin{aligned}
(6.2)\qquad & \langle -\nabla\widetilde{L}_n(\theta_*), X\rangle_g \\
&= \langle \nabla\widetilde{L}_n(\widehat{\theta}) - \nabla\widetilde{L}_n(\theta_*), X\rangle_g \\
&= \langle H_n(\theta_*)(\widehat{\Delta}), X\rangle_g + O(\| \widehat{\Delta} \|^2) + O(\gamma_n \| \widehat{\Delta} \|) \\
&= \langle H(\theta_*;\theta_*)(\widehat{\Delta}), X\rangle_g + \langle [H_n(\theta_*) - H(\theta_*;\theta_*)](\widehat{\Delta}), X\rangle_g + O(\alpha_n^2 + \alpha_n\gamma_n),
\end{aligned}$$

where $H_n(\cdot)(\widehat{\Delta})$ and $H(\cdot;\theta_*)(\widehat{\Delta})$ are the corresponding *covariant derivatives* of $\widetilde{L}_n(\cdot)$ and $L(\cdot;\theta_*)$ in the direction of $\widehat{\Delta}$. By simple calculations based on the expressions in the Appendix (Part IV), there exists a constant $c_{6,\eta} > 0$, such that on $\widetilde{A}_{3,\eta} \cap \widetilde{A}_{4,\eta}$, $\| H_n(\theta_*)(\widehat{\Delta}) - H(\theta_*;\theta_*)(\widehat{\Delta}) \| \leq c_{6,\eta}\alpha_n\gamma_n$. It can be checked using assumptions **A1'** and **C"** that the linear operator $H^{-1}(\theta_*;\theta_*) : \mathcal{T}_{B_*} \oplus \mathbb{R}^r \to \mathcal{T}_{B_*} \oplus \mathbb{R}^r$, is bounded in operator norm (see Part IV of the Appendix). Therefore, using the definition of covariant derivative and inverse of Hessian, from (6.2) we have, on $\widetilde{A}_{3,\eta} \cap \widetilde{A}_{4,\eta}$,

$$\widehat{\Delta} = -H^{-1}(\theta_*;\theta_*)(\nabla \widetilde{L}_n(\theta_*)) + O(\alpha_n\gamma_n) + O(\alpha_n^2). \tag{6.3}$$

Hence, on $\widetilde{A}_{3,\eta} \cap \widetilde{A}_{4,\eta}$, the bound on $\| \widehat{\Delta} \|$ can be improved from $O(\alpha_n)$ to

$$\| \widehat{\Delta} \| = O(\gamma_n\alpha_n + \alpha_n^2). \tag{6.4}$$

We can then repeat exactly the same argument, by using (6.2) to derive (6.3), but now with the bound on $\| \widehat{\Delta} \|$ given by (6.4). Since $M = O(n^a)$ for some $a < 1$, so that $\alpha_n^2 = o(\gamma_n)$, this way we get the more precise expression,

$$\widehat{\Delta} = -H^{-1}(\theta_*;\theta_*)(\nabla \widetilde{L}_n(\theta_*)) + O(\gamma_n^2). \tag{6.5}$$

Moreover, it can be easily verified that,

$$\frac{\partial}{\partial \zeta}\nabla_B L(\theta_*;\theta_*) := \mathbb{E}_{\theta_*}\left[\frac{\partial}{\partial \zeta}\nabla_B \widetilde{L}_n(\theta_*)\right] = 0.$$

Hence, by (A.41) in the Appendix (Part V), the Hessian operator, and its inverse, are "block diagonal", on the parameter space (viewed as a product manifold), with diagonal blocks corresponding to Hessians (inverse Hessians) w.r.t. B and ζ, respectively. This yields (3.3) and (3.4) in Theorem 3.2. Also, (3.5) and (3.6) follow immediately from (6.5).

7. Discussion. In this paper, we have demonstrated the effectiveness of utilizing the geometry of the non-Euclidean parameter space in determining consistency and rates of convergence of the REML estimators of principal components. We first study the REML estimators of eigenvalues and eigenfunctions of the covariance kernel for functional data, estimated from sparse, irregular measurements. The convergence rate of the estimated eigenfunctions is shown to be near-optimal when the number of measurements per curve is bounded and when M, the number of basis functions, varies with n at an appropriate rate (Theorem 2.1 and Corollary 2.1). The technique used in proving Theorem 2.1 is most suitable for dealing with the very sparse case (i.e., number of measurements per curve is bounded). We have also used it to prove consistency for the case where the number of measurements increases slowly with sample size (Corollary 2.2). However, this does not result in the optimal convergence rate. The latter case is more difficult because of the complications of dealing with inverses of random matrices (Σ_i) of growing dimensions. A more delicate analysis, that can handle this issue more efficiently, is likely to give tighter bounds for the average Kullback-Leibler divergence than that obtained in Proposition 4.1. Then it may be possible to extend the current technique to prove optimality of the REML estimators in a broader regime. A variant of the technique used for proving Theorem 2.1 also gives consistency of the REML estimator for functional data in a regime of dense measurements, as well as for a class of high-dimensional Gaussian vectors (Theorems 2.2 and 3.1). In the latter case, we also derive an efficient score representation (Theorem 3.2), which involves determining the intrinsic Fisher information operator and its inverse.

Now we present some conjectures we aim to pursue. First, as discussed earlier, based on the score representation, we conjecture the asymptotic optimality of the REML estimator for the *matrix case*. Secondly, we conjecture that there exists an efficient score representation of the REML estimator in the functional data problem as well. If so, then this estimator is likely to achieve the optimal nonparametric rate (for a broader regime), and may even be asymptotically optimal. This may explain the superior numerical performance of the REML estimator observed by [23]. Thirdly, our results (Theorems 2.2 and 3.1) give a strong indication of an asymptotic equivalence between two classes of problems : statistical inference for functional data with dense measurements; and inference for high-dimensional i.i.d. Gaussian vectors. Finally, in this paper we have not addressed the issue of model selection. A procedure for selection of M and r, based on an approximate leave-one-curve-out cross-validation score, has been proposed and implemented in [23]. This approximation is based on a second order Taylor expansion of the negative log-likelihood at the estimator and it involves the intrinsic Fisher information operator and its inverse. Therefore, based on the analysis presented here, it is conjectured that the approximate CV score thus defined is asymptotically consistent for the class of models considered in this paper.

APPENDIX

Part I : Properties of cubic B-spline basis. In many proofs of this paper, we need to use some properties of the cubic B-spline basis. We state some of them. More details can be found in [7] and [8]. Let

$\tilde{\boldsymbol{\phi}} = (\tilde{\phi}_1, \ldots, \tilde{\phi}_M)^T$ be the (standard) cubic B-spline basis functions on $[0,1]$ with equally spaced knots. Then, the orthonormalized spline functions $\phi_1, \ldots, \phi_M$ are defined through $\boldsymbol{\phi}(t) = G_{\phi,M}^{-1/2}\tilde{\boldsymbol{\phi}}(t)$, where $G_{\phi,M} := ((\int \tilde{\phi}_k(t)\tilde{\phi}_l(t)dt))_{k,l=1}^M$, is the *Gram matrix* of $\tilde{\boldsymbol{\phi}}$. It is known (cf. [6], [4]) that $G_{\phi,M}$ is an $M \times M$ banded matrix, and satisfies,

$$\frac{c_{\phi,0}}{M} I_M \le G_{\phi,M} \le \frac{c_{\phi,1}}{M} I_M \quad \text{for some constants } \; 0 < c_{\phi,0} < c_{\phi,1} < \infty. \tag{A.1}$$

From this, and other properties of cubic B-splines ([8], Chapter 13), we also have the following:

- **S1** $\sup_{t\in[0,1]} \sum_{k=1}^M \phi_k^2(t) \le c_{\phi,2} M$ for some constant $c_{\phi,2} > 0$.
- **S2** For any function $f \in C^{(4)}([0,1])$, we have $\| f - P_{\phi,M}(f) \|_\infty = \| f^{(4)} \|_\infty O(M^{-4})$, where $P_{\phi,M}(f) = \sum_{k=1}^M \langle f, \phi_k \rangle \phi_k$ denotes the projection of f onto $\text{span}\{\phi_1, \ldots, \phi_M\} = \text{span}\{\tilde{\phi}_1, \ldots, \tilde{\phi}_M\}$.

Note that, property **S2** and assumption **A2** imply the existence of orthonormal functions $\{\psi_{*k}\}_{k=1}^r$ of the form

$$(\psi_{*1}(t), \ldots, \psi_{*r}(t)) = (\boldsymbol{\psi}_*(t))^T = B_*^T \boldsymbol{\phi}(t), \quad B_*^T B_* = I_r,$$

which satisfy

$$\max_{1\le k\le r} \| \overline{\psi}_k - \psi_{*k} \|_\infty \le c_{\phi,3} M^{-4} \max_{1\le k\le r} \| \overline{\psi}_k^{(4)} \|_\infty . \tag{A.2}$$

Using these properties we obtain the following approximation to the important quantity $\| \Phi_i \|$, where $\Phi_i = [\boldsymbol{\phi}(T_{i1}) : \ldots : \boldsymbol{\phi}(T_{im_i})]$ and $\| \cdot \|$ denotes the operator norm. This result will be extremely useful in the subsequent analysis.

PROPOSITION A.1. *Given $\eta > 0$, there is an event A_η defined in terms of the design points $\mathbf{T}$, with probability at least $1 - O(n^{-\eta})$, such that on the set A_η,*

$$\| \Phi_i \|^2 \quad \le \quad \overline{m} c_{g,1} + \sqrt{5} c_{\phi,0}^{-1} d_\eta [(M^{3/2} \log n) \vee (M\sqrt{\overline{m}\log n})], \tag{A.3}$$

for some constant $d_\eta > 0$, and $c_{g,1}$ is the constant in condition **B2**. *Furthermore, for all* $\mathbf{T}$, *we have the non-random bound*

$$\| \Phi_i \|^2 \quad \le \quad c_{\phi,2} \overline{m} M, \quad \text{for all} \quad i = 1, \ldots, n. \tag{A.4}$$

PROOF. First, (A.4) follows from the bound **S1**, since

$$\begin{aligned} \| \Phi_i \|^2 &= \| \Phi_i^T \Phi_i \| = \| \Phi_i \Phi_i^T \| \le \| \Phi_i \|_F^2 = \text{tr}\,(\Phi_i^T \Phi_i) \\ &= \sum_{j=1}^{m_i} \sum_{k=1}^M (\phi_k(T_{ij}))^2 \le c_{\phi,2} m_i M \le c_{\phi,2} \overline{m} M, \quad \text{for all } i = 1, \ldots, n. \end{aligned}$$

