

INVERSE REGRESSION FOR LONGITUDINAL DATA

BY CI-REN JIANG^{‡,*} AND WEI YU[§] AND JANE-LING WANG^{¶,†}

Academia Sinica, [‡] *Genentech Inc.*, [§] and *University of California at Davis*[¶]

Sliced inverse regression (Duan and Li (1991), Li (1991)) is an appealing dimension reduction method for regression models with multivariate covariates. It has been extended by Ferré and Yao (2003, 2005) and Hsing and Ren (2009) to functional covariates where the whole trajectories of random functional covariates are completely observed. The focus of this paper is to develop sliced inverse regression for intermittently and sparsely measured longitudinal covariates. We develop asymptotic theory for the new procedure and show, under some regularity conditions, that the estimated directions attain the optimal rate of convergence. Simulation studies and data analysis are also provided to demonstrate the performance of our method.

1. Introduction. Dimension reduction methods have played a central role in statistical modeling with recent interest directed to functional and longitudinal data. In this paper, we focus on the case when the response is a univariate variable Y , but the covariate $X(\cdot)$ is a stochastic process that is observed intermittently over a time interval \mathcal{I} , possibly at a few follow-up times. Such observed data are often termed “longitudinal” data in the literature in contrast to “functional” data, which are observed densely over a period of time, so essentially one may assume that the entire process X is observed.

To motivate our approach, we first consider dimension reduction approaches for a p -dimensional multivariate covariate \mathbf{X} . There are essentially two paradigms. The first adopts a model that reduces the dimensionality of the nonlinear components; this includes projection pursuit regression (Friedman and Stuetzle (1981), Hall (1989)) and additive models (Stone (1985), Hastie and Tibshirani (1990)). Although flexible, both approaches assume a certain additivity structure in the model and dimension reduction is accomplished during the model fitting stage, often through an iterative back-fitting algorithm. In the second type of dimension reduction approach, one separates the dimension reduction stage from the modeling stage, so that

*Supported in part by NSC grant NSC 101-2118-M-001-013-MY2, Taiwan.

†Supported in part by NSF grants DMS-04-04630, DMS-09-06813, and DMS-12-28369.

AMS 2000 subject classifications: Primary 62G05, 62G08; secondary 62G20

Keywords and phrases: covariance operator, dimension reduction, functional data analysis, local polynomial smoothing, regularization, sparse data

model assumptions are not intertwined with effective dimension reduction. This approach was pioneered by Li (1991) and Duan and Li (1991), who proposed the inverse regression or sliced inverse regression (SIR) approach, which assumes that the information about the response contained in the high dimensional covariates can be summarized in a low dimensional subspace.

Specifically,

$$(1.1) \quad Y = f(\beta'_1 \mathbf{X}, \dots, \beta'_k \mathbf{X}, \epsilon),$$

where the β are k unknown but non-random vectors with $k < p$, f is an arbitrary unknown link function on R^{k+1} , and ϵ is a random error independent of \mathbf{X} . An alternative and equivalent form of (1.1) is:

conditional on $\{\beta'_1 \mathbf{X}, \dots, \beta'_k \mathbf{X}\}$, Y is independent of \mathbf{X} .

Hence, under model (1.1), the k -dimensional variables $\{\beta'_1 \mathbf{X}, \dots, \beta'_k \mathbf{X}\}$ capture all the information contained in the original p -dimensional variable \mathbf{X} for predicting Y . These models work well for most practical situations, as the most interesting features of high dimensional data are usually retrievable from low-dimensional projections. Note that because the link function f is unknown, the regression coefficients $\{\beta_1, \dots, \beta_k\}$ are not identifiable. However, the subspace spanned by them is identifiable and is the “effective dimension reduction” (e.d.r.) space. We also call any direction in the e.d.r. space an e.d.r. direction. The goal here is to estimate those directions that span the e.d.r. space. Many approaches have been proposed to estimate those e.d.r. directions since Li’s pioneering work, including Cook and Li (2002) and approaches that are based on higher order moments (Cook and Weisberg, 1991, Yin and Cook, 2002, 2003). We focus here on Li’s SIR approach, due to its simplicity and originality.

Li (1991) showed that under the design condition,

$$(1.2) \quad E(b' \mathbf{X} | \beta'_1 \mathbf{X}, \dots, \beta'_k \mathbf{X}) \text{ is linear in } \beta'_1 \mathbf{X}, \dots, \beta'_k \mathbf{X},$$

for any direction b in R^p , the covariance matrix $\text{cov}[E(\mathbf{X}|Y)]$ is degenerate in any direction which is $\text{cov}(\mathbf{X})$ -orthogonal to the e.d.r. space spanned by $\{\beta_1, \dots, \beta_k\}$. Here α and β are A -orthogonal means that satisfy $\alpha' A \beta = 0$. Therefore, the e.d.r. directions $\{b_j\}_{j=1}^k$ can be located through the generalized eigen-analysis of $\text{cov}[E(\mathbf{X}|Y)]$ with respect to $\text{cov}(\mathbf{X})$:

$$\text{cov}[E(\mathbf{X}|Y)]b_j = \lambda_j \text{cov}(\mathbf{X})b_j.$$

Since this eigen-analysis only involves $E(\mathbf{X}|Y)$, which marginally is a one-dimensional nonparametric regression problem as compared to the orig-

inal regression $E(Y|\mathbf{X})$, a p -dimensional regression problem, we have accomplished the goal of dimension reduction through an “inverse regression”. Once the e.d.r. space is estimated, standard nonparametric smoothing techniques can then be successfully applied to the k -dimensional covariates $\beta_1'\mathbf{X}, \dots, \beta_k'\mathbf{X}$, provided that k is much smaller than p . The goal of dimension reduction is thus achieved.

So far, we have briefly discussed traditional dimension reduction for a multivariate covariate \mathbf{X} . We will next explore this concept for a functional covariate, where \mathbf{X} is replaced by a random function $X(\cdot) \in L_2(\mathcal{I})$ for an interval $\mathcal{I} \subset R$. The space $L_2(\mathcal{I})$ is a collection of Borel measurable real value functions on \mathcal{I} , such that $E(\|X\|^2) = E(\int_{\mathcal{I}} |X(t)|^2 dt) < \infty$. The modified version of dimension reduction model (1.1) for functional data is:

$$(1.3) \quad Y = f(\langle \beta_1, X \rangle, \dots, \langle \beta_k, X \rangle, \epsilon),$$

where the k unknown functions β_1, \dots, β_k are in $L_2(\mathcal{I})$, f is an arbitrary unknown function on R^{k+1} , and ϵ is independent of $X(t)$. The notation $\langle u, v \rangle$, for any $u, v \in L_2(\mathcal{I})$, is defined as $\langle u, v \rangle = \int_{\mathcal{I}} u(t)v(t)dt$.

Ferré and Yao (2003) were the first to consider such a dimension reduction model and to extend SIR to functional data, termed functional SIR. In a subsequent paper (Ferré and Yao, 2005) they replaced the slicing approach for inverse regression by a nonparametric smoothing method. Further refinements and alternative to functional SIR have been proposed in Ferré and Yao (2007), Forzani and Cook (2007), Cook et al. (2010), and Chen et al. (2011). Hsing and Ren (2009) provided a different formulation for the inverse regression method for a scenario where the predictor $X(t)$ is in a reproducing kernel Hilbert space.

Extending SIR to functional data is nontrivial, due to the complication of inverting a covariance operator on $L_2(\mathcal{I})$. An assumption that is essential for all of these works is that complete trajectories for a sample of n random functions X_1, \dots, X_n are fully observed. This assumption, however, is typically not met in longitudinal studies, as subjects can often only be measured at discrete and scattered time points, which may be random and may vary from subject to subject. Thus, while the observed longitudinal data originate from underlying smooth random functions, the observed data have intrinsically different features (Rice (2004), Hall et al. (2006)). Longitudinal data are also often sparsely sampled with very few measurements per subject.

Together with the irregular sampling plan this poses challenges for the extension of SIR to longitudinal data. We will overcome this difficulty by borrowing information across all subjects in the inverse regression step and by applying smoothing to estimate the inverse regression function, $E(X(t)|Y)$.

Moreover, our proposed procedure, described in Section 2.2, although designed for sparse longitudinal data, can accommodate more densely sampled longitudinal data as well. Asymptotic results for the new procedure are presented in Section 2.3, where Theorem 2.1 implies that the e.d.r. space can be estimated at a rate that corresponds to that of one-dimensional smoothing, when the data are sparse. This is the optimal rate attainable for sparse longitudinal data. We also show that the parametric \sqrt{n} -rate can be achieved by our method for densely sampled longitudinal data (or functional data). Thus, our approach not only resolves the difficulty to adapt SIR for longitudinal data but also provides a unified platform for functional SIR that can handle multiple types of sampling frequency for the longitudinal measurements.

The rest of the paper is organized as follows. In the next section, we state the main approaches, the estimating procedure, and the asymptotic properties. A simulation study and an illustrative data analysis are presented in section 3 and section 4, respectively, to demonstrate the effectiveness of the proposed approach. Section 5 contains concluding remarks. The proofs are relegated to an appendix.

2. Main Approaches and Results. A similar condition as (1.2) is needed for functional data:

$$(2.1) \quad \begin{aligned} & E(\langle b, X \rangle | \langle \beta_1, X \rangle, \dots, \langle \beta_k, X \rangle) \text{ is linear in} \\ & \langle \beta_1, X \rangle, \dots, \langle \beta_k, X \rangle, \text{ for any direction } b \text{ in } L_2(\mathcal{I}). \end{aligned}$$

Let $\Gamma(s, t)$ and $\Gamma_e(s, t)$ denote the covariance operators of $X(t)$ and $E(X(t)|Y)$ respectively. Following similar arguments as those in Li (1991), Ferré and Yao (2003) imply that under assumption (2.1) the operator Γ_e is degenerate in any direction Γ -orthogonal to the e.d.r. space. Thus the basis of the e.d.r. space can be recovered through the Γ -orthonormal eigenvectors of Γ_e , associated with the k largest eigenvalues:

$$\Gamma_e \beta_j = \lambda_j \Gamma \beta_j,$$

where $\beta_i' \Gamma \beta_j = 1$, if $i = j$, and 0 otherwise.

Provided that Γ^{-1} exists, one could perform a spectral decomposition of the operator $\Gamma^{-1} \Gamma_e$ (by requiring $\beta_i' \Gamma \beta_j = 1_{\{i=j\}}$, where $1_{\{\cdot\}}$ is the indicator function), or equivalently of the operator $\Gamma^{-1/2} \Gamma_e \Gamma^{-1/2}$ (by requiring the orthogonal eigenvectors to have norm 1) to locate the e.d.r. directions. However, such an approach poses difficulties for functional data, since the compact covariance operator Γ is not invertible in the functional case. A

practical solution to regularize the estimate of $\Gamma(s, t)$ or $\Gamma_e(s, t)$ was proposed by Ferré and Yao (2003, 2005). Here we provide an alternative approach to define $\Gamma^{-1/2}$ that also illuminates the identifiability issue of the e.d.r. space.

2.1. *Identifiability of the e.d.r. Space.* We have mentioned the identifiability issue of the e.d.r. directions, i.e. the individual vectors β_i are not identifiable. However, the goal of the dimension reduction method is not about estimating the individual directions β_1, \dots, β_k in model (1.1) (or $\beta_1(t), \dots, \beta_k(t)$ in model (1.3)), but rather estimating the e.d.r. space spanned by $\{\beta_1, \dots, \beta_k\}$. Since the Γ -orthonormal eigenfunctions of Γ_e are identifiable, with a little abuse of notation we still use the notation β_i to denote those eigenfunctions and regard them as the targeted e.d.r. directions.

