Functional Additive Models

Hans-Georg Müller

Department of Statistics, University of California

One Shield Ave., Davis, CA 95616, USA

Fang Yao

Department of Statistics, University of Toronto

100 St. George Street, Toronto, ON M5S 3G3, Canada

July 2008

Author's Footnote:

Hans-Georg Müller is Professor, Department of Statistics, University of California, Davis, One Shields Avenue, Davis, CA 95616, U.S.A, email: `mueller@wald.ucdavis.edu`. Fang Yao is Associate Professor, Department of Statistics, University of Toronto, 100 St. George Street, Toronto, ON M5S 3G3, Canada, email: `fyao@utstat.toronto.edu`.
We wish to thank two referees and an Associate Editor for helpful remarks that led to an improved version of the paper. This research was supported in part by grants from the National Science Council (NSC95-2118M001-013MY3), the National Science Foundation (DMS-03-54448 and DMS-05-05537), and the Natural Sciences and Engineering Research Council of Canada.

ABSTRACT

In commonly used functional regression models, the regression of a scalar or functional response on the functional predictor is assumed to be linear. This means the response is a linear function of the functional principal component scores of the predictor process. We relax the linearity assumption and propose to replace it by an additive structure. This leads to a more widely applicable and much more flexible framework for functional regression models. The proposed functional additive regression models are suitable for both scalar and functional responses. The regularization needed for effective estimation of the regression parameter function is implemented through a projection on the eigenbasis of the covariance operator of the functional components in the model. The utilization of functional principal components in an additive rather than linear way leads to substantial broadening of the scope of functional regression models and emerges as a natural approach, as the uncorrelatedness of the functional principal components is shown to lead to a straightforward implementation of the functional additive model, just based on a sequence of one-dimensional smoothing steps and without need for backfitting. This facilitates the theoretical analysis, and we establish asymptotic consistency of the estimates of the components of the functional additive model. The empirical performance of the proposed modeling framework and estimation methods is illustrated through simulation studies and in applications to gene expression time course data.

KEY WORDS: Asymptotics, Additive Model, Functional Data Analysis, Functional Regression, Linear Model, Principal Components, Smoothing, Stochastic Processes.

1. INTRODUCTION

A characterizing feature of functional regression models is that either predictor or response or both are functions, where the functional components are typically assumed to be realizations of a stochastic process. The functional linear model is the commonly adopted functional regression model. It has been introduced in its most general form, where both predictor and response are functions, by Ramsay and Dalzell (1991). The case of a functional predictor and scalar response has been the focus of most research to date on the functional linear model, as well as somewhat artificial situations where the functional data are assumed to be observed without noise and on a very dense and regular grid. For this case, Cardot et al. (2003) provided consistency results and introduced a testing procedure. Theory for the case of fixed design and functional response was developed in Cuevas et al. (2002). For a summary of some of these developments, we refer to Ramsay and Silverman (2005).

Several extensions of the basic linear functional regression models have been proposed, often motivated by established analogous extensions of the classical multivariate regression models towards more general regression models. These include generalized functional linear models (James, 2002; Escabias et al., 2004; Cardot and Sarda, 2005; Müller and Stadtmüller, 2005), modifications of functional regression for longitudinal, i.e., sparse, irregular and noisy data (Yao et al., 2005*b*), varying-coefficient functional models (Malfait and Ramsay, 2003; Fan and Zhang, 2000; Fan et al., 2003), wavelet-based functional models (Morris et al., 2003) and multiple-index models (James and Silverman, 2005). Recent asymptotic studies of estimation in functional linear regression models with scalar response and fully observed predictor trajectories include Cai and Hall (2006) and Hall and Horowitz (2007).

It is well known that the estimation of the regression parameter function is an inverse problem and therefore requires regularization. Two main approaches have emerged for the implementation of the regularization step. The first approach is projection onto a finite number of elements of a functional basis, which can be either fixed in advance such as the wavelet or Fourier basis, or may be chosen data-adaptively such as the eigenbasis of the auto-covariance operator of the predictor processes X. If the eigenbasis is used, conveying the advantage of the most parsimonious representation, this approach is based on an initial functional principal component analysis (see, e.g., Rice and Silverman, 1991). No matter which basis is chosen, effective regularization is then obtained by suitably truncating the number of included basis functions. The second approach to regularization is based on penalized likelihood or penalized least squares, and has been implemented for example via splines or ridge regression (Hall and Horowitz, 2007). In this paper we adopt the first approach and express functional regression models in terms of the functional principal components of the predictor, and if applicable, also of the response processes. Since the stochastic part of square integrable stochastic processes can always be equivalently represented by the countable sequence of their functional principal component scores and eigenfunctions, all functional regression models have such a representation, irrespective of their structure. In the previously studied functional linear regression models, the regression of the scalar or functional response on the functional predictors is a linear function of the predictor functional principal component scores, and estimation, inference, asymptotics and extensions of these basic functional linear models are studied within this linear framework.

The purpose of this paper is an analysis of an alternative additive functional regression model. Additive models are attractive as they provide effective dimension and great flexibility in modeling (Hastie and Tibshirani, 1990). While extensions of linear models to single and multiple index models are in place for functional regression, the extension to additive models has proven elusive and this is due to a major challenge: A direct extension, analogous to multivariate data analysis, faces the difficulty that the equivalent to individual predictors in vector regression is the continuum of function values over the entire time domain. Therefore, the predictor set is not countable, and an additive model in these predictors would require uncountably many additive components. We overcome this difficulty by taking as predictors the countable set of functional principal component scores, which can be truncated at an increasing sequence of finitely many predictors, while representing the entire predictor function adequately.

For the case of a scalar response, the combination of functional principal component scores with an additive model, emphasizing applied modeling with readily available software tools, has been demonstrated in concurrent work with applied emphasis (Foutz and Jank, 2008; Liu and Müller, 2008; Sood et al., 2008). As we show, a key to both analysis and implementation of this combination is the uncorrelatedness of the functional predictor scores. The usual implementation of additive models, which must take into account dependencies between predictors, requires backfitting or similarly complex schemes. Because the functional predictor scores are uncorrelated, the additive fitting step can be greatly simplified and requires no more than one-dimensional smoothing steps, separately applied to each predictor score. This has important consequences: Very fast and simple implementation, and simplification which makes it possible to study asymptotic properties of the resulting model, especially in the Gaussian case. The combination of functional principal component score predictors and additive models therefore emerges as a particularly natural and flexible data-adaptive nonparametric framework for functional regression models. We refer to this approach as the Functional Additive Model (FAM).

The paper is organized as follows: Section 2 contains background from functional linear regression. The proposed Functional Additive Models are introduced in Section

3, and issues of fitting these models are the theme of Section 4. Asymptotic consistency properties are presented in Section 5, while Section 6 is devoted to a report on simulation results, followed by a description of an application to regression for gene expression time courses in the *Drosophila* life cycle in Section 7 and concluding remarks in Section 8. Details and assumptions can be found in a separate Appendix, and auxiliary results and proofs in a Supplement.

2. FUNCTIONAL LINEAR MODELS AND EIGENREPRESENTATIONS

We consider regression models where the predictor is a smooth square integrable random function $X(\cdot)$ defined on a domain $\mathcal{S}$, and the response is either a scalar or a random function $Y(\cdot)$ on domain $\mathcal{T}$, with mean functions $EX(s) = \mu_X(s)$ and $EY(t) = \mu_Y(t)$ and covariance functions $\mathrm{cov}(X(s_1), X(s_2)) = G_X(s_1, s_2)$, $\mathrm{cov}(Y(t_1), Y(t_2)) = G_Y(t_1, t_2)$, $s, s_1, s_2 \in \mathcal{S}$ and $t, t_1, t_2 \in \mathcal{T}$, respectively. We denote centered predictor processes by $X^c(s) = X(s) - \mu_X(s)$. The established linear functional regression models with scalar resp. functional response are (Ramsay and Silverman, 2005)

$$E(Y|X) = \mu_Y + \int_{\mathcal{S}} \beta(s) X^c(s)\, ds, \tag{1}$$

$$E\{Y(t)|X\} = \mu_Y(t) + \int_{\mathcal{S}} \beta(s,t) X^c(s)\, ds, \tag{2}$$

where the regression parameter functions β are assumed to be smooth and square integrable. To estimate these functions and thus identify the functional linear model, basis representations such as the eigendecomposition of the functional components in models (1) and (2) are a convenient way to implement the necessary regularization (Yao et al., 2005*b*). Taking advantage of the equivalence between process and countable sequence of functional principal component (FPC) scores, we represent predictor and response processes X and Y in terms of these scores. An implementation of these representations is available through the PACE package which can be found under "programs" at http://www.stat.ucdavis.edu/∼mueller/.

In our sampling model we assume as in Yao et al. (2003) that the functional trajectories are not completely observed but rather are measured on a grid and measurements are contaminated with measurement error. This is a more realistic scenario than the common assumption in functional data analysis that entire trajectories are observed. For our theoretical analysis we assume that the grid of measurements is dense. Empirically, as evidenced by simulations and in applications, the proposed methods also work well for the case of sparse and irregularly observed longitudinal data. Denote by U_{ij} (respectively, V_{il}) the noisy observations made of the random trajectories X_i (respectively, Y_i) at times s_{ij} (resp., t_{il}), where (X_i, Y_i), $i = 1, \ldots, n$, corresponds to an i.i.d. sample of processes (X, Y) (the scalar response case is always included in these considerations). The available observations are contaminated with measurement errors ϵ_{ij} (respectively, ε_{il}), $1 \le j \le n_i$, $1 \le l \le m_i$, $1 \le i \le n$, where n_i and m_i are the numbers of observations from X_i and Y_i. The errors are assumed to be i.i.d., with zero means $E\epsilon_{ij} = 0$, and constant variance $E(\epsilon_{ij}^2) = \sigma_X^2$, respectively, $E\varepsilon_{il} = 0$, $E(\varepsilon_{il}^2) = \sigma_Y^2$, and independent of the FPC scores $\xi_{ik} = \int (X_i(s) - \mu_X(s))\phi_k(s)\, ds$, respectively, $\zeta_{im} = \int (Y_i(t) - \mu_Y(t))\psi_m(t)\, dt$, where ϕ_k, ψ_m are the eigenfunctions of processes X and Y, defined in the Appendix. We note that the FPC scores satisfy $E\xi_{ik} = 0$, $E(\xi_{ik}\xi_{ik'}) = 0$ for $k \ne k'$ and $E(\xi_{ik}^2) = \lambda_k$, respectively, $E\zeta_{im} = 0$, $E(\zeta_{im}\zeta_{im'}) = 0$ for $m \ne m'$ and $E(\zeta_{im}^2) = \rho_m$.

Invoking the population least squares property of conditional expectation, and using the fact that the predictors are uncorrelated, leads to an extension of the representation $\beta_1 = \mathrm{cov}(X, Y)/\mathrm{var}(X)$ of the slope parameter in the simple linear regression model $E(Y|X) = \beta_0 + \beta_1 X$ to the functional case. By solving a "functional population normal equation" (He et al., 2000, 2003), one obtains for scalar resp. functional responses

$$\beta(s) = \sum_{k=1}^{\infty} \frac{E(\xi_k Y)}{E(\xi_k^2)} \phi_k(s), \quad \beta(s,t) = \sum_{k=1}^{\infty} \sum_{m=1}^{\infty} \frac{E(\xi_k \zeta_m)}{E(\xi_k^2)} \phi_k(s) \psi_m(t). \tag{3}$$

Plugging (3) into (1), (2) and observing (23) then leads to

$$E(Y|X) = \mu_Y + \sum_{k=1}^{\infty} b_k \xi_k, \quad \text{with} \quad b_k = \frac{E(\xi_k Y)}{E(\xi_k^2)}, \tag{4}$$

$$E(\zeta_m|X) = \int E(Y(t) - \mu_Y(t)|X) \psi_m(t)\, dt = \sum_{k=1}^{\infty} b_{km} \xi_k, \text{ with } b_{km} = \frac{E(\xi_k \zeta_m)}{E(\xi_k^2)}, \tag{5}$$

for scalar responses Y, respectively, for each FPC score ζ_m of the response process. In practice, when fitting a functional regression model and estimating the regression parameter function β, one needs to regularize by truncating these expansions at a finite number of components K and M.

3. FUNCTIONAL ADDITIVE MODELING

Suppose for the moment that the true FPC scores ξ_{ik} for predictor processes are known. Then the functional linear models (1) and (2) are reduced to a standard linear model with infinitely many of these FPC scores as predictors, as demonstrated in eq. (4) and (5). Moreover, the linear structure in the predictor scores and the uncorrelatedness of the FPC scores then imply that

$$E(Y - \mu_Y|\xi_k) = b_k \xi_k, \quad E(\zeta_m|\xi_k) = b_{km} \xi_k \tag{6}$$

for scalar, respectively, functional response models. Accordingly, the best predictor is the linear predictor, and the regressions on predictor scores are lines through the origin.

