DISCUSSION

of "Inference for density families using functional principal component analysis" by A. Kneip and K. J. Utikal

Jeng-Min Chiou[1] and Hans-Georg Müller[2]

[1] Division of Biostatistics Research, National Health Research Institutes

Taipei 115, Taiwan, R.O.C.

[2]Department of Statistics, University of California, Davis

One Shields Ave., Davis, CA 95616, USA

November 6, 2000

This research was supported in part by NSF Grant DMS-99-71602, NSA Grant MDA 904-99-1-0005, NIH Grant P01-AG08761 and NSC Grant 89-2118-M-194-001. We thank Professor J. Carey for making the medfly fecundity data available to us.

1. INTRODUCTION

Kneip and Utikal (K&U for short) have written a nice and innovative piece. The extension of Functional Data Analysis to cover the case of samples of density functions is a major contribution and can be expected to have many important applications ranging far beyond econometrics. For example, the application of functional principal components analysis (FPCA) in the biological and medical sciences has been proposed in work by Kirkpatrick and Heckman (1989), Staniswalis and Lee (1998) and Müller and Wang (1998), among others.

This method has particular promise for unraveling sources of variability and frailty in dynamic biological processes as reflected in trajectories of mortality or growth curves. These are timely topics in ecology and evolution as biologists try to understand the "degrees of freedom" in a biological system. An extension of FPCA to "eigenzeit analysis", which in addition to principal components and eigenfunctions also includes variation in time scales as a second dimension of variability, was described in Capra and Müller (1997) and illustrated with samples of hazard functions. Eigenzeit analysis not only includes linear additive random effects obtained as principal components, but in addition random time scales and thus nonlinear random effects. In the case of hazard functions this effect may be interpreted as speed of aging. For density functions, it would amount to a random scaling factor.

Obtaining functional principal components via a matrix of pairwise scalar products of the functions in the sample is a particularly nice idea of K&U. This proposal carries

over to all applications of functional principal components. It would be of interest to see in more detail what the advantages are beyond simplicity. Especially an assessment of the computational savings as compared to the traditional method (see Rice and Silverman 1991, Capra and Müller 1997) would be valuable. In the traditional method, one first estimates the covariances on a suitably chosen grid of points and then obtains the spectrum of the resulting covariance matrix. The option to choose the grid points enhances flexibility. It is an open question which method would be better under good data-adaptive choices.

2. FINITE DIMENSIONAL AND INFINITE DIMENSIONAL DATA ANALYSIS

In K&U's approach the principal components are actually fixed parameters and not random variables. In fact, the observed density functions are not considered random. This becomes clear when one realizes that the number of observed functions T is allowed to stay finite in the asymptotic analysis. Only the sample size per density has to grow larger to ensure estimated densities and target densities and in particular their scalar products eventually are close. Therefore we are actually not dealing here with a FDA procedure, a term which refers to procedures for a sample of random functions or other infinite dimensional objects. The level of difficulty of the analysis is informed by this fact: The analysis of infinite dimensional data is "infinitely" harder than that of finite-dimensional data in a certain sense. This is the main reason why to date there are preciously few asymptotic results available for functional data. Arguably, this remains one of the major unsolved asymptotic problems of contemporary statistics.

It seems that instead of discussing methods for a sample of random densities, K&U develop a clever and sophisticated model for doubly indexed data. This model allows for a complex joint distribution, but it remains a distribution of finite dimensional objects. The PCA methodology plays the role of an auxiliary tool in this development, and is not used in the FPCA sense to model a sample of functional data. This might limit the applicability of this specific approach for doing FPCA. In many life sciences studies, a natural assumption is that individuals or cohorts that enter a study are drawn from a larger population, and so it is natural to view the data as a random sample. For example, each observed cohort might generate a random density. The spectrum of the covariance operator then permits insights about the workings of nature, in terms of constraints and variation. In the K&U model, covariance and covariance operator have no meaning as there are no random processes to begin with.

The covariance method for estimating the spectrum of the covariance operator is naturally motivated when the observed densities are considered to be a sample of random functions. This motivation is lost in the non-random situation considered by K&U. For that situation, the scalar product method seems to be much more appropriate. While there is a formal equivalence between these methods for finite-dimensional spectral analysis, does this carry through if $T \to \infty$, the basic asymptopia for functional data analysis? We note that for a random sample of densities, the matrix of scalar products $< f_t - f_\mu, f_s - f_\mu >$ is a random matrix for which all elements have expected value 0 and the same finite variance. Does this indicate problems for the stability and condition of this matrix for the case of a random sample?