In order to prove (A.3), first write

$$\text{(A.5)} \qquad \frac{1}{m_i}\Phi_i\Phi_i^T - \int \boldsymbol{\phi}(t)(\boldsymbol{\phi}(t))^T g(t)dt$$
$$= G_{\phi,M}^{-1/2}\left[\left(\left(\frac{1}{m_i}\sum_{j=1}^{m_i}[\tilde{\phi}_k(T_{ij})\tilde{\phi}_l(T_{ij}) - \mathbb{E}(\tilde{\phi}_k(T_{ij})\tilde{\phi}_l(T_{ij}))]\right)\right)_{k,l=1}^M\right] G_{\phi,M}^{-1/2}.$$

Next, observe that, $\mathbb{E}[\phi_k(T_{i1})\phi_l(T_{i1})]^2 = \int(\tilde{\phi}_k(t))^2(\tilde{\phi}_l(t))^2 g(t)dt = 0$ for $|k-l| > 3$; and is within $[c_{\phi,4}c_{g,0}M^{-1}, c_{\phi,5}c_{g,1}M^{-1}]$, for constants $0 < c_{\phi,4} < c_{\phi,5} < \infty$, if $|k-l| \leq 3$. Then, using the fact that $\max_{1\leq k\leq M} \parallel \tilde{\phi}_k \parallel_\infty$ is bounded, it follows from Bernstein's inequality that the set A_η defined as

$$\text{(A.6)} \quad A_\eta = \left\{\mathbf{T} : \max_{1\leq i\leq n}\max_{1\leq k,l\leq M}\left|\frac{1}{m_i}\sum_{j=1}^{m_i}[\tilde{\phi}_k(T_{ij})\tilde{\phi}_l(T_{ij}) - \mathbb{E}(\tilde{\phi}_k(T_{ij})\tilde{\phi}_l(T_{ij}))]\right| \leq d_{1,\eta}\left(\frac{\log n}{\underline{m}}\right) \vee \sqrt{\frac{\log n}{\underline{m}M}}\right\}$$

has probability at least $1 - O(n^{-\eta})$, for some constant $d_{1,\eta} > 0$. Now, we can bound the Frobenius norm of the matrix in (A.5) by using (A.1) and (A.6), and the fact that the matrix has $O(M)$ nonzero elements. Then using (2.3) we derive (A.3). □

Part II : Proofs for the sparse case.

Proof of Proposition 4.1. The main challenge in the proof of Proposition 4.1 is to efficiently approximate the average Kullback-Leibler divergence. We can express $K(\Sigma_i, \Sigma_{*i})$ as

$$\text{(A.7)} \qquad K(\Sigma_i, \Sigma_{*i}) = \frac{1}{2}\sum_{j=1}^{m}[\lambda_j(R_{*i}) - \log(1 + \lambda_j(R_{*i}))],$$

where $\lambda_j(R_{*i})$ is the j-th largest eigenvalue of $R_{*i} = \Sigma_i^{-1/2}(\Sigma_{*i} - \Sigma_i)\Sigma_i^{-1/2}$. Using the inequality $e^x \geq 1+x$ for $x \in \mathbb{R}$ (so that each term in the summation in (A.7) is nonnegative), and the Taylor series expansion for $\log(1+x)$ for $|x| < 1$, it can be shown that, given $\epsilon > 0$ sufficiently small (but fixed), there exist constants $0 < c_{1,\epsilon} < c_{2,\epsilon} < \infty$ such that for $\parallel R_{*i} \parallel_F \leq \epsilon$,

$$\text{(A.8)} \qquad c_{1,\epsilon} \parallel R_{*i} \parallel_F^2 \leq K(\Sigma_i, \Sigma_{*i}) \leq c_{2,\epsilon} \parallel R_{*i} \parallel_F^2 .$$

Next, observe that

$$\frac{\parallel \Sigma_{*i}^{-1/2}(\Sigma_i - \Sigma_{*i})\Sigma_{*i}^{-1/2} \parallel_F}{1 + \parallel \Sigma_{*i}^{-1/2}(\Sigma_i - \Sigma_{*i})\Sigma_{*i}^{-1/2} \parallel_F} \leq \parallel R_{*i} \parallel_F \leq \frac{\parallel \Sigma_{*i}^{-1/2}(\Sigma_i - \Sigma_{*i})\Sigma_{*i}^{-1/2} \parallel_F}{1 - \parallel \Sigma_{*i}^{-1/2}(\Sigma_i - \Sigma_{*i})\Sigma_{*i}^{-1/2} \parallel_F},$$

whenever $\| \Sigma_{*i}^{-1/2}(\Sigma_i - \Sigma_{*i})\Sigma_{*i}^{-1/2} \|_F < 1$. the proof of Proposition 2.1 can thus be reduced to finding probabilistic bounds for $\frac{1}{n}\sum_{i=1}^n \| \Sigma_{*i}^{-1/2}(\Sigma_i - \Sigma_{*i})\Sigma_{*i}^{-1/2} \|_F^2$.

One difficulty in obtaining those bounds is in handling the inverse of the matrices Σ_{*i}. In order to address that, and some related issues, we use the properties of the cubic spline basis derived in the Appendix (Part I). In the following lemmas we confine ourselves to the restricted parameter space $\Theta(\alpha_n)$, i.e., $(B,\Lambda) \in \Theta(\alpha_n)$.

LEMMA A.1. *Under the assumptions of Theorem 2.1 (for m_i's bounded), or Corollary 2.2 (for m_i's increasing slowly with n),*

$$(1 + d_1\overline{\lambda}_1 r\overline{m})^{-1} \| \Sigma_i - \Sigma_{*i} \|_F \leq \| \Sigma_{*i}^{-1/2}(\Sigma_i - \Sigma_{*i})\Sigma_{*i}^{-1/2} \|_F \leq \| \Sigma_i - \Sigma_{*i} \|_F, \tag{A.9}$$

for some constant $d_1 > 0$, for all $i = 1, \ldots, n$.

PROOF. From condition **A2** and (A.2), it follows that, $\exists\ D_1 > 0$ such that, for all M,

$$\max_{1\leq k\leq r} \| \psi_{*k} \|_\infty \leq D_1 < \infty. \tag{A.10}$$

This, together with the definition of (B_*, Λ_*), leads to the following bound on the eigenvalues of the matrices Σ_{*i}:

$$1 \leq \lambda_{\min}(\Sigma_{*i}) \leq \lambda_{\max}(\Sigma_{*i}) \leq 1 + D_1 r m_i \lambda_{*1} \leq 1 + d_1\overline{\lambda}_1 r\overline{m}, \tag{A.11}$$

for some $d_1 > 0$, for all $i = 1, \ldots, n$, from which (A.9) follows. □

LEMMA A.2. *Under the assumptions of Theorem 2.1 (for m_i's bounded), or Corollary 2.2 (for m_i's increasing slowly with n), given any $\eta > 0$, on the event A_η defined through (A.6) in Proposition A.1, which has probability at least $1 - O(n^{-\eta})$, for sufficiently large n,*

$$\begin{aligned}\max_{1\leq i\leq n} \| \Sigma_i - \Sigma_{*i} \|_F^2 &\\ \leq \left[d_{3,\eta}\left(1 + d_1\left[\left(\frac{M^{3/2}\log n}{\overline{m}}\right) \vee \sqrt{\frac{M^2\log n}{\overline{m}}}\right]\right)\overline{m}^2\alpha_n^2\right] &\wedge \left[d_2 M\overline{m}^2\alpha_n^2\right],\end{aligned} \tag{A.12}$$

where the second bound holds for all $\mathbf{T}$. Here $d_1, d_2, d_{3,\eta} > 0$ are appropriate constants depending on r and $\overline{\lambda}_1$.

PROOF. An upper bound for $\| \Sigma_i - \Sigma_{*i} \|_F$ is obtained by expressing

$I_M = B_* B_*^T + (I_M - B_* B_*^T)$, and then applying the triangle inequality,

$$
\begin{aligned}
\text{(A.13)}\quad \| \Sigma_i - \Sigma_{*i} \|_F \;&=\; \| \Phi_i^T (B\Lambda B^T - B_*\Lambda_* B_*^T)\Phi_i \|_F \\
\leq \;& \| \Phi_i^T B_* (B_*^T B\Lambda B^T B_* - \Lambda_*) B_*^T \Phi_i \|_F \\
& + 2 \| \Phi_i^T B_* B_*^T B\Lambda B^T (I_M - B_* B_*^T)\Phi_i \|_F \\
& + \| \Phi_i^T (I_M - B_* B_*^T) B\Lambda B^T (I_M - B_* B_*^T)\Phi_i \|_F \\
\leq \;& \| \Phi_i^T B_* \|^2 \| B_*^T B\Lambda B^T B_* - \Lambda_* \|_F \\
& + 2 \| \Phi_i^T B_* \| \| \Lambda \| \| B^T (I_M - B_* B_*^T)\Phi \|_F + \| \Lambda \| \| \Phi_i^T (I_M - B_* B_*^T) B \|_F^2 \\
\leq \;& D_1 r \overline{m} \| B_*^T B\Lambda B^T B_* - \Lambda_* \|_F \\
& + \sqrt{d_4 r \overline{m}}\overline{\lambda}_1 \| \Phi_i^T (I_M - B_* B_*^T) B \|_F \,(1 + (D_1 r\overline{m})^{-1/2} \| \Phi_i^T (I_M - B_* B_*^T) B \|_F),
\end{aligned}
$$

for some $d_4 > 1$. For the second inequality we use $\| B_*^T B \| \leq 1$, and for the last inequality we use (A.10) and (A.11). Next, by using (A.4), (4.3) and (2.5), we obtain the (nonrandom) bound

$$
\begin{aligned}
\text{(A.14)}\quad \frac{1}{\overline{m}} \max_{1\leq i\leq n} \| \Phi_i^T (I_M - B_* B_*^T) B \|_F^2 \;&\leq\; c_{\phi,2} M \| (I_M - B_* B_*^T) B \|_F^2 \\
&\leq\; c_{\phi,2} M \alpha_n^2 (1 + o(1)) \;=\; o(1).
\end{aligned}
$$

Then the bound in (A.13) can be majorized by,

$$
D_1 r \overline{m} \| B_*^T B\Lambda B^T B_* - \Lambda_* \|_F + \sqrt{d_4 r\overline{m}}\overline{\lambda}_1 \| \Phi_i \| \| (I_M - B_* B_*^T) B \|_F \,(1 + o(1)).
$$

Using (A.3) to bound $\| \Phi_i \|$, from (A.13), and the definition of $\Theta(\alpha_n)$ together with (4.2) and (4.3), we obtain (A.12). □

LEMMA A.3. *Under the assumptions of Theorem 2.1 (for m_i's bounded), or Corollary 2.2 (for m_i's increasing slowly with n), for any given $\eta > 0$, there is a positive sequence $\varepsilon_{1,n} = o(1)$ (depending on η), and a constant $d_{1,\eta} > 0$, such that*

$$
\mathbb{P}\left[\frac{1}{n} \sum_{i=1}^{n} \| \Sigma_i - \Sigma_{*i} \|_F^2 > d_{1,\eta} \underline{m}^2 \alpha_n^2 (1 - \varepsilon_{1,n}) \right] \geq 1 - O(n^{-(2+2\kappa)Mr - \eta}),
$$

where κ is as in **B1**.