This observation provides a key motivation for the extension of the functional linear model to the functional additive model (FAM). It seems natural to generalize the well-known extension of (generalized) linear to (generalized) additive models (Hastie and Tibshirani, 1990) to the functional case by replacing the linear terms $b_k\xi_k$ resp. $b_{km}\xi_k$ in (6) and (4), (5) by more general relationships. We thus generalize the linear relationship with ξ_k to arbitrary functional relations $f_k(\xi_k)$, where functions $f_k(\cdot)$, $k = 1, 2, \ldots$, respectively, functions $f_{km}(\cdot)$, $k, m = 1, 2, \ldots$, are assumed to be smooth; beyond smoothness, nothing more is needed. This substitution transforms functional linear models (1), (2) to functional additive models with the underlying FPC scores ξ_k as predictors,

$$E(Y|X) = \mu_Y + \sum_{k=1}^{\infty} f_k(\xi_k), \tag{7}$$

$$E(Y(t)|X) = \mu_Y(t) + \sum_{k=1}^{\infty}\sum_{m=1}^{\infty} f_{km}(\xi_k)\psi_m(t), \tag{8}$$

for scalar and functional response cases, respectively. To ensure identifiability, we require furthermore that

$$Ef_k(\xi_k) = 0,\ k = 1, 2, \ldots, \text{ resp.,}\quad Ef_{km}(\xi_k) = 0,\ k = 1, 2, \ldots, m = 1, 2, \ldots. \tag{9}$$

In this model, the linear relationship between the response Y and the predictor FPC scores ξ_k is replaced by an additive relation, which gives rise to a far more flexible and essentially nonparametric model, while avoiding the curse of dimension, which for infinite-dimensional functional data is unsurmountable, if no structure is imposed. Beyond additivity, a second key assumption we make from now on is that the predictor FPC scores ξ_k are independent. Since these scores are always uncorrelated, this assumption is for example satisfied for the case where predictor processes are Gaussian. Then the basic functional additive model assumptions (7), (8) imply

$$E(Y - \mu_Y|\xi_k) = E\{E(Y - \mu_Y|X)|\xi_k\} = E\{\sum_{j=1}^{\infty} f_j(\xi_j)|\xi_k\} = f_k(\xi_k), \tag{10}$$

and for the functional response case analogously

$$E(\zeta_m|\xi_k) = E\{E(\zeta_m|X)|\xi_k\} = E\{\sum_{j=1}^{\infty} f_{jm}(\xi_j)|\xi_k\} = f_{km}(\xi_k). \tag{11}$$

The relations $f_k(\xi_k) = E(Y - \mu_Y|\xi_k)$, $f_{km}(\xi_k) = E(\zeta_m|\xi_k)$, for all $k, m = 1, 2, \ldots$, are a straightforward generalization of the decomposition of functional linear regression into simple linear regressions against the predictor FPC scores as in (6). They are key for surprisingly simple implementations of the functional additive model. While complex iterative procedures are required to fit a regular additive model (backfitting and variants, see, e.g., Hastie and Tibshirani, 1990; Mammen and Park, 2005), representations (10), (11) motivate a straightforward estimation scheme to recover the component functions f_k, respectively f_{km}, by a series of one-dimensional smoothing steps. This not only leads to fast and easily diagnosable procedures for the underlying infinite-dimensional data, but also facilitates asymptotic analysis. The high degree of flexibility and the simplicity of model fitting makes FAM an especially attractive alternative to the special case of the standard functional linear models (1), (2).

4. FITTING OF FUNCTIONAL ADDITIVE MODELS

We begin with an overview of the estimation procedures. In a first step, smooth estimates of the mean and covariance functions for the predictor processes are obtained by scatterplot smoothing. This is followed by a functional principal component (FPC) analysis, which yields estimates $\hat{\phi}_k$ for the eigenfunctions, $\hat{\lambda}_k$ for the eigenvalues, and $\hat{\xi}_{ik}$ for the FPC scores of individual predictor trajectories; some additional details are given in the Appendix. The estimation steps are implemented with the Principal Analysis by Conditional Expectation (PACE) approach, also regarding the choice of the number of included eigenfunctions K through pseudo-AIC (Yao et al., 2005a), available in the PACE package. This was shown to work for densely sampled trajectories and in the Gaussian case in addition to the case of sparse and irregular measurements; it also has been demonstrated to be fairly robust against violations of the Gaussian assumption.

Once these preliminary estimates are in hand, it is straightforward to obtain estimates $\hat{f}_k, \hat{f}_{km}$ of the smooth component functions f_k, respectively, f_{km}. We implement all smoothing steps with local polynomial fitting; other smoothing techniques can be used equally well. For the case of scalar responses, we estimate the functions f_k by fitting a local linear regression to the data $\{\hat{\xi}_{ik}, Y_i\}_{i=1,\ldots,n}$, where $\hat{\xi}_{ik}$ is obtained by (26): Minimizing

$$\sum_{i=1}^{n} K_1(\frac{\hat{\xi}_{ik} - x}{h_k})\{Y_i - \beta_0 - \beta_1(x - \hat{\xi}_{ik})\}^2 \tag{12}$$

with respect to β_0 and β_1, leads to $\hat{f}_k(x) = \hat{\beta}_0(x) - \bar{Y}$, where h_k is the bandwidth used for this smoothing step, and K_1 is a symmetric probability density that serves as kernel function. For the functional response case, the functions f_{mk} are analogously estimated by passing a local linear smoother through the data $\{\hat{\xi}_{ik}, \hat{\zeta}_{im}\}_{i=1,\ldots,n}$ that are obtained by (26), i.e., minimizing

$$\sum_{i=1}^{n} K_1(\frac{\hat{\xi}_{ik} - x}{h_{mk}})\{\hat{\zeta}_{im} - \beta_0 - \beta_1(x - \hat{\xi}_{ik})\}^2 \tag{13}$$

with respect to β_0 and β_1, leading to $\hat{f}_{mk}(x) = \hat{\beta}_0(x)$, where h_{mk} is the bandwidth. Then the fitted version of the functional additive model (7) with scalar response is

$$\widehat{E}(Y|X) = \bar{Y} + \sum_{k=1}^{K} \hat{f}_k(\xi_k). \tag{14}$$

To quantify the strength of the regression relationship, we use a global measure similar to the coefficient of determination R^2 in standard linear regression (Draper and Smith, 1998), with population and sample versions

$$R^2 = 1 - \frac{\sum_{i=1}^{n}\{Y_i - E(Y_i|X_i)\}^2}{\sum_{i=1}(Y_i - \mu_Y)^2}, \qquad \widehat{R}^2 = 1 - \frac{\sum_{i=1}^{n}\{Y_i - \widehat{E}(Y_i|X_i)\}^2}{\sum_{i=1}(Y_i - \bar{Y})^2}, \tag{15}$$

where $E(Y_i|X_i)$ and $\widehat{E}(Y_i|X_i)$ for the ith subject are as in (7) and (14).

Analogously, the fitted FAM for functional responses, based on (8), is

$$\widehat{E}\{Y(t)|X\} = \hat{\mu}_Y(t) + \sum_{m=1}^{M}\sum_{k=1}^{K} \hat{f}_{mk}(\xi_k)\hat{\psi}_m(t), \quad t \in \mathcal{T}, \tag{16}$$

and the strength of the regression for this case can be measured by

$$
\begin{aligned}
R^2 &= 1 - \frac{\sum_{i=1}^{n} \int_{\mathcal{T}} [Y_i(t) - E\{Y_i(t)|X_i\}]^2 \, dt}{\sum_{i=1}^{n} \int_{\mathcal{T}} \{Y_i(t) - \mu_Y(t)\}^2 \, dt}, \\
\widehat{R}^2 &= 1 - \frac{\sum_{i=1}^{n} \sum_{l=2}^{m_i} [V_{il} - \widehat{E}\{Y_i(t_{il})|X_i\}]^2 (t_{il} - t_{i,l-1})}{\sum_{i=1}^{n} \sum_{l=2}^{m_i} \{V_{il} - \mu_Y(t_{il})\}^2 (t_{il} - t_{i,l-1})}, \qquad (17)
\end{aligned}
$$

for population and sample versions, where $E\{Y_i(t)|X_i\}$ and $\widehat{E}(Y_i(t)|X_i)$ for the ith subject are as in (8) and (16), and we assume a dense grid of measurements t_{il} for each subject.

5. THEORETICAL RESULTS

Establishing relevant asymptotic results requires studying the relationship between the true and the estimated FPC scores ξ_{ik} and $\hat{\xi}_{ik}$, ζ_{im} and $\hat{\zeta}_{im}$, $k = 1, \ldots, K$, $m = 1, \ldots, M$, since the estimates of the FAM component functions f_k, respectively, f_{km} need to be based on the estimated scores. Starting with known convergence results for the estimated population components such as mean function, eigenfunction and eigenvalue estimates in model (24) or (25) (see Yao et al., 2005*a*; Hall and Hosseini-Nasab, 2006), a key step in the mathematical analysis is to establish exact upper bounds of $|\hat{\xi}_{ik} - \xi_{ik}|$ and $|\hat{\zeta}_{im} - \zeta_{im}|$, where such bounds are i.i.d. in terms of i or do not depend on i, $i = 1, \ldots, n$ (see Supplement for details). The convergence properties of the estimated additive model components f_k in (7) or f_{mk} in (8) will follow from those upper bounds, as these estimates are obtained by applying a nonparametric smoothing method to $\{\hat{\xi}_{ik}, Y_i\}$ or $\{\hat{\xi}_{ik}, \hat{\zeta}_{im}\}$ for $i = 1, \ldots, n$.

Asymptotic results are obtained for $n \to \infty$, and also the number of included components K, M needs to satisfy $K = K(n) \to \infty$, $M = M(n) \to \infty$, for a genuinely functional (infinite-dimensional) approach. The results concern consistency of estimates (12), (13) and of predicted responses (14), (16), obtained for new predictor processes. Details about the regularity assumptions can be found in the Appendix.

Theorem 1 *Under assumptions (A1.1)-(A5), (C1.1), (C1.2), (C2.1) and (C2.3) (see Appendix), in the scalar response case, for all* $k \geq 1$ *for which* λ_j, $j \leq k$ *are eigenvalues of multiplicity 1,*

$$\hat{f}_k(x) - f_k(x) \xrightarrow{p} 0, \tag{18}$$

for estimates (12). Under the additional assumptions (B1.1)-(B4) and (C2.2), in the functional response case, for all k, m *for which* λ_j, $j \leq k$ *and* ρ_l, $l \leq m$ *are eigenvalues of multiplicity 1,*

$$\hat{f}_{km}(x) - f_{km}(x) \xrightarrow{p} 0, \tag{19}$$

for estimates (13).

Additional results on the rates of convergence of $\tilde{\theta}_k(x) = |\hat{f}_k(x) - f_k(x)|$ and $\tilde{\vartheta}_{mk}(x) = |\hat{f}_{mk}(x) - f_{mk}(x)|$ can be found in the Supplement, eq. (41). Next, we consider consistency of the predictions obtained by applying FAM.

Theorem 2 *Under (A1.1)-(A4), (A6), (A7), (C1.1), (C1.2), (C2.1) and (C2.3), for the scalar response case,*

$$\widehat{E}(Y|X) - E(Y|X) \xrightarrow{p} 0, \tag{20}$$

where $\widehat{E}(Y|X) = \bar{Y} + \sum_{k=1}^{K} \hat{f}_k(\xi_k)$ *as in (12). Under the additional assumptions (B1.1)-(B4), (B5), (B6) and (C2.2), it holds for the functional response case that for all* $t \in \mathcal{T}$,

$$\widehat{E}\{Y(t)|X\} - E\{Y(t)|X\} \xrightarrow{p} 0, \tag{21}$$

where $\widehat{E}\{Y(t)|X\} = \hat{\mu}_Y(t) + \sum_{k=1}^{K} \sum_{m=1}^{M} \hat{f}_{mk}(\xi_k)\hat{\psi}_m(t)$, *with* $\hat{f}_{mk}(\xi_k)$ *as in (13).*

Again, additional results on the rates of convergence of $\theta_n^* = |\widehat{E}(Y|X) - E(Y|X)|$ and $\vartheta_n^* = |\widehat{E}(Y(t)|X) - E(Y(t)|X)|$ are in eq. (42) of the Supplement.

6. SIMULATION STUDIES

Simulation studies were conducted to illustrate the empirical performance of the Functional Additive Models (FAM) (7) with scalar and also (8) with functional response. For both cases, we generated 200 simulation runs, each consisting of a sample of $n = 100$ predictor trajectories X_i, with mean function $\mu_X(s) = s + \sin(s)$, $0 \le s \le 10$, and a covariance function derived from two eigenfunctions, $\phi_1(s) = -\cos(\pi s/10)/\sqrt{5}$, and $\phi_2(s) = \sin(\pi s/10)/\sqrt{5}$, $0 \le s \le 10$. The corresponding eigenvalues were chosen as $\lambda_1 = 4$, $\lambda_2 = 1$ and $\lambda_k = 0$, $k \ge 3$, the measurement errors in (24) as $\epsilon_{ij} \overset{\text{i.i.d.}}{\sim} N(0, 0.5^2)$. We consider two different underlying distributions of the predictor FPC scores: (i) $\xi_{ik} \sim \mathcal{N}(0, \lambda_k)$, (ii) $\xi_{ik} = \sqrt{\lambda_k}(Z_{ik} - 4)/2$, where $Z_{ik} \overset{\text{i.i.d.}}{\sim} \text{Gamma}(4, 1)$, which is a right skewed distribution, $k = 1, 2$.