Another identifiability concern for functional inverse regression is related to the invertibility of the covariance operator Γ . A key issue is how to properly define the unbounded operator $\Gamma^{-1/2}$ so that the e.d.r. space can be estimated. Under the assumption that $E(\|X\|^4) < \infty$, $\Gamma(s, t)$ is a self-adjoint, positive semi-definite, Hilbert-Schmidt operator. Therefore, there exists an orthonormal basis $\{\phi_i(t)\}_{i=1}^\infty$ in $L_2(\mathcal{I})$ such that $\Gamma(s, t)$ and $X(t)$ can be represented as:

$$\Gamma(s, t) = \sum_{i=1}^{\infty} \xi_i \phi_i(s) \phi_i(t),$$

where $\{\xi_i\}$ are the eigenvalues of Γ with corresponding eigenfunctions $\phi_i(t)$, and $\{\xi_i\}$ satisfies $\xi_1 \geq \xi_2 \geq \dots \geq \xi_i \geq \dots \geq 0$, and

$$X(t) = \mu(t) + \sum_{i=1}^{\infty} A_i \phi_i(t),$$

where $\{A_i\}$ are uncorrelated random coefficients with $E(A_i) = 0, E(A_i^2) = \xi_i$.

To define $\Gamma^{-1/2}$ and $\Gamma^{-1/2}\Gamma_e\Gamma^{-1/2}$, we consider two scenarios:

1. If for some m , $\xi_m = 0$, the problem is finite dimensional. We can then use the Moore-Penrose generalized inverse $\Gamma^+(s, t)$ instead of Γ^{-1} , where $\Gamma^+(s, t) = \sum_{i=1}^{m-1} \frac{1}{\xi_i} \phi_i(s) \phi_i(t)$. Hence, $\Gamma^{-1/2}(s, t) \triangleq \sum_{i=1}^{m-1} \frac{1}{\sqrt{\xi_i}} \phi_i(s) \phi_i(t)$.
2. If there are infinitely many positive eigenvalues, then $\lim_{i \rightarrow \infty} \xi_i = 0$ and the Moore-Penrose generalized inverse does not exist any more. Thus, we have to consider another way to define the inverse by restricting the operator to a smaller domain. When the following condition is satisfied

(see Section 1 in the supplement (Jiang et al., 2013) for details):

$$(2.2) \quad \sum_{i,j=1}^{\infty} \frac{E^2\{E(A_i|Y)E(A_j|Y)\}}{\xi_i^2 \xi_j} < \infty,$$

a similar argument as in He et al. (2003) shows that $\Gamma^{-1/2}\Gamma_e\Gamma^{-1/2}$ is well defined on the range space of $\Gamma^{1/2}$, which can be represented as:

$$R_{\Gamma^{1/2}} = \{f \in L_2(\mathcal{I}) : \sum_i \xi_i^{-1} |\langle f, \phi_i \rangle|^2 < \infty, f \perp \ker(\Gamma)\},$$

so that for $f \in R_{\Gamma^{1/2}}$,

$$\Gamma^{-1/2}f \triangleq \sum_{i=1}^{\infty} \xi_i^{-1/2} \langle f, \phi_i \rangle \phi_i.$$

Let $\tilde{\Gamma}^{-1/2} = \Gamma^{-1/2}|_{R_{\Gamma^{1/2}}}$ denote such an inverse operator, then the directions we obtain from $\tilde{\Gamma}^{-1/2}\Gamma_e\tilde{\Gamma}^{-1/2}$ are still in the e.d.r. space, since $R_{\Gamma^{1/2}}$ is a subspace of $L_2(\mathcal{I})$.

REMARK 2.1. Condition (2.2) is only a sufficient condition for $\Gamma^{-1/2}\Gamma_e\Gamma^{-1/2}$ to be well-defined. It originates from another sufficient condition (see Section 1 in the supplement (Jiang et al., 2013) for details), that for all $k \geq 1$,

$$(2.3) \quad \sum_i \frac{1}{\xi_i} \left(\sum_j \frac{E\{E(A_i|Y)E(A_j|Y)\}}{(\xi_i \xi_j)^{1/2}} \langle \phi_j, \eta_k \rangle \right)^2 < \infty,$$

which is weaker than (2.2) as can be seen by employing the Cauchy-Schwarz inequality on the left hand side of (2.3). However, equation (2.3) is not easy to interpret, so we focus on (2.2) for interpretation. Simple calculations lead to

$$\begin{aligned} & \sum_{i,j=1}^{\infty} \frac{E^2\{E(A_i|Y)E(A_j|Y)\}}{\xi_i^2 \xi_j} \\ &= \sum_{i,j=1}^{\infty} \frac{1}{\xi_i} \frac{\text{var}(E(A_i|Y))\text{var}(E(A_j|Y))}{\xi_i \xi_j} \text{corr}^2(E(A_i|Y), E(A_j|Y)). \end{aligned}$$

Hence (2.2) is guaranteed if $\text{var}(E(A_i|Y))/\xi_i$ and the correlations between $E(A_i|Y)$ and $E(A_j|Y)$ decrease fast enough. The first requirement is satisfied if the information on Y carried by A_i decreases fast to zero. The second requirement is also not stringent since A_i and A_j are uncorrelated principal components. Below, we provide an example to illustrate (2.2) and when it does not hold.

EXAMPLE 2.1. Let the process $X(t)$ have mean zero and a Karhunan-Loéve expansion, $X(t) = \sum_i A_i \phi_i(t)$, so $E\{X(t)|Y\} = \sum_i E(A_i|Y)\phi_i(t)$.

- If $\xi_i = E(A_i^2) = 1/i^2$, $\text{var}(E(A_i|Y))/\xi_i = 1/i^2$ and $\text{corr}^2\{E(A_i|Y)E(A_j|Y)\} = 1/\{(i+1)^2(j+1)\}$ for $i \neq j$, then $\sum_{i,j} \frac{E^2\{E(A_i|y)E(A_j|y)\}}{\xi_i^2 \xi_j}$ is finite.
- If $\xi_i = E(A_i^2) = 1/i^2$, $\text{var}(E(A_i|Y))/\xi_i = 1/i$ and $\text{corr}^2\{E(A_i|Y)E(A_j|Y)\} = 1/\{(i+1)^2(j+1)\}$ for $i \neq j$, then $\sum_{i,j} \frac{E^2\{E(A_i|Y)E(A_j|Y)\}}{\xi_i^2 \xi_j} = \infty$.

Having resolved the theoretical difficulty with the inverse problem, in practice, the estimation of $\Gamma^{-1/2}$ will involve some regularization. We discuss this in the next subsection.

2.2. *The Methodology.* In reality, longitudinal data are sampled discretely at times T_{ij} from a collection of trajectories $X_i(t)$, $i = 1, \dots, n$, on a compact interval \mathcal{I} . Following common practice, we assume that the $X_i(t)$ are independent realizations from a smooth random function $X(t)$ in $L_2(\mathcal{I})$. The process $X(t)$ has mean function $\mu(t)$ and covariance operator $\Gamma(s, t)$ and the scalar response Y relates to the process $X(t)$ through the relationship described in (1.3).

Let $X_{ij} = X_i(T_{ij})$ be the j th observation of X_i made at time point T_{ij} , where $i = 1, \dots, n$, and $j = 1, \dots, N_i$. The numbers of observations $\{N_i\}_{i=1}^n$ could be prefixed constants or i.i.d. random variables sampled from N , a discrete random variable with integer values. For generality we assume that they are random variables. The ‘‘observation time points’’ $\{T_{ij}\}$ are all in the compact interval \mathcal{I} and assumed to be i.i.d. realizations of a random variable T . In case the T_{ij} are not random, that is, the data are sampled according to a prefixed schedule, our procedure will still work, as long as these time points are dense in \mathcal{I} . Using vector notation $\mathbf{T}_i = (T_{i1}, \dots, T_{iN_i})$, $\mathbf{X}_i = (X_{i1}, \dots, X_{iN_i})$, we can see that $(\mathbf{T}_i, Y_i, \mathbf{X}_i, N_i)$ are i.i.d., and that the $X_i(t)$ and the T_{ij} are independent of each other even if X_{ij} is correlated with T_{ij} .

To estimate the e.d.r. directions, we adopt the idea of inverse regression. We first construct the estimators $\hat{\Gamma}$ and $\hat{\Gamma}_e$ and then estimate the e.d.r. directions by the eigenfunctions of $\hat{\Gamma}_e$ associated with $\hat{\Gamma}$. The specific steps are:

1. Estimation of Γ_e .

For a given time point t and $Y = y$, denote

$$m(t, y) = E(X(t)|Y = y).$$

We assume that $m(t, y)$ is a smooth function, which can thus be estimated via a two dimensional smoothing method applied to the pooled sample $\{X_{ij}\}$ over $\{T_{ij}, Y_i\}$. While any two dimensional smoother can be employed, we use the local linear regression procedure and derive its asymptotic properties in Lemma 2.1. Specifically, our objective is to minimize

$$(2.4) \quad L(t, y) = \sum_{i=1}^n \sum_{j=1}^{N_i} \{X_{ij} - \alpha_0 - \alpha_1(t - T_{ij}) - \alpha_2(y - Y_i)\}^2 K_2\left(\frac{T_{ij} - t}{h_t}, \frac{Y_i - y}{h_y}\right),$$

where $K_2(\cdot, \cdot)$ denotes a bivariate kernel function as defined in (A.3) in the next section, and h_t, h_y are the bandwidths for t and y respectively. Recall that $\Gamma_e = \text{cov}(E(X(t)|Y))$. Once we have estimated $m(t, y)$ at all points t and y , we can estimate the curve $E(X(t)|Y_i)$ and hence Γ_e by the empirical covariance function:

$$\hat{\Gamma}_e(s, t) = \frac{1}{n} \sum_{i=1}^n \hat{m}(s, Y_i) \hat{m}(t, Y_i) - \frac{1}{n} \sum_{i=1}^n \hat{m}(s, Y_i) \frac{1}{n} \sum_{i=1}^n \hat{m}(t, Y_i).$$

2. Estimation of Γ .

Noting that $\Gamma = E(X(s)X(t)) - E(X(s))E(X(t))$, we need to estimate the mean function $\mu(t) = E(X(t))$ and the cross-product $\phi(s, t) = E(X(s)X(t))$.

The mean function $\mu(t)$ can be estimated by a one-dimensional local linear smoothing method applied to the pooled data $\{X_{ij}\}$ over all of the locations $\{T_{ij}\}$, i.e. finding the solution:

$$(\hat{\alpha}_0, \hat{\alpha}_1) = \arg \min_{(\alpha_0, \alpha_1) \in \mathbb{R}^2} \sum_{i=1}^n \sum_{j=1}^{N_i} \{X_{ij} - \alpha_0 - \alpha_1(t - T_{ij})\}^2 K_1\left(\frac{T_{ij} - t}{h_\mu}\right),$$

where $K_1(\cdot)$ is a univariate kernel function defined in (A.3) in next section. The resulting estimate is $\hat{\mu}(t) = \hat{\alpha}_0$.

