A geometric approach to maximum likelihood estimation of the functional principal components from sparse longitudinal data

Jie Peng* & Debashis Paul

Department of Statistics, University of California, Davis, CA 95616

* Correspondence author: jie@wald.ucdavis.edu

Abstract

In this paper, we consider the problem of estimating the eigenvalues and eigenfunctions of the covariance kernel (i.e., the *functional principal components*) from sparse and irregularly observed longitudinal data. We exploit the smoothness of the eigenfunctions to reduce dimensionality by restricting them to a lower dimensional space of smooth functions. We then approach this problem through a restricted maximum likelihood method. The estimation scheme is based on a Newton-Raphson procedure on the *Stiefel manifold* using the fact that the basis coefficient matrix for representing the eigenfunctions has orthonormal columns. We also address the selection of the number of basis functions, as well as that of the dimension of the covariance kernel by a second order approximation to the leave-one-curve-out cross-validation score that is computationally very efficient. The effectiveness of our procedure is demonstrated by simulation studies and an application to a CD4+ counts data set. In the simulation studies, our method performs well on both estimation and model selection. It also outperforms two existing approaches: one based on a local polynomial smoothing, and another using an EM algorithm.

Keywords : longitudinal data, covariance kernel, functional principal components, Stiefel manifold, Newton-Raphson algorithm, cross-validation.

1 Introduction

In recent years there has been growing interest in the analysis of high dimensional data. One particular subclass consists of data that may be considered as functional data, where the measurements per subject, or replicate, are taken on a finite interval. For data arising in fields such as longitudinal data analysis, chemometrics, econometrics, the functional data analysis viewpoint is becoming increasingly popular (Ferraty and Vieu, 2006). Depending on how the individual curves are measured, one can think of two different scenarios - (i) when the individual 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 levels of concentration or intensity of some substance.

The main goal of this paper is the estimation of the functional principal components of the covariance kernel from sparse, irregularly, observed functional data (scenario (ii)). The eigenfunctions give a nice basis for representing functional data, and hence are very useful in problems related to model building and prediction (see e.g. Cardot, Ferraty and Sarda, 1999, Hall and Horowitz, 2007, Cai and Hall, 2006). Ramsay and Silverman (2005) and Ferraty and Vieu (2006) give an extensive survey of the applications of *functional principal components analysis* (FPCA). Covariance is a positive semidefinite operator. Thus, from statistical as well as aesthetic point of view, it is important that its estimator is also positive semidefinite. In this paper, we shall adopt a *restricted maximum likelihood* framework to obtain a positive semidefinite estimator. In particular, we propose a new computational procedure which utilizes optimization tools recently developed for special manifolds.

We now describe the model for the sparse functional data. Suppose that we observe n independent realizations of an L^2-stochastic process $\{X(t) : t \in [0,1]\}$ at a sequence of points on the interval $[0,1]$, with additive measurement noise. That is, the observed data $\{Y_{ij} : 1 \leq j \leq m_i; 1 \leq i \leq n\}$ can be modeled as:

$$Y_{ij} = X_i(T_{ij}) + \sigma\varepsilon_{ij}, \tag{1}$$

where $\{\varepsilon_{ij}\}$ are i.i.d. with mean 0 and variance 1. Since $X(t)$ is an L^2-stochastic process, by Mercer's theorem (Ash, 1972) there exists a positive semi-definite kernel $C(\cdot,\cdot)$ such that $Cov(X(s), X(t)) =$

$C(s,t)$, and the following a.s. representation holds:

$$X_i(t) = \mu(t) + \sum_{\nu=1}^{\infty} \sqrt{\lambda_\nu}\psi_\nu(t)\xi_{i\nu}, \tag{2}$$

where $\mu(\cdot) = \mathbb{E}(X(\cdot))$ is the mean function; $\lambda_1 \geq \lambda_2 \geq \ldots \geq 0$ are the eigenvalues of $C(\cdot,\cdot)$; $\psi_\nu(\cdot)$ are the corresponding orthonormal eigenfunctions; and the random variables $\{\xi_{i\nu} : \nu \geq 1\}$, for each i, are uncorrelated with zero mean and unit variance. In the observed data model (1), we assume that $\{T_{ij} : j = 1, \ldots, m_i\}$ are randomly sampled from a continuous distribution. In the problems we are interested in, the number of measurements m_i is typically small.

Our estimation procedure is based on the assumption that the covariance kernel C is of finite rank, say r; and the eigenfunctions $\{\psi_\nu\}_{\nu=1}^{r}$ can be closely represented in a known, finite, basis of smooth functions. This results in an orthogonality constraint on the matrix of basis coefficients, say B, as described in Section 2. Specifically, the matrix B lies in a *Stiefel manifold*, that is the space of real valued matrices (of a fixed dimension) with orthonormal columns. Our estimation procedure involves maximization of the log-likelihood under the working assumption of normality, satisfying the orthonormality constraint on B. To implement this, we employ a Newton-Raphson algorithm for optimization on a *Stiefel manifold*, that utilizes its intrinsic Riemannian geometric structure. Moreover, by carefully treating matrix inversions, the resulting estimation procedure is able to efficiently handle different regimes of sparsity of data. The geometric viewpoint has further important implications. Devising a computationally practical and effective model selection procedure is important for such a semi-parametric problem, yet it still remains a challenge (cf. Marron *et al.*, 2004, p. 620). In this paper, we utilize the geometry of the parameter space to derive a second order approximation of the CV score that is computationally very efficient.

Before ending this section, we give a brief overview of the existing literature on FPCA. One approach to FPCA is through a basis representation framework (Section 2). Linear estimation procedures within this framework have been studied by various authors including Cardot (2000), Rice and Wu (2001), and Besse, Cardot and Ferraty (1997). Among non-linear procedures within this framework, James, Hastie and Sugar (2000) propose an EM algorithm to maximize the restricted likelihood (Section 2). However, this EM algorithm results in an estimator not necessarily satisfying the orthonormality constraints, and thus is presumably less efficient than the restricted MLE proposed here. Another approach to FPCA is through kernel smoothing of the data (Boente and Fraiman, 2000). Yao, Müller and Wang (2005) propose to estimate the covariance by local poly-

nomial smoothing. In spite of its nice asymptotic properties (Hall, Müller and Wang, 2006), this method does not ensure a positive semi-definite estimate of the covariance kernel. Moreover, it sometimes results in a negative estimate of the error variance σ^2. Simulation studies presented in Section 5 indicate a significant improvement of the proposed method over both EM and local polynomial approaches, as well as a satisfactory performance in model selection based on the approximate CV score.

The rest of the paper is organized as follows. In Section 2, we describe the restricted maximum likelihood framework. In Section 3, we describe the Newton-Raphson algorithm for finding the restricted maximum likelihood estimator. In Section 4, we derive an approximation to the leave-one-curve-out cross-validation score. Section 5 is devoted to simulation studies. In Section 6, the proposed procedure is applied to a CD4+ counts data set. In Section 7, we outline directions for future research. Technical details are given in the supplementary material (Appendices A–D).

2 Restricted MLE framework

We first describe the basis representation framework. Under some weak conditions on the stochastic processes (like L^2-differentiability of certain order, see, e.g. Ash, 1972), the eigenfunctions have some degree of smoothness. This assumption has been used in various studies, including Boente and Fraiman (2000), Cardot (2000), James *et al.* (2000), Yao *et al.* (2005, 2006), and Hall *et al.* (2006). Smoothness of the eigenfunctions means that they can be well-approximated in some *stable basis* for smooth function classes, e.g. the B-spline basis (Chui, 1987). If in addition, in model (2), we assume that $\lambda_\nu = 0$ for $\nu > r$, for some $r \geq 1$, then we can choose a finite set of linearly independent, L^2 functions $\{\phi_1(\cdot), \ldots, \phi_M(\cdot)\}$ with $M \geq r$, such that eigenfunctions can be modeled as $\psi_\nu(\cdot) = \sum_{k=1}^{M} b_{k\nu}\phi_k(\cdot)$ for $\nu = 1, \ldots, r$. Then, for every t,

$$\boldsymbol{\psi}(t)^T := (\psi_1(t), \ldots, \psi_r(t)) = (\phi_1(t), \ldots, \phi_M(t))B \tag{3}$$

for an $M \times r$ matrix $B = ((b_{k\nu}))$ that satisfies the constraint

$$B^T(\int \boldsymbol{\phi}(t)\boldsymbol{\phi}(t)^T dt)B = \int \boldsymbol{\psi}(t)\boldsymbol{\psi}(t)^T dt = I_r, \tag{4}$$

where $\boldsymbol{\phi}(\cdot) = (\phi_1(\cdot), \ldots, \phi_M(\cdot))^T$. Since the $M \times M$ matrix $\int \boldsymbol{\phi}(t)\boldsymbol{\phi}(t)^T dt$ is known and nonsingular, without loss of generality, hereafter we assume $B^T B = I_r$, by orthonormalizing $\{\phi_1(\cdot), \ldots, \phi_M(\cdot)\}$.

Here, we are assuming a reduced rank model for the covariance kernel as in James *et al.* (2000). The model can be motivated as follows. Suppose that the covariance kernel $C(s,t)$ of the underlying process has the infinite Karhunen-Loéve expansion:

$$C(s,t) = \sum_{k=1}^{\infty} \lambda_k \psi_k(s)\psi_k(t), \tag{5}$$

where $\lambda_1 \geq \lambda_2 \geq \cdots \geq 0$, $\sum_{k=1}^{\infty} \lambda_k < \infty$, and $\{\psi_k\}_{k=1}^{\infty}$ forms a complete orthonormal basis for $L^2[0,1]$. The condition $\sum_{k=1}^{\infty} \lambda_k < \infty$ implies that $\lambda_k \to 0$ as $k \to \infty$. Also, the orthonormality of the eigenfunctions $\{\psi_k\}$ implies that ψ_k typically gets more and more "wiggly" as k increases. One can truncate the series on the right hand side of (5) at some finite $r \geq 1$ to get the *projected covariance kernel*

$$C^r_{proj}(s,t) = \sum_{k=1}^{r} \lambda_k \psi_k(s)\psi_k(t). \tag{6}$$

Note that $\| C - C^r_{proj} \|^2_F = \sum_{k=r+1}^{\infty} \lambda_k^2$. Thus, as long as the eigenvalues decay to zero efficiently fast, even with a relatively small r, the approximation C^r_{proj} only results in a small bias. This motivates us to use a finite rank model as described above. Furthermore, the reduced rank model helps in modeling the eigenfunctions as well, since the eigenfunctions corresponding to larger eigenvalues usually require less complex basis to approximate them well.

Finally, it is worth pointing out some advantages of the reduced rank formulation over the mixed effects model by Rice and Wu (2000). Notice that, in the unconstrained mixed effects model one needs to model the whole covariance kernel. When the observations are sparse, this could lead to an over-parametrization, and it will result in highly variable estimates. Furthermore, if one uses a maximum likelihood approach, the over-parametrization would cause a very rough likelihood surface, with multiple local maxima. Therefore, restricting the rank of the covariance kernel can also be viewed as a form of regularization of the likelihood.

If one assumes Gaussianity of the processes, i.e., $\xi_{i\nu} \overset{i.i.d.}{\sim} N(0,1)$ and $\varepsilon_{ij} \overset{i.i.d.}{\sim} N(0,1)$, and they are independent, then under the assumption (3), the negative log-likelihood of the data, conditional on $\{(m_i, \{T_{ij}\}_{j=1}^{m_i})\}_{i=1}^{n}$ is given by

$$\begin{aligned} -\log L(B, \Lambda, \sigma^2) &= const. + \frac{1}{2}\sum_{i=1}^{n} Tr[(\sigma^2 I_{m_i} + \Phi_i^T B \Lambda B^T \Phi_i)^{-1}(\mathbf{Y}_i - \boldsymbol{\mu}_i)(\mathbf{Y}_i - \boldsymbol{\mu}_i)^T] \\ &\quad + \frac{1}{2}\sum_{i=1}^{n} \log|\sigma^2 I_{m_i} + \Phi_i^T B \Lambda B^T \Phi_i|, \end{aligned} \tag{7}$$

where Λ is the $r \times r$ diagonal matrix of non-zero eigenvalues of $C(\cdot,\cdot)$, $\mathbf{Y}_i = (Y_{i1}, \ldots, Y_{im_i})^T$ and $\boldsymbol{\mu}_i = (\mu(T_{i1}), \ldots, \mu(T_{im_i}))^T$ are $m_i \times 1$ vectors, and $\Phi_i = [\boldsymbol{\phi}(T_{i1}) : \ldots : \boldsymbol{\phi}(T_{im_i})]$ is an $M \times m_i$ matrix. It is clear that the difficulty with the maximum likelihood approach mainly lies in the irregularity of the objective function (7), and the fact that the parameter B has orthonormal constraints (4). Moreover, this is a non-convex optimization problem with respect to the parameters, since the Hessian operator of the objective function is not globally positive definite. Note that, here Gaussianity is simply a working assumption, since (7) is a *bona fide* loss function. It is also assumed throughout that (7) is differentiable with respect to the eigenvalues and eigenfunctions. This in turn depends on the assumption that all the nonzero eigenvalues of the covariance kernel are distinct.