As for the design of number and spacing of the measurement locations at which predictor trajectories were sampled, we considered both dense and also sparse designs, in order to check the robustness of our methods against violations of the dense design assumption. For the dense case, each predictor trajectory was sampled at locations that were uniformly distributed over the domain $[0, 10]$, where the number of measurements was chosen separately and randomly for each predictor trajectory, by selecting a number from $\{30, \ldots, 40\}$ with equal probability. For the more challenging sparse case, the number of measurements was chosen from $\{3, \ldots, 6\}$ with equal probability. For each simulation run, we generated 100 new predictor trajectories X_i^* with measurements U_{ij}^*, taken at the same time points U_{ij} as for the 100 observed predictor trajectories, and 100 associated response variables, respectively, response functions Y_i^*.

For the scalar response case, we generated responses $Y_i = \sum_{k=1}^{2} f_k(\xi_{ik}) + \varepsilon_i$, where f_k are the true component functions that relate the FPC scores ξ_{ik} of the predictor trajectories X_i to the responses Y_i, with errors $\varepsilon_i \overset{\text{i.i.d.}}{\sim} N(0, 0.1)$, so that $\mu_Y = 0$ and $\sigma_Y^2 = 0.1$. We compared the performance of fitting FAM (7) and the functional linear regression model (1) under two situations: (a) the true functions $f_k(x) = x^2 - \lambda_k$ for $k = 1, 2$ were nonlinear; and (b) the true functions $f_1(x) = 2x$ and $f_2(x) = x/2$ were linear. As we demonstrated in (6), the functions f_k become lines through the origin for the case of the functional linear regression model. The representation (6) immediately suggests the estimates $\hat{f}_k(x) = \sum_{i=1}^n (Y_i - \bar{Y})(\hat{\xi}_{ik} - \bar{\hat{\xi}}_{\cdot k})\hat{\lambda}_k^{-1} x$, where $\bar{Y}_i = \sum_{i=1}^n Y_i/n$, $\bar{\hat{\xi}}_{\cdot k} = \sum_{i=1}^n \hat{\xi}_{ik}/n$, $\hat{\lambda}_k$ is the estimate of the kth eigenvalue of the predictor process X, and the $\hat{\xi}_{ik}$ are obtained by the PACE method (26).

For both scalar and functional response cases, FAM was implemented as described in Section 4, including choice of the number of model components for the predictor processes by AIC, and local polynomial smoothing applied to estimate the functions f_k, with one-point-leave-out cross-validation for automatic choice of the smoothing bandwidths. The functional linear model was fitted as described above, and we compared the quality of the prediction of responses for the new subjects.

For the functional response case, we also compared the predictive performance of FAM with that of functional linear regression. For the latter, we estimated the (in this case linear) component functions f_{km} by $\hat{f}_{km}(x) = \sum_{i=1}^n (\hat{\zeta}_{im} - \bar{\zeta}_{\cdot m})(\hat{\xi}_{ik} - \bar{\hat{\xi}}_{\cdot k})\hat{\lambda}_k^{-1} x$. The simulation settings were the same as those for the scalar response case, and in addition functional response trajectories were generated as $Y_i(t) = \mu_Y(t) + \zeta_{i1}\psi_1(t)$, where $\mu_Y(t) = t + \sin(t)$, $\psi_1(t) = -\cos(\pi t/10)/\sqrt{5}$, and ζ_{i1} were the only non-zero FPC scores for the responses, $0 \leq t \leq 10$. To generate the scores ζ_{i1}, as before we considered nonlinear and linear scenarios for the component functions f_{km}, $k = 1, 2$, $m = 1$: (a) the true functions $f_{k1}(x) = x^2 - \lambda_k$ were nonlinear, $k = 1, 2$, i.e., $\zeta_{i1} = \sum_{k=1}^2 (\xi_{ik}^2 - \lambda_k)$; and (b) the true functions $f_{11}(x) = 2x$, $f_{21}(x) = x/2$ were linear, i.e., $\zeta_{i1} = 2\xi_{i1} + \xi_{i2}/2$. The measurement errors ε_{il} in (25) were generated i.i.d. from $\mathcal{N}(0, 0.1)$. The designs for sampling the response trajectories were chosen in the same way as those for sampling the predictor trajectories, with both sparse and dense cases included.

To evaluate the prediction of new responses from future subjects, we generated 100 new predictor and response trajectories X_i^* and Y_i^*, respectively, with measurements U_{ij}^* and V_{il}^* taken at the same time points as U_{ij} and V_{il}, respectively. The quality of the responses was measured in terms of the relative prediction errors (RPE),

$$\mathrm{RPE}_i = \frac{(Y_i^* - \hat{Y}_i^*)^2}{Y_i^{*2}}, \qquad \mathrm{RPE}_{i,f} = \frac{\int_{\mathcal{S}} (Y_i^*(t) - \hat{Y}_i^*(t))^2 dt}{\int_{\mathcal{S}} Y_i^{*2}(t) dt}, \tag{22}$$

for scalar and functional response cases, respectively.

The results for the relative prediction errors when the predictor FPC scores are normal are shown in Table 1 and suggest that FAM leads to similar prediction errors in the sparse and somewhat larger prediction errors in the dense case, as compared to the functional linear approach when the true functions f_k or f_{km} are linear, while FAM improves upon functional linear regression when the underlying component functions are nonlinear. This holds equally for scalar and functional responses.

Similar results emerge for the case of right skewed distributions (Table 2). Again the median losses when using FAM for the case of an underlying linear model are small in the sparse case, while they are now more noticeable in the dense case. The improvements obtained when using FAM for the nonlinear case for both dense and sparse designs are found to persist for the situation with skewed distributions. We note that the performance of FAM appears to be less stable as compared to the linear model in the tails of the error distribution, as evidenced by occasional relatively large values in the 75th percentiles of relative prediction errors. Our conclusion is that, in situations where the signal is not too weak, the losses when using FAM in the linear case are relatively small, while FAM performs better than the linear model in situations when the underlying regression relationship is nonlinear.

7. APPLICATION TO GENE EXPRESSION TIME COURSE DATA

We apply our methods to gene expression profile data where both predictors and responses are functional. Arbeitman et al. (2002) obtained developmental gene expression profiles over the entire lifespan of *Drosophila*, and we apply functional regression to study the relation of the expression profiles for different developmental periods. The genes in the biologically identified group of $n = 21$ "transiently expressed zygotic genes" show early peaks in expression during the embryonal phase and are active in the cellularization phase of the embryo. For this well-defined gene group we study how the expression in the pupa or metamorphosis phase depends on that in the embryo phase. The data consist of 31 measurements during the embryo phase, i.e., the predictor process, and 18 measurements during the pupa phase, i.e., the response process. For one of the genes, data were only available for the embryo phase, and the data for this gene therefore were only used to carry out the FPC analysis for predictor processes. The linearly interpolated gene expression trajectories for both predictor and response processes, as well as their mean functions, are shown in Figure 1, confirming an early peak in the predictor trajectories and displaying quite a bit of variation between genes.

The AIC method selected three components for predictor processes and four for response processes. The corresponding eigenfunctions are displayed in Figure 2; see legend for fraction of variation explained by the corresponding eigenvalues. Of interest are the pairwise scatterplots of all pairings of response FPC scores $\hat{\zeta}_{i1}, \hat{\zeta}_{i2}, \hat{\zeta}_{i3}, \hat{\zeta}_{i4}$ versus the predictor FPC scores $\hat{\xi}_{i1}, \hat{\xi}_{i2}, \hat{\xi}_{i3}$, as shown in Figure 3. Judging from these scatterplots, there exist clear relationships between response and predictor scores; while several of these appear close to linear, for others a linear fit is not good, pointing to the presence of nonlinear relationships. To interpret these relations, one needs to take the shape of both predictor and response eigenfunctions into account. For example, the negative relationship between $\hat{\zeta}_{i1}$ and $\hat{\xi}_{i1}$ implies that sharper initial peaks and lower late embryonal gene expression is coupled with an overall lower pupa phase expression, in the sense that the contribution of the first eigenfunction in the response is reduced.

Interpretation of FAM can be aided by a "principal response plot", displayed in

Figure 4, where one varies one predictor FPC score, while keeping the others fixed at level 0. This can be visualized by looking at a set of predictor functions moving in a certain direction "away" from the mean function of the predictor process, where the direction depends on the chosen eigenfunction; each of these predictor functions, when plugged into FAM, then generates a corresponding response function. The set of these response functions is then jointly visualized with the set of predictor functions in a way that the corresponding predictor/response pairs can be easily identified. This device can be employed for each predictor score. Since these act independently from each other on the response, displaying a series of such plots for all relevant predictor scores then provides a graphical representation of FAM.

The left panels of Figure 4 indicate that shifts in the levels on the right-hand side of the peak of the predictor curves are associated with broad shifts up or down in responses, the middle panels that the size of peak expression in predictors is associated with amplitude shifts in the right half of response curves, and the left panels, that combined time and amplitude shifts of predictor peaks are associated with strong amplitude shifts in responses in a nonlinear way. The principal response plots thus characterize the response changes induced by the modes of variation of the predictors. For proper interpretation, it is helpful to note that the actual sample variation of the predictor scores in the direction of each eigenfunction, as depicted in the top panels of Figure 4, depends on the size of the respective eigenvalue, which corresponds to the variance of the FPC scores. This variation accordingly is much smaller for the second or third eigenfunction, as compared to the first eigenfunction, and therefore the system's response function changes, as depicted in the lower panels, will be realized on increasingly smaller scales for real functional data, as the order of the eigenfunction increases.

The increased bias when fitting functional linear regression to these data is evident in the observed leave-one-subject-out relative prediction errors, i.e., the cross-validated observed version of (22), denoted by $\text{RPE}_{(-i),f}$; mean and percentiles are listed in Table 3. Median and mean of the errors are larger for the functional linear model in comparison with FAM. This finding is in line with the increase in functional R^2 (17) that is obtained for FAM as compared to the functional linear regression model.

8. CONCLUDING REMARKS

The Functional Additive Model (FAM) strikes a fine balance between greatly enhancing flexibility, as compared to the functional linear model, and preventing the curse of dimension incurred by a fully nonparametric approach, due to its sensible structural constraints. This is analogous to the situation in ordinary multiple regression, where additive models do not suffer from the curse of dimension in the way unstructured nonparametric models do. As a reviewer has pointed out, one could also imagine other additive regression models geared towards specific regression relations, where the predictors correspond to suitably chosen functionals of the predictor processes other than the functional principal component scores.

As we have demonstrated, specific advantages of using functional principal component scores in an additive regression model are that (1) such a model is a strict generalization of the functional linear model, which emerges as a special case; (2) asymptotic consistency results can still be derived under truly nonparametric (smoothness) assumptions for all component functions, including mean and covariance functions; (3) implementing the resulting additive model requires not more than applying simple smoothing steps. The necessary smoothing parameters for the component functions can be easily selected in a data-adaptive way. FAM does not impose a largely increased computational burden over the established functional linear regression model, and therefore the added flexibility comes at low additional computational cost. Judging from our simulations, the loss in efficiency against the functional linear regression model when the underlying functional regression is truly linear is quite limited. On the other hand, the increased flexibility of FAM can lead to substantial gains.

APPENDIX

A.1 Eigenrepresentations and Estimating Functional Principal Component Scores

The eigenfunctions $\phi_k, k = 1, 2, \ldots$, for the representation of predictor processes X are the solutions of the equations $(A_{G_X} f)(s) = \lambda f(s)$ on the space of square integrable functions $f \in L^2(\mathcal{S})$ under constraints, where the auto-covariance operator $(A_{G_X} f)(t) = \int f(s) G_X(s,t)\, ds$ is a compact linear integral operator of Hilbert-Schmidt type (Ash and Gardner, 1975). The constraints correspond to orthonormality of the eigenfunctions, i.e., $\int \phi_j(s)\phi_k(s)\, ds = \delta_{jk}$, where $\delta_{jk} = 1$ if $j = k$ and $= 0$ if $j \neq k$. The eigenfunctions ϕ_j are ordered according to the size of the corresponding eigenvalues, $\lambda_1 \geq \lambda_2 \geq \ldots$. Analogously, eigenfunctions and eigenvalues of the response process Y are denoted by ψ_m and ρ_m. We assume that all functions $\{\mu_X, \phi_j\}$, $\{\mu_Y, \psi_k\}$ are smooth (twice continuously differentiable).

One then has well-known representations $G_X(s_1, s_2) = \sum_k \lambda_k \phi_k(s_1)\phi_k(s_2)$ and $G_Y(t_1, t_2) = \sum_m \rho_m \psi_m(t_1)\psi_m(t_2)$ of the covariance functions of X and Y, as well as the Karhunen-Loève expansions for processes X and Y,

$$X(s) = \mu_X(s) + \sum_{j=1}^{\infty} \xi_j \phi_j(s), \quad Y(t) = \mu_Y(t) + \sum_{k=1}^{\infty} \zeta_k \psi_k(t). \tag{23}$$

Since $\{\phi_k,\, k = 1, 2, \ldots\}$ and $\{\psi_m,\, m = 1, 2, \ldots\}$ form orthonormal bases of the respective space of square integrable functions, it follows that the regression parameter functions can also be represented in this basis, i.e., there exist coefficients β_k, β_{km} such that $\beta(s) = \sum_{k=1}^{\infty} \beta_k \phi_k(t)$, resp., $\beta(s,t) = \sum_{k=1}^{\infty} \sum_{m=1}^{\infty} \beta_{km} \phi_k(s)\psi_m(t)$.