To estimate $\phi(s, t)$, a two-dimensional local linear smoother will be applied to the cross-products $\{X_{ij}X_{ik}\}$. Again, with

$$\begin{aligned} (\hat{\alpha}_0, \hat{\alpha}_1, \hat{\alpha}_2) = \operatorname{argmin}_{(\alpha_0, \alpha_1, \alpha_2) \in \mathbb{R}^3} & \sum_{i=1}^n \sum_{j,k=1}^{N_i} \{X_{ij}X_{ik} - \alpha_0 - \alpha_1(t - T_{ij}) - \alpha_2(s - T_{ik})\}^2 \\ & \times K_1\left(\frac{T_{ij} - t}{h_\phi}\right) K_1\left(\frac{T_{ik} - t}{h_\phi}\right), \end{aligned}$$

the estimate is $\hat{\phi}(s, t) = \hat{\alpha}_0$. The estimate for Γ is

$$\hat{\Gamma}(s, t) = \hat{\phi}(s, t) - \hat{\mu}(s)\hat{\mu}(t).$$

3. Estimation of the e.d.r. directions $\beta_j, j = 1, \dots, k$

Once we have estimated both Γ_e and Γ , the e.d.r. directions can be estimated through the eigen-analysis of an estimate of $\Gamma^{-1/2}\Gamma_e\Gamma^{-1/2}$. Since the eigen-analysis for an operator can only be performed in reality on a discrete time grid, the actual implementation involves the following steps:

- (a) Discretize $\hat{\Gamma}_e$ and $\hat{\Gamma}$ on an equally-spaced grid $\{t_1, \dots, t_p\}$ to obtain the $p \times p$ matrices, $\hat{\Gamma}_{e,p}$ and $\hat{\Gamma}_{n,p}$;
- (b) Perform the singular value decomposition on $\hat{\Gamma}_{n,p}$ to compute $\hat{\Gamma}_{n,p}^{-1/2}$;
- (c) Perform the eigenvalue decomposition of $\hat{\Gamma}_{n,p}^{-1/2} \hat{\Gamma}_{e,p} \hat{\Gamma}_{n,p}^{-1/2}$ to obtain the first k eigenfunctions $\hat{\eta}_1, \dots, \hat{\eta}_k$ corresponding to the k largest eigenvalues $\hat{\lambda}_1, \dots, \hat{\lambda}_k$;
- (d) $\hat{\beta}_j = \hat{\Gamma}_{n,p}^{-1/2} \hat{\eta}_j, j = 1, \dots, k$.

Here, $\hat{\eta}_j$ are the estimates for the standardized e.d.r. directions η_j , and $\hat{\beta}_j$ for β_j .

REMARK 2.2. If measurements are taken according to the same time schedule for all subjects, it is natural to use this time schedule as the grid in Step 3(a) above. A guiding principle to select the grid size p for general longitudinal applications is to choose it large enough to reveal the characteristics of the function, but not so large as to generate an excessive computational burden or a large p problem. Our experience has been that there is no significance impact on the performance of the approach unless the choice of p is grossly wrong.

The discretization step in 3(a) provides some level of regularization for the inverse operator, $\hat{\Gamma}_{n,p}^{-1/2}$, in step 3(b). We found that additional regularization through truncating the smallest eigencomponents of $\hat{\Gamma}_{n,p}$ is often helpful. A rule of thumb that has worked well in numerical studies is to retain only the first L eigencomponents that explain a desirable fraction of the variation of $\hat{\Gamma}_{n,p}$. The strategy we adopt is to start with a large fraction, say 0.99, and then decrease it gradually until the global pattern of the estimated direction emerges. Such a scheme also automatically excludes components of $\hat{\Gamma}_{n,p}$ that have negative eigenvalues, so the final covariance estimate will

be positive semi-definite and a square root can be taken using the Moore-Penrose generalized inverse. This approach to reconstruct the covariance estimate provides the optimal projection onto the space of positive semi-definite covariance operator as shown in Hall et al. (2008).

2.3. Asymptotic Properties. We present the consistency and rates of convergence of the estimated covariance operators and the e.d.r. directions in this section. The convergence of the two-dimension local linear estimator of $E(X(t)|Y = y)$ and the convergence of the estimated covariance operators, $\hat{\Gamma}_e$, on bounded intervals are key results and of independent interest (Lemmas 2.1 and 2.2). The main result on the convergence of the e.d.r. directions is presented in Theorem 2.1. All proofs are in the Appendix, except for the proof of Lemma 2.2, which is provided in the supplement (Jiang et al., 2013).

The estimators $\hat{\Gamma}$ and $\hat{\Gamma}_e$ have been constructed by the local linear smoothing method. Therefore, it is natural to make the standard smoothness assumptions on the second derivatives of Γ and Γ_e . Assumed that the data $(\mathbf{T}_i, Y_i, \mathbf{X}_i), i = 1, \dots, n$, have the same distribution, where $\mathbf{T}_i = (T_{i1}, \dots, T_{iN_i})$ and $\mathbf{X}_i = (X_{i1}, \dots, X_{iN_i})$. Notice that (T_{ij}, Y_i) and (T_{ik}, Y_i) are dependent but identically distributed, we assume they have the marginal density $g(t, y)$. The assumption (2.1) and condition (2.2) are assumed to hold throughout the paper. Additional assumptions are listed in (A.1)–(A.8) below.

(A.1) The numbers of observations $\{N_i\}$ are independent random variables, with $\overset{i.i.d.}{\sim} N$, where N is a positive integer random variable with $P(N > 1) > 0$. Here, we will view N as a random function of sample size n , and $N(n)$ may go to ∞ as $n \rightarrow \infty$. Furthermore, we assume that $\sup_{n \rightarrow \infty} \frac{E(N(n)^2)}{(EN(n))^2} < \infty$, and $\sup_{n \rightarrow \infty} \frac{E(N(n)^4)}{(EN(n))^3 EN(n)} < \infty$.

These conditions are automatically satisfied when N_i 's are uniformly bounded, i.e. $N(n)$ is uniformly bounded by a given positive integer M such that $P(N(n) < M) = 1, \forall n$. Therefore, assumption (A.1) is intended for non-sparse data only.

(A.2) As mentioned before, $\{\mathbf{T}_i, Y_i, \mathbf{X}_i, N_i\}$ are independent. Furthermore, $\mathbf{T}_i, Y_i, \mathbf{X}_i$ are independent of N_i .

(A.3) Let $K_2(\cdot, \cdot)$ be the bivariate kernel function, which is compactly supported, symmetric, and Hölder continuous. We further assume that it is

a kernel of order $(|v|, |\kappa|)$, i.e.

$$\sum_{\ell_1 + \ell_2 = \ell} \int \int u^{\ell_1} v^{\ell_2} K_2(u, v) du dv = \begin{cases} 0, & 0 \leq \ell < |\kappa|, \ell \neq |v|, \\ (-1)^{|v|} |v|!, & \ell = |v|, \\ \neq 0, & \ell = |\kappa|. \end{cases}$$

Similarly, the univariate kernel function K_1 can be defined. We say that K_1 is of order (m, k) , if

$$\int u^\ell K_1(u) = \begin{cases} 0, & 0 \leq \ell < k, \ell \neq m, \\ (-1)^m m!, & \ell = m, \\ \neq 0, & \ell = k. \end{cases}$$

Both K_1 and K_2 are square integrable, i.e. $\int K_i^2 du < \infty$, $i = 1, 2$.

In our application, we set $|v| = 0, |\kappa| = 2$ for K_2 and $m = 0, k = 2$ for K_1 , but other order could be used by properly adjusting the results.

(A.4) Without loss of generality, we assume that h_t and h_y have the same order:

$$h_t \sim O(h), \quad h_y \sim O(h).$$

(A.5) The bandwidth h satisfies $\lim_{n \rightarrow \infty} h = 0$, $\lim_{n \rightarrow \infty} h EN(n) < \infty$, $\lim_{n \rightarrow \infty} nh^2 EN(n) = \infty$, and $\lim_{n \rightarrow \infty} nh^6 EN(n) < \infty$.

(A.6) Let $g(t, y)$ be the density function of (T_{ij}, Y_i) and $g_3(s, t, y)$ be the joint density function of (T_{ij}, T_{ik}, Y_i) . Let $\Psi(s, t, y)$ be the covariance of $X(S)$ and $X(T)$ given $S = s, T = t$ and $Y = y$. We assume that g, g_3 and Ψ have continuous and bounded second derivatives and that g is bounded away from zero.

(A.7) The bandwidth h_μ satisfies $\lim_{n \rightarrow \infty} h_\mu = 0$, $\lim_{n \rightarrow \infty} h_\mu EN(n) < \infty$, $\lim_{n \rightarrow \infty} nh_\mu^2 EN(n) = \infty$, and $\lim_{n \rightarrow \infty} nh_\mu^6 EN(n) < \infty$. The bandwidth h_ϕ satisfies $\lim_{n \rightarrow \infty} h_\phi = 0$, $\lim_{n \rightarrow \infty} h_\phi EN(n)^3 < \infty$, $\lim_{n \rightarrow \infty} nh_\phi^2 EN(n)^2 = \infty$, and $\lim_{n \rightarrow \infty} nh_\phi^6 EN(n)^2 < \infty$.

(A.8) $E(\|X\|^4) < \infty$.

LEMMA 2.1. *Under assumptions (A.1)–(A.6), we have:*

$$(2.5) \quad \begin{aligned} & E(\hat{m}(t, y) - m(t, y) | \mathbf{T}_i, Y_i, i = 1, \dots, n) \\ &= \frac{\sigma^2}{2} \text{tr}(H \cdot \mathcal{H}_m(t, y)) + o_p(\text{tr}(H)), \end{aligned}$$

$$(2.6) \quad \begin{aligned} & \text{var}(\hat{m}(t, y) | \mathbf{T}_i, Y_i, i = 1, \dots, n) \\ &= \frac{\Psi(t, t, y)}{nEN(n)|H|^{1/2}} \{R(K_2)g(t, y) + \delta g_3(t, t, y)\} / g^2(t, y) + o_p(1), \end{aligned}$$

where $\sigma^2 = \int v^2 K_2(v) dv$, $H = \text{diag}\{h_t^2, h_y^2\}$, $\mathcal{H}_m(t, y)$ is the second derivative of $m(t, y)$, $R(K_2) = \int K_2^2(u) du$, and $\frac{E[N(n)\{N(n)-1\}]}{E\{N(h)\}h_y} |H|^{1/2} \rightarrow \delta$.

When $EN(n) < \infty$, which is the case for longitudinal data, δ is zero. Thus, the variance (2.6) can be simplified to

$$\text{var}(\hat{m}(t, y) | \mathbf{T}_i, Y_i, i = 1, \dots, n) = \frac{\Psi(t, t, y)}{nEN(n)|H|^{1/2}} R(K_2) / g(t, y) + o_p(1).$$

After estimating $E(X(t)|Y)$, Γ_e can be estimated empirically. Specifically,

$$\hat{\Gamma}_e(s, t) = \frac{1}{n} \sum_{i=1}^n [\hat{m}(s, Y_i) - \frac{1}{n} \sum_{j=1}^n \hat{m}(s, Y_j)] [\hat{m}(t, Y_i) - \frac{1}{n} \sum_{j=1}^n \hat{m}(t, Y_j)].$$

From Lemma 2.1, we obtain:

LEMMA 2.2. *Under the assumptions of Lemma 2.1,*

$$\|\hat{\Gamma}_e(s, t) - \Gamma_e(s, t)\| = O_p\left(\frac{1}{\sqrt{nh^2 EN(n)}} + h^2\right),$$

Here $\|\cdot\|$ denotes the operator norm in L_2 .

Lemma 2.1 and Lemma 2.2 imply that we have the same rate of convergence as the conventional case for smoothing two-dimensional independent data. Thus, the within subject dependency causes technical difficulties but one does not pay a price in the convergence rate.

We also need the convergence of $\hat{\Gamma}$.

LEMMA 2.3. *Under assumptions (A.1)–(A.3), and (A.6)–(A.8),*

$$\|\hat{\Gamma}(s, t) - \Gamma(s, t)\| = O_p\left(\frac{1}{\sqrt{nh_\phi^2 EN(n)}} + \frac{1}{\sqrt{nh_\mu EN(n)}} + (h_\mu + h_\phi)^2\right).$$

An immediate application of Lemma 2.1 – Lemma 2.3 leads to:

LEMMA 2.4. *Under assumptions (A.1)–(A.8),*

$$\begin{aligned} \|\hat{\Gamma}^{-1/2}\hat{\Gamma}_e\hat{\Gamma}^{-1/2} - \Gamma^{-1/2}\Gamma_e\Gamma^{-1/2}\| &= O_p(\|\hat{\Gamma} - \Gamma\| + \|\hat{\Gamma}_e - \Gamma_e\|) \\ &= O_p\left(\frac{1}{\sqrt{nh_T^2EN(n)}} + h_T^2\right), \end{aligned}$$

where $h_T = h + h_\mu + h_\phi$.

