Title
Comments on "Smoothing Spline Models for the Analysis of Nested and Crosed Samples of Curves"

Authors
Fan, Jianqing
Zhang, Jin-Ting

Publication Date
1998-01-28

Discussion of the paper by Brumback and Rice *

Jianqing Fan and Jin-Ting Zhang

January 26, 1998

Brumback and Rice are to be congratulated for this neat and excellent paper on the smoothing spline models for the analysis of nested and crossed samples of curves. Of particularly important are the connections between smoothing spline methods and the mixed effects models. Such connections are not only important in our intuitive understanding of the smoothing spline methods, but also elegant in deriving methods for selecting smoothing parameters.

With modern technology, data can nowadays easily be collected in a form of curves. Fully processing the information contained in the sample curves is a challenging and emerging subject in statistics. The subject has strong connections with traditional longitudinal data analysis (see for example Diggle, Liang and Zeger 1994 and Hand and Crowder 1996). The ideas presented in the Brumback and Rice and this discussion are expected to have strong impact on both functional data analysis and longitudinal data analysis. We welcome the opportunity to make a few comments and to present other simple alternative methods that will be helpful for the future development of the subject.

1 Connections with functional linear models

For given functional data of form $\{(t, X_1(t), \cdots, X_p(t), Y(t))\}$, a natural extension of the classical linear model is to allow coefficients depending on t:

$$Y(t) = X(t)^T \beta(t) + \varepsilon(t), \tag{1}$$

where $X(t) = (X_1(t), \cdots, X_p(t))^T$ is the covariate vector and $\varepsilon(t)$ is the noise that can not be explained by the linear regression. This is a special kind of functional linear models discussed in Ramsay and Silverman (1997) and is called varying-coefficient models by Hastie and Tibshirani (1993). In many applications, the covariates $X(t)$ may be independent of t. In particular, if $X_1(t) \equiv 1$, model (1) includes an intercept term with varying coefficients.

Suppose that we have a random sample of functions from model (1) with possible missing values at time t. Let $\mathbf{X}(t)$ and $\mathbf{Y}(t)$ be the design matrix and the response vector collected at time t.

Jianqing Fan is professor, Department of Statistics, University of California, Los Angels, CA 90095. Jin-Ting Zhang is graduate student, Department of Statistics, University of North Carolina, Chapel Hill, NC 27599-3260. This work was partially supported by NSF grant DMS-9504414 and NSA grant 96-1-0015

Note that the number of observations at each given time is not necessarily the same due to the possibility of missing data. Then, the least-squares estimate of $\beta(t)$ at the time t is given by

$$\hat{\beta}_R(t) = (\mathbf{X}(t)^T\mathbf{X}(t))^{-1}\mathbf{X}(t)^T\mathbf{Y}(t). \tag{2}$$

It follows from the ordinary least-squares theory that the covariance matrix of $\hat{\beta}_R(t)$ is given by

$$\Sigma(t) = \sigma^2(t)(\mathbf{X}(t)^T\mathbf{X}(t))^{-1}, \tag{3}$$

where $\sigma^2(t)$ is the residual variance. The estimator (2) is a raw estimation scheme for $\beta(t)$ and will be refined in the next section. It does illustrate that even with heavily missing data, the coefficient functions $\beta(\cdot)$ are estimable. If the coefficient functions $\beta(t)$ are smooth, which we usually assume, the coefficient functions $\beta(t)$ can still be estimable even though the data points at time t are completely missing since we can pull information from neighboring time points.

The models for nested and crossed samples of curves in the Brumback and Rice paper can be regarded as a functional linear model (1) via introducing some suitable dummy variables. Therefore, the methodology developed for varying coefficient models can directly be applied to the current setting. This also explains why the method of Brumback and Rice can handle so elegantly with a large amount of missing data.

2 A simple alternative method

While the smoothing spline methods are useful for processing the functional data as illustrated by Brumback and Rice, the authors are quite correct that computation burden is extremely heavy. Even for a moderate problem with a sample of 91 progesterone curves measured at 24 distinct time points, one has to blindly invert a matrix of size 2,000 $\times$ 2,000. This requires not only large amount of RAM, but also a very fast CPU. A possible alternative method is the backfitting algorithm, proposed in Hastie and Tibshirani (1994) for varying-coefficient models, which iteratively estimates one coefficient function at a time until convergence. While this method can significantly reduce the use of RAM, its effectiveness in the current context remains to be seen. The key challenge for the backfitting algorithm is that the number of functions to be estimated is large, which is 91 in the current context.