According to (23), we may represent the measurements for predictor trajectories in (1) and both predictor and response trajectories in (2) as

$$U_{ij} = X_i(s_{ij}) + \epsilon_{ij} = \mu_X(s_{ij}) + \sum_{k=1}^{\infty} \xi_{ik}\phi_k(s_{ij}) + \epsilon_{ij},\ s_{ij} \in \mathcal{S},\ 1 \leq i \leq n,\ 1 \leq j \leq n_i, \tag{24}$$

$$V_{il} = Y_i(t_{il}) + \varepsilon_{il} = \mu_Y(t_{il}) + \sum_{m=1}^{\infty} \zeta_{im}\psi_k(t_{il}) + \varepsilon_{il},\ s_{il} \in \mathcal{T},\ 1 \le i \le n,\ 1 \le l \le m_i. \quad (25)$$

More specifically, writing $\boldsymbol{U}_i = (U_{i1}, \ldots, U_{in_i})^T$, $\boldsymbol{\mu}_{X_i} = (\mu_X(s_{i1}), \ldots, \mu_X(s_{in_i}))^T$, and $\boldsymbol{\phi}_{ik} = (\phi_k(s_{i1}), \ldots, \phi_k(s_{in_i}))^T$, the best linear prediction for ξ_{ik} is $\lambda_k \boldsymbol{\phi}_{ik}^T \boldsymbol{\Sigma}_{U_i}^{-1}(\boldsymbol{U}_i - \boldsymbol{\mu}_{X_i})$, where the (j,l) entry of the $n_i \times n_i$ matrix $\boldsymbol{\Sigma}_{U_i}$ is $(\boldsymbol{\Sigma}_{U_i})_{j,l} = G_X(s_{ij}, s_{il}) + \sigma_X^2 \delta_{jl}$, with $\delta_{jl} = 1$, if $j = l$, and $\delta_{jl} = 0$, if $j \ne l$. Then the estimates for the scores ξ_{ik} are obtained by substituting estimates for $\boldsymbol{\mu}_{X_i}$, λ_k and $\boldsymbol{\phi}_{ik}$, $\boldsymbol{\Sigma}_{X_i}$ (see Supplement), obtained from the entire data ensemble, leading to

$$\hat{\xi}_{ik} = \hat{\lambda}_k \hat{\boldsymbol{\phi}}_{ik}^T \widehat{\boldsymbol{\Sigma}}_{U_i}^{-1}(\boldsymbol{U}_i - \hat{\boldsymbol{\mu}}_{X_i}), \quad (26)$$

where the (j,l) element of $\widehat{\boldsymbol{\Sigma}}_{U_i}$ is $(\widehat{\boldsymbol{\Sigma}}_{U_i})_{j,l} = \widehat{G}_X(s_{ij}, s_{il}) + \hat{\sigma}_Y^2(s_{ij})\delta_{jl}$. We note that it follows from results in Müller (2005) that as designs become dense, these best linear (PACE) estimates $\hat{\xi}_{ik}$, $\hat{\zeta}_{im}$ (26) converge to those obtained by the more traditional integration-based estimates

$$\hat{\xi}_{ik}^I = \sum_{j=2}^{n_i} (U_{ij} - \hat{\mu}_X(t_{ij}))\hat{\phi}_k(s_{ij})(s_{ij} - s_{i,j-1}), \hat{\zeta}_{im}^I = \sum_{j=2}^{n_i} (V_{ij} - \hat{\mu}_Y(t_{ij}))\hat{\psi}_k(t_{ij})(t_{ij} - t_{i,j-1}), \quad (27)$$

which are motivated by the definition of the FPC scores as inner products, $\xi_{ik} = \int \{X_i(s) - \mu_X(s)\}\phi_k(s)ds$, $\zeta_{im} = \int \{Y_i(t) - \mu_Y(t)\}\psi_m(t)dt$. Therefore, the PACE estimates and the estimates based on integral approximations can be considered equivalent in the dense design case that we consider here.

A.2 Estimation Procedures

To obtain the FPC scores for predictor and response processes (in case of a functional response), we adopt the PACE (Principal Analysis by Conditional Expectation) methodology (Yao et al., 2005*a*). Estimating the predictor mean function μ_X by local linear scatterplot smoothers, one minimizes

$$\sum_{i=1}^{n} \sum_{j=1}^{n_i} K_1\left(\frac{s_{ij} - s}{b_X}\right)\{U_{ij} - \beta_0^X - \beta_1^X (s - s_{ij})\}^2 \quad (28)$$

with respect to β_0^X, β_1^X, to obtain $\hat{\mu}_X(s) = \hat{\beta}_0^X(s)$. The kernel K_1 is assumed to be a smooth symmetric density function and b_X is a bandwidth. Estimating the covariance function $G_X(s_1, s_2)$ of predictor processes X, define $\mathrm{G}_i^X(s_{ij_1}, s_{ij_2}) = (U_{ij_1} - \hat{\mu}_X(s_{ij_1}))(U_{ij_2} - \hat{\mu}_X(s_{ij_2}))$, and the local linear surface smoother through minimizing

$$\sum_{i=1}^{n} \sum_{1 \le j_1 \ne j_2 \le n_i} K_2\left(\frac{s_{ij_1} - s_1}{h_X}, \frac{s_{ij_2} - s_2}{h_X}\right)\{\mathrm{G}_i^X(s_{ij_1}, s_{ij_2}) - f(\beta^X, (s_1, s_2), (s_{ij_1}, s_{ij_2}))\}^2, \tag{29}$$

where $f(\beta^X, (s_1, s_2), (s_{ij_1}, s_{ij_2})) = \beta_0^X + \beta_{11}^X(s_1 - s_{jl_1}) + \beta_{12}^X(s_2 - s_{ij_2})$, with respect to $\boldsymbol{\beta}^X = (\beta_0^X, \beta_{11}^X, \beta_{12}^X)^T$, yielding $\widehat{G}_X(s_1, s_2) = \hat{\beta}_0^X(s_1, s_2)$. Here, the kernel K_2 is a two-dimensional smooth density with zero mean and finite covariances and h_X is a bandwidth. An essential feature is the omission of the diagonal elements $j_1 = j_2$ which are contaminated with the measurement errors.

For the estimation of the variance of the measurement error σ_X^2, we fit a local quadratic component orthogonal to the diagonal of G_X, and a local linear component in the direction of the diagonal. Denote the diagonal of the resulting surface estimate by $\widetilde{G}_X(s)$, and a local linear smoother focusing on diagonal values $\{G_X(s, s) + \sigma_X^2\}$ by $\widehat{V}_X(s)$, using bandwidth h_X^*. Let $a_0 = \inf\{s : s \in \mathcal{S}\}$, $b_0 = \sup\{s : s \in \mathcal{S}\}$, $|\mathcal{S}| = b_0 - a_0$, $\mathcal{S}_1 = [a_0 + |\mathcal{S}|/4, b_0 - |\mathcal{S}|/4]$. The estimate of σ_X^2 is

$$\hat{\sigma}_X^2 = 2 \int_{\mathcal{S}_1} \{\widehat{V}_X(s) - \widetilde{G}_X(s)\}\, ds / |\mathcal{S}|, \tag{30}$$

if $\hat{\sigma}_X^2 > 0$, and $\hat{\sigma}_X^2 = 0$ otherwise. Estimates of eigenvalues/eigenfunctions $\{\lambda_k, \phi_k\}_{k \ge 1}$ are obtained as numerical solutions $\{\hat{\lambda}_k, \hat{\phi}_k\}_{k \ge 1}$ of suitably discretized eigenequations,

$$\int_{\mathcal{S}} \widehat{G}_X(s_1, s_2)\hat{\phi}_k(s_2) ds_2 = \hat{\lambda}_k \hat{\phi}_k(s_1), \tag{31}$$

with orthonormal constraints on $\{\hat{\phi}_k\}_{k \ge 1}$. These estimates are unique up to a sign change, and a projection on the space of positive definite covariance surfaces is obtained by simply omitting components with non-positive eigenvalues in the final representation. Analogous procedures are applied to response processes Y.

A.3 Assumptions

We compile the necessary assumptions for establishing theoretical properties. Recall $b_X = b_X(n)$, $h_X = h_X(n)$, $h_X^* = h_X^*(n)$, $b_Y = b_Y(n)$, $h_Y = h_Y(n)$, $h_Y^* = h_Y^*(n)$ are the bandwidths for estimating $\hat{\mu}_X$ and $\hat{\mu}_Y$ in (28), $\widehat{G}_X$ and $\widehat{G}_Y$ in (29), and $\widehat{V}_X$ and $\widehat{V}_Y$ in (30). As the number of subjects $n \to \infty$, we require:

(A1.1) $b_X \to 0$, $h_X^* \to 0$, $nb_X^4 \to \infty$, $n{h_X^*}^4 \to \infty$, $nb_X^6 < \infty$, and $n{h_X^*}^6 < \infty$.

(A1.2) $h_X \to 0$, $nh_X^6 \to \infty$, and $nh_X^8 < \infty$.

For FAM with functional responses, analogous requirements are

(B1.1) $b_Y \to 0$, $h_Y^* \to 0$, $nb_Y^4 \to \infty$, $n{h_Y^*}^4 \to \infty$, $nb_Y^6 < \infty$, and $n{h_Y^*}^6 < \infty$.

(B1.2) $h_Y \to 0$, $nh_Y^6 \to \infty$, and $nh_Y^8 < \infty$.

The time points $\{s_{ij}\}_{i=1,\ldots,n;j=1,\ldots,n_i}$ here are considered deterministic. Write the sorted time points across all subjects as $a_0 \le s_{(1)} \le \ldots \le s_{(N_n)} \le b_0$, and $\Delta_X = \max\{s_{(k)} - s_{(k-1)} : k = 1, \ldots, N+1\}$, where $N_n = \sum_{i=1}^n n_i$, $\mathcal{S} = [a_0, b_0]$, $s_{(0)} = a_0$, and $s_{(N+1)} = b_0$. For the ith subject, suppose that the time points s_{ij} have been ordered non-decreasingly. Let $\Delta_{iX} = \max\{s_{ij} - s_{i,j-1} : j = 1, \ldots, n_i + 1\}$, $\Delta_X^* = \max\{\Delta_{iX} : i = 1, \ldots, n\}$, where $s_{i0} = a_0$ and $s_{i,n_i+1} = b_0$, and $\bar{n}_x = n^{-1}\sum_{i=1}^n n_i$. To obtain uniform consistency, we require both the pooled data across all subjects and also the data from each subject to be dense in the time domain $\mathcal{S}$. Assume that

(A2.1) $\Delta_X = O(\min\{n^{-1/2}b_X^{-1}, n^{-1/2}{h_X^*}^{-1}, n^{-1/4}h_X^{-1}\})$.

(A2.2) $\bar{n}_x \to \infty$, $\max\{n_i : i = 1, \ldots, n\} \le C\bar{n}_x$ for some $C > 0$, and $\Delta_X^* = O(1/\bar{n}_x)$.

For the FAM with response process Y and observations $\{t_{il}, V_{il}\}$, $l = 1, \ldots, m_i$, $i = 1, \ldots, n$, we analogously define the quantities Δ_Y, Δ_{iY}, Δ_Y^*, $\bar{m}_y$ and assume that

(B2.1) $\Delta_Y = O(\min\{n^{-1/2}b_Y^{-1}, n^{-1/2}{h_Y^*}^{-1}, n^{-1/4}h_Y^{-1}\})$.

(B2.2) $\bar{m}_y \to \infty$, $\max\{m_i : i = 1, \ldots, n\} \le C\bar{m}_y$ for some $C > 0$, and $\Delta_Y^* = O(1/\bar{m}_y)$.

Denote by $U_i(s) \overset{\text{i.i.d.}}{\sim} U(s)$ the distributions that generate U_{ij} for the ith subject at $s = s_{ij}$. Analogously, let $V_i(t) \overset{\text{i.i.d.}}{\sim} V(t)$ denote the distributions that yield V_{il} for the ith subject at t_{il}. Assume that the fourth moments of $U(s)$ and $V(t)$ are uniformly bounded for all $s \in \mathcal{S}$ and $t \in \mathcal{T}$, respectively; i.e.,

(A3) $\sup_{s\in\mathcal{S}} E[U^4(s)] < \infty$.

(B3) $\sup_{t\in\mathcal{T}} E[V^4(t)] < \infty$.

Denoting the Fourier transforms of kernels K_1 and K_2 by $\kappa_1(t) = \int e^{-iut} K_1(u)du$ and $\kappa_2(t, s) = \int e^{-(iut+ivs)} K_2(u, v)du\, dv$, we require

(C1.1) $\kappa_1(t)$ is absolutely integrable, i.e., $\int |\kappa_1(t)|dt < \infty$, and K_1 is Lipschitz continuous on its compact support.

(C1.2) $\kappa_2(t, s)$ is absolutely integrable; i.e., $\int\int |\kappa_2(t, s)|dtds < \infty$.

In the sequel, let $g_{u1}(u; s)$, $g_{u2}(u_1, u_2; s_1, s_2)$, $g_{v1}(v; t)$ and $g_{v2}(v_1, v_2; t_1, t_2)$ denote the density functions of $U(s)$, $(U(s_1), U(s_2))$, $V(t)$, and $(V(t_1), V(t_2))$, respectively, and p_k and q_m the densities of ξ_k and ζ_m. It is assumed that these density functions satisfy the following regularity conditions:

(C2.1) $(d^2/ds^2)g_{u1}(u; s)$ exists and is uniformly continuous on $\Re \times \mathcal{S}$; $(d^2/(ds_1^{\ell_1} ds_2^{\ell_2}))g_{u2}(u_1, u_2; s_1, s_2)$ exists and is uniformly continuous on $\Re^2 \times \mathcal{S}^2$, for $\ell_1 + \ell_2 = 2$, $0 \le \ell_1, \ell_2 \le 2$.

