Nonparametric Regression Analysis of Longitudinal Data

Version: Sept. 22, 2003

Jane-Ling Wang

Department of Statistics, University of California, Davis, CA 95616, U.S.A.

Email: wang@wald.ucdavis.edu

Abstract. Nonparametric approaches have recently emerged as a flexible way to model longitudinal data. This entry reviews some of the common nonparametric approaches to incorporate time and other covariate effects for longitudinally observed response data. Smoothing procedures are invoked to estimate the associated nonparametric functions, but the choice of smoothers can vary and is often subjective. Both fixed and random effects may be included for vector or longitudinal covariates. A closely related type of data is functional data, where the prevailing approaches to model random effects are through functional principal components analysis and B-splines. Related semi-parametric regression models also play an increasingly important role.

Keywords: Functional data analysis; Scatter plot smoother; Mean curve; Fixed-effects; Random effects; Principal components analysis; Semiparametric regression.

1 Introduction

Longitudinal data involve repeated measurements that are recorded over a period of time on the same subject. The number of measurements for each subject may be different and is denoted by n_i for the i^{th} subject, and there are a total of n subjects in the study. We use $N = \sum_{i=1}^{n} n_i$ to denote the total number of observed measurements on all subjects. The time points at which those measurements were taken are also often different and denoted by $t_{i1}, \cdots, t_{in_i}$. We use $Y_{ij} = Y(t_{ij})$ to denote a measurement for the i^{th} subject at the j^{th} time point, and $\mathbf{Y_i} = (Y_{i1}, \cdots, Y_{in_i})$ to denote the observed vector for the i^{th} subject. This leads to a correlation structure between the repeated measurements within the same subject. Longitudinal data arise commonly in health sciences and engineering research, but different terms have been applied to describe them. They are usually referred to as "longitudinal data" in biomedical applications, where a small number of repeated measurements, n_i, over time per subject is common, and as "functional data" in engineering and biological applications, where n_i is often large. Statistical approaches to analyze such data have also been intrinsically different for longitudinal and functional data. Longitudinal data are treated as vectors, $\mathbf{Y_i}$, with subject specific dimension n_i for the i-th subject, while functional data are regarded as realizations of random processes with smooth paths, $Y(t)$, that are observed at discrete time points. Parametric GEE-based marginal models and parametric random effects models (GLMM) are the predominant approaches for longitudinal data, and non- or semi-parametric approaches are the standard practice to analyze functional data. Recent challenges in the biomedical and biological fields prompted the development of more complex and flexible approaches to model longitudinal data. Nonparametric regression, well known to be more data adaptive and less restrictive than parametric approaches, thus emerged as promising alternative to handle longitudinal data. For readers searching for such nonparametric approaches in the literature, a keyword to include is "functional data" in addition to "longitudinal data". The two books [20] and [21] on functional data analysis provide an excellent introduction to

this topic.

In this entry, we focus on situations where the responses for the experimental subjects are longitudinal data. The covariates can be a baseline vector (X), a time-varying covariate vector ($X(t_{ij})$) which is longitudinal data itself, or a combination of both. Key issues in nonparametric regression for such data include inference for the overall mean and nonparametric fixed effects, and modeling of the within subject covariance structure through nonparametric random effects. We will use the fecundity data set described in the next section to illustrate these issues.

We begin with nonparametric mean function estimation, treating the overall mean as a function on a time interval, $[0, T]$, over which data were recorded for the subjects. This overall mean function is assumed to be smoothed and often referred by researchers as the mean curve.

2 Estimating the Overall Mean as a function of time

We assume that the overall mean function (or mean curve), $\mu(t) = E(Y(t))$, is an unknown but smooth function on $[0, T]$. Hence, $E\ (Y_{ij}) = \mu(t_{ij})$. If we ignore for the moment the within-subject correlations, then the mean function can be regarded as a nonparametric regression function with the regressor being the time variable. A scatter-plot smoother can then be applied to all N observed data points (t_{ij}, Y_{ij}) to estimate the mean function $\mu(\cdot)$. Specifically, the estimate at a particular time t is

$$\hat{\mu}(t) = \sum_{i=1}^{n} \sum_{j=1}^{n_i} w_{ij} Y_{ij}. \tag{1}$$

This is simply a weighted average of all the N measurements, where the weights w_{ij} depend on the design points t_{ij} and the particular smoother. The choice of the smoother can be subjective and common choices include the kernel method ([8]), local polynomials ([6], [14] and [32]) and splines ([22], [23] and [24]).