PROOF. Let $\Delta = B\Lambda B^T - B_*\Lambda_* B_*^T$. Observe that, by definition of $\Theta(\alpha_n)$ (equation (4.1)), and equations (4.2) and (4.3), for large n,

$$
\text{(A.15)}\qquad \| \Delta \|_F^2 \leq c_* \alpha_n^2, \qquad \text{(for some constant } c_* > 0).
$$

First, consider the lower bound

$$
\text{(A.16)}\qquad \| \Sigma_i - \Sigma_{*i} \|_F^2 = \operatorname{tr}[\Phi_i^T \Delta \Phi_i \Phi_i^T \Delta \Phi_i] \geq \sum_{j_1 \neq j_2}^{m_i} [(\phi(T_{ij_1}))^T \Delta \phi(T_{ij_2})]^2.
$$

We will derive an exponential tail bound for $\frac{1}{n}\sum_{i=1}^{n}\sum_{j_1\neq j_2}^{m_i}[(\boldsymbol{\phi}(T_{ij_1}))^T\Delta\boldsymbol{\phi}(T_{ij_2})]^2$. Rewriting the term on the extreme right, and using the fact that $\{T_{ij}\}_{j=1}^{m_i}$ are i.i.d. with density $g : c_{g,0} \leq g \leq c_{g,1}$ (**B2**), we have, for all i,

$$
\begin{aligned}
\text{(A.17)}\qquad & \mathbb{E}\sum_{j_1\neq j_2}^{m_i}[(\boldsymbol{\phi}(T_{ij_1}))^T\Delta\boldsymbol{\phi}(T_{ij_2})]^2 \\
&= m_i(m_i-1)\mathrm{tr}\left(\mathbb{E}[\boldsymbol{\phi}(T_{i1})(\boldsymbol{\phi}(T_{i1}))^T]\Delta\mathbb{E}[\boldsymbol{\phi}(T_{i2})(\boldsymbol{\phi}(T_{i2}))^T]\Delta\right) \\
&\in \left(c_{g,0}^2\underline{m}(\underline{m}-1)\parallel\Delta\parallel_F^2,\;\; c_{g,1}^2\overline{m}(\overline{m}-1)\parallel\Delta\parallel_F^2\right) \\
&\in \left(d_1'\underline{m}^2\alpha_n^2(1+o(1)),\quad d_1''\overline{m}^2\alpha_n^2(1+o(1))\right),
\end{aligned}
$$

for some $d_1'' \geq d_1' > 0$ (whose values depend on $c_{g,0}$, $c_{g,1}$ and the constants appearing in **A1**), where in the last step we use (A.15). The last inequality uses (A.43), (A.44), the definition of $\Theta(\alpha_n)$, and properties (4.2) and (4.3). Notice that, the variance of $\sum_{j_1\neq j_2}^{m_i}[(\boldsymbol{\phi}(T_{ij_1}))^T\Delta\boldsymbol{\phi}(T_{ij_2})]^2$ can be bounded, for sufficiently large n, as

$$
\begin{aligned}
& \max_{1\leq i\leq n}\mathrm{Var}\left(\sum_{j_1\neq j_2}^{m_i}[(\boldsymbol{\phi}(T_{ij_1}))^T\Delta\boldsymbol{\phi}(T_{ij_2})]^2\right) \\
\leq\;& \max_{1\leq i\leq n}\mathbb{E}\left(\parallel\Sigma_i-\Sigma_{*i}\parallel_F^2\sum_{j_1\neq j_2}^{m_i}[(\boldsymbol{\phi}(T_{ij_1}))^T\Delta\boldsymbol{\phi}(T_{ij_2})]^2\right) \\
\leq\;& d_2'M(\overline{m}^2\alpha_n^2)^2 \;=:\; V_{1,n},
\end{aligned}
$$

where $d_2' > 0$ is some constant. In the above, we obtain the first inequality by (A.16), and the second inequality by using (A.12) and (A.17). Next, $\sum_{j_1\neq j_2}^{m_i}[(\boldsymbol{\phi}(T_{ij_1}))^T\Delta\boldsymbol{\phi}(T_{ij_2})]^2$, for $i=1,\ldots,n$, are independent, and bounded by $K_{1,n} := d_4M\overline{m}^2\alpha_n^2$, for a constant $d_4 > 0$ (using (A.12)). Hence, by applying Bernstein's inequality, and noticing that $K_{1,n}\sqrt{\frac{M\log n}{n}} = o(\sqrt{V_{1,n}})$, and $\sqrt{V_{1,n}}\sqrt{\frac{M\log n}{n}} = o(\overline{m}^2\alpha_n^2)$ (by (2.5), or (2.6)), the result follows. □

LEMMA A.4. *Under the assumptions of Theorem 2.1 (for m_i's bounded), or Corollary 2.2 (for m_i's increasing slowly with n), for any given $\eta > 0$, there is a positive sequence $\varepsilon_{2,n} = o(1)$ and a constant $d_{2,\eta} > 0$, such that*

$$
\text{(A.18)}\qquad \mathbb{P}\left[\frac{1}{n}\sum_{i=1}^{n}\parallel\Sigma_i-\Sigma_{*i}\parallel_F^2 < d_{2,\eta}\overline{m}^2\alpha_n^2(1+\varepsilon_{2,n})\right] \geq 1-O(n^{-(2+2\kappa)Mr-\eta}),
$$

where κ is as in **B1**.

PROOF. From the proof of Lemma A.2, especially the inequalities (A.13) and (A.14), it is clear that we only need to provide a sharp upper bound for

$\frac{1}{n}\sum_{i=1}^n \parallel \Phi_i^T(I_M - B_* B_*^T)B \parallel_F^2$. Let $\overline{\Delta} := (I_M - B_* B_*^T)B$. Then from (4.3), for n large enough,

$$\parallel \overline{\Delta} \parallel_F^2 \leq c_* \alpha_n^2 \tag{A.19}$$

for some $c_* > 0$. Then, using (2.3), for all i,

$$\begin{aligned}\mathbb{E} \parallel \Phi_i^T(I_M - B_* B_*^T)B \parallel_F^2 &= \sum_{j=1}^{m_i} \mathrm{tr}\left(\mathbb{E}(\boldsymbol{\phi}(T_{ij})(\boldsymbol{\phi}(T_{ij}))^T)\overline{\Delta}\overline{\Delta}^T\right) \\ &\leq c_{g,1} m_i \mathrm{tr}\,[\overline{\Delta}\overline{\Delta}^T] \leq c_{g,1}\overline{m} \parallel \overline{\Delta} \parallel_F^2 .\end{aligned} \tag{A.20}$$

Combining (A.20) with (A.19), (A.13) and (A.14), we get, for sufficiently large n, and some constant $C > 0$,

$$\frac{1}{n}\sum_{i=1}^n \mathbb{E} \parallel \Sigma_i - \Sigma_{*i} \parallel_F^2 \leq C\, \overline{m}^2 \alpha_n^2$$

Next, using (A.4), (A.14) and (A.20) we have

$$\begin{aligned}\max_{1\leq i\leq n} \mathrm{Var}(\parallel \Phi_i^T \overline{\Delta} \parallel_F^2) &\leq \max_{1\leq i\leq n} \mathbb{E} \parallel \Phi_i^T \overline{\Delta} \parallel_F^4 \\ &\leq c_{\phi,2}\overline{m}M \parallel \overline{\Delta} \parallel_F^2 \mathbb{E} \parallel \Phi_i^T \overline{\Delta} \parallel_F^2 \\ &\leq c_{\phi,2}c_{g,1}\overline{m}^2 M \parallel \overline{\Delta} \parallel_F^4 \\ &\leq C' M \overline{m}^2 \alpha_n^4 (1+\varepsilon_n) \;=:\; V_{2,n},\end{aligned} \tag{A.21}$$

for some positive sequence $\varepsilon_n = o(1)$ and some constant $C' > 0$, where in the last step we used (A.19). Again, using Bernstein's inequality for $\frac{1}{n}\sum_{i=1}^n \parallel \Phi_i^T \overline{\Delta} \parallel_F^2$, which is a sum of independent variables bounded by $K_{2,n} = c_{g,1} M \overline{m} \alpha_n^2(1 + o(1))$, the result follows (checking that, by (2.5), or, (2.6), we have, $K_{2,n}\sqrt{\frac{M\log n}{n}} = o(\sqrt{V_{2,n}})$ and $\sqrt{V_{2,n}}\sqrt{\frac{M\log n}{n}} = o(\overline{m}^2\alpha_n^2))$. □

Proof of Proposition 4.2. Write, $R_i = \overline{\Sigma}_i^{1/2}(\Sigma_i^{-1} - \Sigma_{*i}^{-1})\overline{\Sigma}_i^{1/2}$. We can bound $\parallel R_i \parallel_F$ as

$$\begin{aligned}\parallel R_i \parallel_F &\leq \parallel \overline{\Sigma}_i^{1/2}\Sigma_i^{-1/2} \parallel \parallel \overline{\Sigma}_i^{1/2}\Sigma_{*i}^{-1/2} \parallel \parallel \Sigma_i^{-1/2}\Sigma_{*i}^{1/2} \parallel \parallel \Sigma_{*i}^{-1/2}(\Sigma_i - \Sigma_{*i})\Sigma_{*i}^{-1/2} \parallel_F \\ &\leq \parallel \overline{\Sigma}_i^{1/2}\Sigma_{*i}^{-1/2} \parallel^2 \parallel \Sigma_i^{-1/2}\Sigma_{*i}^{1/2} \parallel^2 \parallel \Sigma_i - \Sigma_{*i} \parallel_F \\ &\leq (1+ \parallel \overline{\Sigma}_i - \Sigma_{*i} \parallel)(1- \parallel \Sigma_i - \Sigma_{*i} \parallel)^{-1} \parallel \Sigma_i - \Sigma_{*i} \parallel_F,\end{aligned}$$

where the third inequality is due to (A.11) and (A.42). Note that, by condition **C**, it follows that $\max_{1\leq i\leq n} \parallel \overline{\Sigma}_i - \Sigma_{*i} \parallel_F \leq C\, \overline{m}\overline{\beta}_n = o(1)$ for some constant $C > 0$. Therefore, applying (A.12), we observe that for $\mathbf{T} \in A_\eta$ with A_η as in (A.6), for large enough n, $\parallel R_i \parallel_F \leq 2 \parallel \Sigma_i - \Sigma_{*i} \parallel_F$. Due to the Gaussianity of the observations, for any symmetric $m_i \times m_i$ matrix A, the random variable $\mathrm{tr}\,(A(S_i - \overline{\Sigma}_i))$ has the same distribution as $\mathrm{tr}\,(D_i(X_i X_i^T - I_{m_i}))$, where D_i is a diagonal matrix of the eigenvalues of $\overline{\Sigma}_i^{-1/2} A \overline{\Sigma}_i^{-1/2}$, and $X_i \sim N(0, I_{m_i})$ are independent. Therefore, using an exponential inequality for a weighted sum of independent χ_1^2 random variables, we have, for $\mathbf{T} \in A_\eta$ and each $(B, \Lambda) \in \Theta(\alpha_n)$,

$$\mathbb{P}_{\mathbf{T}}\left[\left|\frac{1}{n}\sum_{i=1}^{n}\mathrm{tr}\,((\Sigma_i^{-1} - \Sigma_{*i}^{-1})(S_i - \overline{\Sigma}_i))\right| \leq d_{3,\eta}\sqrt{\frac{M\log n}{n}}\left[\frac{1}{n}\sum_{i=1}^{n} \| \Sigma_i - \Sigma_{*i} \|_F^2\right]^{1/2}\right]$$
$$\geq \quad 1 - O(n^{-(2+2\kappa)Mr-\eta}),$$

for a constant $d_{3,\eta} > 0$, where $\mathbb{P}_{\mathbf{T}}$ denotes the conditional probability given $\mathbf{T}$. Therefore, using (A.18) we conclude the proof.