(C2.2) $(d^2/dt^2)g_{v1}(v; t)$ exists and is uniformly continuous on $\Re \times \mathcal{T}$; $(d^2/(dt_1^{\ell_1} dt_2^{\ell_2}))g_{v2}(v_1, v_2; t_1, t_2)$ exists and is uniformly continuous on $\Re^2 \times \mathcal{T}^2$, for $\ell_1 + \ell_2 = 2$, $0 \le \ell_1, \ell_2 \le 2$.

(C2.3) The second derivative $p^{(2)}(x)$ exists and is continuous on $\Re$.

Let $\|f\|_\infty = \sup_{x\in\mathcal{A}} |f(t)|$ for an arbitrary function f with support $\mathcal{A}$, and $\|g\| = \sqrt{\int_{\mathcal{A}} g^2(t)dt}$ for any $g \in L^2(\mathcal{A})$. The following assumptions are needed for Theorem 1.

(A4) $E(\|X'\|_\infty^2) < \infty$, $E(\|X'^2\|_\infty^2) = o(\bar{n}_x)$ and $E(\xi_k^4) < \infty$ for any fixed k.

(B4) $E(\|Y'\|_\infty^2) < \infty$, $E(\|Y'^2\|_\infty^2) = o(\bar{m}_y)$ and $E(\zeta_m^4) < \infty$ for any fixed m.

(A5) $nh_X^4 h_k^2 \to 0$, $nb_X^2 h_k^2 \to 0$ and $\bar{n}_x h_k^2 \to 0$.

For brevity, denote "$\sum_{k=1}^K$" by "$\sum_k$", "$\sum_{m=1}^M$" by "$\sum_m$", "$\sum_{k=1}^K \sum_{m=1}^M$" by "$\sum_{k,m}$", "$\max_{k=1,\ldots,K}$" by "$\max_k$" and "$\max_{m=1,\ldots,M}$" by "$\max_m$" in the following assumptions (A6) and (B5), which are needed for Theorem 2 and are assumed to hold for all fixed ξ_k, $K \le K_0$,

(A6) $\sum_k p_k(\xi_k) h_k^{-1} = o\{\min(n^{1/2} b_X, \bar{n}_x^{1/2})\}, \max_k \|\phi_k \phi_k'\|_\infty = O(\bar{n}_x), \sum_k p_k(\xi_k) \pi_k^x h_k^{-1} = o(n^{1/2} h_X^2), \sum_k E(\xi_k^4) < \infty$.

(B5) For any fixed ξ_k, $t \in \mathcal{T}$, $K \le K_0$, $M \le M_0$,

$$\begin{aligned}
&\textstyle\sum_{k,m} p_k(\xi_k)|\psi_m(t)| h_{mk}^{-1} = o\{\min(n^{1/2} b_X, \bar{n}_x^{1/2})\}, && \max_k \|\phi_k \phi_k'\|_\infty = O(\bar{n}_x),\\
&\textstyle\sum_{k,m} p_k(\xi_k)|\psi_m(t)| = o\{\min(n^{1/2} b_Y, \bar{m}_y^{1/2})\}, && \max_m \|\psi_m \psi_m'\|_\infty = O(\bar{m}_y),\\
&\textstyle\sum_{k,m} p_k(\xi_k) \pi_k^x |\psi_m(t)| h_{mk}^{-1} = o(n^{1/2} h_X^2), && \textstyle\sum_{k,m} p_k(\xi_k) \pi_m^y |\psi_m(t)| = o(n^{1/2} h_Y^2),\\
&\textstyle\sum_{k,m} \pi_m^y |f_{mk}(\xi_k)| = o(n^{1/2} h_Y^2), && \textstyle\sum_m E(\zeta_m^4) < \infty.
\end{aligned}$$

To guarantee the consistency of predictions, we also require that for fixed ξ_k and $t \in \mathcal{T}$,

(A7) $\sum_{k=1}^K [|f_k''(\xi_k)| h_k^2 + n^{-1/2} \{\mathrm{var}(Y|\xi_k)\}^{1/2} p_k^{-1/2}(\xi_k) h_k^{-1/2}] \to 0$.

(B6) $\sum_{k=1}^K \sum_{m=1}^M [|f_{mk}''(\xi_k)\psi_m(t)| h_{mk}^2 + n^{-1/2} \{\mathrm{var}(\zeta_m|\xi_k)\}^{1/2} p_k^{-1/2}(\xi_k) |\psi_m(t)| h_{mk}^{-1/2}] \to 0$.

References

Ash, R. B., and Gardner, M. F. (1975), *Topics in stochastic processes*, New York: Academic Press.

Billingsley, P. (1995), *Probability and measure*, 2nd edn, New York: John Wiley.

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

Cardot, H., Ferraty, F., Mas, A., and Sarda, P. (2003), "Testing hypotheses in the functional linear model," *Scandinavian Journal of Statistics*, 30, 241–255.

Cardot, H., and Sarda, P. (2005), "Estimation in generalized linear models for functional data via penalized likelihood," *Journal of Multivariate Analysis*, 92, 24–41.

Cuevas, A., Febrero, M., and Fraiman, R. (2002), "Linear functional regression: the case of fixed design and functional response," *Canadian Journal of Statistics*, 30, 285–300.

Draper, N. R., and Smith, H. (1998), *Applied Regression Analysis*, New York: Wiley.

Escabias, M., Aguilera, A. M., and Valderrama, M. J. (2004), "Principal component estimation of functional logistic regression: discussion of two different approaches," *Journal of Nonparametric Statistics*, 16, 365 – 384.

Fan, J. Q., Yao, Q. W., and Cai, Z. W. (2003), "Adaptive varying-coefficient linear models," *Journal of the Royal Statistical Society, Series B*, 65, 57 – 80.

Fan, J., and Zhang, J. T. (2000), "Two-step Estimation of Functional Linear Models with Applications to Longitudinal Data," *Journal of the Royal Statistical Society, Series B*, 62, 303–322.

Foutz, N., and Jank, W. (2008), "The wisdom of crowds: pre-release forecasting via functional shape analysis of the online virtual stock market," manuscript.

Hall, P., and Horowitz, J. L. (2007), "Methodology and convergence rates for functional linear regression," *The Annals of Statistics*, 35, 70–91.

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

Hastie, T. J., and Tibshirani, R. J. (1990), *Generalized Additive Models* Chapman and Hall, London.

He, G., Müller, H. G., and Wang, J. L. (2000), "Extending correlation and regression from multivariate to functional data," in *Asymptotics in Statistics and Probability*, ed. M. L. Puri VSP International Science Publishers, pp. 301–315.

He, G., Müller, H. G., and Wang, J. L. (2003), "Functional canonical analysis for square integrable stochastic processes," *Journal of Multivariate Analysis*, 85, 54 – 77.

James, G. (2002), "Generalized linear models with functional predictors," *Journal of the Royal Statistical Society, Series B*, 64, 411–432.

James, G., and Silverman, B. (2005), "Functional adaptive model estimation," *Journal of the American Statistical Association*, 100, 565–576.

Liu, B., and Müller, H. (2008), "Functional data analysis for sparse auction data," in *Statistical Methods in eCommerce Research*, eds. W. Jank, and G. Shmueli Wiley, New York.

Malfait, N., and Ramsay, J. O. (2003), "The historical functional linear model," *Canadian Journal of Statistics*, 31, 115–128.

Mammen, E., and Park, B. U. (2005), "Bandwidth selection for smooth backfitting in additive models," *The Annals of Statistics*, 33, 1260–1294.

Morris, J., Vannucci, M., Brown, P., and Carroll, R. (2003), "Wavelet-based nonparametric modeling of hierarchical functions in colon carcinogenesis (with discussion)," *Journal of the American Statistical Association*, 98, 573–594.

Müller, H. G. (2005), "Functional modelling and classification of longitudinal data," *Scandinavian Journal of Statistics*, 32, 223–240.

Müller, H. G., and Stadtmüller, U. (2005), "Generalized functional linear models," *The Annals of Statistics*, 33, 774–805.

Ramsay, J. O., and Dalzell, C. J. (1991), "Some tools for functional data analysis," *Journal of the Royal Statistical Society, Series B*, 53, 539–572.

Ramsay, J. O., and Silverman, B. W. (2005), *Functional Data Analysis*, 2nd edn, New York: Springer.

Rice, J., and Silverman, B. W. (1991), "Estimating the mean and covariance structure nonparametrically when the data are curves," *Journal of the Royal Statistical Society, Series B*, 53, 233–243.

Sood, A., James, G., and Tellis, G. (2008), "Functional regression: a new model for predicting market penetration of new products," *Marketing Science*, to appear.

Yao, F., and Lee, T. C. M. (2006), "Penalized spline models for functional principal component analysis," *Journal of the Royal Statistical Society, Series B*, 68, 3–25.

Yao, F., Müller, H. G., Clifford, A. J., Dueker, S. R., Follett, J., L., Y., Buchholz, B. A., and Vogel, J. S. (2003), "Shrinkage estimation for functional principal component scores with application to the population kinetics of plasma folate," *Biometrics*, 59, 676–685.

Yao, F., Müller, H. G., and Wang, J. L. (2005*a*), "Functional data analysis for sparse longitudinal data," *Journal of the American Statistical Association*, 100, 577–590.

Yao, F., Müller, H. G., and Wang, J. L. (2005*b*), "Functional linear regression analysis for longitudinal data," *The Annals of Statistics*, 33, 2873–2903.

Table 1: Simulation results for the comparison of predictions obtained by the Functional Additive Model (FAM) and functional linear regression (LIN), for models with scalar response (see (7) for the FAM version, (1) for the linear version) and with functional response ((8) for the FAM version, (2) for the linear version), for both dense and sparse designs. The true component functions are linear (LF) and nonlinear (NLF), and the true FPC scores of the predictor process are generated from normal distributions, as described in Section 6. Simulations were based on 400 Monte Carlo runs with $n = 100$ predictors and responses per sample. Shown in the table are the Monte Carlo estimates of the 25th, 50th and 75th percentiles of the relative prediction error, RPE (22).

Design	Response	Model	True	25th	50th	75th	True	25th	50th	75th
Dense	Scalar	FAM	NLF	.0683	.2983	1.820	LF	.0075	.0431	.3024
		LIN		.6458	1.068	1.870		.0102	.0335	.1362
	Functional	FAM	NLF	.0025	.0086	.0279	LF	.0005	.0012	.0031
		LIN		.0109	.0363	.0705		.0004	.0009	.0019
Sparse	Scalar	FAM	NLF	.1124	.4437	2.884	LF	.0176	.1066	.7149
		LIN		.6614	1.071	1.822		.0270	.1133	.5378
	Functional	FAM	NLF	.0066	.0156	.0432	LF	.0023	.0042	.0090
		LIN		.0138	.0389	.0737		.0023	.0044	.0094

Table 2: Simulation results obtained for the same settings as those in Table 1, except that the true FPC scores of the predictor process are generated from right skewed distributions related to Gamma(4,1), as described in Section 6.

Design	Response	Model	True	25th	50th	75th	True	25th	50th	75th
Dense	Scalar	FAM	NLF	.0163	.0885	1.035	LF	.0014	.0066	.0410
		LIN		.0396	.1827	3.383		.0008	.0030	.0135
	Functional	FAM	NLF	.0028	.0101	.0375	LF	.0005	.0013	.0034
		LIN		.0126	.0417	.0819		.0003	.0005	.0010
Sparse	Scalar	FAM	NLF	.0448	.2775	3.558	LF	.0291	.1511	1.031
		LIN		.1867	.7179	3.642		.0305	.1423	.9316
	Functional	FAM	NLF	.0072	.0146	.0485	LF	.0023	.0041	.0090
		LIN		.0156	.0422	.0803		.0021	.0042	.0091

Table 3: Functional R^2 (17), 25th, 50th and 75th percentiles and mean of the cross-validated observed relative prediction errors, $\text{RPE}_{(-i),f}$ (22), comparing FAM and functional linear regression models for zygotic data.