Standard software such as S-plus can be employed easily to obtain a mean function estimate. The only difference with respect to the standard nonparametric regression setting is that repeated measurements are available for each subject. For this reason, the estimated mean function in (1) can be expected to be consistent if n tends to infinity, provided the time points t_{ij} are spread out over the design interval $[0, T]$.

However, standard nonparametric smoothing methods may need to be adjusted for longitudinal data. An example for such an adjustment is the choice of smoothing (or tuning) parameter, that is required by all smoothing methods. For the popular cross-validation method, it was shown in [22] that a leave-one-subject-out scheme should be employed for longitudinal data rather than the standard leave-one-observation-out scheme. All smoothing procedures require the proper choice of a tuning parameter. This problem is less understood and studied in longitudinal settings, and further research is needed.

Example of Reproductive Fecundity Data: We illustrate here the mean function estimate through a data set collected on 1,000 female Mediterranean fruit flies (medflies) that was analyzed in [6]. Daily egg production, in terms of the number of eggs laid, were recorded individually until death for each of the 1,000 female medflies. This results in a sample of 1,000 longitudinally recorded fecundity curve data, with Y_{ij} = Number of eggs laid on day j by fly i. The goal is to explore the reproductive behavior of medflies through the pattern and modes of variation of these fecundity curves, Y(t). Such information is important because reproduction is considered by evolutionary biologists as the single most important life history trait besides lifetime itself. See [6] and the references therein for details and biological features of the experiment.