Proof of Proposition 4.3. Using Cauchy-Schwarz inequality twice, we can bound the last term in (4.5), which corresponds to *model bias*, as

$$\left|\frac{1}{2n}\sum_{i=1}^{n}\mathrm{tr}\,((\Sigma_i^{-1} - \Sigma_{*i}^{-1})(\overline{\Sigma}_i - \Sigma_{*i}))\right|$$
$$\leq \quad \frac{1}{2}\left[\frac{1}{n}\sum_{i=1}^{n}\| \overline{\Sigma}_i - \Sigma_{*i} \|_F^2\right]^{1/2}\left[\frac{1}{n}\sum_{i=1}^{n}\| \Sigma_i^{-1} - \Sigma_{*i}^{-1} \|_F^2\right]^{1/2}$$
$$\leq \quad \frac{1}{2}\max_{1\leq j\leq n}\| \overline{\Sigma}_j - \Sigma_{*j} \|_F \left[\frac{1}{n}\sum_{i=1}^{n}\| \Sigma_i^{-1} - \Sigma_{*i}^{-1} \|_F^2\right]^{1/2}$$
$$\leq \quad \frac{1}{2}\overline{m}\overline{\beta}_n \max_{1\leq j\leq n}\| \Sigma_j^{-1} \| \| \Sigma_{*j}^{-1} \| \left[\frac{1}{n}\sum_{i=1}^{n}\| \Sigma_i - \Sigma_{*i} \|_F^2\right]^{1/2},$$

where in the last step we used (A.42). Thus, the proof is finished by using condition **C**, (A.11) and (A.18).

Now, to the complete the proof of Theorem 2.1, we give the details of the covering argument.

Proof of Lemma 4.1. Using an expansion analogous to (4.5) and the upper bound in (A.8), and applying Cauchy-Schwarz inequality, we have, for some constants $C_1, C_2 > 0$, on $A_{4,\eta}$ and for $\mathbf{T} \in A_\eta$, for n large enough,

$$
\begin{aligned}
& |L_n(B_1,\Lambda_1) - L_n(B_2,\Lambda_2)| \\
\leq\ & \frac{1}{n}\sum_{i=1}^n \parallel \Sigma_{1,i}^{-1} - \Sigma_{2,i}^{-1} \parallel_F \parallel S_i - \overline{\Sigma}_i \parallel_F \\
& + \frac{1}{n}\sum_{i=1}^n \left[C_1 \parallel \Sigma_{1,i} - \Sigma_{2,i} \parallel_F^2 + \parallel \Sigma_{1,i}^{-1} - \Sigma_{2,i}^{-1} \parallel_F \parallel \overline{\Sigma}_i - \Sigma_{2,i} \parallel_F\right] \\
\leq\ & \max_{1\leq i\leq n} \parallel \Sigma_{1,i}^{-1} - \Sigma_{2,i}^{-1} \parallel_F \parallel \overline{\Sigma}_i \parallel \max_{1\leq i\leq n} \parallel \overline{\Sigma}_i^{-1/2} S_i \overline{\Sigma}_i^{-1/2} - I_{m_i} \parallel_F \\
& \quad + C_2 \max_{1\leq i\leq n} \parallel \Sigma_{1,i} - \Sigma_{2,i} \parallel_F^2 \\
& + \max_{1\leq i\leq n} \parallel \Sigma_{1,i}^{-1} - \Sigma_{2,i}^{-1} \parallel_F (\max_{1\leq i\leq n} \parallel \overline{\Sigma}_i - \Sigma_{*i} \parallel_F + \max_{1\leq i\leq n} \parallel \Sigma_{2,i} - \Sigma_{*i} \parallel_F) \\
\leq\ & d_{4,\eta}[\overline{m}^2 \delta_n \overline{m} \log n + \overline{m}^2 \delta_n^2 + \overline{m}\delta_n(\overline{m}\overline{\beta}_n + \sqrt{M}\overline{m}\alpha_n)] \\
=\ & o(\alpha_n^2).
\end{aligned}
$$

In the last step, we have used Lemma A.2 (for the last term), the identity (A.42) in the Appendix (Part VI), and the fact that $\parallel \Sigma_{k,i}^{-1} \parallel \leq 1$ $(k = 1, 2)$.

Proof of Corollary 2.1. The best rate follows by direct calculation.

The near-optimality of the estimator requires proving that, for an appropriately chosen subclass $\mathcal{C}$ of covariance kernels of rank r, we have the following analog of Theorem 2 of [15]: for any estimator $\{\widehat{\psi}_k\}_{k=1}^r$ of the eigenfunctions $\{\psi_k\}_{k=1}^r$, for n sufficiently large,

$$
\min_{1\leq k\leq r} \sup_{\overline{\Sigma}_0 \in \mathcal{C}} \mathbb{E} \parallel \widehat{\psi}_k - \overline{\psi}_k \parallel_2^2 \geq C n^{-8/9}, \tag{A.22}
$$

for some $C > 0$. Here the parameter space $\mathcal{C}$ consists of covariance kernels of rank r with eigenfunctions satisfying **A1**-**A2**. Moreover, the random design satisfies **B1**-**B2**, with $\overline{m}$ bounded above.

The derivation of the lower bound on the risk involves construction of a finite, "least favorable" parameter set in $\mathcal{C}$ by combining the constructions in [22] (for obtaining lower bounds on risk in high-dimensional PCA) and [15] (for functional data case). This construction is as follows. Let $\phi_1^0, \ldots, \phi_r^0$ be a set of orthonormal functions on $[0, 1]$ which are four times continuously differentiable, with fourth derivative bounded. Let $M_* \asymp n^{1/9}$ be an integer appropriately chosen. Let $\gamma_1, \ldots, \gamma_{M_*}$ be a set of basis functions that are (i) orthonormal on $[0, 1]$, and orthogonal to the set $\{\phi_1^0, \ldots, \phi_r^0\}$; (ii) are four times continuously differentiable and γ_j is supported on an interval of length $O(M_*^{-1})$ around the point $\frac{j}{M_*}$. One particular choice for these functions is to let $\{\phi_k^0\}$ be the translated periodized scaling functions of a wavelet basis at a certain *scale* with adequate degree of smoothness, and to let $\{\gamma_j\}_{j=1}^{M_*}$ be the set of compactly supported, orthonormal, periodized wavelet functions corresponding to the scaling functions. Indeed, then we can choose M_* to be an integer power of 2. Note that, such a basis ($\{\phi_k^0 : 1 \leq k \leq r\} \cup \{\gamma_l : 1 \leq l \leq M_*\}$) has the stability and smoothness property commensurate with the orthonormalized B-spline basis we are using for deriving the REML estimators. Next, let $\overline{\lambda}_1 > \cdots > \overline{\lambda}_r > 0$ be fixed numbers satisfying **A1**. Finally, let us define a covariance kernel $\overline{\Sigma}_0^{(0)}$ as

$$\overline{\Sigma}_0^{(0)}(s,t) = \sum_{k=1}^{r} \overline{\lambda}_k \phi_k^0(s)\phi_k^0(t), \quad s,t \in [0,1]. \tag{A.23}$$

Also, for each fixed j in some index set $\mathcal{F}_0$ (to be specified below), define

$$[\psi_1^{(j)}(s) : \cdots : \psi_r^{(j)}(s)] = \widetilde{\Psi}(s)\overline{B}^{(j)}, \quad s \in [0,1]$$

where $\widetilde{\Psi}(s) = (\phi_1^0(s), \ldots, \phi_r^0(s), \gamma_1(s), \ldots, \gamma_{M^*}(s))$ and $\overline{B}^{(j)}$ is an $(M_*+r)\times r$ matrix with orthonormal columns (to be specified below). Then define

$$\overline{\Sigma}_0^{(j)}(s,t) = \sum_{k=1}^{r} \overline{\lambda}_k \psi_k^{(j)}(s)\psi_k^{(j)}(t), \quad s,t \in [0,1], \tag{A.24}$$

for $j \in \mathcal{F}_0$. We require that $\log|\mathcal{F}_0| \asymp M_* \asymp n^{1/9}$, and $\| \overline{B}^{(j)} - \overline{B}^{(j')} \|_F^2 \asymp n^{-8/9}$, for $j \neq j'$ and $j, j' \in \mathcal{F}_0 \cup \{0\}$. Here $\overline{B}^{(0)}$ is the $(M_* + r) \times r$ matrix of basis coefficients of $\overline{\Sigma}_0^{(0)}$ with columns $\mathbf{e}_k$, the k-th canonical basis vector in $\mathbb{R}^{M_*+r}$.