We propose to minimize (7) directly subject to (4) by a Newton-Raphson algorithm on the Stiefel manifold, whose general form has been developed in Edelman, Arias and Smith (1998). Specifically, the proposed estimator is

$$(\widehat{B}, \widehat{\Lambda}, \widehat{\sigma}^2) = \arg \min_{B \in \mathcal{S}_{M,r}, (\Lambda, \sigma^2) \in \Theta} - \log L(B, \Lambda, \sigma^2),$$

where $\Theta = \mathbb{R}_+^{r+1}$, and $\mathcal{S}_{M,r} := \{A \in \mathbb{R}^{M \times r} : A^T A = I_r\}$ is the Stiefel manifold of $M \times r$ real-valued matrices (with $r \leq M$) with orthonormal columns. The Newton-Raphson procedure involves computation of the intrinsic gradient and Hessian of the objective function (7), and on convergence, it sets its intrinsic gradient to zero. Thus the proposed estimator solves the likelihood equation:

$$\nabla_{(B,\Lambda,\sigma^2)} \log L(B, \Lambda, \sigma^2) = 0.$$

It is noteworthy that while the Newton-Raphson algorithm is not guaranteed to converge, the asymptotic results established in Paul and Peng (2008a) show that the objective function is locally convex in a neighborhood of the truth and hence the Newton-Raphson procedure should work when a reasonable initial estimate is used. We shall discuss the details of this algorithm in Section 3.

3 Estimation procedure

In this section, we describe the Newton-Raphson algorithm for minimizing the loss function (7). In a seminal paper, Edelman *et al.* (1998) derive Newton-Raphson and conjugate gradient algorithms for optimizing functions on Stiefel and Grassman manifolds. As their counterparts in the Euclidean space, these algorithms aim to set the gradient of the objective function (viewed as a function on

the manifold) to zero. In this paper, we propose to use the Newton-Raphson algorithm to find the maximum likelihood estimator, partly due to its faster convergence rate than the gradient based methods (Edelman *et al.*, 1998).

In our setting, the objective is to minimize the loss function (7). For notational simplicity, drop the irrelevant constants and re-write (7) as

$$F(B, \Lambda, \sigma^2) := \sum_{i=1}^{n} \left[F_i^1(B, \Lambda, \sigma^2) + F_i^2(B, \Lambda, \sigma^2) \right], \tag{8}$$

where

$$F_i^1(B, \Lambda, \sigma^2) = Tr[P_i^{-1} \widetilde{\mathbf{Y}}_i \widetilde{\mathbf{Y}}_i^T], \quad F_i^2(B, \Lambda, \sigma^2) = \log |P_i|, \tag{9}$$

with $P_i = \sigma^2 I_{m_i} + \Phi_i^T B \Lambda B^T \Phi_i$, $\widetilde{\mathbf{Y}}_i = \mathbf{Y}_i - \boldsymbol{\mu}_i$, and Φ_i as defined in Section 2. Here we treat $\boldsymbol{\mu}_i$ as known, since we propose to estimate it separately by a local linear method. The parameter spaces for Λ and σ^2 are positive cones in Euclidean spaces and hence convex. The parameter space for B is $\mathcal{S}_{M,r}$, the Stiefel manifold of $M \times r$ matrices with orthonormal columns.

We break each Newton-Raphson updating step into two parts - (a) an update of (Λ, σ^2), keeping B at the current value; and (b) an update of B, setting (Λ, σ^2) at the recently updated value. Thus, the algorithm proceeds by starting at an initial estimate and then cycling through these two steps. For now, assume that the orthonormal basis functions $\{\phi_k\}$ and dimensions M and r ($M \geq r$) are given. Since $\lambda_k > 0$ for all $k = 1, \ldots, r$ and $\sigma^2 > 0$, it is convenient to define $\boldsymbol{\zeta} = \log(\Lambda)$, i.e. $\zeta_k = \log \lambda_k$, and $\tau = \log \sigma^2$, and treat F as a function of $\boldsymbol{\zeta}$ and τ. Note that ζ_k, τ can vary freely over $\mathbb{R}$. The Newton-Raphson step for updating $(\boldsymbol{\zeta}, \tau)$ is straightforward, since the parameter space for $(\boldsymbol{\zeta}, \tau)$ is Euclidean (see Appendix D of the supplementary material for details). In the rest of the paper, we treat $\boldsymbol{\zeta}$ interchangeably as an $r \times r$ matrix and as a $1 \times r$ vector.

We now discuss the Newton-Raphson step for updating B. First, we calculate the *intrinsic* gradient and Hessian of F with respect to B, while treating Λ and σ^2 as fixed. The key fact is that the gradient is a vector field acting on the tangent space of the manifold $\mathcal{S}_{M,r}$, and the Hessian is a bilinear operator acting on the same tangent space. Some basic concepts of Riemannian geometry and the Stiefel manifold are presented in the supplementary material (Appendix A). We use $\mathcal{M}$ to denote $\mathcal{S}_{M,r}$; let $\mathcal{T}_B\mathcal{M}$ denote the tangent space of $\mathcal{M}$ at B . Let F_B denote the usual Euclidean gradient, i.e., $F_B = ((\frac{\partial F}{\partial b_{kl}}))$. For any $\Delta \in \mathcal{T}_B\mathcal{M}$, $F_{BB}(\Delta)$ denotes *the* element of $\mathcal{T}_B\mathcal{M}$ satisfying

$$\langle F_{BB}(\Delta), X \rangle_c = \frac{\partial^2}{\partial s \partial t} F(B + s\Delta + tX) \,|_{s,t=0} \,, \quad \text{for all} \quad X \in \mathcal{T}_B\mathcal{M},$$

where $\langle\ ,\ \rangle_c$ is the *canonical metric* on the Stiefel manifold $\mathcal{M}$.

Proposition 1 : *The gradient of F_i^j with respect to B, and $F_{i,BB}^j(\Delta)$ $(j = 1, 2; i = 1, \cdots n)$ are:*

$$\nabla F_i^j = F_{i,B}^j - B(F_{i,B}^j)^T B, \quad F_{i,BB}^j(\Delta) = H_{i,BB}^j(\Delta) - B(H_{i,BB}^j(\Delta))^T B, \quad \Delta \in \mathcal{T}_B\mathcal{M}, \tag{10}$$

$$\begin{aligned} \text{where,} \quad F_{i,B}^1 &= 2\sigma^{-2}\left[\Phi_i\Phi_i^T BQ_i^{-1}B^T - I_M\right]\Phi_i\widetilde{\mathbf{Y}}_i\widetilde{\mathbf{Y}}_i^T\Phi_i^T BQ_i^{-1}; \\ F_{i,B}^2 &= 2\Phi_i\Phi_i^T BQ_i^{-1}; \end{aligned}$$

$$\begin{aligned} H_{i,BB}^1(\Delta) &= 2\sigma^{-2}\Phi_i\Phi_i^T[\Delta Q_i^{-1}B^T + BQ_i^{-1}\Delta^T - BQ_i^{-1}(\Delta^T\Phi_i\Phi_i^T B + B^T\Phi_i\Phi_i^T\Delta)Q_i^{-1}B^T]\Phi_i\widetilde{\mathbf{Y}}_i\widetilde{\mathbf{Y}}_i^T\Phi_i^T BQ_i^{-1} \\ &\quad +2\sigma^{-2}\left[(\Phi_i\Phi_i^T BQ_i^{-1}B^T - I_M)\Phi_i\widetilde{\mathbf{Y}}_i\widetilde{\mathbf{Y}}_i^T\Phi_i^T(\Delta Q_i^{-1} - BQ_i^{-1}(\Delta^T\Phi_i\Phi_i^T B + B^T\Phi_i\Phi_i^T\Delta)Q_i^{-1})\right]; \\ H_{i,BB}^2(\Delta) &= 2\Phi_i\Phi_i^T\left[\Delta - BQ_i^{-1}(\Delta^T\Phi_i\Phi_i^T B + B^T\Phi_i\Phi_i^T\Delta)\right]Q_i^{-1}. \end{aligned}$$

In the above $Q_i = \sigma^2\Lambda^{-1} + B^T\Phi_i\Phi_i^T B$ is an $r \times r$ positive definite matrix. These formulae are derived from the characterization of the gradient field and Hessian operator for the Stiefel manifold. The expressions differ from that of Euclidean gradient and Hessian due to the curvature of the Riemannian manifold. For example, in (10), ∇F_i^j is the intrinsic gradient whereas F_i^j is the Euclidean gradient.

Note that, the above formulae only involve the inversion of r by r matrices. Thus the proposed procedure can efficiently handle the case of relatively dense measurements, where m_i could be much larger than r. Detailed derivations are given in the supplementary material (Appendix B). In the following, we use H_F to denote the Hessian operator of F and H_F^{-1} to denote its inverse. We then outline **the Newton-Raphson algorithm on $\mathcal{S}_{M,r}$:** Given $B \in \mathcal{S}_{M,r}$,

1. compute the intrinsic gradient $\nabla F\,|_B$ of F at B, given by $\nabla F\,|_B = G := F_B - BF_B^T B$;

2. compute the tangent vector $\Delta := -H_F^{-1}(G)$ at B, by solving the linear system

$$F_{BB}(\Delta) - B\ \text{skew}(F_B^T\Delta) - \ \text{skew}(\Delta F_B^T)B - \frac{1}{2}\Pi\Delta B^T F_B = -G, \tag{11}$$

$$B^T\Delta + \Delta^T B = 0, \tag{12}$$

   where $\Pi = I - BB^T$, and $\text{skew}(X) := (X - X^T)/2$;

3. move from B in the direction Δ to $B(1)$ along the geodesics $B(t) = BM(t) + QN(t)$, where

(i) $QR = (I - BB^T)\Delta$ is the QR-decomposition, so that Q is $M \times r$ with orthonormal columns, and R is $r \times r$, upper triangular;

(ii) $A = B^T\Delta$, and

$$\begin{bmatrix} M(t) \\ N(t) \end{bmatrix} = \exp\left\{ t \begin{bmatrix} A & -R^T \\ R & 0 \end{bmatrix} \right\} \begin{bmatrix} I_r \\ 0 \end{bmatrix};$$

4. set $B = B(1)$, and repeat until convergence. This means that the sup-norm of the gradient G is less than a pre-specified tolerance level.

The formulae of F_B and $F_{BB}(\Delta)$ are given in Proposition 1. In order to update the tangent vector Δ, in step 2 of the algorithm, we need to solve the system of equations given by (11) and (12). These are linear matrix equations and we propose to solve them via vectorization (cf. Muirhead, 1982). This step requires a considerable amount of computational effort since it involves the inversion of an $Mr \times Mr$ matrix. In step 3, the matrix within exponent is a skew-symmetric matrix and so the exponential of that can be calculated easily using the singular value decomposition. Details of these calculations are given in the supplementary material (Appendix B).

In order to apply the Newton-Raphson algorithm, we need to choose a suitable basis for representing the eigenfunctions. In the simulation studies presented in Section 5, we use the orthonormalized cubic B-spline basis with equally spaced knots (Green and Silverman, 1994, p. 157). This is mainly due to the fact that B-splines provide a flexible, localized and stable basis for a wide class of smooth functions (Chui, 1987, de Boor, 1978). Other choices of (stable) basis functions are certainly possible, and can be implemented similarly. In addition, the number of basis functions M and the dimension of the process r need to be specified, which is discussed in Section 4. Given a basis $\{\phi_k(\cdot)\}$ and fixed M, r with $M \geq r$, an initial estimate of Λ and B can be obtained by projecting an initial estimate of the covariance kernel $\widehat{C}(\cdot,\cdot)$ onto $\{\phi_1(\cdot), \cdots, \phi_M(\cdot)\}$ followed by eigen-decomposition.

4 Approximate cross-validation score

One of the key questions pertaining to nonparametric function estimation is the issue of model selection. This, in our context means selecting r, the number of nonzero eigenvalues, and the basis for representing the eigenfunctions. Once we have a scheme for choosing the basis, the second problem boils down to selecting M, the number of basis functions. Various criteria for dealing with this include AIC, BIC, multi-fold cross-validation and leave-one-curve-out cross-validation.

In this paper, we propose to choose $(M,\ r)$ by minimizing an approximation of the leave-one-curve-out cross-validation score

$$CV := \sum_{i=1}^{n} \ell_i(\mathbf{Y}_i, \mathbf{T}_i, \widehat{\Psi}^{(-i)}), \tag{13}$$

where $\mathbf{T}_i = (T_{i1}, \ldots, T_{im_i})$. Here $\Psi = (B, \tau, \boldsymbol{\zeta})$, and $\widehat{\Psi}$ and $\widehat{\Psi}^{(-i)}$ are the estimates of Ψ based on the whole data and data excluding curve i, respectively. In this paper, $\ell_i(\mathbf{Y}_i, \mathbf{T}_i, \Psi) = F_i^1 + F_i^2$ (cf. (8) and (9)). Note that in the Gaussian setting, ℓ_i is proportional to the negative log-likelihood (up to an additive constant) of the i-th curve. Thus, CV defined through (13) is the *empirical predictive Kullback-Leibler risk* . In literature, CV score based on prediction error loss is often used (e.g., Yao *et al.*, 2005). A first order expansion of the difference between the averaged loss under the true and estimated parameters shows that Kullback-Leibler loss correctly scales the difference, while the prediction error loss does not (Paul and Peng, 2008b). Hence the use of the latter is not recommended for the current problem.