From Lemma 2.4, we can see that the optimal convergence rate is achieved when the bandwidth $h_T \sim (nEN(n))^{-1/6}$. This is also the optimal rate for the two-dimension smoothing step involved in both Γ and Γ_e .

Once estimates $\hat{\Gamma}_e$ and $\hat{\Gamma}$ for Γ_e and Γ have been obtained, we proceed to estimate the j th eigenfunction η_j based on: $\hat{\Gamma}^{-1/2}\hat{\Gamma}_e\hat{\Gamma}^{-1/2}\hat{\eta}_j = \hat{\lambda}_j\hat{\eta}_j$. From the perturbation theory for linear operators (Kato (1966), Chapter VIII), we readily obtain:

COROLLARY 2.1. *Under assumptions (A.1)–(A.8) and the assumption that the nonzero eigenvalues $\{\hat{\lambda}_j\}$ are distinct, the eigenvector $\hat{\eta}_j$ satisfies*

$$\|\hat{\eta}_j - \eta_j\| = O_p\left(\frac{1}{\sqrt{nh_T^2EN(n)}} + h_T^2\right),$$

for $1 \leq j \leq k$.

The rate of convergence in Corollary 2.1 is the same as the rate of convergence for the covariance estimates of Γ and Γ_e stated in Lemma 2.2 and Lemma 2.3. These rates correspond to the traditional optimal rate for a two-dimensional smoother, and is a consequence of applying perturbation theory. While such a rate is optimal to estimate the covariance operator, it is not optimal for the estimation of eigenfunctions, which should be estimable at the optimal rate for a one-dimensional smoother. With extra technical work, this is indeed achievable and the result is presented in the next theorem.

THEOREM 2.1. *Under assumptions (A.1)–(A.8), we have*

$$(2.7) \quad \|\hat{\eta}_j - \eta_j\| = O_p\left(\frac{1}{\sqrt{nhEN(n)}} + h^2 + \frac{1}{\sqrt{nh_\phi EN(n)}} + h_\phi^2\right),$$

for $1 \leq j \leq k$. Therefore, three types of optimal convergent rates emerge:

- (i) when $EN(n) < \infty$ (sparse longitudinal data), the optimal rate of $\|\hat{\eta}_j - \eta_j\|$ is $O_p(n^{-2/5})$;
- (ii) when $EN(n)h \rightarrow 0$, but $EN(n) \rightarrow \infty$, the optimal rate of $\|\hat{\eta}_j - \eta_j\|$ is $O_p(n^{-r})$, where $2/5 < r < 1/2$;
- (iii) when $EN(n)h < \infty$ (dense longitudinal data or functional data), the optimal rate of $\|\hat{\eta}_j - \eta_j\|$ is $O_p(n^{-1/2})$.

Note that h and h_ϕ are the bandwidths used in the two dimensional local linear smoothing on $\{X_{ij}(T_{ij}, Y_i)\}$ and $\{X_{ij}X_{ik}\}$, where the optimal order of h and h_ϕ is $(nEN(n))^{-1/6}$ according to Lemmas 2.1 – 2.3. Therefore, (2.7) demonstrates the need of undersmoothing in order to estimate η_j optimally.

3. Simulation Studies. Since our method is applicable to both sparse longitudinal data and functional data, we evaluate its finite sample performance for both types of data through simulations. Without loss of generality, we set the domain interval \mathcal{I} as $[0, 1]$. Let $X(t)$ be a standard Brownian motion on $[0, 1]$, we consider the following model:

$$Y = 3 + \exp(\langle \beta(t), X(t) \rangle) + \epsilon,$$

where $\beta(t) = \sqrt{2}\sin(3\pi t/2)$, and the random error $\epsilon \sim N(0, 0.1^2)$. The standard deviation of ϵ may look small, but the range of $\exp(\langle \beta(t), X(t) \rangle)$ is around $(0.5, 1.5)$, so the signal-noise ratio is about 10.

In each run, n sample trajectories, $\{X_i(t), i = 1, \dots, n\}$, are generated from Brownian motion on $[0, 1]$. This forms the complete data, but for practical implementation we discretized the data to equally spaced 31 time-points, $\{t_0, t_1, \dots, t_{30}\}$, with $t_0 = 0$ and $t_{30} = 1$. Therefore, the actual dense data set is $\{X_{ij} = X_i(t_{ij}), i = 1, \dots, n, j = 1, \dots, 30\}$ along with its response Y_i . To generate the sparse longitudinal data, we randomly selected 2 to 10 observations from $\{t_1, t_2, \dots, t_{30}\}$. This results in the longitudinal data $(X_{i1}, \dots, X_{iN_i})$ for the i th subject at time points $(t_{i1}, \dots, t_{iN_i})$, where N_i follows a uniform distribution on $\{2, 3, \dots, 10\}$. The simulation consists of 100 runs and Table 1 summarizes the numerical findings when n is 100 and 200. As a comparison, we also include the results of the smoothed functional inverse regression approach in Ferré and Yao (2005), which is for complete data.

The first comparison is based on the correlation between $\langle \beta(t), X(t) \rangle$ and $\langle \hat{\beta}(t), X(t) \rangle$, i.e. the correlation between the projection of $X(t)$ on the real e.d.r. direction and that on the estimated e.d.r. direction. Averages of those correlations are reported in the third column of Table 1. The results suggest

that our approach generally produces high correlations and for complete data these are larger than those reported in Ferré and Yao (2005). The remaining comparisons are based on the Integrated Squared Bias (ISB), Integrated Variance (IVAR), and Integrated Mean Square Error (IMSE) (or Mean of Integrated Square Error (MISE)). The Appendix contains details of those definitions. The averages of these statistics over the 100 simulation runs are reported in Columns 4-6 of Table 1. As expected, the results for complete data are better than those for sparse data and the results for larger sample sizes are better. For complete data, our procedure generally led to smaller ISB, IVAR, and IMSE than Ferré and Yao's.

In addition to the above global measures, we plot in Figure 1 the mean function for each of the three β -estimates. The left panel of Figure 1 shows the average of $\hat{\beta}(t)$ -functions (dashed line for complete, dotted line for sparse data and dash-dot line for Ferré and Yao (2005)) when $n = 100$ along with the true $\beta(t)$ (solid line), the right panel provides the same plot for $n = 200$. Figure 1 indicates that bias for our approach is comparable to that reported in Ferré and Yao (2005) when data are observed completely. The bias of our approach is significantly reduced for sparse data when the sample size increases to 200, due to improved estimation of Γ and Γ_e .

Upon the request of a referee, we conducted additional simulations with different sample sizes to check the empirical convergence rate of the standardized e.d.r. directions (η_k) through integrated variance (IVAR). Using the same bandwidths and $N_i = 6$ for all i , the ratios of $\sqrt{\text{IVAR}}$ for two consecutive samples (100 v.s. 200 or 200 v.s. 400) are close to $\sqrt{2}$, which is the square root of the ratio of sample sizes (see the supplement (Jiang et al., 2013).)

4. Data Analysis. The data set contains the record of the lifetimes and daily reproduction of female Medflies, the latter quantified by the number of eggs laid daily for 1000 female Mediterranean fruit flies. Details about the experimental background can be found in Carey et al. (1998). Our goal is to explore the relationship between the early pattern of fecundity, quantified by the number of eggs laid per day until day 20, and mortality for each individual fly. For this reason, we exclude flies that died by day 30 and flies that did not lay any eggs. The remaining 647 flies have an average lifetime (Y) of 43.9 days with a standard deviation of 11.9 days. It is assumed that there is an underlying stochastic predicting process $X(t)$ which quantifies the reproduction pattern and can be characterized as a fecundity curve, that is sampled through the daily egg counts. The numbers of eggs laid in the first 20 days are discrete observations of the function $X(t)$. The objective

of our analysis is to find the e.d.r. directions such that the projection of the fecundity curves onto the resulting e.d.r. space will carry the key information for longevity in the regression $E(Y|X)$.

To test the efficiency of our method and to check the effect of sparse data, we first use the complete information of all 20 days as *complete/dense data*; and then randomly pick N_i points from each fly as our *sparse data*, where N_i is uniformly distributed in $\{2, \dots, 10\}$. We also applied the approach in Ferré and Yao (2005) to the complete data as a comparison.

Figure 2 displays the directions estimated by our approach for both complete and sparse data and by the method in Ferré and Yao (2005) for the complete data only. The directions estimated by Ferré and Yao (2005) are less smooth because $\Gamma(s, t)$ was estimated empirically without smoothing. However, the general trends of these directions are similar to ours except for the first index after 15 days ($t > 15$). The global patterns of the direction estimates by our approach are similar between the two types of data and the difference might be due to the difference in the selected bandwidths (a larger bandwidth is used for sparse data to compensate for the sparsity, and this leads to smoother directions.) The estimated $\beta_1(t)$ indicates that daily reproduction during the period day 4 to day 10 plays an important role in mortality, while the estimated $\beta_2(t)$ shows the effect of daily reproduction from day 10 to day 20.

Since the first two eigenfunctions explain over 90% of the variation for both sparse and complete data, two directions suffice to summarize the information contained in the fecundity data to infer lifetime. We further explore the relation of lifetimes with these two directions by assuming that the error ϵ in model (1.3) is additive but the regression relation is unknown. This unknown bivariate regression function is estimated by a bivariate local linear smoother on the estimated bivariate indices ($\langle \hat{\beta}_1(t), X_i(t) \rangle$ and $\langle \hat{\beta}_2(t), X_i(t) \rangle$). Details regarding the bivariate local linear smoother are provided in the data analysis section of the supplement (Jiang et al., 2013). The estimated regression (link) surfaces are provided in Figure 3, where the indices on the right panel were obtained by using the directions estimated from sparse data but the complete covariate $X(t)$ was used to calculate the indices. This facilitates a comparison on the same platform with the other two plots, where the indices were estimated based on complete data.

Since the estimated e.d.r. directions are not identical the ranges of the resulting indices are slightly different. For both sparse and complete data, lifetimes tend to increase with increasing size of the first index when the second index is held fixed. Lifetime is generally longer when the first index is larger and the second index is close to its average value. Averages of the

square fitted errors are provided in Table 2 and are similar for all three methods but interestingly our approach for sparse data performed slightly better than Ferré and Yao (2005)'s approach based on complete data.

Combining Figures 2 and 3, we find that a fly laying fewer eggs from day 4 to day 10 but making it up later by reaching average number of egg production during the period day 10 to day 20 is expected to live longer. Since egg production is most intense in the early stage (day 4 to 10), this suggests a cost of early reproduction to female Medflies. One plausible explanation is that young Medflies are still fragile and reproduction depletes the needed nutrition for growth.

5. Concluding Remarks. In this paper, we propose a new dimension reduction method for longitudinal data collected over discrete, possibly random time points. There are two key steps: the first is a nonparametric smoothing method to borrow information from sparse longitudinal data to estimate the inverse regression function, $E(X(t)|Y)$; the second is the regularization needed to standardize the longitudinal covariates. The method is simple to implement and effective for dimension reduction, and we establish asymptotic theory. In particular, we achieve the optimal rate of convergence for e.d.r. directions. Although the proposed method is inspired by the difficulties caused by sparse longitudinal data, the approach can also handle dense data both theoretically and practically.

The numerical performance of the new approach is examined in a simulation study, where we compare the estimates from dense (complete) and sparse data. While the results for dense data are better than those for sparse data, the estimates for sparse data still capture the main features of the target function. Further, these estimates are consistent with the smoothed patterns of the estimates by Ferré and Yao (2005), and our new approach has much smaller integrated variance with comparable integrated square bias. We also illustrate the effectiveness of the new dimension reduction approach through the fecundity data of Medflies with survival outcome, as only one index or two indices are needed to summarize the longitudinal covariate information. The high correlation between the results from the complete and sparse data further confirms the ability of the method in borrowing information across the entire sample for the case of sparse data.