The proof of the minimax lower bound is based on an application of *Fano's Lemma* (cf. [25]), which requires computation of the Kullback-Leibler divergence between two specific values of the parameters. In order to apply *Fano's lemma*, we need to choose $\mathcal{F}_0$ and $\overline{B}^{(j)}, j \in \mathcal{F}_0$, such that

$$\frac{\mathrm{ave}_{j\in\mathcal{F}_0}[\sum_{i=1}^{n} \mathbb{E}K(\overline{\Sigma}_i^{(j)}, \overline{\Sigma}_i^{(0)})] + \log 2}{\log|\mathcal{F}_0|} \approx c \in (0,1), \tag{A.25}$$

where $\overline{\Sigma}_i^{(j)}$ denotes the covariance of the observation i given $\{T_{il}\}_{l=1}^{m_i}$ under the model parameterized by $\overline{\Sigma}_0^{(j)}$, and $\mathbb{E}$ denotes expectation with respect to the design points $\mathbf{T}$. Under the assumptions on the design points, using the properties of the basis functions $\{\phi_k^0\}_{=1}^{r}$ and $\{\gamma_k\}_{k=1}^{M_*}$, and the computations carried out in the proof of Proposition 4.1 (in Part II of the Appendix), in particular a nonrandom bound analogous to the second bound appearing in Lemma A.2, it is easy to see that for n large enough (so that $\| \overline{B}^{(j)} - \overline{B}^{(0)} \|_F$ is sufficiently small), we have

$$\frac{1}{n}\sum_{i=1}^{n} K(\overline{\Sigma}_i^{(j)}, \overline{\Sigma}_i^{(0)}) \asymp \frac{1}{n}\sum_{i=1}^{n} \| \overline{\Sigma}_i^{(j)} - \overline{\Sigma}_i^{(0)} \|_F^2 .$$

From this, and the property of the basis used to represent the eigenfunctions, it follows that

$$\text{(A.26)} \qquad \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}K(\overline{\Sigma}_i^{(j)}, \overline{\Sigma}_i^{(0)}) \asymp \| \overline{B}^{(j)} - \overline{B}^{(0)} \|_F^2 .$$

The task remains to construct $\mathcal{F}_0$ and $\overline{B}^{(j)}$ appropriately so that $\mathcal{C}_0 := \{\overline{\Sigma}_0^{(j)} : j \in \mathcal{F}_0\}$ is in $\mathcal{C}$, for n sufficiently large.

Following the proof of Theorem 2 in [22], we first define $M_0 = [\frac{2M_*}{9r}]$. Then define the k-th column of $\overline{B}^{(j)}$ as

$$\text{(A.27)} \qquad \overline{B}_k^{(j)} = \sqrt{1-\delta_k^2}\mathbf{e}_k + \delta_k \sum_{l=1}^{M_*} z_{kl}^{(j)} \mathbf{e}_{r+l}, \quad k = 1, \dots, r,$$

where $z_{kl}^{(j)}$ are appropriately chosen using a "sphere packing" argument (to ensure that $\log|\mathcal{F}_0| \asymp M_*$), and take values in $\{-M_0^{-1/2}, 0, M_0^{-1/2}\}$. Moreover, let S_k be the set of coordinates l such that $z_{kl}^{(j)} \neq 0$ for some $j \in \mathcal{F}_0$. By construction, S_k are disjoint for different $k = 1, \dots, r$, and $|S_k| \sim M_*/r$. Hence,

$$\text{(A.28)} \qquad \psi_k^{(j)} = \sqrt{1-\delta_k^2}\phi_k^0 + \delta_k \sum_{l \in S_k} z_{kl}^{(j)} \gamma_l, \quad k = 1, \dots, r.$$

Furthermore, by the construction of $\{z_{lk}^{(j)}\}$, $\sum_{l \in S_k} |z_{kl}^{(j)}|^2 = 1$, and for any $j \neq j'$ the vectors $\mathbf{z}_k^{(j)} = (z_{kl}^{(j)})_{l \in S_k}$ and $\mathbf{z}_k^{(j')} = (z_{kl}^{(j')})_{l \in S_k}$ satisfy $\| \mathbf{z}_k^{(j)} - \mathbf{z}_k^{(j')} \|_2 \geq 1$. Therefore, from (A.27) it follows that the RHS of (A.26) is of the order δ_k^2, and hence in order that (A.25) is satisfied, we need to choose $\delta_k \sim n^{-4/9} \asymp M_*^{-4}$. It follows immediately from (A.28) that (i) the eigenfunctions $\psi_1^{(j)}, \dots, \psi_r^{(j)}$ are orthonormal, and four times continuously differentiable. Also, since γ_l is centered around l/M_* with a support of the order $O(M_*^{-1})$, it follows that, for only finitely many $l \neq l'$, the support of $\gamma_{l'}$ overlaps with the support of γ_l. Moreover, if $\gamma_l^{(s)}$ denotes the s-th derivative of γ_l, then $\| \gamma_l^{(s)} \|_\infty = O(M_*^{1/2+s})$, for $s = 0, 1, \dots, 4$. Thus, the choice $\delta_k \asymp M_*^{-4}$ ensures that, (ii) for each $k = 1, \dots, r$, the fourth derivative of $\psi_k^{(j)}$ is bounded. Hence, by appropriate choice of the constants, we have that $\mathcal{C}_0 \subset \mathcal{C}$. Finally, arguing as in [22], with an application of the *Fano's Lemma* we conclude (A.22).

Part III : Proof of Proposition 5.1. Using standard arguments, it can be shown that, given $\epsilon > 0$ sufficiently small (but fixed), we have constants $0 < c_{1,\epsilon} < c_{2,\epsilon} < \infty$ such that for $\| \Gamma_*^{-1/2}(\Gamma_* - \Gamma)\Gamma_*^{-1/2} \|_F \leq \epsilon$,

$$\text{(A.29)} \quad c_{1,\epsilon} \| \Gamma_*^{-1/2}(\Gamma_* - \Gamma)\Gamma_*^{-1/2} \|_F^2 \leq K(\Gamma, \Gamma_*) \leq c_{2,\epsilon} \| \Gamma_*^{-1/2}(\Gamma_* - \Gamma)\Gamma_*^{-1/2} \|_F^2 .$$

Thus, it suffices to provide tight bounds for $\parallel \Gamma_*^{-1/2}(\Gamma - \Gamma_*)\Gamma_*^{-1/2} \parallel_F$. We introduce some notations first. Define, $G = \frac{\sigma^2}{s}\Lambda^{-1} + I_r$, $G_* = \frac{\sigma^2}{s}\Lambda_*^{-1} + I_r$ and $\Delta = B\Lambda B^T - B_*\Lambda_* B_*^T$. Then,

$$\Gamma^{-1} = \frac{1}{\sigma^2}(I_M - B(\frac{\sigma^2}{s}\Lambda^{-1} + I_r)^{-1}B^T) = \frac{1}{\sigma^2}(I_M - BG^{-1}B^T),$$

and

$$\Gamma_*^{-1} = \frac{1}{\sigma^2}(I_M - B_*G_*^{-1}B_*^T).$$

Moreover, due to **A1'**, there exist constants, $c_3, c_4 > 0$, such that,

$$c_3\left(\frac{\sigma^2}{s}\right) \le \sigma_{min}(I_r - G_*^{-1}) \le \sigma_{max}(I_r - G_*^{-1}) \le c_4\left(\frac{\sigma^2}{s}\right).$$

We express $I_M - B_*G_*^{-1}B_*^T$ as $(I_M - B_*B_*^T) + B_*(I_r - G_*^{-1})B_*^T$. Then we can express

$$(\sigma^2/s)^2 \parallel \Gamma_*^{-1/2}(\Gamma - \Gamma_*)\Gamma_*^{-1/2} \parallel_F^2 = \sigma^4 \parallel \Gamma_*^{-1/2}\Delta\Gamma_*^{-1/2} \parallel_F^2$$

as

$$\begin{aligned}
\text{(A.30)} \quad & \text{tr}\,[(I_M - B_*G_*^{-1}B_*^T)\Delta(I_M - B_*G_*^{-1}B_*^T)\Delta] \\
= \quad & \text{tr}\,[(I_M - B_*B_*^T)B\Lambda B^T(I_M - B_*B_*^T)B\Lambda B^T] \\
& +2\text{tr}\,[B_*(I_r - G_*^{-1})B_*^T B\Lambda B^T(I_M - B_*B_*^T)B\Lambda B^T] \\
& +\text{tr}\,[B_*(I_r - G_*^{-1})B_*^T\Delta B_*(I_r - G_*^{-1})B_*^T\Delta] \\
= \quad & \parallel (I_M - B_*B_*^T)B\Lambda B^T(I_M - B_*B_*^T) \parallel_F^2 \\
& +2\text{tr}\,[(I_r - G_*^{-1})^{1/2}B_*^T B\Lambda(B^T(I_M - B_*B_*^T)B)\Lambda B^T B_*(I_r - G_*^{-1})^{1/2}] \\
& + \parallel (I_r - G_*^{-1})^{1/2}(B_*^T B\Lambda B^T B_* - \Lambda_*)(I_r - G_*^{-1})^{1/2} \parallel_F^2 \\
\ge \quad & 2c_3\lambda_r^2(\sigma_{min}(B_*^T B))^2\frac{\sigma^2}{s} \parallel (I_M - B_*B_*^T)B \parallel_F^2 \\
& +c_3^2\left(\frac{\sigma^2}{s}\right)^2 \parallel B_*^T B\Lambda B^T B_* - \Lambda_* \parallel_F^2 \\
\ge \quad & c_4\left(\frac{\sigma^2}{s}\right) \parallel (I_M - B_*B_*^T)B \parallel_F^2 + c_3^2\left(\frac{\sigma^2}{s}\right)^2 \parallel B_*^T B\Lambda B^T B_* - \Lambda_* \parallel_F^2
\end{aligned}$$

for constants $c_3, c_4 > 0$. Now, since $(B, \Lambda) \in \widetilde{\Theta}(\alpha_n)$, where $B = \mathbf{exp}(1, B_*A_U + C_U)$ it follows that $\parallel A_U \parallel_F \le \alpha_n\sqrt{\frac{s}{\sigma^2}}$ and $\parallel C_U \parallel_F \le \alpha_n$. Moreover, from (4.2), and using the fact that $A_U = -A_U^T$, we have,

$$\begin{aligned}
& B_*^T B\Lambda B^T B_* - \Lambda_* \\
= \quad & D + (A_U\Lambda - \Lambda A_U) + O(\parallel A_U \parallel_F^2 + \parallel D \parallel_F^2 + \parallel U \parallel_F (\parallel A_U \parallel_F + \parallel C_U \parallel_F)).
\end{aligned}$$

Since $A_U\Lambda - \Lambda A_U$ is symmetric, has zeros on the diagonal, and its Frobenius norm is bounded below by $\min_{1\leq j<k\leq r}(\lambda_j - \lambda_k) \parallel A_U \parallel_F$, and D is diagonal, it follows that for some constant $c_6 > 0$,

$$\parallel B_*^T B\Lambda B^T B_* - \Lambda_* \parallel_F^2 \geq c_6(\parallel D \parallel_F^2 + \parallel A_U \parallel_F^2) - O((\frac{s}{\sigma^2})^{3/2}\alpha_n^3).$$

From this, and using (4.3) to approximate the first term in (A.30), it follows that for some constant $c_7 > 0$,

$$\begin{aligned}
&\parallel (I_M - B_* G_{*i}^{-1} B_*^T)^{1/2} \Delta (I_M - B_* G_{*i}^{-1} B_*^T)^{1/2} \parallel_F^2 \\
\geq\ & c_7 \left(\frac{\sigma^2}{s}\right)\left[\parallel C_U \parallel_F^2 + \frac{\sigma^2}{s} \parallel A_U \parallel_F^2 + \frac{\sigma^2}{s} \parallel D \parallel_F^2 - O(\sqrt{\frac{s}{\sigma^2}}\alpha_n^3)\right] \\
=\ & c_7 \left(\frac{\sigma^2}{s}\right)\alpha_n^2(1 - o(1)).
\end{aligned} \tag{A.31}$$

The last equality is because $\alpha_n\sqrt{\frac{s}{\sigma^2}} = o(1)$. Also, it is easy to show now that, for some $c_8 > 0$

$$\parallel (I_M - B_* G_*^{-1} B_*^T)^{1/2} \Delta (I_M - B_* G_*^{-1} B_*^T)^{1/2} \parallel_F^2 \leq c_8 \left(\frac{\sigma^2}{s}\right)\alpha_n^2(1 + o(1)). \tag{A.32}$$

Hence, from (A.31) and (A.32), it follows that, there are constants $c_9, c_{10} > 0$ such that, for sufficiently large n,

$$c_9\alpha_n\sqrt{\frac{s}{\sigma^2}} \leq \parallel \Gamma_*^{-1/2}(\Gamma - \Gamma_*)\Gamma_*^{-1/2} \parallel_F \leq c_{10}\alpha_n\sqrt{\frac{s}{\sigma^2}},$$

which, together with (A.29), proves (5.2).