It is well known that, the computational cost to get CV is prohibitive, which necessitates the use of efficient approximations. Our method of approximation, which is similar in spirit to the approach taken by Burman (1990) in the context of fitting generalized additive models, is based on the observation that, $\widehat{\Psi}$, and $\{\widehat{\Psi}^{(-i)}\}_{i=1}^n$ satisfy

$$\sum_{i=1}^{n} \nabla \ell_i(\mathbf{Y}_i, \mathbf{T}_i, \widehat{\Psi}) = 0, \tag{14}$$

$$\sum_{j:j\neq i} \nabla \ell_j(\mathbf{Y}_j, \mathbf{T}_j, \widehat{\Psi}^{(-i)}) = 0, \quad i = 1, \ldots, n. \tag{15}$$

Here $\{\ell_i\}_{i=1}^n$ are functions on the product space $\widetilde{\mathcal{M}} = \mathcal{M} \times \mathbb{R}^{r+1}$, where $\mathcal{M}$ is the Stiefel manifold with the canonical metric g, to be denoted by $\langle\ ,\ \rangle_c$. The parameter space $\mathbb{R}^{r+1}$ refers to $\{(\tau,\ \underset{r\times r}{\boldsymbol{\zeta}}\ ) : \tau \in \mathbb{R}, \zeta_k \in \mathbb{R},\ k = 1, \ldots, r\}$, with Euclidean metric. $\nabla \ell_i$ denotes the gradient of ℓ_i viewed as a vector field on the product manifold.

The main idea for our approximation scheme is the observation that for each $i = 1, \ldots, n$, the "leave curve i out" estimate $\widehat{\Psi}^{(-i)}$ is a perturbation of the estimate $\widehat{\Psi}$ based on the whole data. Thus, one can expand the left hand side of (15) around $\widehat{\Psi}$ to obtain an approximation of $\widehat{\Psi}^{(-i)}$. This approximation is used to obtain a second order approximation to the cross-validation score given by (13). Hereafter, we shall use $\ell_j(\widehat{\Psi})$, and $\ell_j(\widehat{\Psi}^{(-i)})$ to denote $\ell_j(\mathbf{Y}_j, \mathbf{T}_j, \widehat{\Psi})$ and $\ell_j(\mathbf{Y}_j, \mathbf{T}_j, \widehat{\Psi}^{(-i)})$,

respectively, for $1 \leq i, j \leq n$. Let $\nabla_B \ell_i$ and $\nabla_B^2 \ell_i$ denote gradient and Hessian of ℓ_i with respect to B, and $\nabla_{(\tau,\zeta)} \ell_i$ and $\nabla^2_{(\tau,\zeta)} \ell_i$ denote gradient and Hessian of ℓ_i with respect to $(\tau, \boldsymbol{\zeta})$. Since the parameter $(\tau, \boldsymbol{\zeta}) \in \mathbb{R}^{r+1}$, $\nabla_{(\tau,\zeta)} \ell_i$ is an $(r+1) \times 1$ vector and $\nabla^2_{(\tau,\zeta)} \ell_i$ is an $(r+1) \times (r+1)$ matrix. As mentioned before, $\nabla_B \ell_i$ is a tangent vector and $\nabla_B^2 \ell_i$ is a bilinear operator on the tangent space $\mathcal{T}_B \mathcal{M}$. The expression of the approximate CV score is given in the following proposition.

Proposition 2 : *Denote the Hessian operator of $\sum_j \ell_j$ with respect to B and $(\tau, \boldsymbol{\zeta})$ by $\mathbf{H}_B$ and $\mathbf{H}_{(\tau,\zeta)}$, respectively. Then a second order approximation to the CV score is given by*

$$\begin{aligned}
\widetilde{CV} \;:=\; & \sum_{i=1}^{n} \ell_i(\widehat{\Psi}) + \sum_{i=1}^{n} \langle \nabla_{(\tau,\zeta)} \ell_i(\widehat{\Psi}), [\mathbf{H}_{(\tau,\zeta)}(\widehat{\Psi})]^{-1} \nabla_{(\tau,\zeta)} \ell_i(\widehat{\Psi}) \rangle \\
& + \sum_{i=1}^{n} \langle \nabla_B \ell_i(\widehat{\Psi}), [\mathbf{H}_B(\widehat{\Psi})]^{-1} \nabla_B \ell_i(\widehat{\Psi}) \rangle_c \\
& + \frac{3}{2} \sum_{i=1}^{n} \langle \nabla^2_{(\tau,\zeta)} \ell_i(\widehat{\Psi}) [\mathbf{H}_{(\tau,\zeta)}(\widehat{\Psi})]^{-1} \nabla_{(\tau,\zeta)} \ell_i(\widehat{\Psi}), [\mathbf{H}_{(\tau,\zeta)}(\widehat{\Psi})]^{-1} \nabla_{(\tau,\zeta)} \ell_i(\widehat{\Psi}) \rangle \\
& + \frac{3}{2} \sum_{i=1}^{n} \nabla_B^2 \ell_i(\widehat{\Psi}) ([\mathbf{H}_B(\widehat{\Psi})]^{-1} \nabla_B \ell_i(\widehat{\Psi}), [\mathbf{H}_B(\widehat{\Psi})]^{-1} \nabla_B \ell_i(\widehat{\Psi})). \qquad (16)
\end{aligned}$$

Observe that, in order to obtain the estimate $\widehat{\Psi}$ using the Newton-Raphson algorithm, we need to compute the objects $\nabla_B \ell_i$, $\nabla_{(\tau,\zeta)} \ell_i$, $\nabla_B^2 \ell_i$, $\nabla^2_{(\tau,\zeta)} \ell_i$, $\mathbf{H}_B$, and $\mathbf{H}_{(\tau,\zeta)}$ at each step. Indeed, since the Newton-Raphson procedure aims to solve (14), whenever the procedure converges, we immediately have these objects evaluated at $\widehat{\Psi}$. Therefore, the additional computational cost for computing $\widetilde{CV}$ is negligible. This provides huge computational advantage in comparison to the usual leave-one-curve-out CV score approach.

The details of the derivation of $\widetilde{CV}$ are given in the supplementary material (Appendix C). Here we elucidate the main steps leading to (16). We introduce some notations first. Let $\delta_\tau^i = \widehat{\tau}^{(-i)} - \widehat{\tau}$, $\boldsymbol{\delta}_\zeta^i = \widehat{\boldsymbol{\zeta}}^{(-i)} - \widehat{\boldsymbol{\zeta}}$ (a $1 \times r$ vector), and $\Delta_i = \dot{\gamma}(0) \in \mathcal{T}_{\widehat{B}} \mathcal{M}$, with $\gamma(t)$ a geodesic on $(\mathcal{M}, g)$ starting at $\gamma(0) = \widehat{B}$, and ending at $\gamma(1) = \widehat{B}^{(-i)}$. Note that, Δ_i is an element of the tangent space at $\widehat{B}$. The Hessian $\nabla^2 \ell_i$ with respect to $\Psi = (B, \tau, \boldsymbol{\zeta})$ is approximated by ignoring the mixed-derivative terms $\nabla_{(\tau,\zeta)}(\nabla_B \ell_i)$ and $\nabla_B(\nabla_{(\tau,\zeta)} \ell_i)$. This simplifies the calculation considerably and allows us to treat the terms involving approximation of $\widehat{B}^{(-i)}$ (keeping $(\tau, \boldsymbol{\zeta})$ fixed at $(\widehat{\tau}, \widehat{\boldsymbol{\zeta}})$) and that of $(\widehat{\tau}^{(-i)}, \widehat{\boldsymbol{\zeta}}^{(-i)})$ (keeping B fixed at $\widehat{B}$) separately. Thus, a second order expansion of the CV score around $\widehat{\Psi}$ becomes

$$CV \approx \sum_{i=1}^{n} \ell_i(\widehat{\Psi}) + \Big[\sum_{i=1}^{n} \langle \nabla_{(\tau,\zeta)} \ell_i(\widehat{\Psi}), (\delta_\tau^i, \boldsymbol{\delta}_\zeta^i)^T \rangle + \frac{1}{2} \sum_{i=1}^{n} \langle [\nabla^2_{(\tau,\zeta)} \ell_i(\widehat{\Psi})](\delta_\tau^i, \boldsymbol{\delta}_\zeta^i)^T, (\delta_\tau^i, \boldsymbol{\delta}_\zeta^i)^T \rangle \Big]$$
$$+ \Big[\sum_{i=1}^{n} \langle \nabla_B \ell_i(\widehat{\Psi}), \Delta_i \rangle_c + \frac{1}{2} \sum_{i=1}^{n} \nabla_B^2 \ell_i(\widehat{\Psi})(\Delta_i, \Delta_i) \Big]. \quad (17)$$

We can make the above approximation rigorous by quantifying the errors in the approximation (Paul and Peng, 2008a). In order to get first order approximations to the second and third terms in (17), we shall use equations (14) and (15). These equations separate into two sets of equations involving the gradients $\nabla_{(\tau,\zeta)}\ell_i$ and $\nabla_B \ell_i$, respectively. The treatment of the former does not require any extra concept beyond regular matrix algebra, whereas the treatment of the latter requires Riemannian geometric concepts.

The above approach to approximating the CV score can also be applied in various other problems and/or with other types of smooth loss functions. To illustrate this, we consider the ordinary regression setting:

$$y_i = \mathbf{x}_i^T \beta + \epsilon_i, \quad i = 1, \ldots, n.$$

There, the GCV (generalized cross validation) score (cf. Golub, Heath and Wahba, 1979) is given by $GCV = \sum_{i=1}^n (y_i - \mathbf{x}_i^T \widehat{\beta})^2/(1-p/n)^2$ where p is the dimension of the regressor $\mathbf{x}$. In contrast, the above approximation will yield $\widetilde{CV} = \sum_{i=1}^n (y_i - \mathbf{x}_i^T \widehat{\beta})^2(1 + 2h_{ii} + 3h_{ii}^2)$, where $h_{ii} = \mathbf{x}_i^T (X^T X)^{-1} \mathbf{x}_i$, with $X^T = (\mathbf{x}_1^T, \ldots, \mathbf{x}_n^T)$. Note that the latter is a second order (term-wise) approximation to the usual PRESS criterion (equivalently, the CV score): $CV = \sum_{i=1}^n (y_i - \mathbf{x}_i^T \widehat{\beta})^2/(1-h_{ii})^2$.

Regarding comparison of the second order approximation with a first order approximation such as an AIC-type or a GCV-type approximation (though we are not aware of any GCV criterion for the current problem), it can be said that the second order approximation derived here is likely to have smaller bias as an estimate of the expected predictive Kullback-Leibler loss. On the other hand, the use of empirical Hessian operator, as opposed to the expected Hessian operator as in AIC-type approximations (Stone, 1977), means that $\widetilde{CV}$ will tend to have higher variability. However, it is unclear if an AIC-type approximation is appropriate in the current context since its implicit assumptions about asymptotic behavior of the Hessian operator (or, empirical Fisher information operator) have not been verified. One can also employ multi-fold cross validation for model selection, which has clear computational advantage over the leave-one-out cross validation, and is also more stable. However, multi-fold cross validation with moderately large number of groups is computa-

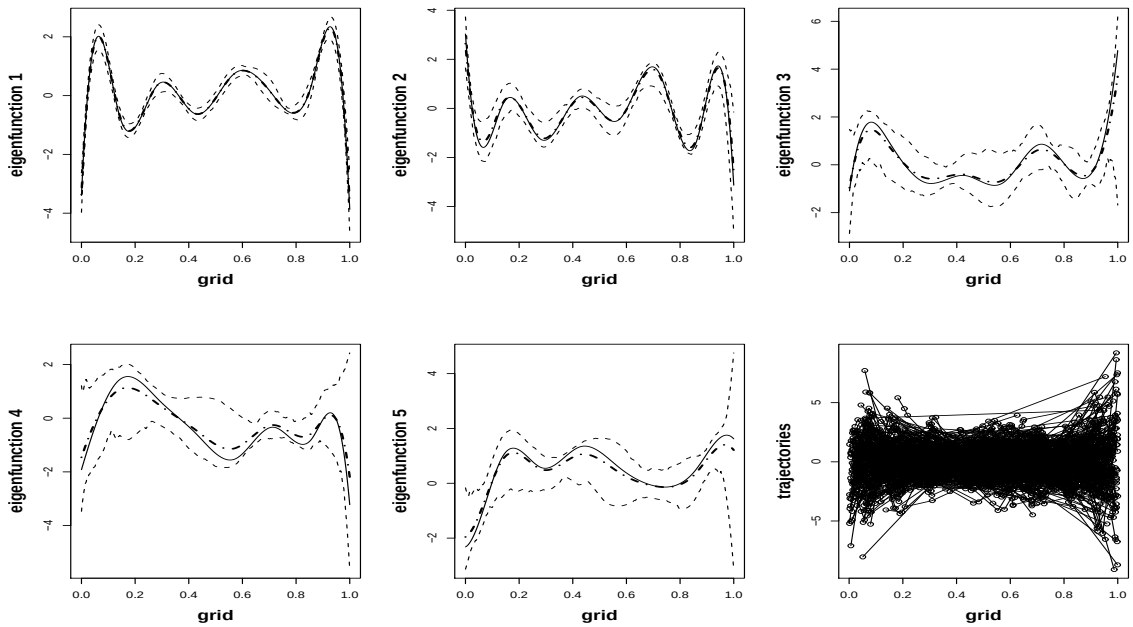
tionally much more expensive than $\widetilde{CV}$. On the other hand, as shown by Burman (1989), without additional correction terms it can have large bias if the number of groups is small. Thus, $\widetilde{CV}$ balances the need of stability and computational efficiency as a model selection criterion for this type of semi-parametric problems.