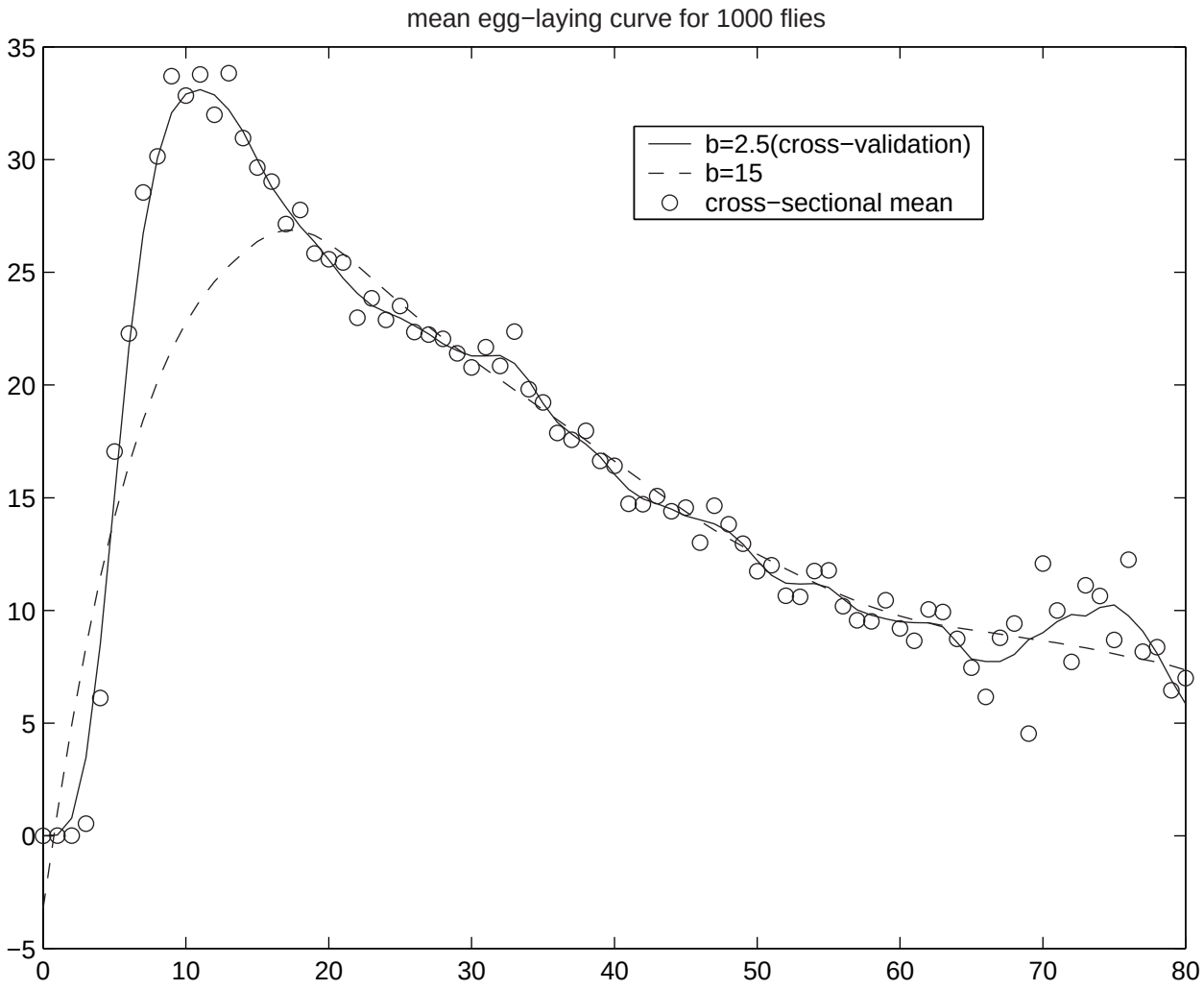

Figure 1: Mean egg-laying curves of 1,000 female Mediterranean fruit flies. (a) daily cross-sectional means of flies alive at the beginning of the day (circles) (b) smoothed mean curve with fixed bandwidth $b = 15$ (dash line) (c) smoothed mean curve with cross-validated bandwidth choice $b = 2.5$ (solid line).

Figure 1 provides the mean function estimates at various bandwidths based on a local linear scatter-plot smoother. Details of the procedure, including the leave-one-subject-out cross-validated bandwidth choice are available in [6]. Because of the large number of subjects (there are 1,000 flies) in the study and the dense recording per subject (repeated measurements were available daily), cross-validation selected a very small bandwidth at 2.5 days. The mean curve based on the larger bandwidth 15 is smoother but has larger bias than the mean curve based on the smaller bandwidth 2.5, as is expected for any nonparametric smoothing procedure.

The smoothing weights for the scatter-plot smoothing procedure in Figure 1 were determined by the choice of the local linear smoother and the corresponding bandwidths following the standard practice in nonparametric regression. The within-subject correlation structure was not incorporated. An intriguing question for longitudinal data is how to effectively adjust the weights w_{ij} in (1) in the smoothing step to reflect the within- subject correlation structure of $\mathbf{Y_i}$. This was cleverly demonstrated for the case of smoothing splines in [29], and for local polynomial smoothers in [26] and [27]. It was shown that the asymptotic variance of the mean function estimators can be minimized if the weights are selected properly. However, these optimal weights require the use of the true within-subject correlation structure and do not necessarily minimize the asymptotic bias. The bias issue is more elusive and has not been resolved.

3 Nonparametric Fixed-effects Covariates

The procedures and discussion in the previous section apply directly to covariates other than time. To estimate the regression function $E(\mathbf{Y_i}|\mathbf{X_i})$, corresponding to fixed-effects of covariates $\mathbf{X_i}$, simply replace in the scatter-plot smoother the t_{ij} by X_{ij} for time-varying covariates, and by $\mathbf{X_i}$ for vector covariates. Another framework mimicking the marginal approach of GEE (Generalized Estimating Equations) can be found in [14], [26], [27] and

[29]; where $E(Y_{ij}|X_{ij}) = \mu_{ij} = h(g(X_{ij}))$, with h a known and differentiable link function, and g an unknown smooth function. Here, the covariance structure of the response $\mathbf{Y}_i$ is also assumed to be a function of the means μ_{ij}, as suggested in the generalized linear model setting. The asymptotic variance of the minimum variance estimate, $\hat{g}(\cdot)$, was derived in [29] for smoothing splines and in [27] for local polynomial estimators. Additional results on semiparametric marginal models are also available in these two papers and [15] and [18].

4 Nonparametric Random-Effects

The overall mean function in Section 2 represents the population average, but individual trajectories may vary due to subject effects which also contribute to correlated repeated measurements within the same subject. Subject-specific random-effects can be added for example by assuming:

$$Y_i(t) = \mu(t) + v_i(t) + e_i(t), \quad i = 1, \cdots, n, \tag{2}$$

where μ is the unknown overall mean function, v_i are unknown subject-specific random effects reflecting the individual variation from the overall mean function, and e_i are measurement errors independent of v_i. The random effects are often regarded as realizations of a mean zero random process with smooth paths. It is thus expected that both the smoothed mean and random effects functions in (2) can be approximated using some basis functions. A B-spline basis, such as cubic splines, is a common choice and was proposed independently in [23] and [24], resulting in the following mixed-effects model:

$$Y_i(t) = \sum_{k=1}^{K} \beta_k B_k(t) + \sum_{k=1}^{K} b_k B_k(t) + e_i(t). \tag{3}$$

Here β_k are coefficients, b_k are random variables with mean zero, $B_k(\cdot)$ is a basis of spline functions on $[0, T]$, and $e(t)$ is the measurement error. Consequently, the first summand yields the population mean function and corresponds to a fixed effect, while the second summand represents the random effects attributed to subject variations and describes the

within subject correlation structure. It is possible to use separate bases for random and fixed effects in (3). If we further assume normality for b_k and e_i, then model (3) becomes a linear mixed-effects model and thus can be fitted using either S-PLUS LME or SAS PROC MIXED. This computational advantage is an attractive feature of the B-spline approach. However, it requires the choice of the spline basis and the number of basis functions K, which in turn involves fairly complex choices of the number and location of knots. Cross-validation procedures or information based criteria such AIC and BIC are among the suggestions in [23] for the choice of knots, but the issue is elusive and remains unsettled.