Part IV : Score representation in the matrix case. First, define the *canonical metric* on the tangent space $\mathcal{T}_B \oplus \mathbb{R}^r$ of the parameter space $\tilde{\Omega} = S_{M,r} \otimes \mathbb{R}^r$ (for $\theta = (B, \zeta)$) by

$$\langle X, Y\rangle_g = \langle X_B, Y_B\rangle_c + \langle X_\zeta, Y_\zeta\rangle, \quad \text{for} \quad X_B, Y_B \in \mathcal{T}_B,\ X_\zeta, Y_\zeta \in \mathbb{R}^r,$$

where $\langle X_B, Y_B\rangle_c = \mathrm{tr}\,(X_B^T(I_M - \frac{1}{2}BB^T)Y_B)$ is the canonical metric on $\mathcal{S}_{M,r}$ and $\langle X_\zeta, Y_\zeta\rangle = \mathrm{tr}\,(X_\zeta^T Y_\zeta)$ is the usual Euclidean metric. Next, for an arbitrary θ, write $\widetilde{L}_n(\theta) = F_n^1(\theta) + F_n^2(\theta)$, where $F_n^1(\theta) = \mathrm{tr}\,(\Gamma^{-1}\widetilde{S})$ and $F_n^2(\theta) = \log|\Gamma| = \log|I_r + e^\zeta| = \log|I_r + \Lambda|$. Similarly, we write $L(\theta;\theta_*) = F^1(\theta;\theta_*) + F^2(\theta;\theta_*)$, where $F^1(\theta;\theta_*) = \mathrm{tr}\,(\Gamma^{-1}\Gamma_*)$ and $F^2(\theta;\theta_*) = F_n^2(\theta)$. Below, we shall only give expressions for gradient and Hessian of $F_n^1(\cdot)$ and $F_n^2(\cdot)$, since the gradient and Hessian of $F^1(\cdot;\theta_*)$ and $F^2(\cdot;\theta_*)$ follow from these (by replacing $\widetilde{S}$ with Γ_*).

Gradient and Hessian. From Appendices B and D of [23], we obtain expressions for the gradient and Hessian of $F_n^1(\cdot)$ and $F_n^2(\cdot)$. We mainly follow the notations used there. Let, $P := P(\theta) = I_M + B\Lambda B^T$. Then $P^{-1} = I_M - BQ^{-1}B^T$, where

$$Q := Q(\theta) = \Lambda^{-1} + B^T B = \Lambda^{-1} + I_r \implies Q^{-1} = \Lambda(I_r + \Lambda)^{-1}.$$

The fact that Q is independent of B is of importance in the calculations throughout. Use $F^1_{n,B}(\cdot)$ to denote the Euclidean gradient of $F^1_n(\cdot)$ w.r.t. B. It is easy to see that $F^1_{n,B}(\theta) = -2\widetilde{S}BQ^{-1}$. Then, under the canonical metric the intrinsic gradient is given by

$$\nabla_B F^1_n(\theta) = F^1_{n,B}(\theta) - B(F^1_{n,B}(\theta))^T B = 2[BQ^{-1}B^T\widetilde{S}B - \widetilde{S}BQ^{-1}].$$

Since $F^2_n(\theta)$ does not involve B, the Euclidean gradient $F^2_{n,B}(\theta) = 0$, and hence $\nabla_B F^2_n(\theta) = 0$. Therefore,

$$\nabla_B \widetilde{L}_n(\theta) = \nabla_B F^1_n(\theta) = 2[BQ^{-1}B^T\widetilde{S}B - \widetilde{S}BQ^{-1}]. \tag{A.33}$$

Next, for $X_B \in \mathcal{T}_B$, let $G^1_{n,BB}(\cdot)(X_B)$ be the Euclidean Hessian operator of $F^1_n(\cdot)$ evaluated at X_B. It is computed as

$$G^1_{n,BB}(\theta)(X_B) = -2\widetilde{S}X_BQ^{-1}.$$

The Hessian operator of $\widetilde{L}_n(\cdot)$ w.r.t. B, equals the Hessian operator of $F^1_n(\cdot)$ w.r.t. B. For $X_B, Y_B \in \mathcal{T}_B$, it is given by

$$\begin{aligned}(\text{A.34}) \quad & H_{n,B}(\theta)(X_B, Y_B) \\ = \; & \text{tr}\left(Y_B^T G^1_{n,BB}(\theta)(X_B)\right) + \frac{1}{2}\text{tr}\left[\left((F^1_{n,B}(\theta))^T X_B B^T + B^T X_B (F^1_{n,B}(\theta))^T\right) Y_B\right] \\ & -\frac{1}{2}\text{tr}\left[\left(B^T F^1_{n,B}(\theta) + (F^1_{n,B}(\theta))^T B\right) X_B^T (I_M - BB^T) Y_B\right].\end{aligned}$$

For computing gradient and Hessian with respect to ζ, we only need to compute first and second derivatives of the function $\widetilde{L}_n(\cdot)$ (equivalently, of $F^1_n(\cdot)$ and $F^2_n(\cdot)$). Using calculations carried out in Appendix D of [23], and the identity $P^{-1}B_k = (1 + \lambda_k)^{-1}B_k$ where B_k is the k-th column of B, $1 \le k \le r$, we have,

$$\begin{aligned}\frac{\partial F^1_n}{\partial \zeta_k}(\theta) &= -e^{\zeta_k} B_k^T P^{-1}\widetilde{S}P^{-1}B_k = -\frac{\lambda_k}{(1+\lambda_k)^2} B_k^T \widetilde{S} B_k, \\ \frac{\partial F^2_n}{\partial \zeta_k}(\theta) &= e^{\zeta_k} B_k^T P^{-1} B_k = \frac{\lambda_k}{1+\lambda_k}.\end{aligned}$$

Thus

$$\nabla_\zeta \widetilde{L}_n(\theta) = \text{diag}\left(\frac{\lambda_k}{(1+\lambda_k)^2}(1 + \lambda_k - B_k^T\widetilde{S}B_k)\right)_{k=1}^r.$$

Since $B_k^T P^{-1} B_l = 0$, for $1 \leq k \neq l \leq r$, it follows that $\frac{\partial^2 F_n^i}{\partial \zeta_k \partial \zeta_l}(\theta) = 0$ for $k \neq l$, $i = 1, 2$. Also,

$$\frac{\partial^2 F_n^1}{\partial \zeta_k^2}(\theta) = e^{\zeta_k} B_k^T P^{-1} \widetilde{S} P^{-1} B_k \left[2 e^{\zeta_k} (B_k^T P^{-1} B_k) - 1\right] = \frac{\lambda_k(\lambda_k - 1)}{(1+\lambda_k)^3} B_k^T \widetilde{S} B_k,$$

$$\frac{\partial^2 F_n^2}{\partial \zeta_k^2}(\theta) = e^{\zeta_k} B_k^T P^{-1} B_k \left[1 - e^{\zeta_k} (B_k^T P^{-1} B_k)\right] = \frac{\lambda_k}{(1+\lambda_k)^2}.$$

Thus, the Hessian operator of $\widetilde{L}_n(\cdot)$ w.r.t. ζ is given by

$$H_{n,\zeta}(\theta) = \mathrm{diag}\left(\frac{\lambda_k}{(1+\lambda_k)^3}\left((\lambda_k - 1) B_k^T \widetilde{S} B_k + (1+\lambda_k)\right)\right)_{k=1}^r. \tag{A.35}$$

Boundedness and inversion of $H(\theta_*;\theta_*)$. As discussed in Section 6, the Hessian operator $H(\theta_*;\theta_*)$ is "block diagonal". So, we only need to show the boundedness and calculate the inverse of $H_B(\theta_*;\theta_*)$ and $H_\zeta(\theta_*;\theta_*)$. First note that, from (A.35) we have,

$$\begin{aligned} H_\zeta(\theta_*;\theta_*) &= \mathrm{diag}\left(\frac{\lambda_{*k}}{(1+\lambda_{*k})^3}\left((\lambda_{*k} - 1) B_{*k}^T \Gamma_* B_{*k} + (1+\lambda_{*k})\right)\right)_{k=1}^r \\ &= \Lambda_*^2 (I_r + \Lambda_*)^{-2}, \end{aligned}$$

which is clearly positive definite with eigenvalues bounded away from 0 and ∞, due to conditions **A1'** and **C"**.