5 Simulation

In this section, we conduct two simulation studies. The first study is focused on the estimation accuracy of the proposed method (henceforth, `Newton`) and comparing it with two existing procedures: the local polynomial method (henceforth, `loc`) (Yao *et al.*, 2005), and the EM algorithm (henceforth, `EM`) (James *et al.*, 2000). The second study aims to illustrate the usefulness of the model selection approach described in Section 4. All data are generated under model (1) with Gaussian principal component scores $\{\xi_{i\nu}\}$ (equation (2)). For all settings, $\mu(t) \equiv 0$, and its estimate $\widehat{\mu}(t)$, obtained by a local linear smoothing, is subtracted from the observations before estimating the other model parameters. The number of measurements m_i are i.i.d. $\sim$ uniform$\{2, \cdots, 10\}$; the measurement points for the ith subject $\{T_{ij} : j = 1, \cdots, m_i\}$ are i.i.d. $\sim$ uniform$[0, 1]$. For both `Newton` and `EM`, cubic B-splines with equally spaced knots are used as basis functions. `loc` and `EM` are used to obtain two different sets of initial estimates for `Newton`. The resulting estimates by `Newton` are therefore denoted by `New.loc` and `New.EM`, respectively. For `EM`, only initial value of σ is needed. Since the result is rather robust to this choice (James *et al.*, 2000), it is set to be one. For `loc`, bandwidths are selected by the `h.select()` function in the `R` package `sm`, with `method="cv"`.

In the first study, data are generated under the settings referred as `easy` and `practical`, respectively. For the `easy` case, there are three non-zero eigenvalues: $1, 0.66, 0.517$ and the eigenfunctions are represented by the cubic B-splines with $M = 5$ equally spaced knots. In contrast, the more complex `practical` case has five non-zero eigenvalues: $1, 0.66, 0.517, 0.435, 0.381$ and $M = 10$ equally spaced knots (Figure 1). Independent Gaussian errors with $\sigma^2 = 1/16$ is added to each observation. Three different sample sizes $n = 100,\ n = 500,\ n = 1000$ are considered for the `practical` case, as well as $n = 50$ for the `easy` case. 100 independent replicates are generated for each case. In the first study, the true r is used by all three methods. Note that, the estimation of covariance kernel by `loc` does not rely on either M or r. For a given r, the first r eigenfunctions and eigenvalues of the estimated covariance $\widehat{C}(\cdot, \cdot)$ (using the optimal bandwidth) are used. For `Newton` and `EM`, a number of different values of M, including the truth, are used to fit the model. For `Newton`, we report

Figure 1: `Practical` case with $n = 500$. True and estimated eigenfunctions (panels one to five): true eigenfunctions (solid); Point-wise average of estimated eigenfunctions by `New.EM` (dash-dots); Point-wise 0.95 and 0.05 quantiles of estimated eigenfunctions by `New.EM` (dash); Sample trajectories (panel six).



the number of converged replicates (cf. Section 3) for each M (Table 2). As we shall see, lack of convergence of `Newton` is primarily caused by poor initial estimates. Therefore, it is fair to compare all three methods on the converged replicates only. The performance of these three methods under the true model is summarized in Table 1. For the estimation of eigenfunctions, mean integrated squared error (MISE) is used as a measure of accuracy.

As can be seen from Table 1, for most cases, the MISE corresponding to `Newton` (`New.loc`, `New.EM`) shows a good risk behavior. To give a visual illustration, in Figure 1, we plot the point-wise average of estimated eigenfunctions over all converged replicates, as well as the point-wise 0.95 and 0.05 quantiles, under the true model: $r = 5, M = 10$ for the `practical` case with $n = 500$. As can be seen from this figure, the average is very close to the truth, meaning only small biases. The width between two quantiles is fairly narrow meaning small variations, except for large variances at the boundaries for higher order eigenfunctions (that is, the eigenfunctions corresponding to smaller eigenvalues). Properties of the B-spline basis ensure that as long as the eigenfunctions are sufficiently smooth, the estimates have small bias, as demonstrated above. However, Since the

Table 1: Estimation of the eigenfunctions under the true model. Mean integrated squared error (MISE) over converged replicates is reported for each method under the true model ($r = 5, M = 10$ for `practical` case and $r = 3, M = 5$ for `easy` case). Numbers in the parenthesis are the standard deviations of the integrated squared error over converged replicates. "Reduction" stands for the percentage change of MISE of `New.loc` (or `New.EM`) over `loc` (or `EM`).

		Practical case: n=100					Easy case: n=50				
Method		ψ_1	ψ_2	ψ_3	ψ_4	ψ_5	ψ_1	ψ_2	ψ_3		
New.loc	MISE*	1.31	1.39	1.46	1.51	1.46	0.353	0.629	0.537		
	(Sd)	(0.54)	(0.42)	(0.38)	(0.37)	(0.35)	(0.435)	(0.577)	(0.576)		
loc	MISE*	1.24	1.52	1.43	1.51	1.55	0.786	0.978	0.907		
	(Sd)	(0.50)	(0.36)	(0.46)	(0.39)	(0.38)	(0.591)	(0.596)	(0.516)		
	Reduction (%)	-6.0	9.1	-2.1	-0.3	5.8	55.1	35.7	40.8		
New.EM	MISE	0.247	0.683	1.075	0.996	0.846	0.260	0.592	0.462		
	(Sd)	(0.312)	(0.495)	(0.531)	(0.559)	(0.554)	(0.353)	(0.566)	(0.541)		
EM	MISE	0.310	0.780	1.071	0.961	0.848	0.309	0.701	0.539		
	(Sd)	(0.365)	(0.546)	(0.507)	(0.504)	(0.527)	(0.429)	(0.570)	(0.531)		
	Reduction (%)	20.3	12.4	-0.4	-3.6	0.2	15.9	15.5	14.3		
		Practical case: n=500					Practical case: n=1000				
Method		ψ_1	ψ_2	ψ_3	ψ_4	ψ_5	ψ_1	ψ_2	ψ_3	ψ_4	ψ_5
New.loc	MISE	0.035	0.195	0.463	0.556	0.343	0.015	0.067	0.169	0.228	0.146
	(Sd)	(0.025)	(0.347)	(0.532)	(0.531)	(0.404)	(0.014)	(0.134)	(0.226)	(0.259)	(0.203)
loc	MISE	0.434	1.059	1.143	1.211	1.127	0.224	0.813	1.069	1.035	0.930
	(Sd)	(0.387)	(0.502)	(0.514)	(0.536)	(0.523)	(0.137)	(0.523)	(0.466)	(0.580)	(0.541)
	Reduction (%)	91.9	81.6	59.5	54.1	69.6	93.3	91.8	84.2	78.0	84.3
New.EM	MISE	0.036	0.172	0.396	0.498	0.332	0.016	0.063	0.145	0.232	0.172
	(Sd)	(0.031)	(0.288)	(0.432)	(0.509)	(0.420)	(0.014)	(0.122)	(0.193)	(0.337)	(0.307)
EM	MISE	0.054	0.231	0.578	0.626	0.401	0.033	0.144	0.309	0.370	0.250
	(Sd)	(0.046)	(0.263)	(0.469)	(0.517)	(0.463)	(0.039)	(0.204)	(0.330)	(0.416)	(0.359)
	Reduction (%)	33.3	25.5	31.5	20.4	17.2	51.5	56.3	53.1	37.3	31.2

* for `practical` case with $n = 100$, there is no convergence for `New.loc` under the true model, so the results over all 100 replicates are reported here.

higher order eigenfunctions tend to be more oscillating, effectively there is much less data near the boundaries for their estimation. Moreover, it is well known that basis representation approaches that do not choose the basis functions adaptively have relatively large variability at the boundaries when estimating a function from noisy data. To overcome this limitation, we could incorporate in the basis representation knowledge about the behavior of the processes at the boundaries whenever such information is available. The proposed method can be modified to use adaptively chosen basis functions. Although this is beyond the scope of this paper, it is a future direction of our research.

In comparison with `loc` and `EM`, `Newton` performs better in terms of MISE for eigenfunctions, except for the `practical` case with $n = 100$ (Table 1). In that case, `New.EM` works considerably better than `EM` for the two leading eigenfunctions, and works comparably as `EM` for the other three eigenfunctions. As for `New.loc`, since the initial estimates by `loc` are very poor, `New.loc` has trouble to converge. Thus we report the results based on all 100 replicates there and they are much worse than either `New.EM` or `EM`. For all other cases, the reduction (in percentage) in MISE by `Newton` varies from 35% to as high as 95% compared to `loc`; and 15% to around 55% compared to `EM`. Moreover, there are greater improvements by `Newton` with larger sample sizes. In general, there is also a significant amount of reduction in mean squared error in estimation of eigenvalues and error variance by the `Newton` method compared to the other two methods, especially to `loc` (data

Table 2: Selection of the number of basis functions M. Reported are number of converged replicates of `Newton` under each M; and number of times of each M being selected by the approximate CV score.

Model	Method	Number of converged replicates				Frequency of being selected			
		$M=4$	$\mathbf{M=5}$	$M=6$	$M=7$	$M=4$	$\mathbf{M=5}$	$M=6$	$M=7$
`Easy`	`New.loc`	82	72	67	55	9	54	17	11
($n=50$)	`New.EM`	99	96	97	94	0	74	12	14
		$M=5$	$\mathbf{M=10}$	$M=15$	$M=20$	$M=5$	$\mathbf{M=10}$	$M=15$	$M=20$
`Practical`	`New.loc`	40	0	0	0	40	0	0	0
($n=100$)	`New.EM`	87	82	29	1	9	79	7	1
		$M=5$	$\mathbf{M=10}$	$M=15$	$M=20$	$M=5$	$\mathbf{M=10}$	$M=15$	$M=20$
`Practical`	`New.loc`	53	46	21	7	25	46	8	1
($n=500$)	`New.EM`	96	93	94	88	0	93	2	5
		$M=5$	$\mathbf{M=10}$	$M=15$	$M=20$	$M=5$	$\mathbf{M=10}$	$M=15$	$M=20$
`Practical`	`New.loc`	53	77	43	28	7	77	4	5
($n=1000$)	`New.EM`	98	97	98	92	0	97	0	3

not shown). One problem with `loc` is that, it gives highly variable, and sometimes even negative, estimates of σ^2. Actually, for all the simulations we have done, at least around 20% replicates result in a negative estimate. Although here we only report results under the true model, results under an adequate model (that is a model corresponding to an M which is at least as large as the true M) are similar. As can be seen from Table 2, as long as `Newton` converges for the true model, it is selected most of the time by the approximate CV score (16). Moreover, a model corresponding to an inadequate M is rarely selected unless it is the only model under which `Newton` converges. We also observe that `New.loc` often suffers from lack of convergence, especially when sample size is small and the true model is complex (Table 2). This is mainly due to the poor initial estimates by `loc`. However, among the converged replicates, the performance of `New.loc` is not much worse than that of `New.EM`, especially for the leading eigenfunctions.

In summary, we observe satisfactory performance of `Newton` in terms of estimation, as well as improvements of `Newton` over the two alternative methods, especially over `loc`. These improvements pertain to both average and standard deviation of the measures of accuracy, and they increase with the sample size. We also find out that, for `Newton`, good initial estimates are important mainly for the convergence of the procedure. As long as the estimates converge, they do not depend heavily on the initial estimates. We have also studied more settings. One of them has eigenfunctions not exactly representable by B-spline basis. The general picture remains the same as here. We also studied the impact of error variance and error distribution. These simulations show that all three methods are quite robust with respect to these two aspects. Due to space limitations, these results are not reported here.

In order to study the selection of M and r simultaneously, we conduct the second simulation

Table 3: Model selection : `Hybrid` case, $n = 500$, $\sigma^2 = 1/16$, Gaussian noise

	New.loc							
	I. number of converged replicates							
r	$M = 8$	$M = 9$	$\mathbf{M = 10}$	$M = 11$	$M = 12$	$M = 13$	$M = 14$	$M = 15$
2	85	83	92	87	87	85	81	79
3	51	85	95	93	91	86	89	91
4	15	37	46	23	26	20	22	15
5	4	9	7	4	5	3	1	0
6	0	1	1	1	0	0	0	0
	II. frequencies of models selected (by FEV for κ = 0.995,0.99,0.95)							
r	$M = 8$	$M = 9$	$\mathbf{M = 10}$	$M = 11$	$M = 12$	$M = 13$	$M = 14$	$M = 15$
2	1	0	0 (0,0,0)	0	0	0	0	0
3	0	0	52 (66,74,93)	0	0	0	0	1
4	0	0	37 (26,18,0)	0	0	1	2	0
5	0	0	3 (1,1,0)	0	0	0	0	0
6	0	0	1 (0,0,0)	0	0	0	0	0
	New.EM							
	I. number of converged replicates							
r	$M = 8$	$M = 9$	$\mathbf{M = 10}$	$M = 11$	$M = 12$	$M = 13$	$M = 14$	$M = 15$
2	99	98	100	99	98	99	99	99
3	98	98	99	100	100	100	100	99
4	95	94	99	98	98	92	95	95
5	94	95	95	95	95	87	90	83
6	82	91	60	84	75	73	68	59
	II. frequencies of models selected (by FEV for κ = 0.995,0.99,0.95)							
r	$M = 8$	$M = 9$	$\mathbf{M = 10}$	$M = 11$	$M = 12$	$M = 13$	$M = 14$	$M = 15$
2	0	0	0 (0,0,0)	0	0	0	0	0
3	0	0	1 (1,1,97)	0	0	0	0	0
4	0	0	3 (48,95,0)	0	0	2	0	0
5	0	0	58 (48,1,0)	0	0	0	0	0
6	0	0	35 (0,0,0)	1	0	0	0	0

study, in which there are three leading eigenvalues $(1, 0.66, 0.52)$, and a fourth eigenvalue which is comparable to the error variance ($\lambda_4 = 0.07$). Additionally, there are 6 smaller eigenvalues $(9.47 \times 10^{-3}, 1.28 \times 10^{-3}, 1.74 \times 10^{-4}, 2.35 \times 10^{-5}, 3.18 \times 10^{-6}, 4.30 \times 10^{-7})$. Thus, under this setting, the first three eigenvalues explain 96.4% total variability in the signal, and the first four eigenvalues explain 99.5%. The corresponding orthonormal eigenfunctions are represented in a cubic B-spline basis with $M = 10$ equally spaced knots. Data are generated with sample size $n = 500$ and Gaussian noises with $\sigma^2 = 1/16$. This setting is referred as the `hybrid` case. We fit models with $M = 8, 9, \ldots, 15$, and $r = 2, \ldots, 6$. In this setting, our aim is to get an idea about the typical sizes (meaning (M, r)) of models selected. The result is summarized in Table 3.