The B-spline basis approach may have difficulties for a data set that requires a large number of basis functions. This is because the degrees of freedom involved in (3) may be too small or even negative for sparse data. One solution is to use instead a local basis such as local polynomials in model (3) as only a low degree polynomial is needed locally to fit the data, and often a linear polynomial suffices. This is explored in [32]. The trade-off is computation time as smoothing is done locally at each point while the B-spline approach is a global smoothing procedure.

Another remedy for the B-spline procedure is proposed in [10] based on the reduced rank procedure that involves the use of principal components. This approach aims at reducing the actual number of parameters needed in the nonparametric mixed-effects model (3) so that one can increase the degrees of freedom. However, it often involves a compromise in terms of computational feasibility and model flexibility. An alternative is the principal component analysis approach, a commonly used method for functional data which can be adapted to longitudinal data.

Principal Component Analysis Approach

Simply put, this approach just replaces the pre-specified basis B_k in (3) by the eigenfunctions of the covariance operator of the response. This is the essence of principal component analysis and results in a data-adaptive basis that can effectively reduce the number of basis

functions needed to model the random effects. They are effective dimension reduction tools for longitudinal data. The concept is similar to the problem to find the best K-dimensional linear model for stochastic processes, and has been extended in [4] and [22] to the case of functional or longitudinal data and termed functional principal components analysis.

Here the response functions $Y_i(t)$ are considered realizations of smooth L^2-process with mean $\mu(t)$ and covariance function $cov(Y(s), Y(t)) = \gamma(s,t)$. The covariance function γ allows a spectral decomposition into orthonormal eigenfunctions $\rho_k(\cdot)$

$$\gamma(s,t) = \sum_{k=1}^{\infty} \lambda_k \rho_k(s) \rho_k(t), \tag{4}$$

with ordered nonnegative eigenvalues $\lambda_1 \geq \lambda_2 \geq \ldots \geq 0$.

Let $< \cdot, \cdot >$ denote the inner product in L^2 space. The Karhunen-Loève representation for a randomly selected curve is

$$Y(t) = \mu(t) + \sum_{k=1}^{\infty} A_k \rho_k(t), \tag{5}$$

where the random variables a_k correspond to the principal component scores, and are given by

$$A_k =< \rho_k, Y - \mu > . \tag{6}$$

The principal components A_k in (6) are uncorrelated random variables with

$$E(A_k) = 0, \quad var(A_k) = \lambda_k, \quad \sum_{k=1}^{\infty} \lambda_k < \infty,$$

i.e., the k-th eigenvalue in (4) corresponds to the variance of the k-th principal component as in the multivariate case. The principal components A_k and basis functions ρ_k in (5) can be interpreted as defining the variation of the stochastic process about its mean function, and $A_1\rho_1$ explains the maximum amount of variation in Y among all functions which involve a single real-valued random variable. Similarly, the function $A_2\rho_2$ explains the maximum additional amount of process variation which is unexplained by $A_1\rho_1$, and so forth for $k = 3, 4, \ldots$.

Methods to estimate the eigenfunctions and principal components are described in [22] based on smoothing splines. The leave-one-subject-out cross-validation method to select the number of eigen-basis is also first proposed there. Theoretical properties of the functional principal components estimates can be found in [19] under the hypothetical assumption that the entire process $Y(t)$ is observable. Similar theoretical results for another functional principal components approach based on kernel smoothers are provided in [2].

Compared to the nonparametric mixed-effects model in (3), functional principal component analysis is more data adaptive and therefore typically requires fewer basis functions. It also has the advantage to allow direct interpretations in terms of modes of variation of the underlying process and is favored by biologists to explore the covariance structure of the data. Although functional principal component analysis is not yet available on standard statistical packages, it is not difficult to write code in either MATLAB or S-PLUS to perform this analysis.