An asymptotic model for infinite dimensional data would target the spectrum of the covariance operator for inference and would contain a sequence of non-random eigenfunctions determined exclusively by the moment structure of the underlying processes. The number L of assumed eigenfunctions in the data model cannot be fixed, but would in most cases have to increase in analogy to the bandwidth sequences in ordinary curve smoothing. Bandwidths have to shrink to zero to achieve unbiasedness, except for cases where the function to be estimated is not truly infinite-dimensional. If this function is for example linear then local linear fitting with a fixed bandwidth can provide unbiased estimates. An asymptotic analysis for the case $L, T \to \infty$ might perhaps borrow some ideas from the asymptotics of regression modeling with increasing number of predictors (see, for example, He and Shao 2000).

3. AN ILLUSTRATION OF SMOOTH RANDOM EFFECTS DENSITY MODELING

We provide here an illustration of the methods in Chiou and Müller (2000), adapted to a small sample of random densities. An extension of FPCA is introduced, assuming that the principal components, viewed as random effects obtained from a sample of random curves, depend on a one- or higher-dimensional predictor variable. More precisely, the model is, in K&U's notation, and for the case of random density functions,

$$f_{\mathbf{U},\mathbf{V}}(s) = f_{\mu,\mathbf{U}}(s) + \sum_{r=1}^{\infty} \theta_r(\mathbf{V}) g_r(s),$$

where $(\mathbf{U}, \mathbf{V})$ is a covariate vector. This model combines ideas from FPCA with varying coefficient modeling (see, e.g., Hoover et al. 1998 and Fan and Zhang 2000).

For the case of multivariate predictors, our model includes a dimension reduction step which is implemented using the Quasi-Likelihood with Unknown link and variance function Estimation (QLUE) procedure, a semiparametric single index model (Chiou and Müller 1998). In this approach, both link and variance function are assumed to be smooth and are estimated nonparametrically, while the components of the projection vector defining the single index constitute the parametric part. Of special interest in this functional regression model are the conditional link functions for the random effects, also referred to as principal component functions. These are given by

$$\eta_r(\mathbf{v}) = E\left(\theta_r(\mathbf{V})|\mathbf{V} = \mathbf{v}\right),$$

and are assumed to be smooth functions of a single index.

Considering a simplified version where $\mathbf{U}$ is omitted and $\mathbf{V} = V$ is a univariate covariate, we apply this model to log densities. Our model then is, using L components,

$$\log f_V(s) = \mu(s) + \sum_{r=1}^{L} \theta_r(V) g_r(s).$$

This model is implemented in an iterative fashion, iterating over updates of the link functions and of the means (in the multidimensional case, also the QLUE single index steps will be updated).

In Chiou and Müller (2000), bandwidths for smoothing steps and dimension L are chosen by the criterion of "leave-one-curve-out cross-validation". This idea goes back to Rice and Silverman (1991) and is also mentioned (however not implemented) in

K&U. For the special case of density estimation, one could minimize integrated squared prediction error (PE) with respect to bandwidth b and number of components L, where

$$PE(b, L) = \sum_t \left(\int \left[\hat{f}_t(u) - \left(\hat{f}_\mu^{(-t)}(u) + \sum_{r=1}^{L} \hat{\eta}_r^{(-t)}(V_t) \hat{g}_r^{(-t)}(u) \right) du \right]^2 \right).$$

,

and V_t is the covariate for density f_t.

As an illustration, we apply this model to data on the relationship between number of eggs laid (outcome) and lifetime (predictor) for a sample of 936 medflies. Details on data and experiment are given in Carey et al. (1998). The idea is to relate variability and distribution of reproductive success for the flies, measured by total number of eggs laid, to the lifetime of the fly. How does the distribution of the reproductive success measure (number of eggs laid) vary with lifetime and what changes occur in density shapes? This is a situation where the densities could also be assumed to be non-random as in the model of K&U. The data are shown in Figure 1. We obtain 18 densities by partitioning the lifetime axis into 18 subintervals, each containing about 50 flies.

These 18 estimated densities are shown in Figure 2; we use density estimation via local linear fitting, prebinning the data with a very small bin width; one can show that this is equivalent to kernel estimation with boundary modified kernels if the bin widths are small enough. The eigenfunctions for the log-densities are in Figure 3, for the choice $L = 2$, and the estimated random effects link functions or principal component functions are shown in Figure 4. These graphs indicate that component 2 is increasing to about

30 days and then is fairly constant while component 1 declines and stabilizes after day 40. A continuous shift from a density foimnated by the first to a density dominated by the second eigenfunction (with a later peak) is completed by day 40.