We find that, for `New.EM`, $M = 10$ and $r = 5$ or 6 are the preferred models by the proposed approximate CV score; however, for `New.loc`, $M = 10$ and $r = 3$ or 4 are the ones selected most often. The latter is mainly due to lack of convergence of `New.loc` for $r \geq 4$. Therefore, we will focus on the results of `New.EM` hereafter. For models with $r = 5$ and $r = 6$, the small eigenvalues (the fifth one and/or the sixth one) are estimated to be reasonably small by `New.EM` (data not shown). We then use the standard procedure of FEV (*fraction of explained variance*) on the selected model to further prune down the value of r: for every model (M^*, r^*) selected by the CV criterion, we choose

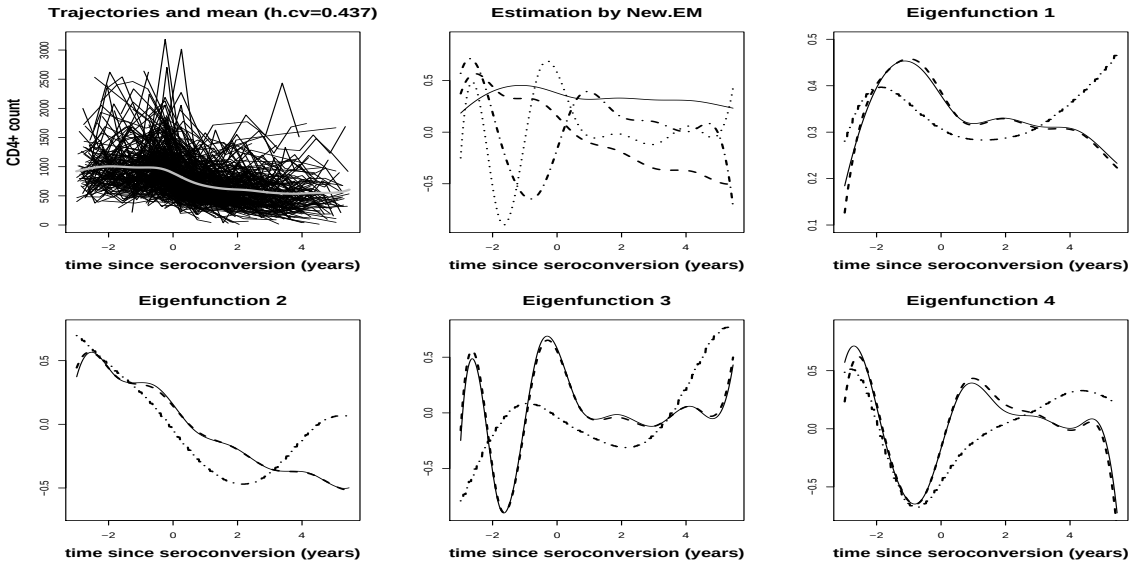
the smallest index $\overline{r}$ for which the ratio $\sum_{\nu=1}^{\overline{r}} \widehat{\lambda}_\nu / \sum_{\nu=1}^{r*} \widehat{\lambda}_\nu$ exceeds a threshold κ. In this study, we consider $\kappa = 0.995, 0.99, 0.95$. The results of the model selection using this additional FEV criterion are reported in Table 3 in the parentheses. As can be seen there, the additional FEV criterion gives very reasonable model selection results, which also indicates that the small eigenvalues are indeed estimated to be reasonably small.

In summary, the approximate CV score (16) is very effective in selecting the correct M– the number of basis functions needed to represent the eigenfunctions. It has the tendency to select slightly larger than necessary r. However, in those selected models, the `Newton` estimates of the small or zero eigenvalues are relatively small. Therefore, the model selection results are not going to be very misleading and an additional FEV criterion can be applied to select a smaller model (in terms of r).

All codes for the simulation studies are written in `R` language and running under `R` version 2.4.1 on a machine with Pentium Duo core, CPU 2 GHz and 2 GB RAM. The code for `EM` is provided by professor G. James at USC via personal communication. For the `practical` case with $n = 500$, on average, it takes around 20 seconds for `loc` to fit the model for each replicate. Also, it takes 43 seconds for `EM` to fit the model (with $M = 10$), and an additional 71 seconds for `New.EM`. Thus in total `New.EM` takes 114 seconds for each replicate on average. One way to speed up `Newton` is to update the Hessian operator in every several steps. By doing so (updating Hessian in every five steps), the whole `New.EM` procedure takes about 80 seconds (including initial estimates by `EM` and calculating the approximate CV score), while resulting an estimate almost identical to the one which updates Hessian in every step. The R package `fpca` is available on `cran`. An implementation in a more efficient language such as `C++` is currently being pursued.

It has been our experience that the initial steps of the Newton-Raphson algorithm are usually too large, which leads to instability in the Hessian and occasional lack of convergence. One way to address this problem is to choose smaller step sizes in the beginning. We have already implemented this in our algorithm where the step size is chosen to be 0.5 in the first few steps. A data-driven choice of the step size, using e.g., *Armijo's rule* (cf. Dussault, 2000), is under consideration. Another possibility is to implement the *conjugate gradient* method and use it in the first few steps of the optimization procedure. After a few steps of conjugate gradient procedure, we intend to use the Newton-Raphson method. This, we hope, would ensure more stable initial steps of the estimation procedure, and a faster convergence later on due to the use of Newton-Raphson. A similar recommendation has been made, although for different problems, by Boyd and Vandenberghe (2004).

Figure 2: `CD4+ counts` data: estimated mean and eigenfunctions under the selected model ($M = 10, r = 4$). First panel: sample trajectories (thin dark) and estimated mean function (thick light); Second panel: estimated eigenfunctions by `New.EM`: $\widehat{\psi}_1$(solid), $\widehat{\psi}_2$ (dash), $\widehat{\psi}_3$ (dots), $\widehat{\psi}_4$ (dash-dots); Third to sixth panels: estimated eigenfunctions by `loc` (dash-dots), `New.EM` (solid) and `EM` (dash).



6 Application

In this section, we analyze the data on *CD4+ cell number counts* collected as part of the Multicenter AIDS Cohort Study (MACS) (Kaslow *et al.*, 1987). The data is from Diggle, Heagerty, Liang and Zeger (2002) (`http://www.maths.lancs.ac.uk/~diggle/lda/Datasets/lda.dat`). It consists of 2376 measurements of CD4+ cell counts against time since seroconversion (time when HIV becomes detectable which is used as zero on the time line) for 369 infected men enrolled in the study. Five patients with only one measurement each were removed from our analysis. For the rest 364 subjects, the number of measurements varies between 2 and 12, with a median of 6 and a standard deviation of 2.66. The time span of the study is about 8.5 years (covering about three years before seroconversion and 5.5 years after that). The goal of our analysis is to understand the variability of CD4+ counts as a function of time since seroconversion. This is expected to provide useful insights into the dynamics of the process. This data set has been analyzed using various approaches, including varying coefficient models (Fan and Zhang, 2000, Wu and Chiang, 2000), functional principal component approach (Yao *et al.*, 2005) and parametric random effects models (Diggle *et al.*, 2002).

In our analysis, four methods: `EM`, `New.EM`, `loc` and `New.loc` are used. Several different models, with M taking values $5, 10, 15, 20$, and r taking values $2, \ldots, 6$ are considered. The model with $M = 10, r = 4$ results in the smallest $\widetilde{CV}$ score and thus is selected. Figure 2 shows the estimated eigenfunctions under the selected model.

Under the selected model, `New.EM` and `EM` result in quite similar estimates for both eigenvalues and eigenfunctions, whereas the estimates of `loc` are very different. Since `New.loc` fails to converge under the selected model, its estimates are not reported here. Moreover, based on our experience, this is an indicator that the corresponding results by `loc` might not be altogether reliable either. Comparing with the estimated noise variance ($\widehat{\sigma}^2 = 38,411$), the results of `New.EM` suggest that among the 4 non-negligible eigenvalues, two of which are large ($\widehat{\lambda}_1 = 473,417,\ \widehat{\lambda}_2 = 208,201$), and the other two are relatively small ($\widehat{\lambda}_3 = 53,254,\ \widehat{\lambda}_4 = 24,582$). The similarity between the results by `EM` and `New.EM` in this case is mainly due to the fast decay of the eigenvalues, which results in a relatively simple model. This is in contrast with most simulation results reported in the previous section where the rate of decay is much slower, rendering `New.EM` more advantageous compared to `EM`.

Next, we give an interpretation of the shape of the eigenfunctions. The first eigenfunction is rather flat compared to the other three eigenfunctions (Figure 2, panel two). This means that it mainly captures the baseline variability in the CD4+ cell counts from one subject to another. This is consistent with the random effects model proposed in Diggle *et al.* (2002) (page 108-113). It is also noticeable that the second eigenfunction has a shape similar to that of the mean function (Figure 2, panels one and four). The shapes of the first two eigenfunctions, and the fact that their corresponding eigenvalues are relatively large, seem to indicate that a simple linear dynamical model, with random initial conditions, may be employed in studying the dynamics of CD4+ cell counts. This observation is also consistent with the implication by the time-lagged graphs used in Diggle *et al.* (2002, Fig. 3.13, p. 47). The third and fourth eigenvalues are comparable in magnitude to the error variance, and the corresponding eigenfunctions have somewhat similar shapes. They correspond to the contrast in variability between early and late stages of the disease.

7 Discussion

In this paper, we presented a method that utilizes the intrinsic geometry of the parameter space explicitly to obtain the estimate in a non-regular problem, that of estimating eigenfunctions and

eigenvalues of the covariance kernel when the data are only observed at sparse and irregular time points. We did comparative studies with two other estimation procedures by James *et al.* (2000) and Yao *et al.* (2005). We presented a model selection approach based on the minimization of an approximate cross-validation score. Based on our simulation studies, we have found that the proposed geometric approach works well for both estimation and model selection. Moreover, its performance is in general better than that of the other two methods. We also looked at a real-data example to see how our method captures the variability in the data.

In Paul and Peng (2008a), consistency of the proposed estimator has been established under suitable regularity conditions on the covariance kernel. Furthermore, the estimators of the eigenfunctions achieve the optimal nonparametric rate up to a factor of $\log n$ under the l^2 loss. The key components of this asymptotic analysis are : (i) utilization of the geometry of the tangent space of the manifold; and (ii) analysis of the expected loss function.

Finally, we present two other applications of the method studied in this paper. There are many statistical problems with (part of) the parameters having orthornormality constraints. If we have, (i) explicit form and smoothness of the loss function; (ii) the ability to compute the intrinsic gradient and Hessian of the loss function, we can adopt a similar approach for estimation and model selection.

In typical spatio-temporal problems, the domain D of the observations is a subset of $\mathbb{R}^2$ rather than $\mathbb{R}^1$, and the random processes $\{X(\mathbf{s}, t) : \mathbf{s} \in D, t \in \mathbb{Z}\}$ may be temporally correlated. Here our interest is in the estimation of the spatial covariance kernel assuming a relatively simple temporal dependence structure. Since the stations at which the observations are taken are typically discrete and sparse, for example, when the measurements are on some pollutant in the atmosphere, this is a natural extension of the FPCA problem studied here.

Another problem relates to incorporation of covariate effects in the analysis of longitudinal data. For example, Cardot (2006) studies a model where the covariance of $X(\cdot)$ conditioned on a covariate W has the expansion : $C^w(s,t) := \mathrm{Cov}(X(s), X(t)|W = w) = \sum_{\nu \geq 1} \lambda_\nu(w)\psi_\nu(s,w)\psi_\nu(t,w)$. The author proposes a kernel-based nonparametric approach for estimating the eigenvalues and eigenfunctions (now dependent on w). In practice this method would require dense measurements. A modification of our method can easily handle the case, even for sparse measurements, when the eigenvalues are considered to be simple parametric functions of w, and eigenfunctions do not depend on w. One such model : $\lambda_\nu(w) := \alpha_\nu e^{w^T \beta_\nu}$, $\nu = 1, \ldots, r$, for parameters $(\alpha_1, \beta_1), \ldots, (\alpha_r, \beta_r)$, and assuming $\alpha_\nu = 0$ and $\beta_\nu = 0$ for $\nu > r$. This model captures the variability in amplitude of the eigenfunctions in the sample curves as a function of the covariate. The conditional likelihood of the

data $\{(\{Y_{ij}\}_{j=1}^{m_i}, W_i) : i = 1, \ldots, n\}$ is explicit and can be maximized using a modification of our procedure.

Acknowledgement

The authors thank the associate editor and two referees for their helpful comments. Peng and Paul are partially supported by grant DMS-0806128 from the National Science Foundation.

References

1. Ash, R. B. (1972) *Real Analysis and Probability*, Academic Press.

2. Besse, P., Cardot, H. and Ferraty, F. (1997) Simultaneous nonparametric regression of unbalanced longitudinal data. *Computational Statistics and Data Analysis* **24**, 255-270.

3. Boente, G. and Fraiman, R. (2000) Kernel-based functional principal components analysis. *Statistics and Probability Letters* **48**, 335-345.