Example of Fecundity data: To incorporate subject specific random effects of the reproduction process of medflies, we perform principal component analysis on the fecundity data. As in [6], we restricted the analysis to the first 50 days of lifetime due to high variability of the fecundity curves beyond day 50. This avoids the eigen-analysis being dominated by erratic tail behavior of the data, and provides more sensible analysis. See [6] for more details. The rest of the principal component analysis presented below is based on the first 50 days of daily egg counts for the 167 flies that lived beyond day 50. The egg-laying curves $Y_i(t)$, $i = 1, \ldots, 167$, are considered as realizations of a stochastic process on the interval $T = [0, 50]$.

We applied the eigen-analysis based on local linear smoothing as described in [6]. The optimal number of principal components based on AIC is 9, but there is little gain after 4 components as these four components explain 95.88% of the total variation of the fecundity data. The eigen-functions corresponding to the largest four eigen-values of the covariance function, $\gamma(s, t)$, are shown in Figure 2. This example demonstrates how principal component analysis effectively reduces the dimension of the data from an infinite-dimensional curve to a

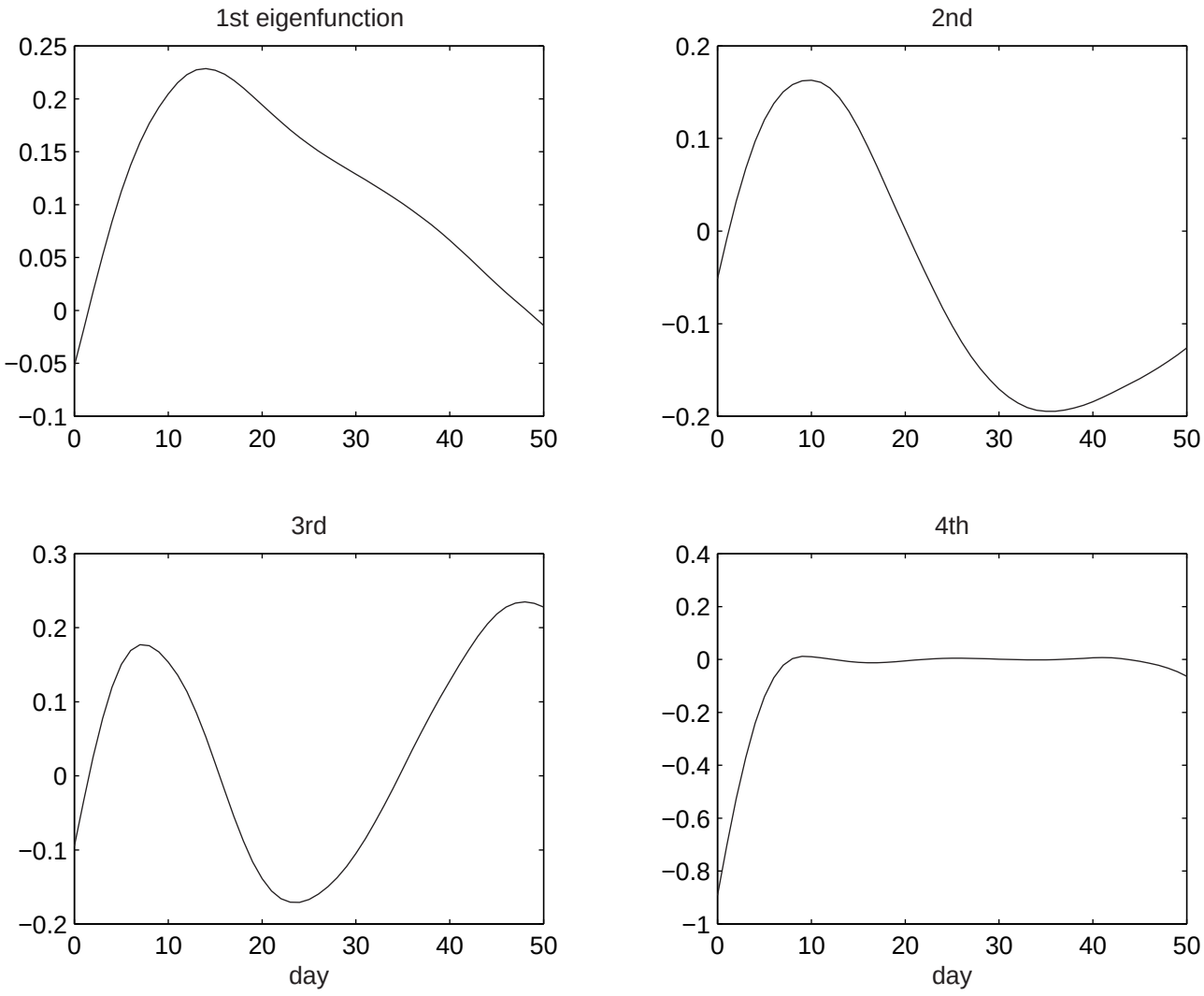

Figure 2: First four eigenfunctions for egg-laying data. The fraction of variation explained by each of these components are: 0.6183, 0.2090, 0.0779, 0.0536 respectively.