	25th	50th	75th	Mean	R^2
FAM	.0506	.0776	.1662	.1301	0.19
LIN	.0479	.0891	.1727	.1374	0.16

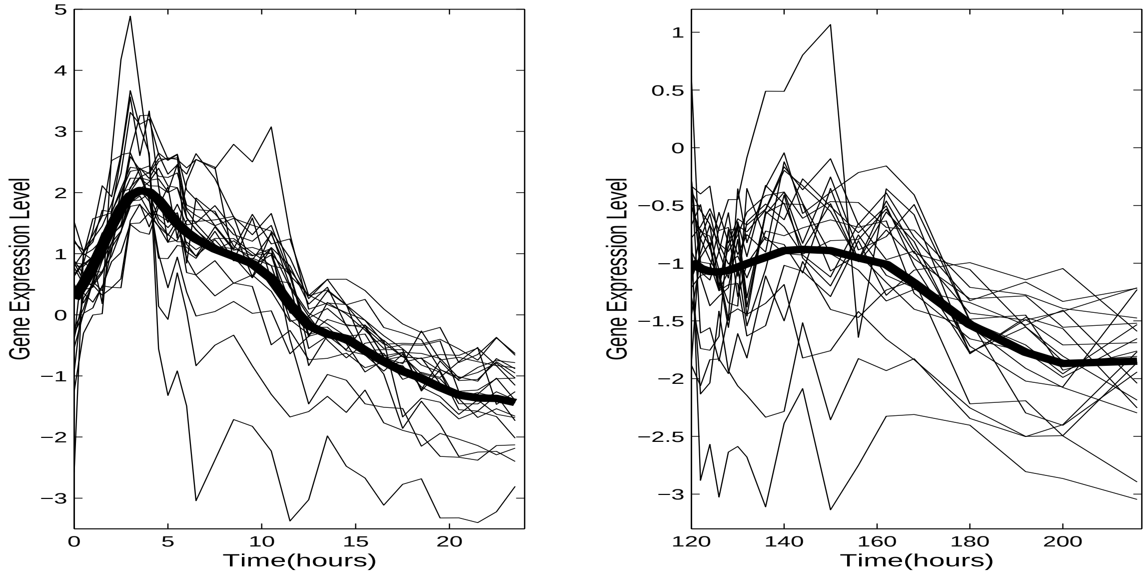

Figure 1: Observed (thin curves) gene expression levels and smoothed estimates of the mean functions (thick curves) for embryo phase (left) and pupa phase (right), for zygotic data.

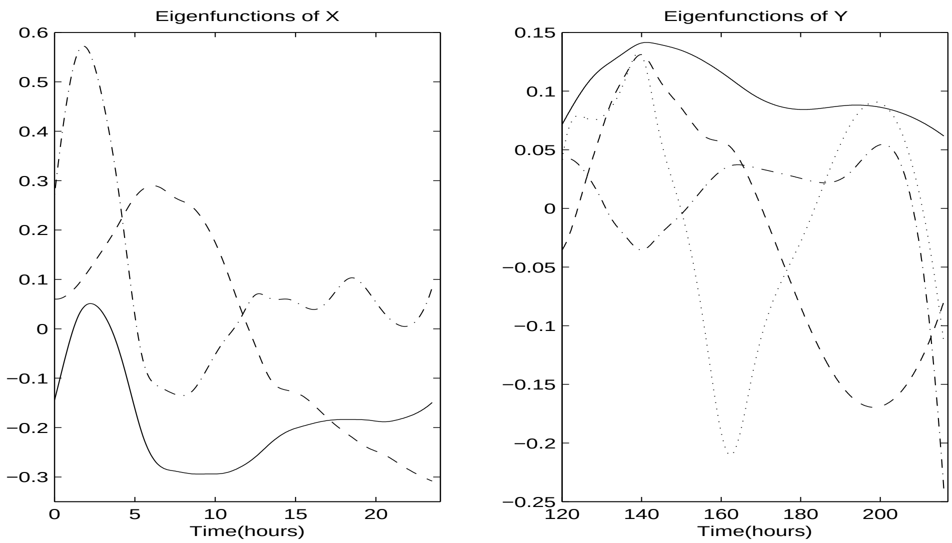

Figure 2: Left: Estimates of the first 3 eigenfunctions for predictor trajectories from the embryo phase (predictor process), $K = 3$ components selected by AIC, accounting for 80.69% (first component, solid), 13.65% (second component, dashed) and 2.71% (third component, dash-dot) of the total variation. Right: Estimates of the first 4 eigenfunctions of the pupa (metamorphosis) phase (response process), $M = 4$ components selected by AIC, accounting for 80.47% (first component, solid), 8.89% (second component, dashed), 4.2% (third component, dash-dot) and 1.55% (fourth component, dotted) of the total variation.

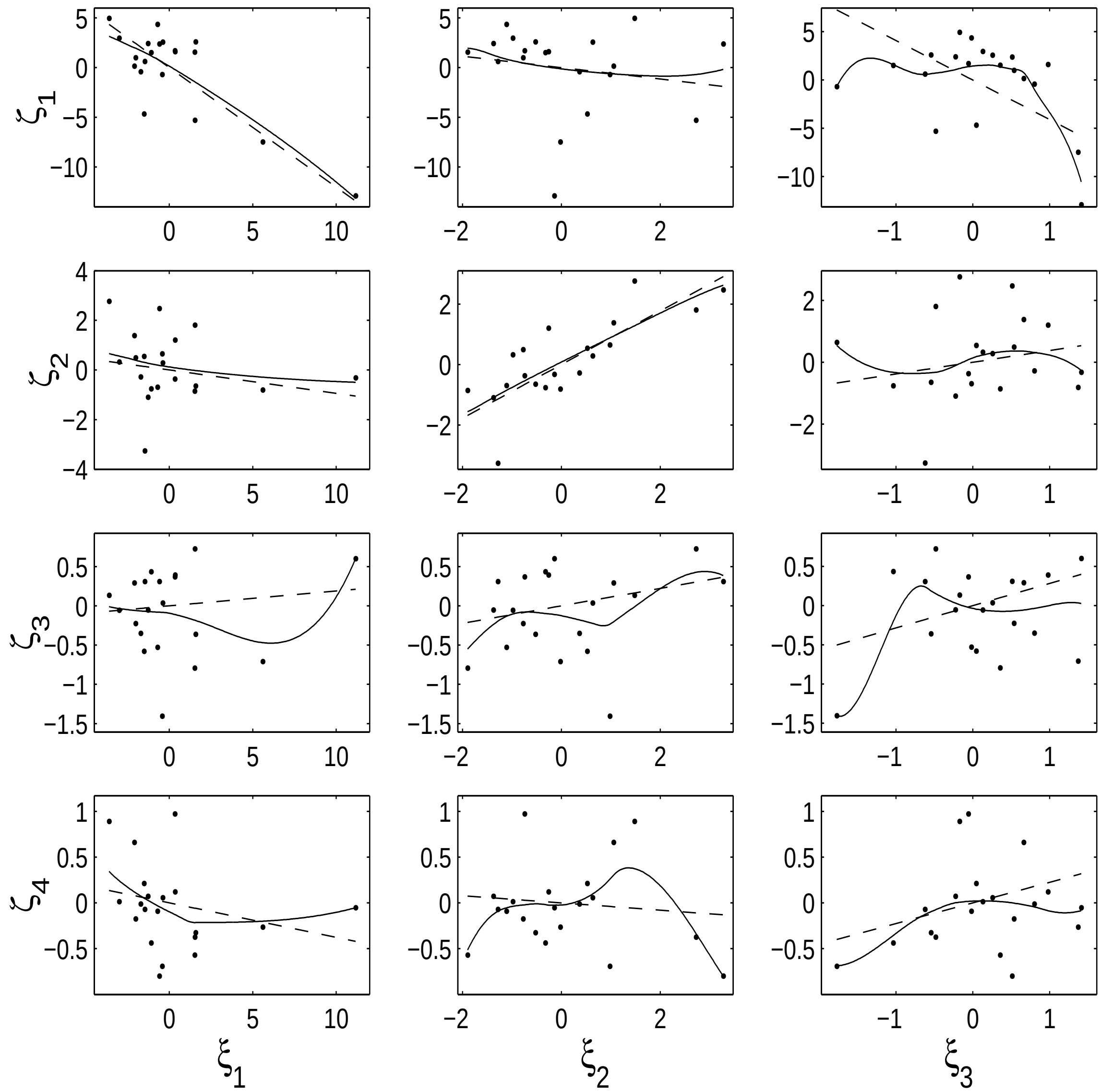

Figure 3: Scatterplots (dots), local polynomial (solid) and linear (dashed) estimates for the regressions of estimated FPC scores of the pupa phase (responses, y-axis) versus those for the embryo phase (predictors, x-axis). The FPC scores of the embryo phase expressions are arranged from left to right ($\xi_k, k = 1, 2, 3$) and the FPC scores of the pupa phase expressions are arranged from top to bottom ($\zeta_m, m = 1, 2, 3, 4$), for zygotic data.

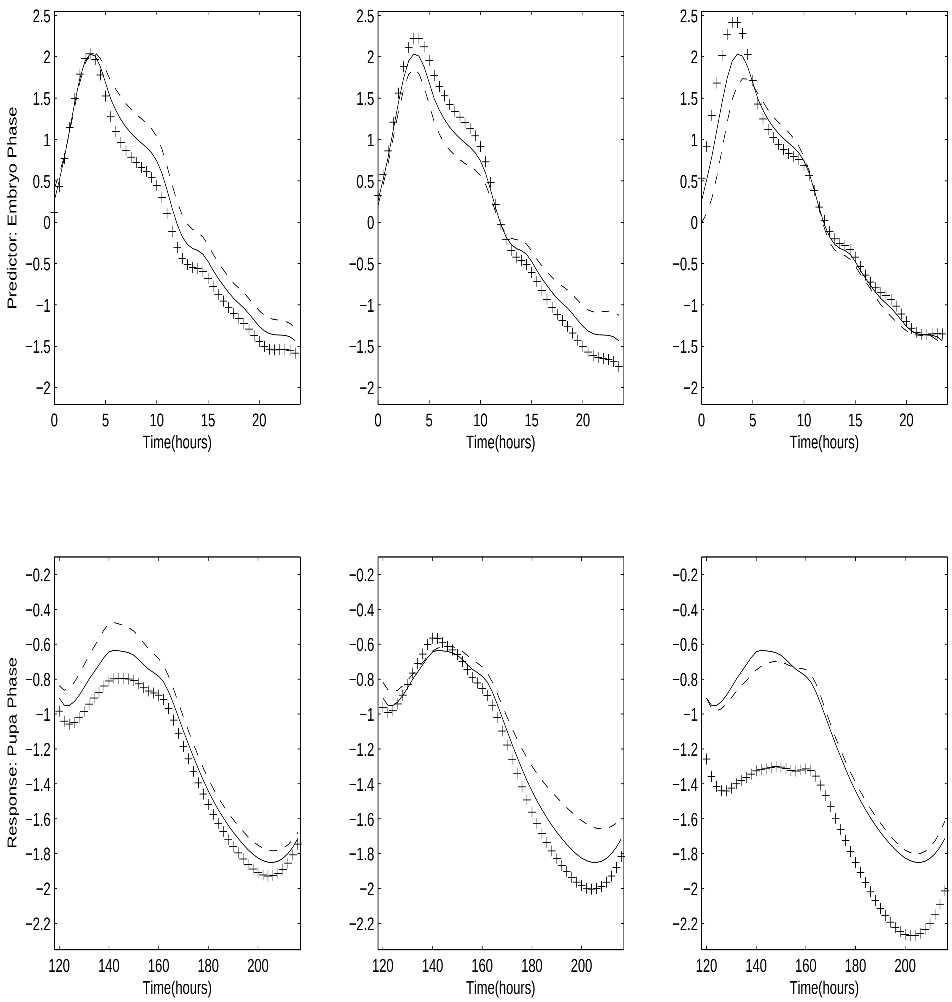

Figure 4: Top panels: Predictor trajectories $\hat{\mu}_X(s) + \alpha\hat{\phi}_k(s)$, for $\alpha = 0$ (solid), $\alpha = 1$ (dashed) and $\alpha = -1$ ($_{+++}$) for $k = 1$ (left panels), $k = 2$ (middle panels) and $k = 3$ (right panels). Bottom panels: Corresponding response trajectories $\hat{\mu}_Y(t) + \sum_{m=1}^{M} \left\{\hat{f}_{km}(\alpha) + \sum_{\ell\neq k} \hat{f}_{\ell m}(0)\right\}\hat{\psi}_m(t)$, for zygotic data.

Supplement for "Functional Additive Models"

1. NOTATIONS AND AUXILIARY RESULTS

Covariance operators are denoted by $\mathcal{G}_X$, $\widehat{\mathcal{G}}_X$, generated by kernels G_X, $\widehat{G}_X$; i.e., $\mathcal{G}_X(f) = \int_\mathcal{T} G_X(s,t)f(s)ds$, $\widehat{\mathcal{G}}_X(f) = \int_\mathcal{T} \widehat{G}_X(s,t)f(s)ds$ for any $f \in L^2(\mathcal{T})$. Define

$$\begin{aligned} &D_X = \textstyle\int_{\mathcal{T}^2}\{\widehat{G}_X(s,t) - G_X(s,t)\}^2 dsdt, && \delta_k^x = \min_{1\le j\le k}(\lambda_j - \lambda_{j+1}), \\ &K_0 = \inf\{j \ge 1 : \lambda_j - \lambda_{j+1} \le 2D_X\} - 1, && \pi_k^x = 1/\lambda_k + 1/\delta_k^x. \end{aligned} \tag{32}$$

Let $K = K(n)$ denote the numbers of leading eigenfunctions included to approximate X as sample size n varies; i.e., $\widehat{X}_i(s) = \hat{\mu}_X(s) + \sum_{k=1}^K \hat{\xi}_{ik}\hat{\phi}_k(s)$. Analogously, define the quantities $\mathcal{G}_Y$, $\widehat{\mathcal{G}}_Y$, D_Y, δ_m^y π_m^y, M_0 and M for the process Y, for the case of functional responses. The following lemma gives the weak uniform convergence rates for the estimators of the FPCs, setting the stage for the subsequent developments. The proof is in Section 2.