4. Boyd, S. and Vandenberghe, L. (2004) *Convex Optimization*. Cambridge University Press.

5. Burman, P. (1989) : A comparative study of ordinary cross-validation, v-fold cross-validation and the repeated learning-testing methods. *Biometrika* **76**, 503-514.

6. Burman, P. (1990) : Estimation of generalized additive models. *Journal of Multivariate Analysis* **32**, 230-255.

7. Cai, T. and Hall, P. (2006) Prediction in functional linear regression. *Annals of Statistics* **34**, 2159-2179.

8. Cardot, H., Ferraty F. and Sarda P. (1999) Functional Linear Model. *Statistics and Probability Letters* **45**, 11-22.

9. Cardot, H. (2000) Nonparametric estimation of smoothed principal components analysis of sampled noisy functions. *Journal of Nonparametric Statistics* **12**, 503-538.

10. Cardot, H. (2006) Conditional functional principal components analysis. *Scandinavian Journal of Statistics* **33**, 317-335.

11. Chui, C. (1987) *Multivariate Splines*. SIAM.

12. de Boor, C. (1978) *A Practical Guide to Splines*. Springer-Verlag, New York.

13. Diggle, P. J., Heagerty, P., Liang, K.-Y., and Zeger, S. L. (2002) *Analysis of Longitudinal Data, 2nd. Edition*. Oxford University Press.

14. Dussault, J. P. (2000) : Convergence of implementable descent algorithms for unconstrained problems. *Journal of Optimization Theory and Applications* **104**, 739-745.

15. 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.

16. Fan, J. and Zhang, J. T. (2000) Two-step estimation of functional linear models with applications to longitudinal data. *Journal of Royal Statistical Society, Series B* **62**, 303-322.

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

18. Golub, G. H., Heath, M. and Wahba, G. (1979) : Generalized cross validation as a method for choosing a good ridge parameter. *Technometrics* **21**, 215-224.

19. Green, P. J. and Silverman, B. W. (1994) *Nonparametric Regression and Generalized Linear Models : A Roughness Penalty Approach*. Chapman & Hall/CRC.

20. Hall, P. and Horowitz, J. L. (2007) Methodology and convergence rates for functional linear regression. *Annals of Statistics* **35**, 41-69.

21. 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.

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

23. Kaslow R. A., Ostrow D. G., Detels R., Phair J. P., Polk B. F., Rinaldo C. R. (1987) The Multicenter AIDS Cohort Study: rationale, organization, and selected characteristics of the participants. *American Journal of Epidemiology* **126**(2), 310-318.

24. Lee, J. M. (1997) *Riemannian Manifolds: An Introduction to Curvature*, Springer.

25. Marron, S. J., Müller, H.-G., Rice, J., Wang, J.-L., Wang, N. and Wang, Y. (2004) Discussion of nonparametric and semiparametric regression. *Statistica Sinica* **14**, 615-629.

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

27. Paul, D. and Peng, J. (2008a) Consistency of restricted maximum likelihood estimators of principal components. To appear in *Annals of Statistics* (arXiv:0805.0465v1).

28. Paul, D. and Peng, J. (2008b) Principal components analysis for sparsely observed correlated functional data using a kernel smoothing approach. *Technical report* (arXiv:0807.1106).

29. Rice, J. A. and Wu, C. O. (2001) Nonparametric mixed effects models for unequally sampled noisy curves. *Biometrics* **57**, 253-259.

30. Ramsay, J. and Silverman, B. W. (2005) *Functional Data Analysis, 2nd Edition.* Springer.

31. Stone, M. (1977) : An asymptotic equivalence of choice of model by cross validation and Akaike's information criterion. *Journal of Royal Statistical Society, Series B* **39**, 44-47.

32. Wahba, G. (1985) : A comparison of GCV and GML for choosing the smoothing parameter in the generalized spline smoothing problem. *Annals of Statistics* **13**, 1378-1402.

33. Wu, C. and Chiang, C. (2000) Kernel smoothing on varying coefficient models with longitudinal dependent variables. *Statistica Sinica* **10**, 433-456.

34. 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.

35. Yao, F., Müller, H.-G. and Wang, J.-L. (2006) Functional linear regression for longitudinal data. *Annals of Statistics* **33**, 2873-2903.

Supplementary Material

Appendix A : Some Riemannian geometric concepts

Let $(\mathcal{M}, g)$ be a smooth manifold with Riemannian metric g. Denote the tangent space of $\mathcal{M}$ at $p \in \mathcal{M}$ by $\mathcal{T}_p\mathcal{M}$. We shall first give some basic definitions related to the work presented in this article (see, e.g., Lee, 1997).

Gradient and Hessian of a function

- **Gradient :** Let $f : \mathcal{M} \to \mathbb{R}$ be a smooth function. Then ∇f, the *gradient* of f, is a *vector field* on $\mathcal{M}$ defined by the following: for any $X \in \mathcal{T}\mathcal{M}$, (i.e., a vector field on $\mathcal{M}$), $\langle \nabla f, X \rangle_g = X(f)$, where $X(f)$ is the *directional derivative* of f w.r.t. X.

- **Covariant derivative :** (also known as *Riemannian connection*) : Let $X, Y \in \mathcal{T}\mathcal{M}$ be two vector fields on $\mathcal{M}$. Then the vector field $\overline{\nabla}_Y X \in \mathcal{T}\mathcal{M}$ is called the *covariant derivative of X in the direction of Y* if the operator $\overline{\nabla}$ satisfies the following properties: (i) bi-linearity; (ii) Leibnitz rule; (iii) preserving metric; and (iv) symmetry.

- **Hessian operator:** For a smooth function $f : \mathcal{M} \to \mathbb{R}$, $H_f : \mathcal{T}\mathcal{M} \times \mathcal{T}\mathcal{M} \to \mathbb{R}$ is the bi-linear form defined as: $H_f(Y, X) = \langle \overline{\nabla}_Y(\nabla f), X \rangle_g$, $X, Y \in \mathcal{T}\mathcal{M}$. Note that, by definition, H_f is bi-linear and symmetric. For notational simplicity, sometimes we also write $\overline{\nabla}_Y(\nabla f)$ as $H_f(Y)$.

- **Inverse of Hessian :** For $X \in \mathcal{T}\mathcal{M}$, and a smooth function $f : \mathcal{M} \to \mathbb{R}$, $H_f^{-1}(X) \in \mathcal{T}\mathcal{M}$ is defined as the vector field satisfying: for $\forall\ \Delta \in \mathcal{T}\mathcal{M}$, $H_f(H_f^{-1}(X), \Delta) = \langle X, \Delta \rangle_g$.

Some facts about Stiefel manifold

The manifold $\mathcal{M} = \{B \in \mathbb{R}^{M \times r} : B^T B = I_r\}$ is known as the *Steifel manifold* in $\mathbb{R}^{M \times r}$. Here we present some basic facts about this manifold which are necessary for implementing the proposed method. A more detailed description is given in Edelman *et al.* (1998).

- **Tangent space :** $\mathcal{T}_B\mathcal{M} = \{\Delta \in \mathbb{R}^{M \times r} : B^T \Delta$ is skew-symmetric $\}$.

- **Canonical metric :** For $\Delta_1, \Delta_2 \in \mathcal{T}_B\mathcal{M}$ with $B \in \mathcal{M}$, the *canonical metric* (a Riemannian metric on $\mathcal{M}$) is defined as

$$\langle \Delta_1, \Delta_2 \rangle_c = Tr(\Delta_1^T (I - \frac{1}{2} BB^T) \Delta_2).$$

- **Gradient :** For a smooth function $f : \mathcal{M} \to \mathbb{R}$,

$$\nabla f \,|_B = f_B - B f_B^T B,$$

where f_B is the usual Euclidean gradient of f defined through $(f_B)_{ij} = \frac{\partial f}{\partial B_{ij}}$.

- **Hessian operator :** (derived from the geodesic equation): For $\Delta_1, \Delta_2 \in \mathcal{T}_B\mathcal{M}$,

$$H_f(\Delta_1, \Delta_2) \,|_B = f_{BB}(\Delta_1, \Delta_2) + \frac{1}{2} Tr\left[(f_B^T \Delta_1 B^T + B^T \Delta_1 f_B^T) \Delta_2 \right] - \frac{1}{2} Tr\left[(B^T f_B + f_B^T B) \Delta_1^T \Pi \Delta_2 \right],$$

where $\Pi = I - BB^T$.

- **Inverse of Hessian :** For $\Delta, G \in \mathcal{T}_B\mathcal{M}$, the equation $\Delta = H_f^{-1}(G)$ means that Δ is the solution of

$$f_{BB}(\Delta) - B \text{ skew}(f_B^T \Delta) - \text{ skew}(\Delta f_B^T) B - \frac{1}{2} \Pi \Delta B^T f_B = G,$$

subject to the condition that $B^T \Delta$ is skew-symmetric, i.e., $B^T \Delta + \Delta^T B = 0$, where $f_{BB}(\Delta) \in \mathcal{T}_B\mathcal{M}$ such that

$$\langle f_{BB}(\Delta), X \rangle_c = f_{BB}(\Delta, X) = Tr(\Delta^T f_{BB} X) \qquad \forall\ X \in \mathcal{T}_B\mathcal{M}.$$

This implies that $f_{BB}(\Delta) = H(\Delta) - BH^T(\Delta)B$, where $H(\Delta) = f_{BB}^T \Delta$. Here $\text{skew}(X) = \frac{1}{2}(X - X^T)$.

Appendix B : Detailed Calculations

Proof of Proposition 1

We use the following lemmas (cf. Muirhead, 1982) repeatedly in our computations in this subsection.

Lemma 1 : *Let $P = I_p + AC$ where A is $p \times q$, C is $q \times p$. Then* $\det(P) = |I_p + AC| = |I_q + CA|$.

Lemma 2 : *Let A be $p \times p$ and E be $q \times q$, both nonsingular, matrices. If $P = A + CED$, for any $p \times q$ matrix C and any $q \times p$ matrix D, then*

$$P^{-1} = (A + CED)^{-1} = A^{-1}[A - CQ^{-1}D]A^{-1}, \qquad \text{where} \qquad Q = E^{-1} + DA^{-1}C.$$

Remark : If $A = I_p$ and $q < p$, then $P^{-1} = I_p - CQ^{-1}D$ where Q is $q \times q$.

By *Lemma 1*,

$$|P_i| = \sigma^{2m_i}|I_r + \sigma^{-2}\Lambda B^T \Phi_i \Phi_i^T B| = \sigma^{2(m_i - r)}|\Lambda||\sigma^2 \Lambda^{-1} + B^T \Phi_i \Phi_i^T B| = \sigma^{2(m_i-r)}|\Lambda||Q_i|, \quad (18)$$

where $Q_i = \sigma^2 \Lambda^{-1} + B^T \Phi_i \Phi_i^T B$ is an $r \times r$ positive semi-definite matrix. Also, by *Lemma 2*

$$P_i^{-1} = \sigma^{-2} I_{m_i} - \sigma^{-4} \Phi_i^T B(\Lambda^{-1} + \sigma^{-2} B^T \Phi_i \Phi_i^T B)^{-1} B^T \Phi_i = \sigma^{-2} \left[I_{m_i} - \Phi_i^T B Q_i^{-1} B^T \Phi_i \right]. \quad (19)$$

For simplifying notations, we shall drop the subscript i from functions F_i^1 and F_i^2. Using (19)

$$\begin{aligned} F^1(B) = Tr(P_i^{-1} \widetilde{\mathbf{Y}}_i \widetilde{\mathbf{Y}}_i^T) &= \sigma^{-2} Tr(\widetilde{\mathbf{Y}}_i \widetilde{\mathbf{Y}}_i^T) - \sigma^{-2} Tr(\Phi_i^T B Q_i^{-1} B^T \Phi_i \widetilde{\mathbf{Y}}_i \widetilde{\mathbf{Y}}_i^T) \\ &= \sigma^{-2} Tr(\widetilde{\mathbf{Y}}_i \widetilde{\mathbf{Y}}_i^T) - \sigma^{-2} Tr(B Q_i^{-1} B^T \Phi_i \widetilde{\mathbf{Y}}_i \widetilde{\mathbf{Y}}_i^T \Phi_i^T). \end{aligned}$$

Similarly by (18),

$$F^2(B) = \log |P_i| = \log(\sigma^{2(m_i - r)}|\Lambda|) + \log |Q_i|.$$

Let $B(t) = B + t\Delta$. Then $\frac{dQ_i(t)}{dt} |_{t=0} = \Delta^T \Phi_i \Phi_i^T B + B^T \Phi_i \Phi_i^T \Delta$, so that

$$\begin{aligned} \langle F_B^1, \Delta \rangle &= \frac{dF^1(B(t))}{dt} |_{t=0} \\ &= -\sigma^{-2} Tr \left[(\Delta Q_i^{-1} B^T + B Q_i^{-1} \Delta^T - B Q_i^{-1} \frac{dQ_i}{dt} |_{t=0} \, Q_i^{-1} B^T) \Phi_i \widetilde{\mathbf{Y}}_i \widetilde{\mathbf{Y}}_i^T \Phi_i^T \right] \\ &= -2\sigma^{-2} Tr \left[(\Phi_i \widetilde{\mathbf{Y}}_i \widetilde{\mathbf{Y}}_i^T \Phi_i^T B Q_i^{-1} - \Phi_i \Phi_i^T B Q_i^{-1} B^T \Phi_i \widetilde{\mathbf{Y}}_i \widetilde{\mathbf{Y}}_i^T \Phi_i^T B Q_i^{-1}) \Delta^T \right]. \quad (20) \end{aligned}$$