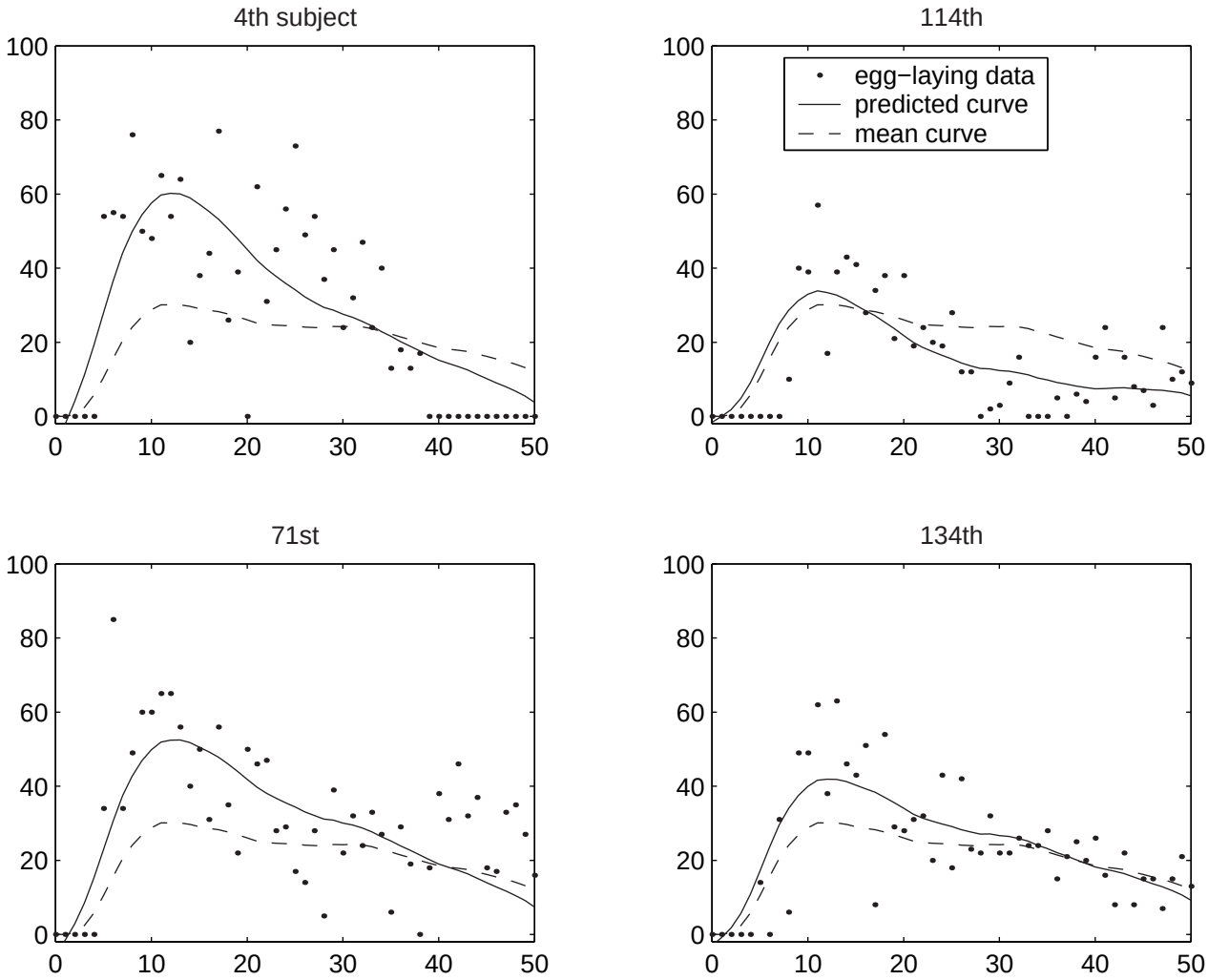

Figure 3: Observed data, predicted egg-laying curves and mean egg-laying curves for four randomly selected subjects.

few components. In fact, selecting two components may suffice as they explain already 82% of the total variation based on Figure 2. To see whether one can reasonably predict the shape of the fecundity curve using just two principal components, we proceed to fit the fecundity curves of four randomly selected flies using the Karhunen-Loève representation (5) with two components.