Next, we show that $H_B(\theta_*;\theta_*)(X,X) \geq C\langle X, X\rangle_c$, for some $C > 0$, for all $X \in \mathcal{T}_{B_*}$. Define $F_B^1(\theta_*;\theta_*) = \mathbb{E}_{\theta_*} F_{n,B}^1(\theta_*)$ and $G_{BB}^1(\theta_*;\theta_*) = \mathbb{E}_{\theta_*} G_{n,BB}^1(\theta_*)$. Note that $H_B(\theta_*;\theta_*)$ is obtained by replacing $F_{n,B}^1(\theta_*)$ and $G_{n,BB}^1(\theta_*)$ by $F_B^1(\theta_*)$ and $G_{BB}^1(\theta_*;\theta_*)$, respectively, in (A.34). Observe that,

$$F_B^1(\theta_*) = -2\Gamma_* B_* Q_*^{-1} = -2 B_*(I_r + \Lambda_*) Q_*^{-1} = -2 B_* \Lambda_*,$$

where $Q_* = Q(\theta_*) = \Lambda_*^{-1}(I_r + \Lambda_*)$, and we have used the fact that $\Gamma_* B_* = B_*(I_r + \Lambda_*)$. For notational simplicity we use F_B^1 to denote $F_B^1(\theta_*)$. Note that, for $X \in \mathcal{T}_{B_*}$, $X = B_* A_X + (I - B_* B_*^T) C_X$, where $A_X = -A_X^T \in \mathbb{R}^{r\times r}$, and $C_X \in \mathbb{R}^{M\times r}$. Using this representation, for any $X, Y \in \mathcal{T}_{B_*}$, we have

$$\begin{aligned} &\frac{1}{2}\mathrm{tr}\left[\left((F_B^1)^T X B_*^T + B_*^T X (F_B^1)^T\right) Y\right] \\ = \; & -\mathrm{tr}\left[\Lambda_* B_*^T X B_*^T Y + B_*^T X \Lambda_* B_*^T Y\right] \\ = \; & \mathrm{tr}\left[\Lambda_* X^T B_* B_*^T Y + B_*^T X \Lambda_* Y^T B_*\right] \;=\; 2\mathrm{tr}\left[\Lambda_* X^T B_* B_*^T Y\right], \end{aligned} \tag{A.36}$$

and

$$\begin{aligned} &-\frac{1}{2}\mathrm{tr}\left[\left(B_*^T (F_B^1) + (F_B^1)^T B_*\right) X^T \left(I_M - B_* B_*^T\right) Y\right] \\ = \; & \mathrm{tr}\left[(B_*^T B_* \Lambda_* + \Lambda_* B_*^T B_*) X^T (I_M - B_* B_*^T) Y\right] \\ = \; & 2\mathrm{tr}\left[\Lambda_* X^T (I_M - B_* B_*^T) Y\right]. \end{aligned} \tag{A.37}$$

Next, notice that, for $X \in \mathcal{T}_{B_*}$, $G^1_{BB}(\theta_*;\theta_*)(X) = -2\Gamma_* X Q_*^{-1}$. Therefore,

$$\begin{aligned}(A.38) \qquad \mathrm{tr}\left[Y^T G^1_{BB}(\theta_*;\theta_*)(X)\right] &= -2\mathrm{tr}\left[Y^T \Gamma_* X Q_*^{-1}\right] \\ = -2\mathrm{tr}\left[Y^T(I_M - B_*B_*^T)XQ_*^{-1}\right] &- 2\mathrm{tr}\left[Y^T B_*(I_r+\Lambda)B_*^T X Q_*^{-1}\right].\end{aligned}$$

Now, combining (A.36), (A.37) and (A.38), and using the definition of $H_B(\theta_*;\theta_*)$, and the facts that $I_r - Q_*^{-1} = (I_r+\Lambda_*)^{-1}$, $X^T B_* = A_X^T = -A_X$ and $B_*^T Y = A_Y = -A_Y^T$, after some simple algebra we have,

$$\begin{aligned}(A.39) \qquad & H_B(\theta_*;\theta_*)(X,Y) \\ = \;& 2\left(\mathrm{tr}\left[X^T B_*\Lambda_* B_*^T Y(I_r+\Lambda_*)^{-1}\right] - \mathrm{tr}\left[X^T B_* B_*^T Y Q_*^{-1}\right]\right) \\ & +2\mathrm{tr}\left[\Lambda_*^2(I_r+\Lambda_*)^{-1}X^T(I_r - B_*B_*^T)Y\right]\end{aligned}$$

Again, since $A_X = -A_X^T$ and $A_Y = -A_Y^T$, denoting by $A_{X,ij}$ and $A_{Y,ij}$, the (i,j)-th element of A_X and A_Y respectively, we have

$$\begin{aligned}(A.40) \qquad & \mathrm{tr}\left[X^T B_*\Lambda_* B_*^T Y(I+\Lambda_*)^{-1}\right] - \mathrm{tr}\left[X^T B_* B_*^T Y Q_*^{-1}\right] \\ = \;& -\mathrm{tr}\left[A_X\left(\Lambda_* A_Y(I_r+\Lambda_*) - A_Y\Lambda_*(I_r+\Lambda_*)^{-1}\right)\right] \\ = \;& -\sum_{i=1}^r\sum_{j=1}^r A_{X,ij}\left(\frac{\lambda_{*j}}{1+\lambda_{*i}}A_{Y,ji} - \frac{\lambda_{*i}}{1+\lambda_{*i}}A_{Y,ji}\right) \\ = \;& \sum_{i=1}^r\sum_{j=1}^r A_{X,ij}A_{Y,ij}\left(\frac{\lambda_{*j}-\lambda_{*i}}{1+\lambda_{*i}}\right) \\ = \;& \sum_{i=1}^{j-1}\sum_{j=i+1}^r A_{X,ij}A_{Y,ij}\left[(\lambda_{*i}-\lambda_{*j})\left(\frac{1}{1+\lambda_{*j}} - \frac{1}{1+\lambda_{*i}}\right)\right] \\ = \;& \sum_{i=1}^{j-1}\sum_{j=i+1}^r A_{X,ij}A_{Y,ij}\frac{(\lambda_{*i}-\lambda_{*j})^2}{(1+\lambda_{*i})(1+\lambda_{*j})} \\ = \;& \frac{1}{2}\sum_{i=1}^r\sum_{j=1}^r A_{X,ij}A_{Y,ij}\frac{(\lambda_{*i}-\lambda_{*j})^2}{(1+\lambda_{*i})(1+\lambda_{*j})}.\end{aligned}$$

Since $\min_{1\le k\ne k'\le r}(\lambda_{*k}-\lambda_{*k'})^2(1+\lambda_{*k})^{-2} \ge C_{1*}$, and $\lambda_{*r} \ge C_{*2}$, for some constants $C_{1*}, C_{*2} > 0$ (value depending on $\bar{c}_1$ and $\bar{c}_2$ appearing in **A1'**), it follows from (A.39) and (A.40) that for $X \in \mathcal{T}_{B_*}$,

$$\begin{aligned}H_B(\theta_*;\theta_*)(X,X) &\ge C_{*1}\mathrm{tr}\left(X^T B_* B_*^T X\right) + 2C_{*2}\mathrm{tr}\left(X^T(I_M - B_*B_*^T)X\right) \\ &\ge C_{*3}\mathrm{tr}\left(X^T X\right),\end{aligned}$$

where $C_{*3} = \min\{C_{*1}, 2C_{*2}\}$. This proves that $H_B(\theta_*;\theta_*)(X,X)$ is bounded below in the Euclidean norm and hence in the canonical metric because of the norm equivalence. An upper bound follows similarly.

Proof of Corollary 3.1. From (A.39) and (A.40), we can derive an explicit expression of $H_B^{-1}(\theta_*;\theta_*)$. Note that $H_B^{-1}(\theta_*;\theta_*)(X)$ is defined as

$$H_B\left(\theta_*;\theta_*)(H_B^{-1}(\theta_*;\theta_*)(X),Y\right) = \langle X,Y\rangle_c, \quad \text{for any } Y \in \mathcal{T}_{B_*}.$$

Therefore, for $X \in \mathcal{T}_{B*}$ and $A_X = B_*^T X$,

$$\begin{aligned} & H_B^{-1}(\theta_*;\theta_*)(X) \\ = \quad & \frac{1}{2} B_* \left(\left(\frac{(1+\lambda_{*i})(1+\lambda_{*j})}{(\lambda_{*i}-\lambda_{*j})^2} A_{X,ij}\right)\right) + \frac{1}{2}(I_M - B_* B_*^T) X \Lambda_*^{-2}(I_r + \Lambda_*). \end{aligned}$$

Using this, we can now get an explicit expression for $H_B^{-1}(\theta_*;\theta_*)(\nabla_B \widetilde{L}_n(\theta_*))$. From (A.33), we have

$$\begin{aligned} B_*^T \nabla_B \widetilde{L}_n(\theta_*) &= 2\left[Q_*^{-1} B_*^T \widetilde{S} B_* - B_*^T \widetilde{S} B_* Q_*^{-1}\right] \\ &= 2\left(\left(B_{*i}^T \widetilde{S} B_{*j}\left(\frac{\lambda_{*i}}{1+\lambda_{*i}} - \frac{\lambda_{*j}}{1+\lambda_{*j}}\right)\right)\right) \\ &= 2\left(\left(\frac{(\lambda_{*i}-\lambda_{*j})}{(1+\lambda_{*i})(1+\lambda_{*j})} B_{*i}^T \widetilde{S} B_{*j}\right)\right). \end{aligned}$$

Also,

$$(I_M - B_* B_*^T)\nabla_B \widetilde{L}_n(B_*,\Lambda_*) = -2(I_M - B_* B_*^T)\widetilde{S} B_* Q_*^{-1}.$$

Thus, it follows that

$$\begin{aligned} & -H_B^{-1}(\theta_*;\theta_*)\left(\nabla_B \widetilde{L}_n(\theta_*)\right) \\ = \quad & -\left(\left(\frac{1}{(\lambda_{*i}-\lambda_{*j})} B_{*i}^T \widetilde{S} B_{*j}\right)\right) + (I_M - B_* B_*^T)\widetilde{S} B_* \Lambda_*^{-1} \\ = \quad & -\left[\mathbf{R}_1 \widetilde{S} B_{*1} : \cdots : \mathbf{R}_r \widetilde{S} B_{*r}\right]. \end{aligned}$$

Part V : Gradient and Hessian on product manifolds. In this section, we give a brief outline of the intrinsic geometry associated with the product manifold of two Riemannian manifolds, and as an application we consider the manifold $\mathcal{S}_{M,r} \otimes \mathbb{R}^r$, which is the parameter space for (B,ζ) in our problem.

Product Manifolds. Consider two Riemannian manifolds: $\mathcal{M},\mathcal{N}$ with metrics g_M and g_N, respectively. The product manifold $\mathcal{P}$ of $\mathcal{M},\mathcal{N}$ is then defined as:

$$\mathcal{P} := \mathcal{M} \otimes \mathcal{N} = \{(x,y) : x \in \mathcal{M}, y \in \mathcal{N}\}$$

with the tangent space at a point $p = (x,y) \in \mathcal{P}$,

$$\mathcal{T}_p\mathcal{P} := \mathcal{T}_x\mathcal{M} \oplus \mathcal{T}_y\mathcal{N}$$

where $\mathcal{T}_x\mathcal{M}, \mathcal{T}_y\mathcal{N}$ are tangent spaces of $\mathcal{M}, \mathcal{N}$ at points x, y, respectively. The Riemannian metric g on the tangent space $\mathcal{TP}$ is naturally defined as

$$\langle T_1, T_2\rangle_g := \langle \xi_1, \xi_2\rangle_{g_M} + \langle \eta_1, \eta_2\rangle_{g_N},$$

where $T_i = (\xi_i, \eta_i) \in \mathcal{TP}$, with $\xi_i \in \mathcal{TM}$ and $\eta_i \in \mathcal{TN}$ $(i = 1, 2)$.