In practice, one needs to choose k , the dimension of the e.d.r. space. For the Medfly data, we adopted an ad hoc approach to subjectively select k based on the fraction of variance explained by the first few dominant eigen-components. For functional data, a more formal procedure to use a criterion to measure the quality of the estimates for the e.d.r. space has been pro-

posed in Ferré and Yao (2005) as a model selection tool. Another approach based on sequential χ^2 -tests was investigated in Li and Hsing (2010). The choice of k for sparse data would be an interesting topic for future research.

APPENDIX A: DEFINITIONS

1. Integrated square bias (ISB)

$$ISB = \int (E\hat{\beta}(t) - \beta(t))^2 dt \approx \sum_{j=1}^{N_p-1} \left(\bar{\hat{\beta}}(t_j) - \beta(t_j) \right)^2 (t_{j+1} - t_j),$$

where $\bar{\hat{\beta}}(t) = \frac{1}{N_s} \sum_{i=1}^{N_s} \hat{\beta}_i(t)$, and N_p is the number of points used to approximate the integral. In the simulation study, we used $N_s = 100$ and 200 and $N_p = 31$.

2. Integrated variance (IVAR)

$$IVAR = \int E[\hat{\beta}(t)]^2 - (E\hat{\beta}(t))^2 dt \approx \sum_{j=1}^{N_p-1} \left(\bar{\hat{\beta}^2}(t_j) - [\bar{\hat{\beta}}(t_j)]^2 \right) (t_{j+1} - t_j),$$

where $\bar{\hat{\beta}^2}(t) = \frac{1}{N_s} \sum_{i=1}^{N_s} [\hat{\beta}_i(t)]^2$.

3. IMSE and MISE

$$\begin{aligned} IMSE &= \int E(\hat{\beta}(t) - \beta(t))^2 dt \approx \int \frac{1}{N_s} \sum_{i=1}^{N_s} (\hat{\beta}_i(t) - \beta(t))^2 dt \\ &\approx \frac{1}{N_s} \sum_{i=1}^{N_s} \sum_{j=1}^{N_p-1} (\hat{\beta}_i(t_j) - \beta(t_j))^2 (t_{j+1} - t_j) = MISE. \end{aligned}$$

APPENDIX B: PROOFS

For simplicity of notation, we let $\sum_{i,j}$ stand for $\sum_{i=1}^n \sum_{j=1}^{N_i}$ and rewrite formula (2.4) in the main paper as

$$(B.1) \quad (\hat{m}(\mathbf{z}), \hat{\boldsymbol{\alpha}}^T) = \underset{\alpha_0, \boldsymbol{\alpha}}{\operatorname{argmin}} \sum_{i=1}^n \sum_{j=1}^{N_i} \{X_{ij} - \alpha_0 - \boldsymbol{\alpha}^T \mathbf{Z}_{ij}\}^2 K_H(\mathbf{Z}_{ij} - \mathbf{z}),$$

where $\mathbf{z} = (t, y)^T$, $\mathbf{Z}_{ij} = (T_{ij}, Y_i)^T$, and $K_H(\mathbf{u}) = |H|^{-1/2} K(H^{-1/2} \mathbf{u})$, i.e. $K_H(\mathbf{Z}_{ij} - \mathbf{z}) = \frac{1}{h_t h_y} K_2\left(\frac{T_{ij}-t}{h_t}, \frac{Y_i-y}{h_y}\right)$, $H = \operatorname{diag}(h_t^2, h_y^2) \sim h^2 I_{2 \times 2}$ (from assumption (A.4)) and $\boldsymbol{\alpha} = (\alpha_1, \alpha_2)^T$. Here “ \sim ” means “is of the same order as”.

PROOF OF LEMMA 2.1. In order to setup the matrix-vector format of (B.1), we define:

$$\begin{aligned} X &= (\mathbf{X}_1^T, \dots, \mathbf{X}_n^T)^T = (X_{11}, \dots, X_{1N_1}, \dots, X_{n1}, \dots, X_{nN_n})^T, \\ Z &= (\mathbf{Z}_1^T, \dots, \mathbf{Z}_n^T)^T = (Z_{11}, \dots, Z_{1N_1}, \dots, Z_{n1}, \dots, Z_{nN_n})^T, \end{aligned}$$

where $\mathbf{X}_i = (X_{i1}, \dots, X_{iN_i})^T$, and $\mathbf{Z}_i = (Z_{i1}, \dots, Z_{iN_i})^T$.

Thus, define

$$Z_t = \begin{pmatrix} 1 & (\mathbf{z} - \mathbf{Z}_1) \\ 1 & (\mathbf{z} - \mathbf{Z}_2) \\ \vdots & \vdots \\ 1 & (\mathbf{z} - \mathbf{Z}_n) \end{pmatrix},$$

and

$$W = \text{diag}\{K_H(\mathbf{Z}_1 - \mathbf{z}), \dots, K_H(\mathbf{Z}_n - \mathbf{z})\},$$

the estimate $\hat{m}(t, Y) = \hat{m}(\mathbf{z})$ is the weighted least square solution of (B.1) b

$$(B.2) \quad \hat{m}(\mathbf{z}) = \hat{\alpha}_0 = e_1^T (Z_t^T W Z_t)^{-1} Z_t^T W X,$$

where $e_1 = [1, 0, 0]^T$.

We next consider the bias and variance of $\hat{m}(\mathbf{z})$ in two steps.

Step 1: the bias of $\hat{m}(\mathbf{z})$

$$E(\hat{\alpha}_0 | \mathbf{Z}_i, i = 1, \dots, n) = e_1^T (Z_t^T W Z_t)^{-1} Z_t^T W M,$$

where $M = (m(\mathbf{Z}_1), \dots, m(\mathbf{Z}_n))^T$. Apply Taylor expansion of M at \mathbf{z} ,

$$M = Z_t \begin{pmatrix} m(\mathbf{z}) \\ D_m(\mathbf{z}) \end{pmatrix} + \frac{1}{2} Q_m(\mathbf{z}) + R_m(\mathbf{z}),$$

where D_m and Q_m denote the first and second derivative respectively, and $R_m(\mathbf{z})$ is the reminder term. Thus,

$$\begin{aligned} E(\hat{\alpha}_0 | \mathbf{Z}_i, i = 1, \dots, n) &= e_1^T (Z_t^T W Z_t)^{-1} Z_t^T W Z_t \begin{pmatrix} m(\mathbf{z}) \\ D_m(\mathbf{z}) \end{pmatrix} \\ &\quad + e_1^T (Z_t^T W Z_t)^{-1} Z_t^T W \left\{ \frac{1}{2} Q_m(\mathbf{z}) + R_m(\mathbf{z}) \right\} \\ (B.3) \quad &= m(\mathbf{z}) + \frac{1}{2} e_1^T (Z_t^T W Z_t)^{-1} Z_t^T W \{ Q_m(\mathbf{z}) + 2R_m(\mathbf{z}) \}. \end{aligned}$$

We first show that

$$(B.4) \quad Z_t^T W Z_t = \begin{pmatrix} \sum_{i,j} K_H(\mathbf{Z}_{ij} - \mathbf{z}) & \sum_{i,j} K_H(\mathbf{Z}_{ij} - \mathbf{z})(\mathbf{Z}_{ij} - \mathbf{z})^T \\ \sum_{i,j} K_H(\mathbf{Z}_{ij} - \mathbf{z})(\mathbf{Z}_{ij} - \mathbf{z}) & \sum_{i,j} K_H(\mathbf{Z}_{ij} - \mathbf{z})(\mathbf{Z}_{ij} - \mathbf{z})(\mathbf{Z}_{ij} - \mathbf{z})^T \end{pmatrix},$$

is of the order nEN and find its leading term. Since $\{(N_i, \mathbf{T}_i, Y_i), i = 1, \dots, n\}$ and $\{\sum_{j=1}^{N_i} K_H(\mathbf{Z}_{ij} - \mathbf{z}), i = 1, \dots, n\}$ are i.i.d. and $N_i = N_i(n)$ and $Z_{ij} = Z_{ij}(n)$ are functions of n , we will apply classical limit theorems to the triangular array $\sum_{i,j} K_H(\mathbf{Z}_{ij} - \mathbf{z})$ in the first entry of (B.4).

The expectation of $\sum_{i,j} K_H(\mathbf{Z}_{ij} - \mathbf{z})$ is:

$$\begin{aligned} nE(E(\sum_{j=1}^N K_H(\mathbf{Z}_j - \mathbf{z})|N)) &= (nEN) \cdot E(K_H(\mathbf{Z} - \mathbf{z})) \\ &= (nEN) \cdot \int K(\mathbf{v})g(\mathbf{z} + H^{1/2}\mathbf{v})d\mathbf{v} \\ &= (nEN) \cdot (g(\mathbf{z}) + o_p(1)), \end{aligned}$$

where the last step follows from the Taylor expansion of $g(\mathbf{z} + H^{1/2}\mathbf{v})$ at \mathbf{z} .

The variance of $\sum_{i,j} K_H(\mathbf{Z}_{ij} - \mathbf{z})$ is:

$$\begin{aligned} &\sum_{i=1}^n \text{var}(\sum_{j=1}^{N_i} K_H(\mathbf{Z}_{ij} - \mathbf{z})) \\ &= n\{E(\sum_{j=1}^N \sum_{k=1}^N K_H(\mathbf{Z}_j - \mathbf{z})K_H(\mathbf{Z}_k - \mathbf{z})) - (E(\sum_{j=1}^N K_H(\mathbf{Z}_j - \mathbf{z}))^2)\} \\ &= I_1 - \mu_n^2/n, \end{aligned}$$

where

$$\begin{aligned} I_1 &= n\{E(\sum_{j=1}^N (K_H(\mathbf{Z}_j - \mathbf{z}))^2 + \sum_{j \neq k} K_H(\mathbf{Z}_j - \mathbf{z})K_H(\mathbf{Z}_k - \mathbf{z}))\} \\ &= n\{E(E(\sum_{j=1}^N (K_H(\mathbf{Z}_j - \mathbf{z}))^2|N)) + E(E(\sum_{j \neq k} K_H(\mathbf{Z}_j - \mathbf{z})K_H(\mathbf{Z}_k - \mathbf{z})|N))\} \\ &= (nEN) \cdot E(K_H(\mathbf{Z} - \mathbf{z}))^2 + (nEN(N-1)) \cdot E(K_H(\mathbf{Z} - \mathbf{z})K_H(\mathbf{Z}' - \mathbf{z})). \end{aligned}$$

Condition (A.1) implies that $\text{var}(\sum_{i,j} K_H(\mathbf{Z}_{ij} - \mathbf{z}))/nEN \rightarrow 0$, hence,

$$\frac{\sum_{i,j} K_H(\mathbf{Z}_{ij} - \mathbf{z}) - E(\sum_{i,j} K_H(\mathbf{Z}_{ij} - \mathbf{z}))}{nEN} \rightarrow 0 \quad \text{in probability,}$$

i.e.

$$\sum_{i,j} K_H(\mathbf{Z}_{ij} - \mathbf{z}) = (nEN) \cdot (g(\mathbf{z}) + o_p(1)).$$

The mean of the other entries in the second row of (B.4) can be handled similarly with

$$\begin{aligned} \frac{1}{nEN} E\left\{\sum_{i,j} K_H(\mathbf{Z}_{ij} - \mathbf{z})(\mathbf{Z}_{ij} - \mathbf{z})\right\} &= EK_H(\mathbf{u} - \mathbf{z})(\mathbf{u} - \mathbf{z}) \\ &= H\nabla g \int K_2(\mathbf{v})\mathbf{v}\mathbf{v}^T d\mathbf{v} + o_p(H \cdot \mathbf{1}), \end{aligned}$$

where $\mathbf{1}$ is a column vector of length 2 with all entries equal to 1, and

$$\begin{aligned} \frac{1}{nEN} E\left\{\sum_{i,j} K_H(\mathbf{Z}_{ij} - \mathbf{z})(\mathbf{Z}_{ij} - \mathbf{z})(\mathbf{Z}_{ij} - \mathbf{z})^T\right\} \\ &= EK_H(\mathbf{u} - \mathbf{z})(\mathbf{u} - \mathbf{z})(\mathbf{u} - \mathbf{z})^T \\ &= Hg(\mathbf{z}) + o_p(H). \end{aligned}$$