Lemma 1 *Under (A1.1)-(A3), (C1.1), (C1.2) and (C2.1),*

$$\sup_{t\in\mathcal{S}} |\hat{\mu}_X(s) - \mu_X(s)| = O_p(\frac{1}{\sqrt{n}b_X}), \quad \sup_{s_1,s_2\in\mathcal{S}} |\widehat{G}_X(s_1,s_2) - G_X(s_1,s_2)| = O_p(\frac{1}{\sqrt{n}h_X^2}), \tag{33}$$

and as a consequence, $\hat{\sigma}_X^2 - \sigma_X^2 = O_p(n^{-1/2}h_X^{-2} + n^{-1/2}{h_X^*}^{-1})$. *Considering eigenvalues* λ_k *of multiplicity one,* $\hat{\phi}_k$ *can be chosen such that*

$$P(\sup_{1\le k\le K_0} |\hat{\lambda}_k - \lambda_k| \le D_X) = 1, \quad \sup_{s\in\mathcal{S}} |\hat{\phi}_k(s) - \phi_k(s)| = O_p(\frac{\pi_k^x}{\sqrt{n}h_X^2}), \quad k = 1, \ldots, K_0, \tag{34}$$

where D_X, π_k^x *and* K_0 *are defined in (32).*

Analogously, under (B1.1)-(B3), (C1.1), (C1.2) and (C2.2),

$$\sup_{t\in\mathcal{T}} |\hat{\mu}_Y(t) - \mu_Y(t)| = O_p(\frac{1}{\sqrt{n}b_Y}), \quad \sup_{t_1,t_2\in\mathcal{T}} |\widehat{G}_Y(t_1,t_2) - G_Y(t_1,t_2)| = O_p(\frac{1}{\sqrt{n}h_Y^2}), \tag{35}$$

and as a consequence, $\hat{\sigma}_Y^2 - \sigma_Y^2 = O_p(n^{-1/2}h_Y^{-2} + n^{-1/2}{h_Y^*}^{-1})$. *Considering eigenvalues* ρ_m *of multiplicity one,* $\hat{\psi}_m$ *can be chosen such that*

$$P(\sup_{1\le m\le M_0} |\hat{\rho}_m - \rho_m| \le D_Y) = 1, \sup_{t\in\mathcal{T}} |\hat{\psi}_m(t) - \psi_m(t)| = O_p(\frac{\pi_k^y}{\sqrt{n}h_Y^2}), \; m = 1, \ldots, M_0, \tag{36}$$

where D_Y, π_k^y and M_0 are defined analogously to (32) for process Y.

Recall that $\|f\|_\infty = \sup_{x\in\mathcal{A}} |f(t)|$ for an arbitrary function f with support $\mathcal{A}$, and $\|g\| = \sqrt{\int_{\mathcal{A}} g^2(t)dt}$ for any $g \in L^2(\mathcal{A})$ and define

$$\begin{aligned}
\theta_{ik}^{(1)} &= c_1\|X_i\| + c_2\|X_iX_i'\|_\infty\Delta_X^* + c_3, & Z_k^{(1)} &= \sup_{s\in\mathcal{S}} |\hat{\phi}_k(s) - \phi_k(s)|,\\
\theta_{ik}^{(2)} &= 1 + \|\phi_k\phi_k'\|_\infty\Delta_X^*, & Z_k^{(2)} &= \sup_{s\in\mathcal{S}} |\hat{\mu}_X(s) - \mu_X(s)|,\\
\theta_{ik}^{(3)} &= c_4\|X_i\|_\infty + c_5\|X_i'\|_\infty + c_6, & Z_k^{(3)} &= \|\phi_k'\|_\infty\Delta_X^*,\\
\theta_{ik}^{(4)} &= |\textstyle\sum_{j=2}^{n_i} \epsilon_{ij}\phi_k(s_{ij})(s_{ij} - s_{i,j-1})|, & Z_k^{(4)} &\equiv 1,\\
\theta_{ik}^{(5)} &= \textstyle\sum_{j=2}^{n_i} |\epsilon_{ij}|(s_{ij} - s_{i,j-1}), & Z_k^{(5)} &\equiv Z_k^{(1)},
\end{aligned} \tag{37}$$

for some positive constants $a_1, \ldots, c_6$ that do not depend on i or k. Similarly, define corresponding quantities for the process Y as follows,

$$\begin{aligned}
\vartheta_{im}^{(1)} &= d_1\|Y_i\| + d_2\|Y_iY_i'\|_\infty\Delta_Y^* + d_3, & Q_m^{(1)} &= \sup_{t\in\mathcal{T}} |\hat{\psi}_m(t) - \psi_m(t)|,\\
\vartheta_{im}^{(2)} &= 1 + \|\psi_m\psi_m'\|_\infty\Delta_Y^*, & Q_m^{(2)} &= \sup_{t\in\mathcal{T}} |\hat{\mu}_Y(t) - \mu_Y(t)|,\\
\vartheta_{im}^{(3)} &= d_4\|Y_i\|_\infty + d_5\|Y_i'\|_\infty + d_6, & Q_m^{(3)} &= \|\psi_m'\|_\infty\Delta_Y^*,\\
\vartheta_{im}^{(4)} &= |\textstyle\sum_{l=2}^{m_i} \varepsilon_{il}\psi_m(t_{il})(t_{il} - t_{i,l-1})|, & Q_m^{(4)} &\equiv 1,\\
\vartheta_{im}^{(5)} &= \textstyle\sum_{l=2}^{m_i} |\varepsilon_{il}|(t_{il} - t_{i,l-1}), & Q_m^{(5)} &\equiv Q_m^{(1)},
\end{aligned} \tag{38}$$

for some positive constants $d_1, \ldots, d_6$ that do not depend on i or m. We note that the subscripts are mainly for notational convenience and do not necessarily reflect dependence on these indices. Note that in (37), $\theta_{ik}^{(1)}$, $\theta_{ik}^{(3)}$, $\theta_{ik}^{(5)}$ in fact do not depend on k and $\theta_{ik}^{(2)}$ does not depend on i, while $Z_k^{(3)}$ is deterministic and $Z_k^{(4)}$ is a constant. More importantly, we emphasize that $\theta_{ik}^{(\ell)}$ are i.i.d. over i ($\ell = 1, 3, 4, 5$) or free of i ($\ell = 2$), and that the $Z_k^{(\ell)}$ do not depend on i.

The next lemma is critical for the subsequent developments, providing exact upper bounds for the estimation errors $|\hat{\xi}_{ik}^I - \xi_{ik}|$ and $|\hat{\zeta}_{im}^I - \zeta_{im}|$, for the FPC estimates $\hat{\xi}_{ik}^I$, $\hat{\zeta}_{im}^I$ (27).

Lemma 2 *For* $\theta_{ik}^{(\ell)}$, $Z_k^{(\ell)}$, $\vartheta_{im}^{(\ell)}$ *and* $Q_m^{(\ell)}$ *as defined in (37) and (38),*

$$|\hat{\xi}_{ik}^I - \xi_{ik}| \le \sum_{\ell=1}^{5} \theta_{ik}^{(\ell)} Z_k^{(\ell)}, \qquad |\hat{\zeta}_{im}^I - \zeta_{im}| \le \sum_{\ell=1}^{5} \vartheta_{im}^{(\ell)} Q_m^{(\ell)}. \tag{39}$$

The proof is in Section 2. In the sequel we suppress the superscript I in the FPC estimates $\hat{\xi}_{ik}^I$ and $\hat{\zeta}_{im}^I$.

Recall that the sequences of bandwidths h_k and h_{mk} are employed to obtain the estimates $\hat{f}_k$ and $\hat{f}_{mk}$ for the regression functions f_k and f_{mk}, and that the density of ξ_k is denoted by p_k. Define

$$\begin{aligned} \theta_k(x) &= p_k(x)\{\frac{\pi_k^x}{\sqrt{n}h_X^2} + \frac{1}{\sqrt{n}b_X} + \sqrt{\Delta_X^*}\}, \\ \vartheta_{mk}(x) &= p_k(x)\{\frac{\pi_m^y}{\sqrt{n}h_Y^2} + \frac{1}{\sqrt{n}b_Y} + \sqrt{\Delta_Y^*}\}. \end{aligned} \tag{40}$$

The weak convergence rates $\tilde{\theta}_k$ and $\tilde{\vartheta}_{mk}$ of the regression function estimators $\hat{f}_k(x)$ and $\hat{f}_{mk}(x)$ (see Theorem 1) are as follows,

$$\begin{aligned} \tilde{\theta}_k(x) &= \frac{\theta_k(x)}{h_k} + \frac{1}{2}|f_k''(x)|h_k^2 + \sqrt{\frac{\mathrm{var}(Y|x)\|K_1\|^2}{p_k(x)nh_k}}, \\ \tilde{\vartheta}_{mk}(x) &= \frac{\theta_k(x)}{h_{mk}} + \vartheta_{mk}(x) + \frac{1}{2}|f_{mk}''(x)|h_{mk}^2 + \sqrt{\frac{\mathrm{var}(\zeta_m|x)\|K_1\|^2}{p_k(x)nh_{mk}}}. \end{aligned} \tag{41}$$

Considering the predictions $\widehat{E}(Y|X)$ for the scalar response case and $\widehat{E}\{Y(t)|X\}$ for the functional response case, the numbers of eigenfunctions K and M used for approximating the infinite dimensional processes X and Y generally tend to infinity as the sample size n increases. We require $K \le K_0$ and $M \le M_0$ in (A6). Since it follows from (35) that $K_0 \to \infty$, as long as all eigenvalues λ_j are of multiplicity 1, and analogously for M_0, this is not a strong restriction. Denote the set of positive integers by $\mathcal{N}$ and $\mathcal{N}_k = \{1, \ldots, k\}$. Convergence rates θ_n^* and ϑ_n^* for the predictions (20) and

(21) are as follows,

$$\theta_n^* = \sum_{k=1}^K \{ \frac{\theta_k(\xi_k)}{h_k} + \frac{1}{2}|f_k''(\xi_k)|h_k^2 + \sqrt{\frac{\mathrm{var}(Y|\xi_k)\|K_1\|^2}{p_k(\xi_k)nh_k}} \} + \Big| \sum_{k\geq K+1} f_k(\xi_k) \Big|, \tag{42}$$

$$\vartheta_n^* = \sum_{k=1}^K \sum_{m=1}^M \{ (\frac{\theta_k(\xi_k)}{h_{mk}} + \vartheta_{mk}(\xi_k))|\psi_m(t)| + \frac{1}{2}|f_{mk}''(\xi_k)|\psi_m(t)|h_{mk}^2 + \sqrt{\frac{\mathrm{var}(\zeta_m|\xi_k)\|K_1\|^2}{p_k(\xi_k)nh_k}}|\psi_m(t)| + \frac{\pi_m^y|f_{mk}(\xi_k)|}{\sqrt{n}h_Y^2} \} + \Big| \sum_{(k,m)\in\mathcal{N}^2\backslash\mathcal{N}_K\times\mathcal{N}_M} f_{mk}(\xi_k)\psi_m(t) \Big|,$$

where θ_k and ϑ_k are defined in (40) and we note that ϑ_n^* depends on t.

2. PROOFS

Proof of Lemma 1. It is sufficient to show (33) and (34). The weak convergence results (33) for $\hat{\mu}_X$ and $\widehat{G}_X(s_1, s_2)$ have been derived in Lemma 2 of Yao and Lee (2006). Theorem 1 of Hall and Hosseini-Nasab (2006) implies the first equation of (34) for the estimated eigenvalues $\hat{\lambda}_k$. Assuming $\lambda_k > 0$ without loss of generality, we have

$$\begin{aligned}
&|\hat{\lambda}_k\hat{\phi}_k(s) - \lambda_k\phi_k(s)| = |\int_{\mathcal{S}} \widehat{G}_X(t,s)\hat{\phi}_k(t)dt - \int_{\mathcal{S}} G_X(t,s)\phi_k(t)dt| \\
&\leq \int_{\mathcal{S}} |\widehat{G}_X(t,s) - G_X(t,s)| \cdot |\hat{\phi}_k(t)|dt + \int_{\mathcal{S}} |G_X(t,s)| \cdot |\hat{\phi}_k(t) - \phi_k(t)|dt \\
&\leq \sqrt{\int_{\mathcal{S}} (\widehat{G}_X(t,s) - G_X(t,s))^2 dt} + \sqrt{\int_{\mathcal{S}} G_X^2(t,s)dt}\|\hat{\phi}_k - \phi_k\|,
\end{aligned}$$

and $|\hat{\lambda}_k\hat{\phi}_k(s)/\lambda_k - \phi_k(s)| = O_p\{D_X(1/\lambda_k + 1/\delta_k^x)\}$ uniformly in $s \in \mathcal{S}$ for $1 \leq k \leq K_0$, where K_0 is defined in (32). The second equation of (34) follows immediately. $\square$

Proof of Lemma 2. Let

$$\hat{\eta}_{ik} = \sum_{j=2}^{n_i} \{X_i(s_{ij}) - \hat{\mu}(s_{ij})\}\hat{\phi}_k(s_{ij})(s_{ij} - s_{i,j-1}),$$

$$\tilde{\eta}_{ik} = \sum_{j=2}^{n_i} \{X_i(s_{ij}) - \mu(s_{ij})\}\phi_k(s_{ij})(s_{ij} - s_{i,j-1}),$$

$$\hat{\tau}_{ik} = \sum_{j=2}^{n_i} \epsilon_{ij}\hat{\phi}_k(s_{ij})(s_{ij} - s_{i,j-1}), \qquad \tilde{\tau}_{ik} = \sum_{j=2}^{n_i} \epsilon_{ij}\phi_k(s_{ij})(s_{ij} - s_{i,j-1}).$$