By the above definition of the product manifold $\mathcal{P}$, the intrinsic gradient and Hessian of a smooth function f defined on $\mathcal{P}$ are as follows:

- Gradient:
$$\nabla f = (\nabla_{\mathcal{M}} f_{\mathcal{M}}, \nabla_{\mathcal{N}} f_{\mathcal{N}}),$$
where $f_{\mathcal{M}}$ ($f_{\mathcal{N}}$) is f viewed as a function on $\mathcal{M}$ ($\mathcal{N}$); and $\nabla_{\mathcal{M}}$ ($\nabla_{\mathcal{N}}$) denotes the gradient operator for functions defined on $\mathcal{M}$ ($\mathcal{N}$).
- Hessian: for $T_i = (\xi_i, \eta_i) \in \mathcal{TP}$ $(i = 1, 2)$,
$$\begin{aligned} H_f(T_1, T_2) &= H_{f_{\mathcal{M}}}(\xi_1, \xi_2) + \langle \nabla_{\mathcal{N}} \langle \nabla_{\mathcal{M}} f_{\mathcal{M}}, \xi_1\rangle_{g_M}, \eta_2\rangle_{g_N} \\ &+ \langle \nabla_{\mathcal{M}} \langle \nabla_{\mathcal{N}} f_{\mathcal{N}}, \eta_1\rangle_{g_N}, \xi_2\rangle_{g_M} + H_{f_{\mathcal{N}}}(\eta_1, \eta_2). \end{aligned}$$
The above expression is derived from the bi-linearity of the Hessian operator and its definition. Also note that
$$\langle \nabla_{\mathcal{N}} \langle \nabla_{\mathcal{M}} f_{\mathcal{M}}, \xi_1\rangle_{g_M}, \eta_2\rangle_{g_N} = \langle \nabla_{\mathcal{M}} \langle \nabla_{\mathcal{N}} f_{\mathcal{N}}, \eta_2\rangle_{g_N}, \xi_1\rangle_{g_M}.$$

Application to the product of a Stiefel manifold and an Euclidean space. Consider the special case: $\mathcal{M} = \mathcal{S}_{M,r}$ with the canonical metric $\langle \cdot, \cdot\rangle_c$, and $\mathcal{N} = \mathbb{R}^d$ with Euclidean metric. For a point $p = (B, x)$ on the product manifold $\mathcal{P}$, the tangent space is

$$\mathcal{T}_p\mathcal{P} = \mathcal{T}_B\mathcal{M} \oplus \mathcal{T}_x\mathcal{N},$$

where

$$\mathcal{T}_B\mathcal{M} = \{\Delta \in \mathbb{R}^{M\times r} : B^T\Delta = -\Delta^T B\}, \qquad \text{and} \qquad \mathcal{T}_x\mathcal{N} = \mathbb{R}^d.$$

For a smooth function f defined on the product space $\mathcal{P}$:

- Gradient (at p):
$$\nabla f|_p = \left(\nabla_{\mathcal{M}} f, \frac{\partial f}{\partial x}\right)\Big|_p,$$
where $\nabla_{\mathcal{M}} f\,|_p = f_B - B f_B^T B$ (with $f_B = \frac{\partial f}{\partial B}$).
- Hessian operator (at p): for $T = (\Delta, a)$, and for$X = (X_B, \eta) \in \mathcal{T}_p\mathcal{P}$,
$$\begin{aligned} &H_f(T, X)|_p \\ &= H_{f_{\mathcal{M}}}(\Delta, X_B)|_p + \langle \frac{\partial}{\partial x}\langle \nabla_{\mathcal{M}} f, \Delta\rangle_c, \eta\rangle + \langle \frac{\partial}{\partial x}\langle \nabla_{\mathcal{M}} f, X_B\rangle_c, a\rangle + a^T \frac{\partial^2 f}{\partial x^2}\eta, \end{aligned} \tag{A.41}$$

where

$$\begin{aligned} & H_{f_{\mathcal{M}}}(\mathit{\Delta}, X_B)\,|_p \\ = \;& f_{BB}(\mathit{\Delta}, X_B) + \frac{1}{2} Tr\left[(f_B^T \mathit{\Delta} B^T + B^T \mathit{\Delta} f_B^T) X_B \right] \\ & - \frac{1}{2} Tr\left[(B^T f_B + f_B^T B) \mathit{\Delta}^T \mathit{\Pi} X_B \right], \end{aligned}$$

with $\mathit{\Pi} = I - BB^T$.

- Inverse of Hessian operator (at p): for $G \in \mathcal{T}_p\mathcal{P}$, $T = H_f^{-1}(G)|_p$ is defined as: $T = (\Delta, a) \in \mathcal{T}_p\mathcal{P}$ such that for any $X = (X_B, \eta) \in \mathcal{T}_p\mathcal{P}$ the following equation is satisfied

$$H_f(T, X)|_p = \langle G, X \rangle_g \quad .$$

Part VI : Some inequalities involving matrices. In this paper we make frequent use of the following matrix inequalities:

- For any A, B, and with $\lambda_{min}(B)$ denoting the smallest eigenvalue of B

$$\| AB \|_F \leq \| A \|_F \| B \|, \quad \text{and} \quad \| AB \|_F \geq \| A \|_F \lambda_{min}(B),$$

where the last inequality holds for B positive definite. Also, if A and B are invertible then

$$A^{-1} - B^{-1} = A^{-1}(B - A)B^{-1} = B^{-1}(B - A)A^{-1}. \tag{A.42}$$

- *Weilandt's inequality ([16]):* For symmetric $p \times p$ matrices A, B with eigenvalue sequences $\lambda_1(A) \geq \cdots \geq \lambda_p(A)$ and $\lambda_1(B) \geq \cdots \geq \lambda_p(B)$, respectively,

$$\sum_{i=1}^{p} |\lambda_i(A) - \lambda_i(B)|^2 \leq \| A - B \|_F^2 \tag{A.43}$$

- *Eigenvector perturbation ([21]):* Let A be a $p \times p$ positive semidefinite matrix, with j-th largest eigenvalue $\lambda_j(A)$ with corresponding eigenvector $\mathbf{p}_j$, and $\tau_j := \max\{(\lambda_{j-1}(A) - \lambda_j(A))^{-1}, (\lambda_j(A) - \lambda_{j+1}(A))^{-1}\}$ is bounded (we take $\lambda_0(A) = \infty$ and $\lambda_{p+1}(A) = 0$). Let B be a symmetric matrix. If $\mathbf{q}_j$ denotes the eigenvector of $A + B$ corresponding to the j-th largest eigenvalue (which is of multiplicity 1, for $\| B \|$ small enough, by (A.43)), then (assuming without loss of generality $\mathbf{q}_j^T \mathbf{p}_j > 0$),

$$\| \mathbf{q}_j - \mathbf{p}_j \| \leq 5 \, \frac{\| B \|}{\tau_j} + 4 \left(\frac{\| B \|}{\tau_j} \right)^2 . \tag{A.44}$$

ACKNOWLEDGEMENTS

We thank two anonymous reviewers for their comments which helped in improving the presentation of the material.

REFERENCES

[1] Antoniadis, A. and Sapatinas, T. (2007). Estimation and inference in functional mixed-effects models. *Computational Statistics and Data Analysis* **51**, 4793–4813.

[2] Bickel, P. J. and Levina, E. (2007). Covariance regularization by thresholding. *Technical report #744*, Department of Statistics, UC Berkeley. *Annals of Statistics.* To appear.

[3] Bickel, P. J. and Levina, E. (2008). Regularized estimation of large covariance matrices. *Annals of Statistics* **36**, 199–227.

[4] Burman, P. (1985). A data dependent approach to density estimation. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete* **69**, 609–628.

[5] Chen, A. and Bickel, P. J. (2006). Efficient independent component analysis. *Annals of Statistics* **34**, 2825–2855.

[6] de Boor, C. (1974). Bounding the error in spline interpolation. *SIAM Review* **16**, 531–544.

[7] de Boor, C. (1978). *A Practical Guide to Splines.* Springer–Verlag.

[8] deVore, R. A. and Lorentz, G. G. (1993). *Constructive Approximation.* Springer.

[9] El Karoui, N. (2007). Operator norm consistent estimation of large dimensional sparse covariance matrices. *Annals of Statistics.* To appear.

[10] Davidson, K. R. and Szarek, S. (2001). Local operator theory, random matrices and Banach spaces. In *"Handbook on the Geometry of Banach spaces"*, (**V. 1**, Johnson, W. B., Lendenstrauss, J. eds.), 317–366, Elsevier Science.

[11] Edelman, A., Arias, T. A. and Smith, S. T. (1998). The geometry of algorithms with orthogonality constraints, *SIAM Journal on Matrix Analysis and Applications* **20**, 303–353.

[12] Fan, J., Fan, Y. and Lv., J. (2006). High dimensional covariance matrix estimation using a factor model. *Journal of Econometrics.* To appear.

[13] Ferraty, F. and Vieu, P. (2006). *Nonparametric Functional Data Analysis : Theory and Practice.* Springer.

[14] Hall, P. and Hosseini-Nasab, M. (2006). On properties of functional principal components analysis. *Journal of the Royal Statistical Society, Series B* **68**, 109–126.

[15] Hall, P., Müller, H.-G. and Wang, J.-L. (2006). Properties of principal component methods for functional and longitudinal data analysis. *Annals of Statistics* **34**, 1493–1517.

[16] Horn, R. A. and Johnson, C. R. (1994). *Topics in Matrix Analysis.* Cambridge University Press.

[17] James, G. M., Hastie, T. J. and Sugar, C. A. (2000). Principal component models for sparse functional data. *Biometrika*, **87**, 587–602.

[18] Kato, T. (1980). *Perturbation Theory of Linear Operators.* Springer–Verlag.

[19] Muirhead, R. J. (1982). *Aspects of Multivariate Statistical Theory*, John Wiley & Sons.

[20] Oller, J. M. and Corcuera, J. M. (1995). Intrinsic analysis of statistical estimation. *Annals of Statistics* **23**, 1562–1581.

[21] Paul, D. (2005). *Nonparametric Estimation of Principal Components.* Ph. D. Thesis. Stanford University.

[22] PAUL, D. and JOHNSTONE, I. M. (2007). Augmented sparse principal component analysis for high dimensional data. *Working Paper*. (http://anson.ucdavis.edu/∼debashis/techrep/augmented-spca.pdf)

[23] PENG, J. and PAUL, D. (2007). A geometric approach to maximum likelihood estimation of covariance kernel from sparse irregular longitudinal data. *Technical Report*. arXiv:0710.5343v1 [stat.ME]. (also at http://anson.ucdavis.edu/∼jie/pd-cov-likelihood-technical.pdf)

[24] RAMSAY, J. and SILVERMAN, B. W. (2005). *Functional Data Analysis, 2nd Edition*. Springer.

[25] YANG, A. and BARRON, A. (1999). Information-theoretic determination of minimax rates of convergence. *Annals of Statistics* **27**, 1564–1599.

[26] YAO, F., MÜLLER, H.-G. and WANG, J.-L. (2005). Functional data analysis for sparse longitudinal data. *Journal of the American Statistical Association* **100**, 577–590.

DEPARTMENT OF STATISTICS
UNIVERSITY OF CALIFORNIA
DAVIS, CA 95616
E-MAIL: debashis@wald.ucdavis.edu
jie@wald.ucdavis.edu