Figure 3 exhibits the observed and predicted egg-laying profiles of individual flies, as well as the overall mean curve. The overall mean curve would be the predicted curve when no random effects are included. As can be seen form Figure 3, the functional PCA approach seems to fit the curves reasonably well and has much less bias than the overall mean curve. We have thus demonstrated the effectiveness of the nonparametric principal components analysis approach for longitudinal data.

Although the fecundity data illustrated here were sampled at the same time points (daily in this case), the procedure can as well be applied to longitudinal data sampled at irregular time points.

5 Nonparametric Mixed-Effects Models with Covariates

The procedures in the previous section allow to handle the time effects for a sample of individuals from the same population. When there are other covariates $\mathbf{X}$ that affect the longitudinal response data, one needs to incorporate these covariate effects in addition to the time effects. This has been explored for vector covariates in [6] by taking the conditional expectation with respect to $\mathbf{X}$ on both sides of (5). As a result, we have the following model:

$$E(Y_i(t)|\mathbf{X_i}) = \mu(t) + \sum_{k=1}^{\infty} E(A_{ik}|\mathbf{X_i})\rho_k(t). \tag{7}$$

Procedures to estimate all the components including the mean function, the eigenfunctions and the conditional principal components $E(A_{ik}|(X_i))$ can be found in [6], which also includes a semiparametric index model to tackle situations when the vector $\mathbf{X}$ is high dimensional. It is interesting to note here that the fixed-effect of a covariate in (7) is derived from the

unconditional principal components A_{ik} and eigenfunctions $\rho_k(t)$. This is because although A_{ik} has overall mean zero, the conditional mean $E(A_{ik}|\mathbf{X_i})$ in (7) is not zero and thus contributes to the fixed-effects. Additional random effects can then be added to the model through $b_{ik} = A_{ik} - E(A_{ik}|\mathbf{X_i})$ to reach the following nonparametric mixed-effects model:

$$Y_i(t) = \mu(t) + \sum_{k=1}^{\infty} E(A_{ik}|\mathbf{X_i})\rho_k(t) + \sum_{k=1}^{\infty} b_{ik}\rho_k(t).$$

Other types of mixed-effects functional PCA regression models include [3] and [25], and more parsimonious semiparametric mixed-effects models in [28] and [33].

6 Other Non- and Semi-parametric Regression Approaches

So far, we discussed briefly a few nonparametric approaches for longitudinal/functional data. There are many other non- and semiparametric alternatives. One of them is the Generalized Additive Model (GAM) of the form:

$$E(Y_i(t)) = \beta_0 + \sum_{k=1}^{P} g_k(X_{itk}) + e_i(t), \tag{8}$$

where $(Y_i(t), X_{it1}, \cdots, X_{itP})$ is observed at time t for the i^{th} subject with P-dimensional covariates denoted by X_{itj} at time t_j. The functions g_k in (8) are unknown smooth functions. Backfitting algorithms were proposed in [1] and [30], and inference procedures studied in [16].

When the data exhibit a common shape or structure, a smooth curve g can be used to model this common shape with individual responses adjusted by some parametric transformation of the common curve. This is referred to as self-modelling regression (SEMOR) in [13] and studied in [12] and [17] among others. A more general semiparametric model which includes SEMOR is recently proposed in [11].

Another approach which emerged recently to model longitudinal data is the varying-coefficients model of the form:

$$Y_i(t) = \mu(t) + \sum_{k=1}^{K} \beta(t)X_i(t) + e_i(t).$$

This was first applied to longitudinal data in [9], and subsequently studied in [7], [5] among others. See the review of this model in [31] for details.

References

[1] Berhane, K. and Tibshirani, R. J. (1998) Generalized additive models for longitudinal data. *Canadian Journal of Statistics*, **26**, 517-535.

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

[3] Capra, W. B. and Müller, H. G. (1997) An accelerated time model for response curves. *J. Am. Statist. Ass.*, **92**, 72-83.

[4] Castro, P. E., Lawton, W. H., and Sylvestre, E. A. (1986) Principal modes of variation for processes with continuous sample curves. *Technometrics*, **28**, 329-337.

[5] Chiang, C. T., Rice, J. A. and Wu, C. O. (2001) Smoothing spline estimation for varying coefficient models with repeatedly measured dependent variables. *J. Am. Statist. Ass.*, **96**, 605-619.

[6] Chiou, J. M., Müller, H. G. and Wang, J. L. (2002). Functional quasi-likelihood regression models with smooth random effects. *Journal of Royal Statistical Society, Series B*, **65**, 405-423.