We now offer a simple and powerful alternative method. The raw estimator (2) is an unbiased estimator of coefficient functions $\beta(t)$. Note that when there are no sufficient data points at a given time t, one can use some neighboring data to run the regression. For nested samples of curves, one can also use averages of observations at certain level of nesting to impute missing data. The point is to keep the bias of the raw estimate (2) small. The raw estimator is not effective because it can be augmented from data in neighboring time points. An effective way to aggregate information from neighboring points is smoothing. For each component of $\hat{\beta}_R(\cdot)$, one can smooth the raw estimate along the time axis, resulting in a refined and smoothed estimator $\hat{\beta}_S(\cdot)$. In this step of smoothing, any smoothing method such as kernel, spline or local polynomial method can be employed. Since the estimate $\hat{\beta}_R(\cdot)$ is heteroscedastic with the variance matrix given in (3), one can easily incorporate this information into the smoothing step. The idea above is proposed and studied in Fan and Zhang (1998) for longitudinal data analysis. Intuitively, this two-step method does seem to use neighboring information fully and hence is expected to be effective. This is indeed shown in Fan and Zhang (1998).

One advantage of the above smoothing method is the computation expediency. One can easily use the existing software to compute the estimated coefficient functions. For each given t, one first applies an ANOVA or regression routing to compute $\hat{\beta}_R(t)$ and then uses an existing smoothing procedure to obtain $\hat{\beta}_S(t)$. This can easily be implemented in languages such as SAS and Splus with little programming effort. For the progesterone data analyzed in the Brumback and Rice paper, Figure 1 presents some of the estimated functions. The raw estimates of $\beta(t)$ are indicated by dots while the refined estimates are presented as solid curves. Since the estimated curves are linear in $\hat{\beta}_R(t)$ whose variance is given in (3), the pointwise standard errors for $\hat{\beta}_S(t)$ can be computed. These standard errors are depicted as dashed curves in Figure 1 where twice of the standard errors above and below the estimated curves are presented.

Another advantage of the above two-step procedure is that one is allowed to use different smoothing parameters for different coefficient functions. This flexibility is important when the coefficient functions admit different degrees of smoothness. Plots of the raw estimates can assist one to choose appropriate amount of smoothing for each coefficient function.

3 Other useful methods

The coefficient functions in the functional linear models (1) can be estimated by various methods. In the longitudinal data setting, Hoover, Rice, Wu and Yang (1997) proposed to use a smoothing spline approach and noted that it outperforms a simple version of kernel estimator because the latter uses only one smoothing parameter for all components of the estimated coefficient functions. Realizing such a drawback, Fan and Zhang (1997) presented a two-step approach for varying coefficient models. They showed that when the coefficient functions admit different degrees of smoothness, the simple version of kernel method fails to produce optimal estimators for all coefficient functions, while the two-step procedure can overcome this problem and permits one to use different bandwidths for different coefficient functions. Their idea can also be applied to the present setting as follows.

First of all, use a local polynomial method with a small bandwidth h_0 to obtain an initial estimate $\hat{\beta}_0(t)$ of the coefficient functions. This step aims at obtaining a consistent estimator with a negligible bias. For each given t_0, since all covariates and responses in the time interval $t_0 \pm h_0$ are used to run a linear regression, many sampled curves are allowed to have missing values and one can still have enough data points to perform a local regression. We now describe the second step. In the second step, each coefficient function is estimated separately by treating the initial estimators as if they were population coefficient functions. Suppose that we wish to estimate the r^{th} component. Let

$$Y_r^*(t_{il}) = Y(t_{il}) - \sum_{s=1, s\neq r}^{p} X_s(t_{il})\hat{\beta}_{0,s}(t_{il}), \quad l = 1, \cdots, n_i; i = 1, \cdots, N,$$

where $\{t_{il}\}$ are the locations at which the i^{th} observed curve is sampled and $\hat{\beta}_{0,s}(\cdot)$ is the s^{th} component of the initial estimator $\hat{\beta}_0(\cdot)$. Then, by (1), we have the approximate one-dimensional nonparametric regression problem as follows:

$$Y_r^*(t_{il}) = X_r(t_{il})\beta_r(t_{il}) + \epsilon(t_{il}), l = 1, \cdots, n_i; i = 1, \cdots, N.$$

Therefore, one can use a one-dimensional nonparametric regression technique with a smoothing parameter h_r to obtain a refined estimator of $\beta_r(\cdot)$. For details of the approach, see Fan and Zhang (1997, 1998). Clearly, this two-step estimator allows one to use different smoothing parameters for estimation of different components of coefficient functions.