The variances of these entries in the second row of (B.4) can be dealt with as the variance of $\sum_{i,j} K_H(\mathbf{Z}_{ij} - \mathbf{z})$, so we omit the details here and summarize the findings for the rate of convergence and leading terms as:

$$\frac{1}{nEN} Z_t^T W Z_t = \begin{pmatrix} g(\mathbf{z}) + o_p(1) & (H\nabla g)^T + o_p((H \cdot \mathbf{1})^T) \\ H\nabla g + o_p(H \cdot \mathbf{1}) & Hg(\mathbf{z}) + o_p(H) \end{pmatrix}.$$

We next consider the inverse $(\frac{1}{nEN} Z_t^T W Z_t)^{-1}$. To locate the leading terms, we apply the well-known formula for a matrix inverse in block form,

$$\begin{pmatrix} a & \alpha^T \\ \alpha & B \end{pmatrix}^{-1} = \begin{pmatrix} \frac{1}{a} + \frac{1}{a^2} \alpha^T (B - \frac{\alpha\alpha^T}{a})^{-1} \alpha & -\frac{1}{a} \alpha^T (B - \frac{\alpha\alpha^T}{a})^{-1} \\ -(B - \frac{\alpha\alpha^T}{a})^{-1} \alpha \frac{1}{a} & (B - \frac{\alpha\alpha^T}{a})^{-1} \end{pmatrix}$$

where a is a scalar, α is a column vector and B is a submatrix. Applying this formula to the matrix $\frac{1}{nEN} Z_t^T W Z_t$, we first obtain

$$\begin{aligned} (B - \frac{\alpha\alpha^T}{a})^{-1} &= \{Hg(\mathbf{z}) + o_p(H) - \frac{(H\nabla g + o_p(H \cdot \mathbf{1}))(H\nabla g + o_p(H \cdot \mathbf{1}))^T}{g(\mathbf{z}) + o_p(1)}\}^{-1} \\ &= (Hg(\mathbf{z}))^{-1} + o_p(H^{-1}). \end{aligned}$$

Thus,

$$\begin{aligned} -(B - \frac{\alpha\alpha^T}{a})^{-1} \alpha &= -((Hg(\mathbf{z}))^{-1} + o_p(H^{-1})) \frac{H\nabla g + o_p(H \cdot \mathbf{1})}{g(\mathbf{z}) + o_p(1)} \\ &= -\frac{\nabla g}{g^2(\mathbf{z})} + o_p(1), \end{aligned}$$

and the first entry of $(\frac{1}{nEN}Z_t^T W Z_t)^{-1}$ becomes

$$\begin{aligned} & \frac{1}{a} + \frac{1}{a^2}\alpha^T(B - \frac{\alpha\alpha^T}{a})^{-1}\alpha \\ &= \frac{1}{g(\mathbf{z}) + o_p(1)} \left[1 + \frac{1}{g(\mathbf{z}) + o_p(1)} (H\nabla g + o_p(H \cdot \mathbf{1}))^T (\frac{\nabla g}{g^2(\mathbf{z})} + o_p(1)) \right] \\ &= \frac{1}{g(\mathbf{z})} + o_p(1). \end{aligned}$$

Therefore, we obtain

$$(B.5) \quad \left(\frac{1}{nEN}Z_t^T W Z_t \right)^{-1} = \begin{pmatrix} g^{-1}(\mathbf{z}) + o_p(1) & -\frac{(\nabla g)^T}{g^2(\mathbf{z})} + o_p(\mathbf{1}^T) \\ -\frac{\nabla g}{g^2(\mathbf{z})} + o_p(\mathbf{1}) & (Hg(\mathbf{z}))^{-1} + o_p(H^{-1}) \end{pmatrix}.$$

Finally, we consider the rate of convergence and the leading term of $Z_t^T W Q_m(\mathbf{z})$ in (B.3). Since

$$Q_m(\mathbf{z}) = [(\mathbf{Z}_{11} - \mathbf{z})^T H_m(\mathbf{z})(\mathbf{Z}_{11} - \mathbf{z}), \dots, (\mathbf{Z}_{nN_n} - \mathbf{z})^T H_m(\mathbf{z})(\mathbf{Z}_{nN_n} - \mathbf{z})]^T,$$

we have

$$(B.6) \quad \frac{1}{nEN}Z_t^T W Q_m(\mathbf{z}) = \left(\frac{1}{nEN} \sum_{i,j} K_H(\mathbf{Z}_{ij} - \mathbf{z})(\mathbf{Z}_{ij} - \mathbf{z})^T H_m(\mathbf{z})(\mathbf{Z}_{ij} - \mathbf{z}) \right. \\ \left. - \frac{1}{nEN} \sum_{i,j} K_H(\mathbf{Z}_{ij} - \mathbf{z})(\mathbf{Z}_{ij} - \mathbf{z})^T H_m(\mathbf{z})(\mathbf{Z}_{ij} - \mathbf{z})(\mathbf{Z}_{ij} - \mathbf{z}) \right).$$

The second term in (B.6) is $O_p(H^{3/2} \cdot \mathbf{1})$, hence

$$\begin{aligned} & E(\hat{m}(\mathbf{z}) - m(\mathbf{z}) | \mathbf{T}_i, Y_i, i = 1, \dots, n) \\ &= \frac{1}{2}g^{-1}(\mathbf{z})E\left[\frac{1}{nEN} \sum_{i,j} K_H(\mathbf{Z}_{ij} - \mathbf{z})(\mathbf{Z}_{ij} - \mathbf{z})^T H_m(\mathbf{z})(\mathbf{Z}_{ij} - \mathbf{z})\right] \\ &+ \left[-\frac{(\nabla g)^T}{g^2(\mathbf{z})} + o_p(\mathbf{1}^T)\right]O_p(H^{3/2} \cdot \mathbf{1}) \\ &= \frac{1}{2}g^{-1}(\mathbf{z}) \int K_2(\mathbf{v})\mathbf{v}^T H^{1/2} H_m(\mathbf{z}) H^{1/2} \mathbf{v} g(\mathbf{z} + H^{1/2}\mathbf{v}) d\mathbf{v} + o_p(\text{tr}(H)) \\ &= \frac{\sigma^2}{2} \text{tr}(H H_m(\mathbf{z})) + o_p(\text{tr}(H)). \end{aligned}$$

Step 2: order of the variance (2.6). From (B.2), we know that

$$(B.7) \quad \text{var}(\hat{m}(\mathbf{z}) | \mathbf{T}_i, Y_i, i = 1, \dots, n) = e_1^T (Z_t^T W Z_t)^{-1} Z_t W \Sigma W Z_t (Z_t^T W Z_t)^{-1} e_1,$$

where $\Sigma = \text{var}(X | \mathbf{T}_i, Y_i, i = 1, \dots, n) = \text{diag}\{\Sigma_1, \dots, \Sigma_n\}$, with $\Sigma_i = \text{var}(\mathbf{X}_i | \mathbf{T}_i, Y_i) = \{\sigma_{ijk}\}$.

Let $\sum_{i,j,k}$ stand for $\sum_{i=1}^n \sum_{j=1}^{N_i} \sum_{k=1}^{N_i}$ and define

$$(B.8) \quad \left(\frac{1}{nEN} \right) Z_t^T W \Sigma W Z_t \triangleq \begin{pmatrix} d_{11} & d_{12} \\ d_{21} & d_{22} \end{pmatrix},$$

where

$$\begin{aligned} d_{11} &= \left(\frac{1}{nEN} \right) \sum_{i,j,k} K_H(\mathbf{Z}_{ij} - \mathbf{z}) K_H(\mathbf{Z}_{ik} - \mathbf{z}) \sigma_{ijk}, \\ d_{12} &= \left(\frac{1}{nEN} \right) \sum_{i,j,k} K_H(\mathbf{Z}_{ij} - \mathbf{z}) K_H(\mathbf{Z}_{ik} - \mathbf{z}) (\mathbf{Z}_{ik} - \mathbf{z})^T \sigma_{ijk}, \\ d_{21} &= \left(\frac{1}{nEN} \right) \sum_{i,j,k} K_H(\mathbf{Z}_{ij} - \mathbf{z}) K_H(\mathbf{Z}_{ik} - \mathbf{z}) (\mathbf{Z}_{ik} - \mathbf{z}) \sigma_{ijk}, \\ d_{22} &= \left(\frac{1}{nEN} \right) \sum_{i,j,k} K_H(\mathbf{Z}_{ij} - \mathbf{z}) K_H(\mathbf{Z}_{ik} - \mathbf{z}) (\mathbf{Z}_{ij} - \mathbf{z}) (\mathbf{Z}_{ik} - \mathbf{z})^T \sigma_{ijk}, \end{aligned}$$

and $\sigma_{ijk} = \text{cov}(X_{ij}, X_{ik} | \mathbf{T}_i, Y_i, i = 1, \dots, n)$.

The first entry of (B.8) is:

$$\begin{aligned} d_{11} &= E \left[\frac{1}{nEN} \sum_{i,j,k} K_H(\mathbf{Z}_{ij} - \mathbf{z}) K_H(\mathbf{Z}_{ik} - \mathbf{z}) \sigma_{ijk} \right] + o_p(1) \\ &= E \left[\frac{1}{nEN} \sum_{i,j} K_H(\mathbf{Z}_{ij} - \mathbf{z}) K_H(\mathbf{Z}_{ij} - \mathbf{z}) \Psi(T_{ij}, T_{ij}, Y_i) \right] \\ &\quad + E \left[\frac{1}{nEN} \sum_{i=1}^n \sum_{j \neq k}^{N_i} K_H(\mathbf{Z}_{ij} - \mathbf{z}) K_H(\mathbf{Z}_{ik} - \mathbf{z}) \Psi(T_{ij}, T_{ik}, Y_i) \right] + o_p(1) \\ &= I_1 + I_2 + o_p(1), \end{aligned}$$

where

$$\begin{aligned} I_1 &= E[K_H^2(U - \mathbf{z}) \Psi(T, T, Y)] \\ &= |H|^{-1/2} \int \int K_2^2(\mathbf{u}) \Psi(t + h_t t_1, t + h_t t_1, y + h_y y_1) g(\mathbf{z} + H^{1/2} \mathbf{u}) d\mathbf{u} + o_p(|H|^{1/2}) \\ &= |H|^{-1/2} R(K_2) \Psi(t, t, y) g(\mathbf{z}) + o_p(|H|^{-1/2}), \end{aligned}$$

with $R(K_2) = \int K_2^2(\mathbf{u}) d\mathbf{u}$, and

$$\begin{aligned} I_2 &= \frac{EN(N-1)}{EN} \int K_H(\mathbf{u} - \mathbf{z}) K_H(\mathbf{v} - \mathbf{z}) \Psi(s_1, t_1, y_1) g_3(s_1, t_1, y_1) ds_1 dt_1 dy_1 \\ &= \frac{EN(N-1)}{ENh_y} \Psi(t, t, y) g_3(t, t, y) + o_p(|H|^{-1/2}), \end{aligned}$$

where $\mathbf{u} = (s_1, y_1)^T$, $\mathbf{v} = (t_1, y_1)^T$, g_3 is the joint distribution of T_{ij} , T_{ik} and Y_i , as defined in (A.6).

Thus, $d_{11} = |H|^{-1/2} \Psi(t, t, y) \{R(K_2)g(\mathbf{z}) + \delta g_3(t, t, y)\} + o_p(|H|^{-1/2})$, where $\delta = \frac{EN(N-1)}{ENh_y} |H|^{1/2}$. The limit of δ exists when Assumptions (A.1) and (A.5) hold.