Thus,

$$\begin{aligned} F_B^1 &= -2\sigma^{-2} \left[\Phi_i \widetilde{\mathbf{Y}}_i \widetilde{\mathbf{Y}}_i^T \Phi_i^T B Q_i^{-1} - \Phi_i \Phi_i^T B Q_i^{-1} B^T \Phi_i \widetilde{\mathbf{Y}}_i \widetilde{\mathbf{Y}}_i^T \Phi_i^T B Q_i^{-1} \right] \\ &= 2\sigma^{-2} \left[\Phi_i \Phi_i^T B Q_i^{-1} B^T - I_M \right] \Phi_i \widetilde{\mathbf{Y}}_i \widetilde{\mathbf{Y}}_i^T \Phi_i^T B Q_i^{-1} \quad (21) \end{aligned}$$

Similarly,

$$\begin{aligned}\langle F_B^2, \Delta\rangle = \frac{dF^2(B(t))}{dt}\mid_{t=0} &= Tr\left(Q_i^{-1}\frac{dQ_i}{dt}\mid_{t=0}\right)\\ &= Tr(Q_i^{-1}(\Delta^T\Phi_i\Phi_i^TB + B^T\Phi_i\Phi_i^T\Delta))\\ &= 2Tr(Q_i^{-1}B^T\Phi_i\Phi_i^T\Delta). \qquad (22)\end{aligned}$$

Thus,

$$F_B^2 = 2\Phi_i\Phi_i^TBQ_i^{-1}. \qquad (23)$$

Let $B(t,s) = B + t\Delta + sX$. Then using (20),

$$\begin{aligned}F_{BB}^1(\Delta, X) &= \langle F_{BB}^1(\Delta), X\rangle_c = \langle H_{BB}^1(\Delta), X\rangle\\ &= \frac{\partial}{\partial t}\frac{\partial}{\partial s}F^1(B(t,s))\mid_{s,t=0}\\ &= 2\sigma^{-2}Tr\left[\frac{\partial}{\partial t}(\Phi_i\Phi_i^TB(t,0)Q_i(t)^{-1}B(t,0)^T - I_M)\mid_{t=0}\ \Phi_i\widetilde{\mathbf{Y}}_i\widetilde{\mathbf{Y}}_i^T\Phi_i^TBQ_i^{-1}X^T\right]\\ &\quad + 2\sigma^{-2}Tr\left[(\Phi_i\Phi_i^TBQ_i^{-1}B^T - I_M)\Phi_i\widetilde{\mathbf{Y}}_i\widetilde{\mathbf{Y}}_i^T\Phi_i^T\frac{\partial}{\partial t}(B(t,0)Q_i(t)^{-1})\mid_{t=0}\ X^T\right].\end{aligned}$$

Note that

$$\frac{\partial}{\partial t}(B(t,0)Q_i(t)^{-1})\mid_{t=0} = \Delta Q_i^{-1} - BQ_i^{-1}(\Delta^T\Phi_i\Phi_i^TB + B^T\Phi_i\Phi_i^T\Delta)Q_i^{-1},$$

and

$$\frac{\partial}{\partial t}(B(t,0)Q_i(t)^{-1}B(t,0)^T)\mid_{t=0} = \Delta Q_i^{-1}B^T + BQ_i^{-1}\Delta^T - BQ_i^{-1}(\Delta^T\Phi_i\Phi_i^TB + B^T\Phi_i\Phi_i^T\Delta)Q_i^{-1}B^T.$$

Thus,

$$\begin{aligned}H_{BB}^1(\Delta) &= 2\sigma^{-2}\Phi_i\Phi_i^T\left[\Delta Q_i^{-1}B^T + BQ_i^{-1}\Delta^T - BQ_i^{-1}(\Delta^T\Phi_i\Phi_i^TB + B^T\Phi_i\Phi_i^T\Delta)Q_i^{-1}B^T\right]\Phi_i\widetilde{\mathbf{Y}}_i\widetilde{\mathbf{Y}}_i^T\Phi_i^TBQ_i^{-1}\\ &\quad + 2\sigma^{-2}\left[(\Phi_i\Phi_i^TBQ_i^{-1}B^T - I_M)\Phi_i\widetilde{\mathbf{Y}}_i\widetilde{\mathbf{Y}}_i^T\Phi_i^T(\Delta Q_i^{-1} - BQ_i^{-1}(\Delta^T\Phi_i\Phi_i^TB + B^T\Phi_i\Phi_i^T\Delta)Q_i^{-1})\right].\end{aligned}$$

Similarly, using (22),

$$\begin{aligned} F^2_{BB}(\Delta, X) &= \langle F^2_{BB}(\Delta), X\rangle_c = \langle H^2_{BB}(\Delta), X\rangle \\ &= \frac{\partial}{\partial t}\frac{\partial}{\partial s}F^2(B(t,s))\mid_{s,t=0} \\ &= \frac{\partial}{\partial t}[2Tr(Q_i(t)^{-1}B(t,0)^T\,\Phi_i\,\Phi_i^T X)]\mid_{t=0} \\ &= 2Tr\left[(-Q_i^{-1}\frac{dQ_i(t)}{dt}\mid_{t=0} Q_i^{-1}B^T + Q_i^{-1}\Delta^T)\,\Phi_i\,\Phi_i^T X\right] \\ &= 2Tr\left[(-Q_i^{-1}(\Delta^T\,\Phi_i\,\Phi_i^T B + B^T\,\Phi_i\,\Phi_i^T\Delta)Q_i^{-1}B^T + Q_i^{-1}\Delta^T)\,\Phi_i\,\Phi_i^T X\right]. \end{aligned}$$

From this,

$$\begin{aligned} H^2_{BB}(\Delta) &= 2\left[-Q_i^{-1}(\Delta^T\,\Phi_i\,\Phi_i^T B + B^T\,\Phi_i\,\Phi_i^T\Delta)Q_i^{-1}B^T + Q_i^{-1}\Delta^T)\,\Phi_i\,\Phi_i^T\right]^T \\ &= 2\,\Phi_i\,\Phi_i^T\left[\Delta - BQ_i^{-1}(\Delta^T\,\Phi_i\,\Phi_i^T B + B^T\,\Phi_i\,\Phi_i^T\Delta)\right]Q_i^{-1}. \end{aligned}$$

Exponential of skew-symmetric matrices

Let $X = -X^T$ be a $p \times p$ matrix. Want to compute $\exp(tX) := \sum_{k=0}^{\infty}\frac{t^k}{k!}X^k$ for $t \in \mathbb{R}$, where $X^0 = I$. Let the SVD of X be given by $X = UDV^T$, where $U^TU = V^TV = I_p$, and D is diagonal. So, $X^2 = XX = -XX^T = -UDV^TVDU^T = -UD^2U^T$. This also shows that all the eigenvalues of X are purely imaginary. Using the facts that $D^0 = I_p$; $X^{2k} = (X^2)^k = (-1)^k(UD^2U^T)^k = (-1)^kUD^{2k}U^T$; and $X^{2k+1} = (-1)^kUD^{2k}U^TUDV^T = (-1)^kUD^{2k+1}V^T$, we have

$$\begin{aligned} \exp(tX) &= U\left[\sum_{k=0}^{\infty}\frac{(-t)^k}{(2k)!}D^{2k}\right]U^T + U\left[\sum_{k=0}^{\infty}\frac{(-t)^k}{(2k+1)!}D^{2k+1}\right]V^T \\ &= U\cos(tD)U^T + U\sin(tD)V^T, \end{aligned}$$

where $\cos(tD) = diag((\cos(td_{jj}))_{j=1}^p)$ and $\sin(tD) = diag((\sin(td_{jj}))_{j=1}^p)$, if $D = diag((d_{jj})_{j=1}^p)$.

Vectorization of matrix equations

A general form of the equation in the $M \times r$ matrix Δ is given by

$$L = A\Delta + \Delta K + C\Delta D + E\Delta^T F,$$

where L is $M \times r$, A is $M \times M$, K is $r \times r$, C is $M \times M$, D is $r \times r$, E is $M \times r$, and F is $M \times r$. Vectorization of this equation using the vec operation means that $\text{vec}(L)$ is given by

$$\begin{aligned} & \text{vec}(A\Delta) + \text{vec}(\Delta K) + \text{vec}(C\Delta D) + \text{vec}(E\Delta^T F) \\ = \quad & \left[(I_r \otimes A) + (K^T \otimes I_M) + (D^T \otimes C) + (F^T \otimes E)P_{M,r}\right] \text{vec}(\Delta), \end{aligned} \tag{24}$$

where, $\otimes$ denotes the Kronecker product, and we have used the following properties of the vec operator (Muirhead, 1982): (i) $\text{vec}(KXC) = (C^T \otimes K)\text{vec}(X)$; (ii) $\text{vec}(X^T) = P_{m,n}\text{vec}(X)$. Here X is $m \times n$, K is $r \times m$, C is $n \times s$, and $P_{m,n}$ is an appropriate $mn \times mn$ permutation matrix.

Appendix C : Derivation of $\widetilde{CV}$ (16)

For now, in (15), considering only the part corresponding to the gradient w.r.t. B and expanding it around $\widehat{\Psi}$, while approximating $(\widehat{\tau}^{(-i)}, \widehat{\boldsymbol{\zeta}}^{(-i)})$ by $(\widehat{\tau}, \widehat{\boldsymbol{\zeta}})$, we have (for notational simplicity, write $\ell_j(\widehat{B})$ to denote $\ell_j(\widehat{\Psi})$)

$$0 = \sum_{j \neq i} \nabla_B \ell_j(\widehat{\Psi}^{(-i)}) \approx \sum_{j \neq i} \nabla_B \ell_j(\widehat{B}) + \sum_{j \neq i} \overline{\nabla}_{\Delta_i}(\nabla_B \ell_j(\widehat{B})), \tag{25}$$

where $\overline{\nabla}_{\Delta_i}(\nabla_B \ell_j)$ is the *covariant derivative* of $\nabla_B \ell_j$ in the direction of Δ_i. Now, substituting (14) in (25), we get

$$0 \approx -\nabla_B \ell_i(\widehat{B}) + \overline{\nabla}_{\Delta_i}[\sum_{j \neq i} \nabla_B \ell_j(\widehat{B})]. \tag{26}$$

Then for any $X \in \mathcal{T}_{\widehat{B}}\mathcal{M}$,

$$\langle \overline{\nabla}_{\Delta_i}(\sum_{j \neq i} \nabla_B \ell_j(\widehat{B})), X \rangle_c = [\sum_{j \neq i} \nabla_B^2 \ell_j(\widehat{B})](\Delta_i, X) \approx \langle \nabla_B \ell_i(\widehat{B}), X \rangle_c.$$

Thus by the definition of the Hessian inverse operator, $\Delta_i \approx [\sum_{j\neq i} \nabla_B^2 \ell_j(\widehat{B})]^{-1}(\nabla_B \ell_i(\widehat{B}))$. This, together with (26), leads to the approximation of Δ_i,

$$\begin{aligned}
\Delta_i &\approx [\sum_{j\neq i} \nabla_B^2 \ell_j(\widehat{B})]^{-1}(\nabla_B \ell_i(\widehat{B})) = [\sum_j \nabla_B^2 \ell_j(\widehat{B}) - \nabla_B^2 \ell_i(\widehat{B})]^{-1}(\nabla_B \ell_i(\widehat{B})) \\
&\approx \left[I + [\sum_j \nabla_B^2 \ell_j(\widehat{B})]^{-1}\nabla_B^2 \ell_i(\widehat{B})\right][\sum_j \nabla_B^2 \ell_j(\widehat{B})]^{-1}(\nabla_B \ell_i(\widehat{B})) \\
&= \left[I + [\mathbf{H}_B(\widehat{B})]^{-1}\nabla_B^2 \ell_i(\widehat{B})\right][\mathbf{H}_B(\widehat{B})]^{-1}(\nabla_B \ell_i(\widehat{B})), \qquad (27)
\end{aligned}$$

where $\mathbf{H}_B = \sum_j \nabla_B^2 \ell_j$. Note that the last approximation is because, for linear operators A and C such that A is invertible and $\parallel A^{-1}C \parallel$ is small, we have $(A - C)^{-1} = (I - A^{-1}C)^{-1}A^{-1} \approx (I + A^{-1}C)A^{-1}$. Now, for $X \in \mathcal{T}_{\widehat{B}}\mathcal{M}$, by definition of Hessian,

$$\begin{aligned}
&\langle \sum_{j=1}^n \overline{\nabla}_{\Delta_i}(\nabla_B \ell_j(\widehat{B})), X\rangle_c = [\nabla_B^2(\sum_{j=1}^n \ell_j(\widehat{B}))](\Delta_i, X) \\
\approx\ & \mathbf{H}_B(\widehat{B})\left([\mathbf{H}_B(\widehat{B})]^{-1}(\nabla_B \ell_i(\widehat{B})) + [(\mathbf{H}_B(\widehat{B})]^{-1}\nabla_B^2 \ell_i(\widehat{B})[\mathbf{H}_B(\widehat{B})]^{-1}(\nabla_B \ell_i(\widehat{B})), X\right) \\
=\ & \langle \nabla_B \ell_i(\widehat{B}), X\rangle_c + \langle \nabla_B^2(\ell_i(\widehat{B}))[\mathbf{H}_B(\widehat{B})]^{-1}(\nabla_B \ell_i(\widehat{B})), X\rangle_c, \qquad (28)
\end{aligned}$$

where, by definition, $\nabla_B^2(\ell_i(\widehat{B}))(\gamma) = \overline{\nabla}_\gamma(\nabla_B \ell_i(\widehat{B}))$, for $\gamma \in \mathcal{T}_{\widehat{B}}\mathcal{M}$. In the first approximation of (28), we have used the approximation (27), and the last step follows from the definition of Hessian inverse and linearity of the Hessian. From (25) we also have,

$$\overline{\nabla}_{\Delta_i}(\nabla_B \ell_i(\widehat{B})) \approx -\nabla_B \ell_i(\widehat{B}) + \sum_{j=1}^n \overline{\nabla}_{\Delta_i}(\nabla_B \ell_j(\widehat{B})). \qquad (29)$$