Noting $\hat{\xi}_{ik} = \hat{\eta}_{ik} + \hat{\tau}_{ik}$, one finds

$$|\hat{\xi}_{ik} - \xi_{ik}| \leq \{|\hat{\eta}_{ik} - \tilde{\eta}_{ik}| + |\tilde{\eta}_{ik} - \xi_{ik}| + |\hat{\tau}_{ik}|\}. \tag{43}$$

Without loss of generality, assume $\|\phi_k\|_\infty \geq 1$, $\|\phi_k'\|_\infty \geq 1$, $\|X_i\|_\infty \geq 1$ and $\|X_i'\|_\infty \geq 1$. For $\theta_{ik}^{(\ell)}$ and $Z_k^{(\ell)}$ (37), $\ell = 1, \ldots, 5$, the first term on the r.h.s. of (43) is bounded by

$$\begin{aligned}
&\{\sum_{j=2}^{n_i}[|X_i(s_{ij}) - \hat{\mu}(s_{ij})| \cdot |\hat{\phi}_k(s_{ij}) - \phi_k(s_{ij})| + |\hat{\mu}(s_{ij}) - \mu(s_{ij})| \cdot |\phi_k(s_{ij})|](s_{ij} - s_{i,j-1})\} \\
&\leq \{\sum_{j=1}^{n_i}[|X_i(s_{ij})| + |\mu(s_{ij})| + 1]^2(s_{ij} - s_{i,j-1})\}^{1/2}\{\sum_{j=2}^{n_i}[\hat{\phi}_k(s_{ij}) - \phi_k(s_{ij})]^2(s_{ij} - s_{i,j-1})\}^{1/2} \\
&\quad + \{\sum_{j=1}^{n_i}[\hat{\mu}(s_{ij}) - \mu(s_{ij})]^2(s_{ij} - s_{i,j-1})\}^{1/2}\{\sum_{j=2}^{n_i}\phi_k^2(s_{ij})(s_{ij} - s_{i,j-1})\}^{1/2} \\
&\leq \theta_{ik}^{(1)}Z_k^{(1)} + \theta_{ik}^{(2)}Z_k^{(2)}.
\end{aligned}$$

The second term on the r.h.s. of (43) has the upper bound

$$|\tilde{\eta}_{ij} - \xi_{ik}| \leq \|(X_i + \mu)'\phi_k + (X_i + \mu)\phi_k'\|_\infty \Delta_X^* \leq \theta_{ik}^{(3)}Z_k^{(3)}.$$

From the above, the third term on the r.h.s. of (43) is bounded by $(\theta_{ik}^{(4)}Z_k^{(4)} + \theta_{ik}^{(5)}Z_k^{(5)})$.

□

Proof of Theorem 1. For simplicity, denote "$\sum_{i=1}^n$" by "$\sum_i$", $w_i = K_1\{(x - \xi_{ik})/h_k\}/(nh_k)$, $\hat{w}_i = K_1\{(x - \hat{\xi}_{ik})/h_k\}/(nh_k)$, and write $\theta_k = \theta_k(x)$. From (12), the local linear estimator $\hat{f}_k(x)$ of the regression function $f_k(x)$ can be explicitly written as

$$\hat{f}_k(t) = \frac{\sum_i \hat{w}_i Y_i}{\sum_i \hat{w}_i} - \frac{\sum_i \hat{w}_i(\hat{\xi}_{ik} - x)}{\sum_i \hat{w}_i}\hat{f}_k'(x), \tag{44}$$

where

$$\hat{f}_k'(x) = \frac{\sum_i \hat{w}_i(\hat{\xi}_{ik} - x)Y_i - \{\sum_i \hat{w}_i(\hat{\xi}_{ik} - x)\sum_i \hat{w}_i Y_i\}/\sum_i \hat{w}_i}{\sum_i \hat{w}_i(\hat{\xi}_{ik} - x)^2 - \{\sum_i \hat{w}_i(\hat{\xi}_{ik} - x)\}^2/\sum_i \hat{w}_i}. \tag{45}$$

Let $\tilde{f}_k(x)$ be a hypothetical estimator, obtained by substituting the true values w_i and ξ_{ik} for $\hat{w}_i$, $\hat{\xi}_{ik}$ in (44) and (45). To evaluate $|\hat{f}_k(x) - \tilde{f}_k(x)|$, one has to quantify the orders of the differences

$$\begin{aligned} D_1 &= \sum_i (\hat{w}_i - w_i), \quad D_2 = \sum_i (\hat{w}_i - w_i) Y_i, \\ D_3 &= \sum_i (\hat{w}_i \hat{\xi}_{ik} - w_i \xi_{ik}), \quad D_4 = \sum_i (\hat{w}_i \hat{\xi}_{ik}^2 - w_i \xi_{ik}^2). \end{aligned}$$

Considering D_1, without loss of generality, assume the compact support of K_1 is $[-1, 1]$. Since K_1 is Lipschitz continuous on its support,

$$D_1 \le \frac{c}{nh_k^2} \sum_i |\hat{\xi}_{ik} - \xi_{ik}| \{ I(|x - \xi_{ik}| \le h_k) + I(|x - \hat{\xi}_{ik}| \le h_k) \}, \tag{46}$$

for some $c > 0$, where $I(\cdot)$ is an indicator function. Lemma 2 implies for the first term on the r.h.s. of (46)

$$\frac{1}{nh_k^2} \sum_i |\hat{\xi}_{ik} - \xi_{ik}| I(|x - \xi_{ik}| \le h_k) \le \sum_{\ell=1}^{5} Z_k^{(\ell)} \frac{1}{nh_k^2} \sum_i \theta_{ik}^{(\ell)} I(|x - \xi_{ik}| \le h_k).$$

Applying the central limit theorem for a random number of summands (Billingsley, 1995, page 380), observing $\sum_i I(|x - \xi_{ik}| \le h_k)/(nh_k) \xrightarrow{p} 2p_k(x)$, one finds

$$\frac{1}{nh_k} \sum_i \theta_{ik}^{(\ell)} I(|x - \xi_{ik}| \le h_k) \xrightarrow{p} 2p_k(x) E(\theta_{ik}^{(\ell)}), \tag{47}$$

provided that $E(\theta_{ik}^{(\ell)}) < \infty$ for $\ell = 1, \ldots, 5$. Note that $E\theta_{ik}^{(1)} < \infty$, $E\theta_{ik}^{(3)} < \infty$ by (A4), $E\theta_{ik}^{(4)} \le 2\sigma_X \sqrt{\Delta_X^*}$ and $E\theta_{ik}^{(5)} \le |\mathcal{S}| \sigma_X$ by the Cauchy-Schwarz inequality. Then

$$\begin{aligned} Z_k^{(1)} \frac{1}{nh_k^2} \sum_i \theta_{ik}^{(1)} I(|x - \xi_{ik}| \le h_k) &= O_p\{ \frac{\pi_k^x}{\sqrt{n} h_X^2 h_k} p_k(x) \}, \\ Z_k^{(2)} \frac{1}{nh_k^2} \sum_i \theta_{ik}^{(2)} I(|x - \xi_{ik}| \le h_k) &= O_p\{ \frac{1}{\sqrt{n} b_X h_k} p_k(x) \}, \\ Z_k^{(3)} \frac{1}{nh_k^2} \sum_i \theta_{ik}^{(3)} I(|x - \xi_{ik}| \le h_k) &= O_p\{ \frac{\|\phi_k\|_\infty \Delta_X^*}{h_k} p_k(x) \}, \\ Z_k^{(4)} \frac{1}{nh_k^2} \sum_i \theta_{ik}^{(4)} I(|x - \xi_{ik}| \le h_k) &= O_p\{ \frac{\sqrt{\Delta_X^*}}{h_k} p_k(x) \}, \\ Z_k^{(5)} \frac{1}{nh_k^2} \sum_i \theta_{ik}^{(5)} I(|x - \xi_{ik}| \le h_k) &= O_p\{ \frac{\pi_k^x}{\sqrt{n} h_X^2 h_k} p_k(x) \}. \end{aligned} \tag{48}$$

We now obtain $(nh_k^2)^{-1}\sum_i|\hat{\xi}_{ik}-\xi_{ik}|I(|x-\xi_{ik}|\leq h_k)=O_p(\theta_k h_k^{-1})$. The asymptotic rate of the second term can be derived analogously, observing

$$\frac{1}{nh_k}\sum_i I(|x-\hat{\xi}_{ik}|\leq h_k)\leq\frac{1}{nh_k}\sum_i\{I(|x-\xi_{ik}|\leq 2h_k)+I(\sum_{\ell=1}^5\theta_{ik}^{(\ell)}Z_k^{(\ell)}>h_k)\}\xrightarrow{p}4p_k(x),$$

leading to $(nh_k^2)^{-1}\sum_i|\hat{\xi}_{ik}-\xi_{ik}|I(|x-\hat{\xi}_{ik}|\leq h_k)=O_p(\theta_k h_k^{-1})$. Then $D_1=O_p(\theta_k h_k^{-1})$ follows.

Analogously, one shows $D_2=O_p(\theta_k h_k^{-1})$, applying the Cauchy-Schwarz inequality for $\theta_{ik}^{(\ell)}$, $\ell=1,3$, and observing the independence between Y_i and $\theta_{ik}^{(\ell)}$ for $\ell=2,4,5$, given the moment condition (A4). For D_3, observe

$$D_3=\sum_i\{(\hat{w}_i-w_i)\xi_{ik}+(\hat{w}_i-w_i)(\hat{\xi}_{ik}-\xi_{ik})+w_i(\hat{\xi}_{ik}-\xi_{ik})\}\equiv D_{31}+D_{32}+D_{33}.$$

Then $D_{31}=O_p(\theta_k h_k^{-1})$, analogously to D_1. It is easy to see that $D_{32}=o_p(D_{31})$. Since $D_{33}\leq c\sum_{\ell=1}^5 Z_k^{(\ell)}(nh_k)^{-1}\sum_i\theta_{ik}^{(\ell)}I(|x-\xi_{ik}|\leq h_k)$ for some $c>0$, one also has $D_{33}=o_p(D_{31})$. This results in $D_3=O_p(\theta_k h_k^{-1})$. Observing $|\hat{\xi}_{ik}^2-\xi_{ik}^2|\leq|\hat{\xi}_{ik}-\xi_{ik}|\cdot|\xi_{ik}|+(\hat{\xi}_{ik}-\xi_{ik})^2$, one can show $D_4=O_p(\theta_k h_k^{-1})$, using similar arguments as for D_3, and $E\xi_{ik}^4<\infty$ from (A4). Combining the results for D_ℓ, $\ell=1,\ldots,4$, and applying Slutsky's Theorem leads to $|\hat{f}_k(x)-\tilde{f}_k(x)|=O_p(\theta_k h_k^{-1})$. Using (A5), and applying standard asymptotic results for the hypothetical local linear smoother $\tilde{f}_k(x)$ completes the proof of (18).

To derive (19), additionally one only needs to consider $\sum_i(\hat{w}_i\hat{\zeta}_{im}-w_i\zeta_{im})=\sum_i(\{\hat{w}_i-w_i)\zeta_{im}+(\hat{w}_i-w_i)(\hat{\zeta}_{im}-\zeta_{im})+w_i(\hat{\zeta}_{im}-\zeta_{im})\}$, where the third term yields an extra term of order $O_p(\vartheta_{mk})$ by observing

$$|\sum_i w_i(\hat{\zeta}_{im}-\zeta_{im})|\leq\sum_{\ell=1}^5 Q_m^{(\ell)}\sum_i w_i\vartheta_{im}^{(\ell)}\leq\frac{1}{nh_{mk}}\sum_{\ell=1}^5 Q_m^{(\ell)}\sum_i\vartheta_{im}^{(\ell)}I(|x-\xi_{ik}|\leq h_{mk})$$

Similar arguments as above complete this derivation. $\square$

Proof of Theorem 2. Using (A7), the derivation of θ_n^* in (42) is straightforward, following the above arguments. To obtain (21), note that

$$
\begin{aligned}
& \widehat{E}\{Y(t)|X\} - E\{Y(t)|X\} \\
\leq & \sum_{k=1}^{K}\sum_{m=1}^{M} |\hat{f}_{mk}(\xi_k)\hat{\psi}_m(t) - f_{mk}(\xi_k)\psi_m(t)| + |\sum_{k\geq K+1}\sum_{m\geq M+1} f_{mk}(\xi_k)\psi_m(t)| \\
\leq & \sum_{k=1}^{K}\sum_{m=1}^{M} [|\hat{f}_{mk}(\xi_k) - f_{mk}(\xi_k)|\{|\psi_m(t)| + |\hat{\psi}_m(t) - \psi_m(t)|\} + |f_{mk}(\xi_k)| \cdot |\hat{\psi}_m(t) - \psi_m(t)|] \\
& +| \sum_{(k,m)\in\mathcal{N}^2\setminus\mathcal{N}_K\times\mathcal{N}_M} f_{mk}(\xi_k)\psi_m(t)|.
\end{aligned}
$$

This implies the convergence rate ϑ_n^* in (42).