Similarly, we can obtain d_{21} and d_{22} as follows

$$\begin{aligned} d_{21} &= O_p(|H|^{-1/2}H) + o_p(|H|^{-1/2}H), \\ d_{22} &= |H|^{-1/2}H \int K_2^2(\mathbf{u})\mathbf{u}\mathbf{u}^T d\mathbf{u} \Psi(t, t, y)g(\mathbf{z}) + o_p(|H|^{-1/2}H), \end{aligned}$$

thus (B.8) is obtained.

Finally, we obtain (2.6) in the main paper by plugging the results of (B.5) and (B.8) into (B.7). □

PROOF OF LEMMA 2.2. Due to space limitation, the proof is provided in the supplement (Jiang et al., 2013) of this paper. □

PROOF OF LEMMA 2.3. Similar steps as in Lemma 2.1 and Lemma 2.2 can be adopted and we omit the details. □

PROOF OF THEOREM 2.1. Recall that $\hat{\eta}_j$ is the estimate of the j th standardized e.d.r. direction, and satisfies the equation $\hat{\Gamma}^{-1/2} \hat{\Gamma}_e \hat{\Gamma}^{-1/2} \hat{\eta}_j = \hat{\lambda}_j \hat{\eta}_j$. By the definition of $\Gamma^{-1/2}$ in Section 2.1, $\Gamma^{-1/2} \Gamma_e \Gamma^{-1/2}$ is a nonnegative symmetric Hilbert-Schmidt operator, which can be interpreted as the kernel of a linear mapping on $L_2(\mathcal{I})$. As for Γ , a spectral decomposition of $\Gamma^{-1/2} \Gamma_e \Gamma^{-1/2}$ can be achieved from Mercer's theorem as

$$\Gamma^{-1/2} \Gamma_e \Gamma^{-1/2}(s, t) = \sum_{j=1}^k \lambda_j \eta_j(s) \eta_j(t),$$

where $\lambda_1 > \lambda_2 > \dots > \lambda_k > 0$, and $\{\eta_j\}_{j=1}^k \cup \{\eta_j\}_{j=k+1}^\infty$ will generate a complete orthogonal basis of $L_2(\mathcal{I})$.

Once we estimate Γ and Γ_e by $\hat{\Gamma}$ and $\hat{\Gamma}_e$, the operator $\hat{\Gamma}^{-1/2} \hat{\Gamma}_e \hat{\Gamma}^{-1/2}$, which is symmetric and Hilbert-Schmidt, has the empirical expansion:

$$\hat{\Gamma}^{-1/2} \hat{\Gamma}_e \hat{\Gamma}^{-1/2}(s, t) = \sum_{j=1}^k \hat{\lambda}_j \hat{\eta}_j(s) \hat{\eta}_j(t).$$

Since $\{\eta_j\}_{j=1}^\infty$ is a complete orthogonal basis, $\hat{\eta}_j$ may be written as $\hat{\eta}_j = \sum_{\ell \geq 1} \hat{a}_{j\ell} \eta_\ell$. Let

$$\Delta = \hat{\Gamma}^{-1/2} \hat{\Gamma}_e \hat{\Gamma}^{-1/2}(s, t) - \Gamma^{-1/2} \Gamma_e \Gamma^{-1/2}(s, t).$$

Following similar arguments as in Lemma 1 and Lemma 2 of Hall et al. (2006), we can arrive at:

$$(B.9) \quad \|\hat{\eta}_j - \eta_j\|^2 = D_{j2} + \lambda_j^{-2} \|\Delta\|_{(j)}^2 - \lambda_j^{-2} \left(\int_{\mathcal{I} \times \mathcal{I}} \Delta \eta_j \eta_j \right)^2 + O_p(\|\Delta\| \|\Delta\|_{(j)}^2),$$

where

$$\begin{aligned} D_{j2} &= \sum_{\ell: \ell \neq j} \{(\lambda_j - \lambda_\ell)^{-2} - \lambda_j^{-2}\} \left(\int_{\mathcal{I} \times \mathcal{I}} \Delta \eta_j \eta_\ell \right)^2; \\ \|\Delta\|_{(j)}^2 &\triangleq \int_{\mathcal{I}} \left\{ \int_{\mathcal{I}} \Delta(s, t) \eta_j(t) dt \right\}^2 ds \\ &= \sum_{\ell=1}^{\infty} \left(\int_{\mathcal{I} \times \mathcal{I}} \Delta \eta_j \eta_\ell \right)^2. \end{aligned}$$

As λ_j are distinct, $D_{j2} = O_p(\|\Delta\|_{(j)}^2)$. Thus, in order to see the asymptotic performance of $\hat{\eta}_j$, it suffices to evaluate $\|\Delta\|_{(j)}$ and $\int_{\mathcal{I} \times \mathcal{I}} \Delta \eta_j \eta_j$.

Consider $\|\Delta\|_{(j)}$. Using the sandwich technique and the fact that Γ and Γ_e are continuous operators on compact support $\mathcal{I} \times \mathcal{I}$, we have

$$(B.10) \quad \begin{aligned} \|\Delta\|_{(j)}^2 &= \|\hat{\Gamma}^{-1/2} \hat{\Gamma}_e \hat{\Gamma}^{-1/2}(s, t) - \Gamma^{-1/2} \Gamma_e \Gamma^{-1/2}(s, t)\|_{(j)}^2 \\ &= O_p(\|\hat{\Gamma} - \Gamma\|_{(j)}^2 + \|\hat{\Gamma}_e - \Gamma_e\|_{(j)}^2). \end{aligned}$$

Similar arguments as in the proof of Theorem 1 in Hall et al. (2006) imply

$$\|\hat{\Gamma} - \Gamma\|_{(j)}^2 = O_p\left(\frac{1}{nh_\phi EN} + h_\phi^4\right).$$

Hereafter, for simplicity, we assume that $EX(t) = 0$ so that

$$\Gamma_e = E\{E(X(s)|Y)E(X(t)|Y)\} = E(m(s, Y)m(t, Y)).$$

From the estimation procedure,

$$(B.11) \quad \begin{aligned} \hat{\Gamma}_e - \Gamma_e &= \frac{1}{n} \sum_{i=1}^n \hat{m}(s, Y_i) \hat{m}(t, Y_i) - \Gamma_e \\ &= \left\{ \frac{1}{n} \sum_{i=1}^n \hat{m}(s, Y_i) \hat{m}(t, Y_i) - \frac{1}{n} \sum_{i=1}^n m(s, Y_i) m(t, Y_i) \right\} + \left\{ \frac{1}{n} \sum_{i=1}^n m(s, Y_i) m(t, Y_i) - \Gamma_e \right\} \\ &\triangleq \Delta_1 + \Delta_2, \end{aligned}$$

where $\Delta_2 = \frac{1}{n} \sum_{i=1}^n m(s, Y_i)m(t, Y_i) - \Gamma_e = O_p(\frac{1}{\sqrt{n}})$.

Therefore,

$$(B.12) \quad \|\hat{\Gamma}_e - \Gamma_e\|_{(j)}^2 \leq \|\Delta_1\|_{(j)}^2 + \|\Delta_2\|_{(j)}^2 = \|\Delta_1\|_{(j)}^2 + O_p(\frac{1}{n}).$$

Since $\Delta_1 = \frac{1}{n} \sum_{i=1}^n \{\hat{m}(s, Y_i)\hat{m}(t, Y_i) - m(s, Y_i)m(t, Y_i)\}$, it suffices to consider $\|\delta\|_{(j)}^2$, where

$$(B.13) \quad \delta = \hat{m}(s, y)\hat{m}(t, y) - m(s, y)m(t, y).$$

A Taylor expansion on $X_{ik}(T_{ik}, Y_i)$ at the point (t, y) leads to:

$$(B.14) \quad \begin{aligned} X_{ik}(T_{ik}, Y_i) &= X_i(t, y) + (t - T_{ik})X_i^{1,t}(t, y) + (y - Y_i)X_i^{1,y}(t, y) \\ &\quad + \frac{1}{2}(t - T_{ik}, y - Y_i) \nabla (\nabla X_i(t_{ik}, y_i))(t - T_{ik}, y - Y_i)^T + O((t - T_{ik}, y - Y_i)^3) \\ &= X_{ik}^{[1]} + X_{ik}^{[2]} + X_{ik}^{[3]} + X_{ik}^{[4]}, \end{aligned}$$

where $X_i^{1,\cdot}$ is defined as the first derivative of $X_i(t, y)$ with respect to the corresponding variable, ∇ is the gradient of $X_i(t, y)$, t_{ik} and y_i are between t and T_{ik} , and y and Y_i , respectively. Note that $X_{ik}^{[1]}(t, y)$ is $X_i(t, y)$, $X_{ik}^{[2]} = (t - T_{ik})X_i^{1,t}(t, y)$, $X_{ik}^{[3]} = (y - Y_i)X_i^{1,y}(t, y)$, and $X_{ik}^{[4]}$ is the remaining terms in (B.14). Although we do not claim that t has to be close to T_{ik} and y to Y_i in the Taylor expansion (B.14), only those $\{X_{ik}\}$, whose corresponding (T_{ik}, Y_i) satisfy $|t - T_{ik}| \leq h_t$ and $|y - Y_i| \leq h_y$, will contribute to the estimation when the kernel weights of local linear smoother are applied. This provides the correct order of the Taylor expansion (B.14), to be elaborated below.

As defined above, $X_{ik}^{[2]}$ and $X_{ik}^{[3]}$ have the linear term of $t - T_{ik}$ and $y - Y_i$ whose order is h , and $X_{ik}^{[4]}$ contains the quadratic terms with order h^2 . Applying the local linear smoother in (B.1) to (B.14), we obtain

$$(B.15) \quad \hat{m}(t, y) = \hat{m}^{[1]}(t, y) + \hat{m}^{[2]}(t, y) + \hat{m}^{[3]}(t, y) + \hat{m}^{[4]}(t, y),$$

where $\hat{m}^{[\ell]}$, $\ell = 1, \dots, 4$ are the corresponding smoothers on $\{X_{ik}^{[\ell]}\}$.

Let E' denote expectation conditioned on $\{(\mathbf{T}_i, Y_i), i = 1, \dots, n\}$. Com-

binning (B.13) to (B.15), we deduce that

$$\begin{aligned}
 (B.16) \quad \delta &= \hat{m}(s, y)\hat{m}(t, y) - E'\{\hat{m}(s, y)\hat{m}(t, y)\} + E'\{\hat{m}(s, y)\hat{m}(t, y)\} - m(s, y)m(t, y) \\
 &= (\hat{m}^{[1]}(s, y) + \cdots + \hat{m}^{[4]}(s, y))(\hat{m}^{[1]}(t, y) + \cdots + \hat{m}^{[4]}(t, y)) \\
 &\quad - E'\{(\hat{m}^{[1]}(s, y) + \cdots + \hat{m}^{[4]}(s, y))(\hat{m}^{[1]}(t, y) + \cdots + \hat{m}^{[4]}(t, y))\} \\
 &\quad + E'\{\hat{m}(s, y)\hat{m}(t, y)\} - m(s, y)m(t, y) \\
 &= \delta_1 + \delta_2 + \delta_3 + \delta_4 + \delta_5 + \delta_0,
 \end{aligned}$$

where

$$\begin{aligned}
 (B.17) \quad \delta_0 &= E'\{\hat{m}(s, y)\hat{m}(t, y)\} - m(s, y)m(t, y), \\
 \delta_1 &= \hat{m}^{[1]}(s, y)\hat{m}^{[1]}(t, y) - E'\{\hat{m}^{[1]}(s, y)\hat{m}^{[1]}(t, y)\}, \\
 \delta_2 &= \sum -E' \sum, \text{ where } \sum \text{ is the sum of all the linear terms of } (t - T_{ik}, y - Y_i), \\
 \delta_3 &= \sum -E' \sum, \text{ where } \sum \text{ is the sum of all the quadratic terms of } (t - T_{ik}, y - Y_i), \\
 \delta_4 &= \sum -E' \sum, \text{ where } \sum \text{ is the sum of all the cubic terms of } (t - T_{ik}, y - Y_i), \\
 \delta_5 &= \sum -E' \sum, \text{ where } \sum \text{ is the sum of all the quartic terms of } (t - T_{ik}, y - Y_i),
 \end{aligned}$$

Standard arguments in the proof of Lemma 2.1 can be applied to δ_0 to show that δ_0 has the same convergence rate as the bias of $\hat{m}(t, y)$, i.e. $O_p(h^2)$.