Various methods for estimating coefficient functions have been proposed. These include the simple version of kernel method, smoothing spline methods and two versions of the two-step methods presented above. Their relative efficiencies need more careful analysis and study. It is clear, however, the estimator $\hat{\beta}_S(t)$ is the easiest one to compute among all competitors.

4 Hypothesis testing on coefficient functions

Statistical inferences for the coefficient functions in the functional linear models are interesting and challenging. For example, one is interested in testing if the difference between the effects of the conceptive and nonconceptive groups is statistically significant. The dimensionality for hypothesis testing in functional linear models is high since whole functions are now compared. Special cares are needed in order to enhance the power of a testing procedure. To derive the distribution of an estimator or a test statistic, one needs to model and estimate the covariance structure of the noise process $\varepsilon(t)$. The dimensionality of this covariance matrix can be very high and some modeling on the covariance structure is desirable.

A naive testing statistic is to use raw estimate directly for hypothesis testing. For example, to compare if there is any statistical difference between the conceptive and nonconceptive groups, a naive test statistic is

$$X_R^2 = \sum_{t=1}^{T} \{\hat{\beta}_{R1}(t) - \hat{\beta}_{R2}(t)\}^2, \tag{4}$$

where $\hat{\beta}_{R1}(t)$ and $\hat{\beta}_{R2}(t)$ denote the raw estimates of the conceptive and nonconceptive effects, respectively. As pointed out by Fan (1996), such a type of test statistics accumulate large stochastic noises and decrease discrimability power. An improved version is to use

$$X_S^2 = \sum_{t=1}^{T} \{\hat{\beta}_{S1}(t) - \hat{\beta}_{S2}(t)\}^2, \tag{5}$$

where the subscript S indicates a smoothed version of the raw estimates. The role of smoothing can easily be seen in a simplified setting — a Gaussian white noise model discussed in Fan (1996), where he used truncated Fourier series as a smoothing device, resulting in an adaptive Neyman test. Such a test statistic is extended to compare multiple sets of curves in Fan and Lin (1996) and hence can also be applied to the current setting. The sampling distributions of X_S^2 and the adaptive Neyman test in the current setting remain unknown. The bootstrap methods outlined in the Brumback and Rice paper may provide useful device for estimating the null distribution of the test statistics so that the P-values of the testing procedures can be obtained.

5 Smoothing splines are functional BLUPs

In his insightful comments, Speed (1991) showed that smoothing splines are BLUPs at each design point. We would like to show further that the smoothing splines are BLUPs not only at each point (including nondesign points) but also in a global Mean Integrated Squared Error (MISE) sense.

Let $X(t) = (X_1(t), X_2(t), \cdots, X_p(t))^T$ and $Z(t) = (Z_1(t), \cdots, Z_q(t))^T$ be covariate functions and $Y(t)$ the response function. Suppose that we wish to model covariates $X(t)$ as fixed effects and $Z(t)$ as random effects. This leads to the mixed effects model

$$Y(t) = X(t)^T \beta + Z(t)^T u + \epsilon(t), \tag{6}$$

where u is an unobserved random vector with mean zero and covariance matrix $\sigma^2 G$ and is independent of the noise process $\varepsilon(t)$ with covariance function $E\{\varepsilon(s)\varepsilon(t)\} = \sigma^2 \rho(s,t)$. We will call this model a functional mixed effects model, since the covariates and the response are now functions rather than vectors in conventional mixed effects models.

An estimator $\hat{f}(\cdot)$ is called a functional Best Linear Unbiased Predictor of the random function $f(\cdot) = X(\cdot)^T \beta + Z(\cdot)^T u$ if it is an unbiased linear (in observed responses) estimator which has the smallest MISE

$$E \int \{\hat{f}(t) - f(t)\}^2 w(t) dt \tag{7}$$

among all unbiased linear estimators, where w is a nonnegative weight function.