[7] Fan, J. and Zhang, J. T. (2000) Statistical estimation in varying-coefficient models. *Annals of Statistics.*, **27**, 1491-1518.

[8] Hart, J. P. and Wehrly, T. E. (1986) Kernel regression estimation using repeated measurements data. *J. Am. Statist. Ass.*, **81**, 1080-1088.

[9] Hoover, D., Rice, J., Wu, C. and Yang, L.-P. (1998) Nonparametric smoothing estimates of time-varying coefficient models with longitudinal data. *Biometrika*, **85**, 809-822.

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

[11] Ke, C. and Wang, Y. (2001) Semiparametric nonlinear mixed-effects models and their applications. *J. Am. Statist. Ass.*, **96**, 1272-1298.

[12] Kneip, A. and Gasser, T. (1988) Convergence and consistency results for self-modeling nonlinear regression. *Annals of Statistics*, **16**, 82-112.

[13] Lawton, W. H., Sylvestre, E. A. and Maggio,M. S.(1972) Self-modeling nonlinear regression. *Technometrics*, **14**, 513-532.

[14] Lin, X. and Carroll, R. J. (2000) Nonaprametric function estimation for clustered data when the predictor is measured without/with error. *J. Am. Statist. Assoc.*, **95**, 520-534.

[15] Lin, X. and Carroll, R. J. (2001) Semiparametric regression for clustered data using generalized estimating equations. *J. Am. Statist. Assoc.* **96**, 1045-1056.

[16] Lin, X. and Zhang, D. W. (1999) Inference in generalized addiitve mixed models by using smoothing splines. *Journal of Royal Statistical Society, Series B*, **61**, 381-400.

[17] Linsdtrom, M. J. (1995) Self-modeling with random shift and sacle parameters and a free-knots spline shape function. *Statistics in Medicine*, **14**, 2009-2021.

[18] Moyeed, R. A. and Diggle, P. J. (1994) Rates of convergence in semi-parametric modelling of longitudinal data. *Australian Journal of Statistics*, **36**, 75-93.

[19] Pezzulli, S. and Silverman, B. W. (1993) Some properties of smoothed principal components analysis for functional data. *Comput. Statist.*, **8**, 1-16.

[20] Ramsay, J. O. and Silverman, B. W. (1997) *Functional Data Analysis.* New York: Springer.

[21] Ramsay, J. O. and Silverman, B. W. (2002) *Applied Functional Data Analysis.* New York: Springer.

[22] Rice, J. A. and Silverman, B. W. (1991) Estimating the mean and covariance structure nonparametrically when the data are curves. *J. R. Statist. Soc.* **B**, **53**, 233-243.

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

[24] Shi, M., Weiss, R. E., and Taylor, J. (1996) An analysis of paediatric CD4 counts for

acquired Immune Deficiency Syndrome using flexible random curves. *Applied Statistics*, **45**, 151-163.

[25] Staniswalis, J. G. and Lee, J. J. (1998) Nonparametric regression analysis of longitudinal data. *J. Am. Statist. Ass.*, **93**, 1403-1418.

[26] Wang, N.(2003) Marginal nonparametric kernel regression accounting for within-subject correlation. *Biometrika*, In Press.

[27] Wang, N., Carroll, R. J. and Lin, X.(2003) Efficient semiparametric marginal estimation for longitudinal/clustered data. manuscript.

[28] Wang, Y.(1998) Mixed-effects smoothing spline analysis of variance. *Journal of Royal Statistical Society, Ser. B.*, **60**, 159-174.

[29] Welch, A., Lin, X. and Carroll, R. J. (2002) Marginal longitudinal nonparametric regression: Locality and efficiency of spline and kernel methods. *J. Am. Statist. Ass.*, **97**, 482-493.

[30] Wild, C. J. and Yee, T. W. (1996) Additive extension to generalized estimating equation methods. *Journal of Royal Statistical Society, Ser. B*, **58**, 711-725.

[31] Wu, C. and Yu, K. F. (2002) Nonparametric Varying-Coefficient Models for the Analysis of Longitudinal Data. *International Statistical Institute Review*, **70**, 373-393.

[32] Wu, H. and Zhang, J. T.(2002) Local polynomial mixed-effects models for longitudinal data. *J. Am. Statist. Ass.*, **97**, 883-897.

[33] Zhang, D., Lin, X., Raz, J. and Sowers, M.(1998) Semiparametric stochastic mixed models for longitudinal data. *J. Am. Statist. Ass.*, **93**, 710-719.