Substituting (28) in (29), we then have the approximation

$$\nabla_B^2 \ell_i(\widehat{B})(\Delta_i) = \overline{\nabla}_{\Delta_i}(\nabla_B \ell_i(\widehat{B})) \approx \nabla_B^2 \ell_i(\widehat{B})[\mathbf{H}_B(\widehat{B})]^{-1}(\nabla_B \ell_i(\widehat{B})). \qquad (30)$$

Using (27) and (30), and ignoring terms higher than the second order, we have

$$\begin{aligned}
& \sum_{i=1}^{n}\langle\nabla_B\ell_i(\widehat{B}),\Delta_i\rangle_c+\frac{1}{2}\sum_{i=1}^{n}\nabla_B^2\ell_i(\widehat{B})(\Delta_i,\Delta_i) \\
\approx & \left[\sum_{i=1}^{n}\langle\nabla_B\ell_i(\widehat{B}),[\mathbf{H}_B(\widehat{B})]^{-1}(\nabla_B\ell_i(\widehat{B}))\rangle_c+\sum_{i=1}^{n}\langle\nabla_B\ell_i(\widehat{B}),[\mathbf{H}_B(\widehat{B})]^{-1}\nabla_B^2\ell_i(\widehat{B})[\mathbf{H}_B(\widehat{B})]^{-1}(\nabla_B\ell_i(\widehat{B}))\rangle_c\right] \\
& +\frac{1}{2}\sum_{i=1}^{n}\langle[\mathbf{H}_B(\widehat{B})]^{-1}(\nabla_B\ell_i(\widehat{B})),\nabla_B^2\ell_i(\widehat{B})[\mathbf{H}_B(\widehat{B})]^{-1}(\nabla_B\ell_i(\widehat{B}))\rangle_c \\
= & \sum_{i=1}^{n}\langle\nabla_B\ell_i(\widehat{B}),[\mathbf{H}_B(\widehat{B})]^{-1}\nabla_B\ell_i(\widehat{B})\rangle_c+\frac{3}{2}\sum_{i=1}^{n}\nabla_B^2\ell_i(\widehat{B})([\mathbf{H}_B(\widehat{B})]^{-1}\nabla_B\ell_i(\widehat{B}),[\mathbf{H}_B(\widehat{B})]^{-1}\nabla_B\ell_i(\widehat{B})). \quad (31)
\end{aligned}$$

Here we give brief justifications for the steps in (31). The first approximation follows from the definition of Hessian, and the approximation of Δ_i by (27). The last equation follows from: (i) by definition of Hessian, applied to $\nabla_B^2\ell_i(\widehat{B})$, the term on the third line equals

$$\frac{1}{2}\sum_{i=1}^{n}\nabla_B^2\ell_i(\widehat{B})([\mathbf{H}_B(\widehat{B})]^{-1}(\nabla_B\ell_i(\widehat{B})),[\mathbf{H}_B(\widehat{B})]^{-1}(\nabla_B\ell_i(\widehat{B})));$$

and (ii) the second term on the second line equals the same term as above, except for the factor $\frac{1}{2}$, by definition of Hessian^{-1}, now applied to $[\mathbf{H}_B(\widehat{B})]^{-1}$.

Using very similar (but conceptually much simpler) arguments, we also have the second order approximation for the terms involving τ,ζ, and combining it with (17) and (31), we have the approximate CV score given by (16).

Appendix D : Gradients and Hessians with respect to ζ and τ

Define $H_i=\Phi_i^TB$, $i=1,\ldots,n$. Then

$$P_i=\sigma^2I_{m_i}+\Phi_i^TB\Lambda B^T\Phi_i=\sigma^2I_{m_i}+H_i\Lambda H_i^T=e^\tau I_{m_i}+H_i\exp(\zeta)H_i^T,$$

$$Q_i=\sigma^2\Lambda^{-1}+B^T\Phi_i\Phi_i^TB=\sigma^2\Lambda^{-1}+H_i^TH_i=e^\tau\exp(-\zeta)+H_i^TH_i.$$

and re-writing (19):

$$P_i^{-1}=\sigma^{-2}I_{m_i}-\sigma^{-4}H_i(\Lambda^{-1}+\sigma^{-2}H_i^TH_i)^{-1}H_i^T=e^{-\tau}\left[I_{m_i}-H_iQ_i^{-1}H_i^T\right].$$

For future uses, we calculate the following derivatives.

$$\frac{\partial P_i}{\partial \tau} = \frac{\partial}{\partial \tau}[e^\tau I_{m_i} + H_i \exp(\zeta) H_i^T] = e^\tau I_{m_i}. \tag{32}$$

Let $H_{ik} = \Phi_i^T B_k$, $k = 1, \ldots, r$ where B_k is the k-th column of B. We shall use the following fact

$$\frac{\partial P_i}{\partial \zeta_k} = \frac{\partial}{\partial \zeta_k}[e^\tau I_{m_i} + \sum_{k=1}^{r} e^{\zeta_k} H_{ik} H_{ik}^T] = e^{\zeta_k} H_{ik} H_{ik}^T. \tag{33}$$

In the following, we shall drop the subscript i from the functions F_i^1 and F_i^2, and treat the latter as functions of (τ, ζ).

Gradient of F^1 and F^2

By direct computations we have,

$$\begin{aligned} \frac{\partial F^1}{\partial \tau} = \frac{\partial}{\partial \tau} Tr[P_i^{-1} \widetilde{Y}_i \widetilde{Y}_i^T] &= -Tr\left[P_i^{-1}\left(\frac{\partial P_i}{\partial \tau}\right) P_i^{-1} \widetilde{Y}_i \widetilde{Y}_i^T\right] \\ &= -e^\tau Tr[P_i^{-2} \widetilde{Y}_i \widetilde{Y}_i^T] = -e^\tau \widetilde{Y}_i^T P_i^{-2} \widetilde{Y}_i; \quad \text{(by (32))}, \end{aligned} \tag{34}$$

$$\begin{aligned} \frac{\partial F^1}{\partial \zeta_k} &= \frac{\partial}{\partial \zeta_k} Tr[P_i^{-1} \widetilde{Y}_i \widetilde{Y}_i^T] = -Tr\left[P_i^{-1}\left(\frac{\partial P_i}{\partial \zeta_k}\right) P_i^{-1} \widetilde{Y}_i \widetilde{Y}_i^T\right] \\ &= -e^{\zeta_k} Tr[P_i^{-1} H_{ik} H_{ik}^T P_i^{-1} \widetilde{Y}_i \widetilde{Y}_i^T] = -e^{\zeta_k} (H_{ik}^T P_i^{-1} \widetilde{Y}_i)^2, \quad \text{(by (33))}. \end{aligned} \tag{35}$$

Also,

$$\frac{\partial F^2}{\partial \tau} = \frac{\partial}{\partial \tau} \log |P_i| = Tr\left[P_i^{-1}\left(\frac{\partial P_i}{\partial \tau}\right)\right] = e^\tau Tr(P_i^{-1}), \quad \text{(by (32))}, \tag{36}$$

$$\frac{\partial F^2}{\partial \zeta_k} = \frac{\partial}{\partial \zeta_k} \log |P_i| = Tr\left[P_i^{-1}\left(\frac{\partial P_i}{\partial \zeta_k}\right)\right] = e^{\zeta_k} H_{ik}^T P_i^{-1} H_{ik}, \quad \text{(by (33))}. \tag{37}$$

Hessian of F^1

From (34),

$$\begin{aligned} \frac{\partial^2 F^1}{\partial \tau^2} &= \frac{\partial}{\partial \tau}\left[-e^\tau \widetilde{Y}_i^T P_i^{-2} \widetilde{Y}_i\right] \\ &= -e^\tau \widetilde{Y}_i P_i^{-2} \widetilde{Y}_i + e^\tau \widetilde{Y}_i^T P_i^{-1}\left(\frac{\partial P_i}{\partial \tau}\right) P_i^{-2} \widetilde{Y}_i + \widetilde{Y}_i^T P_i^{-2}\left(\frac{\partial P_i}{\partial \tau}\right) P_i^{-1} \widetilde{Y}_i \\ &= e^\tau \widetilde{Y}_i^T [2e^\tau P_i^{-3} - P_i^{-2}] \widetilde{Y}_i; \quad \text{(by (32))}. \end{aligned} \tag{38}$$

From (35),

$$\begin{aligned}\frac{\partial^2 F^1}{\partial\tau\partial\zeta_k} &= \frac{\partial}{\partial\tau}\left[-e^{\zeta_k}(H_{ik}^T P_i^{-1}\widetilde{Y}_i)^2\right] \\ &= 2e^{\zeta_k}(\widetilde{Y}_i^T P_i^{-1} H_{ik})\left[H_{ik}^T P_i^{-1}\left(\frac{\partial P_i}{\partial\tau}\right)P_i^{-1}\widetilde{Y}_i\right] \\ &= 2e^{\zeta_k+\tau}\widetilde{Y}_i^T P_i^{-1} H_{ik} H_{ik}^T P_i^{-2}\widetilde{Y}_i; \quad \text{(by (32))}. \end{aligned} \tag{39}$$

Again using (35), and denoting by δ_{kl} the indicator of $\{k=l\}$,

$$\begin{aligned}\frac{\partial^2 F^1}{\partial\zeta_l\partial\zeta_k} &= \frac{\partial}{\partial\zeta_l}\left[-e^{\zeta_k}(H_{ik}^T P_i^{-1}\widetilde{Y}_i)^2\right] \\ &= -\delta_{kl}e^{\zeta_k}(H_{ik}^T P_i^{-1}\widetilde{Y}_i)^2 + 2e^{\zeta_k}(\widetilde{Y}_i^T P_i^{-1} H_{ik})\left[H_{ik}^T P_i^{-1}\left(\frac{\partial P_i}{\partial\zeta_l}\right)P_i^{-1}\widetilde{Y}_i\right] \\ &= -\delta_{kl}e^{\zeta_k}(H_{ik}^T P_i^{-1}\widetilde{Y}_i)^2 + 2e^{\zeta_k+\zeta_l}\widetilde{Y}_i^T P_i^{-1} H_{ik} H_{ik}^T P_i^{-1} H_{il} H_{il}^T P_i^{-1}\widetilde{Y}_i, \quad \text{(by (33))} \\ &= \begin{cases} 2e^{\zeta_k+\zeta_l}(H_{ik}^T P_i^{-1}\widetilde{Y}_i)(H_{il}^T P_i^{-1}\widetilde{Y}_i)(H_{ik}^T P_i^{-1} H_{il}) & \text{if } k\neq l \\ e^{\zeta_k}(H_{ik}^T P_i^{-1}\widetilde{Y}_i)^2\left[2e^{\zeta_k}(H_{ik}^T P_i^{-1} H_{ik})-1\right] & \text{if } k=l; \end{cases} \end{aligned} \tag{40}$$

Hessian of F^2

From (36),

$$\frac{\partial^2 F^1}{\partial\tau^2} = \frac{\partial}{\partial\tau}\left[e^\tau Tr(P_i^{-1})\right] = e^\tau Tr(P_i^{-1}) - e^\tau Tr\left[P_i^{-1}\left(\frac{\partial P_i}{\partial\tau}\right)P_i^{-1}\right] = e^\tau[Tr(P_i^{-1}) - e^\tau Tr(P_i^{-2})]; \tag{41}$$

by (32)). From (37),

$$\frac{\partial^2 F^2}{\partial\tau\partial\zeta_k} = \frac{\partial}{\partial\tau}[e^{\zeta_k} H_{ik}^T P_i^{-1} H_{ik}] = -e^{\zeta_k} H_{ik}^T P_i^{-1}\left(\frac{\partial P_i}{\partial\tau}\right)P_i^{-1} H_{ik} = -e^{\zeta_k+\tau} H_{ik}^T P_i^{-2} H_{ik}; \tag{42}$$

again by (32). Finally,

$$\begin{aligned}\frac{\partial^2 F^2}{\partial\zeta_l\zeta_k} &= \frac{\partial}{\partial\zeta_l}\left[e^{\zeta_k} H_{ik}^T P_i^{-1} H_{ik}\right] \\ &= \delta_{kl}e^{\zeta_k} H_{ik}^T P_i^{-1} H_{ik} - e^{\zeta_k} H_{ik}^T P_i^{-1}\left(\frac{\partial P_i}{\partial\zeta_l}\right)P_i^{-1} H_{ik} \\ &= \delta_{kl}e^{\zeta_k} H_{ik}^T P_i^{-1} H_{ik} - e^{\zeta_k+\zeta_l}(H_{ik}^T P_i^{-1} H_{il})^2, \quad \text{(by (33))} \\ &= \begin{cases} -e^{\zeta_k+\zeta_l}(H_{ik}^T P_i^{-1} H_{il})^2 & \text{if } k\neq l \\ e^{\zeta_k} H_{ik} P_i^{-1} H_{ik}[1 - e^{\zeta_k} H_{ik}^T P_i^{-1} H_{ik}] & \text{if } k=l. \end{cases} \end{aligned} \tag{43}$$