In practice, data from the functional mixed effects model (6) are only observed at discrete points $t_1 \le t_2 \le \cdots \le t_n$, resulting in the data vector $\mathbf{Y} = (Y(t_1), \cdots, Y(t_n))^T$. Let $\mathbf{X} = (X(t_1), \cdots, X(t_n))^T$ and $\mathbf{Z} = (Z(t_1), Z(t_2), \cdots, Z(t_n))^T$. These data matrices satisfy the usual mixed effects model

$$\mathbf{Y} = \mathbf{X}\beta + \mathbf{Z}u + \varepsilon \tag{8}$$

where $\varepsilon = (\epsilon(t_1), \cdots, \epsilon(t_n))^T$ whose covariance matrix is denoted by $\sigma^2 R$. The best linear unbiased estimator (BLUE) of β and the BLUP of u for model (8) are respectively given by

$$\hat{\beta} = (X^T V^{-1} X)^{-1} X^T V^{-1} Y,$$

and

$$\hat{u} = (Z^T R^{-1} Z + G^{-1})^{-1} Z^T R^{-1} (Y - X\hat{\beta}),$$

where $V = ZRZ^T + G$. They minimize the Mahalanobis distance between $(u^T, \{Y - X\beta - Zu\}^T)^T$ and 0. For details, see Robinson (1991).

For the functional mixed effects model (6), a nature predictor of the random function $f(t)$ is given by

$$\hat{f}(t) = X(t)^T \hat{\beta} + Z(t)^T \hat{u}.$$

Clearly, it is a linear unbiased predictor of $f(\cdot)$. We now show that it is a functional BLUP. For any given constant vectors $\mathbf{c}$ and $\mathbf{d}$, according to Robinson (1991), $\mathbf{c}^T\hat{\beta} + \mathbf{d}^T\hat{u}$ is a BLUP of the random variable $\mathbf{c}^T\hat{\beta} + \mathbf{d}^T\hat{u}$. Therefore, for any fixed t_0, $\hat{f}(t_0)$ is a BLUP of the random variable $f(t_0)$ in the mean square sense. Since the last conclusion holds for any t_0, it follows easily that the function $\hat{f}(\cdot)$ is a functional BLUP of the random function $f(\cdot)$ in the MISE sense.

From the above discussions, it is clear that the smoothing spline solution to the penalized least-squares problem (2) of the Brumback and Rice paper is a functional BLUP in the functional mixed effects model (6) with

$$X(t)^T = (B_1(t), B_2(t)), Z(t)^T = (B_3(t), \cdots, B_n(t))UD^{-1/2},$$

and $G = \sigma^2 I_{n-2}$ and $\rho(s,t) = I(s=t)$, where the notation introduced in Section 2 of the Brumback and Rice paper is adopted. Similarly, by Theorems 1 and 2 as well as their proofs in the Brumback and Rice paper, one can easily see the smoothing spline solutions to the penalized least squares problems given in (14) and (18) of the Brumback and Rice paper are also BLUP functions.

Additional References

Diggle, P.J., Liang, K.Y. and Zeger, S.L. (1994), *Analysis of Longitudinal Data*, Oxford: Clarendon Press.

Fan, J. (1996), "Test of significance based on wavelet thresholding and Neyman's truncation", *Journal of American Statistical Association*, 91, 674–688.

Fan, J. and Lin, S.K. (1996), "Test of significance when data are curves", revised for *Journal of American Statistical Association.*

Fan, J. and Zhang, W. (1997), "Statistical estimation in varying coefficients models", *Unpublished manuscript.*

Fan, J. and Zhang, J.T. (1998), "Functional linear models for longitudinal data", *Unpublished manuscript.*

Hand, D. and Crowder, M. (1996), *Practical Longitudinal Data Analysis*, Chapman and Hall, London.

Hastie, T. J. and Tibshirani, R.J.(1993), "Varying-coefficients models", *Journal of Royal Statistical Society B*, **55** 757-796.

Hoover, D.R., Rice, J.A., Wu, C.O. and Yang, L.P.(1997), "Nonparametric smoothing estimates of time-varying coefficient models with longitudinal data", *Biometrika*, to appear.