Based on the results and proofs in Lemma 2.1 and Lemma 2.2, the same claims as in *Step* (iii) of Hall et al. (2006) can be made for $\{\delta_\ell\}$, $\ell = 1, \dots, 5$. As a result, and from (B.16) and (B.17), we have

$$(B.18) \quad \|\delta\|_{(j)}^2 = E'\|\delta_1\|_{(j)}^2 + h^4.$$

Furthermore, if we define

$$E'\|\delta_1\|_{(j)}^2 = \int_{\mathcal{I}} \int \int_{\mathcal{I}^2} E'\{\delta_1(s, t_1, y)\delta_1(s, t_2, y)\}\eta_j(t_1)\eta_j(t_2)dt_1dt_2ds,$$

and apply the properties of two dimensional linear smoother and the same steps as in *Step* (v) of Hall et al. (2006), we obtain

$$(B.19) \quad E'\|\delta_1\|_{(j)}^2 \sim O_p\left(\frac{1}{nhEN}\right).$$

We have now shown

$$\|\Delta_1\|_{(j)}^2 = O_p\left(\frac{1}{nhEN}\right) + h^4.$$

This and (B.10) - (B.12) imply

$$\|\Delta\|_{(j)}^2 = O_p\left(\frac{1}{nhEN} + h^4\right) + O_p\left(\frac{1}{nh_\phi EN} + h_\phi^4\right).$$

Finally, we consider $\int_{\mathcal{I} \times \mathcal{I}} \Delta \eta_j \eta_j$ which can be dominated by $|\int_{\mathcal{I} \times \mathcal{I}} \delta_0 \eta_j \eta_j| + |\int_{\mathcal{I} \times \mathcal{I}} (\hat{\Gamma} - \Gamma) \eta_j \eta_j|$, and hence is of order $O_p(h^2) + O_p(h_\phi^2)$. Equation (2.7) now follows from (B.9).

Next, we will discuss the optimal convergent rates under three different sampling plans. If $EN < \infty$ (longitudinal data), it is obvious that the optimal convergent rate $n^{-2/5}$ is achieved when $h \sim n^{-1/5}$. When $EN \rightarrow \infty$, we assume that $ENh \sim n^{-\tau}$, where $0 \leq \tau < 1/5$ (because of Assumption (A.5)). Simple calculations show that the optimal rate n^{-r} is achieved when $h \sim n^{-(1-\tau)/4}$ and $r = -(1-\tau)/2$. Thus, the proof completes. \square

ACKNOWLEDGEMENTS

The authors would like to express gratitude for the insightful comments of three referees, the associate editor, and the Editor.

SUPPLEMENTARY MATERIAL

Supplement to ‘Inverse Regression for Longitudinal Data’
(doi: COMPLETED BY THE TYPESETTER; .pdf). We provide additional supporting information for section 2.1, for simulation studies and for data analysis.

REFERENCES

- Carey, J. R., Müller, H. G., Wang, J. L., and Chiou, J. M. (1998). Relationship of age patterns of fecundity to mortality, longevity, and lifetime reproduction in a large cohort of mediterranean fruit fly females. *J. Gerontology: Biological Sciences*, 53A:B245–B251.
- Chen, D., Hall, P., and Müller, H.-G. (2011). Single and multiple index functional regression models with nonparametric link. *Ann. Statist.*, 39:1720–1747.
- Cook, R. and Weisberg, S. (1991). Discussion of sliced inverse regression for dimension reduction. *J. Amer. Statist. Assoc.*, 86:316–342.
- Cook, R. D., Forzani, L., and Yao, A. F. (2010). Necessary and sufficient conditions for consistency of a method for smoothed functional inverse regression. *Statist. Sinica*, 20:235–238.
- Cook, R. D. and Li, B. (2002). Dimension reduction for conditional mean in regression. *Ann. Statist.*, 30(2):455–474.
- Duan, N. and Li, K. (1991). Slicing regression: a link free regression method. *Ann. Statist.*, 19:505–530.

- Ferré, L. and Yao, A. (2003). Functional sliced inverse regression analysis. *Statistics*, 37(6):475–488.
- Ferré, L. and Yao, A. (2005). Smoothed functional inverse regression. *Statist. Sinica*, 15:665–683.
- Ferré, L. and Yao, A. F. (2007). Reply to the paper by liliana forzani and r. dennis cook: “a note on smoothed functional inverse function”. *Statist. Sinica*, 17:1683–1687.
- Forzani, L. and Cook, R. D. (2007). A note on smoothed functional inverse regression. *Statist. Sinica*, 17:1677–1681.
- Friedman, J. and Stuetzle, W. (1981). Projection pursuit regressions. *J. Amer. Statist. Assoc.*, 76:817–823.
- Hall, P. (1989). On projection pursuit regressions. *Ann. Statist.*, 17:573–588.
- Hall, P., Müller, H. G., and Wang, J. (2006). Properties of principal component methods for functional and longitudinal data analysis. *Ann. Statist.*, 34:3:1493–1517.
- Hall, P., Müller, H.-G., and Yao, F. (2008). Modeling sparse generalized longitudinal observations via latent gaussian processes. *J. R. Statist. Soc. B*, 70:703–723.
- Hastie, T. and Tibshirani, R. (1990). *Generalized Additive Models*. Chapman and Hall.
- He, G., Müller, H. G., and Wang, J. (2003). Functional canonical analysis for square integrable stochastic processes. *J. Multi. Anal.*, 85:54–77.
- Hsing, T. and Ren, H. (2009). An rkhs formulation of the inverse regression dimension-reduction problem. *Ann. Statist.*, 37:726–755.
- Jiang, C.-R., Yu, W., and Wang, J.-L. (2013). Supplement to “inverse regression for longitudinal data”.
- Kato, T. (1966). *Perturbation Theory for Linear Operators*. Berlin : Springer-Verlag.
- Li, K. (1991). Sliced inverse regression for dimension reduction. *J. Amer. Statist. Assoc.*, 86:316–342.
- Li, Y. and Hsing, T. (2010). Deciding the dimension of effective dimension reduction space for functional and high-dimensional data. *Ann. Statist.*, 38:3028–3062.
- Rice, J. (2004). Functional and longitudinal data analysis: perspectives on smoothing. *Statist. Sinica*, 14:631–648.
- Stone, C. J. (1985). Additive regression and other nonparametric models. *Ann. Statist.*, 13:689–705.
- Yin, X. and Cook, R. (2002). Dimension reduction for the conditional kth moment in regression. *J.R. Statis. Sco. B*, 64, Part2:159–175.
- Yin, X. and Cook, R. D. (2003). Estimating central subspaces via inverse third moments. *Biometrika*, 90:113–125.

CI-REN JIANG
 INSTITUTE OF STATISTICAL SCIENCE
 ACADEMIA SINICA
 TAIPEI, 115
 TAIWAN
 E-MAIL: cirenjiang@stat.sinica.edu.tw

WEI YU
 GENENTECH, INC
 1 DNA WAY
 SOUTH SAN FRANCISCO, CA 94080
 USA
 E-MAIL: yu.wei@gene.com

JANE-LING WANG
 DEPARTMENT OF STATISTICS
 UNIVERSITY OF CALIFORNIA
 DAVIS, CA 95616
 USA
 E-MAIL: jlwang.ucdavis@gmail.com

TABLE 1

Simulation comparison of FY (Ferré and Yao (2005)) for complete data and our procedures for both complete and sparse data. The comparison is based on the averages of correlations, ISB, IVAR and IMSE in 100 simulation runs.

n	Data type	Correlation	ISB	IVAR	IMSE
100	FY (Complete)	0.7159	0.0114	0.2008	0.2123
	Complete	0.9912	0.0043	0.0084	0.0127
	Sparse	0.8831	0.0583	0.2823	0.3406
200	FY (Complete)	0.8218	0.0024	0.0837	0.0861
	Complete	0.9921	0.0024	0.0092	0.0116
	Sparse	0.9438	0.0274	0.1602	0.1876

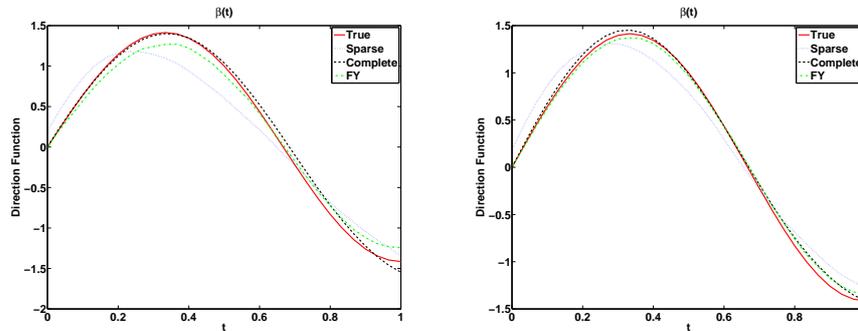


FIG 1. Simulation comparison of the average estimates of $\beta(t)$ for the three methods in Table 1. The left panel shows the average of n estimate ($\hat{\beta}(t)$) for various methods vs the target ($\beta(t)$) for $n=100$, and the right panel for $n=200$.

TABLE 2

Average of the square fitted errors, $\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2$, of the fecundity data, based on three different methods: Ferré and Yao (2005), our method with complete data, and our method with sparse data

Method	Complete:FY	Complete	Sparse
Fitted Error	134.28	134.05	134.13

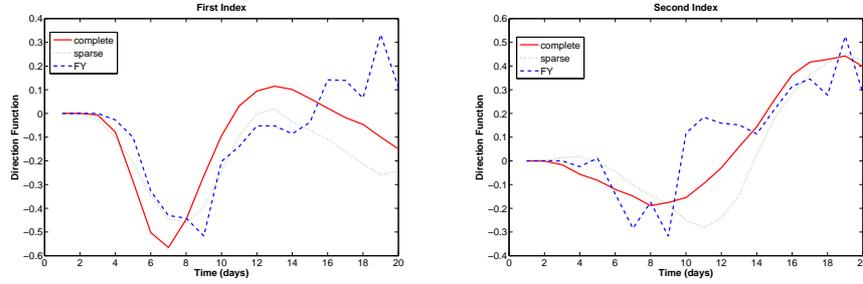


FIG 2. Estimated $\beta_1(t)$ and $\beta_2(t)$ from complete and sparse fecundity data.

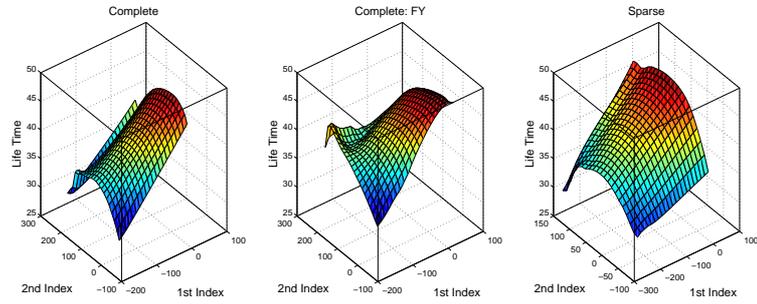


FIG 3. Estimated link functions for the fecundity data: the left panel shows the regression surface when the directions were estimated by the approach in Ferré and Yao (2005) with complete data; the middle panel shows the regression surface when the directions were estimated by our approach with complete data; the right shows the regression surface when the directions were estimated by our approach with sparse data, but the indices were calculated from the true complete covariate.