Ramsay, J.O. and Silverman, B.W. (1997), "Functional Data Analysis", Springer-Verlag, Berlin.

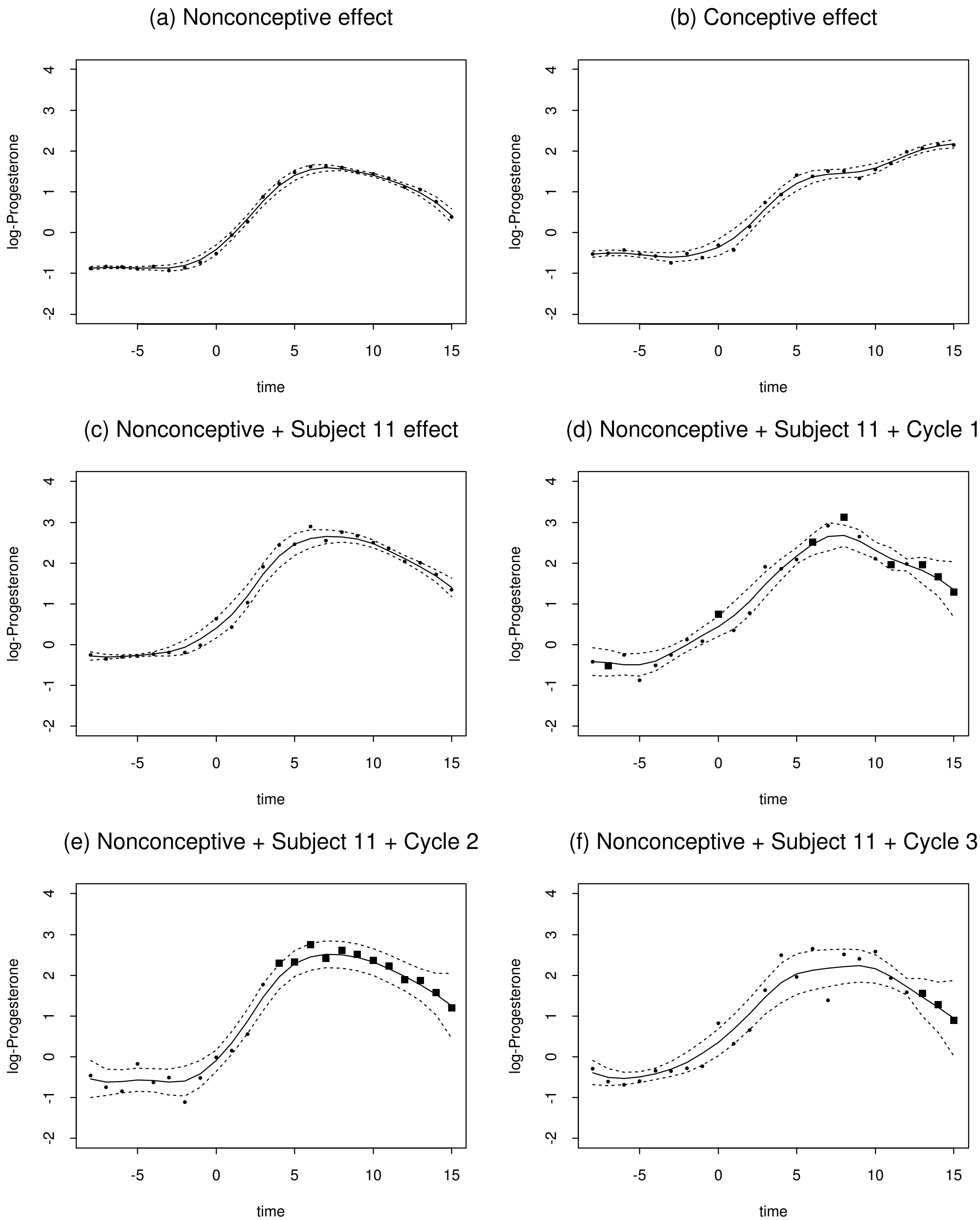

Figure 1: Results of applying the method of Fan and Zhang (1998) to the progesterone data. Solid curves – estimator $\hat{\beta}_S(\cdot)$, dots — estimate $\hat{\beta}_R(\cdot)$; solid squares — imputed data.