Finally, putting everything together, we obtain the fitted surface

$$\hat{f}(s,v) = \tilde{f}(s,v) \Big/ \int \tilde{f}(u,v)du,$$

where

$$\tilde{f}(s,v) = \exp\left[\hat{\mu}(s) + \sum_{r=1}^{L} \hat{\eta}_r(v)\hat{g}_r(s)\right].$$

This resulting surface is shown in Figure 5. The modal ridge clearly visible in Figure 5 highlights the shifting pattern of the total eggs distribution with increasing lifetime. This brief analysis provides another glimpse of the wide applicability and scope of the proposal of K&U to use FPCA models for families of density functions.

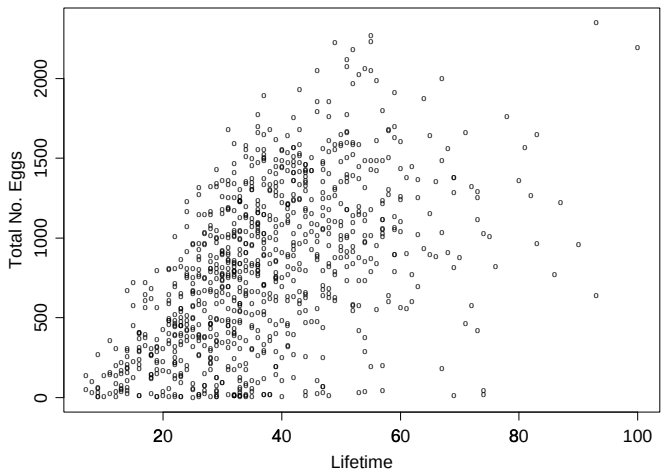

Figure 1: Scatterplot of total number of eggs versus lifetimes for 936 female medflies.

Figure 2: Sample of 18 individually estimated density curves. The method used is density estimation via regression smoothing (bandwidth is 180).

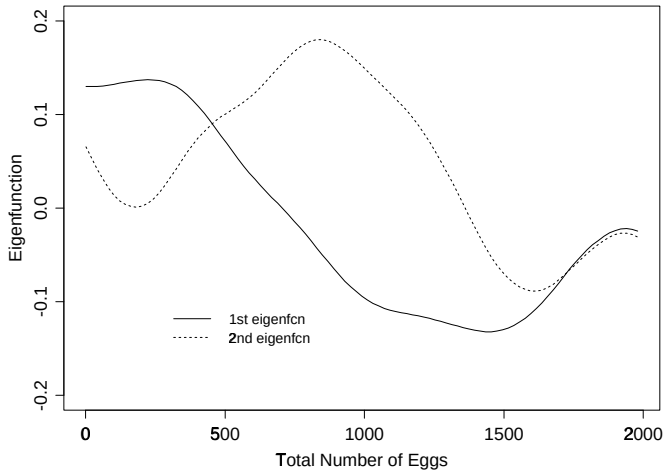

Figure 3: The first two estimated eigenfunctions corresponding to the first two largest eigenvalues for the log-densities (bandwidth is 180).

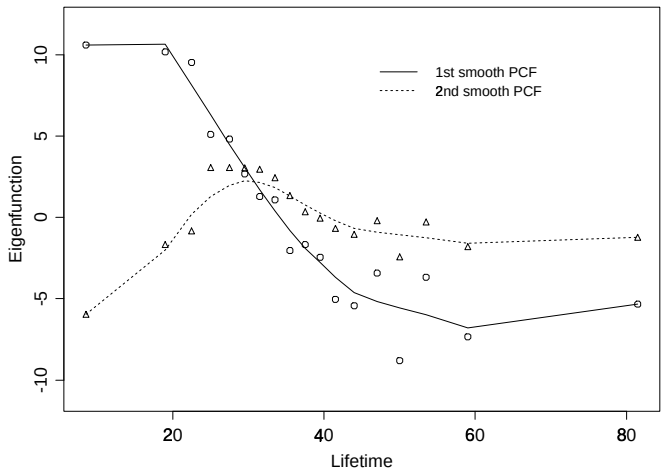

Figure 4: The first two estimated principal component functions corresponding to the first two largest eigenvalues (CV bandwidth).

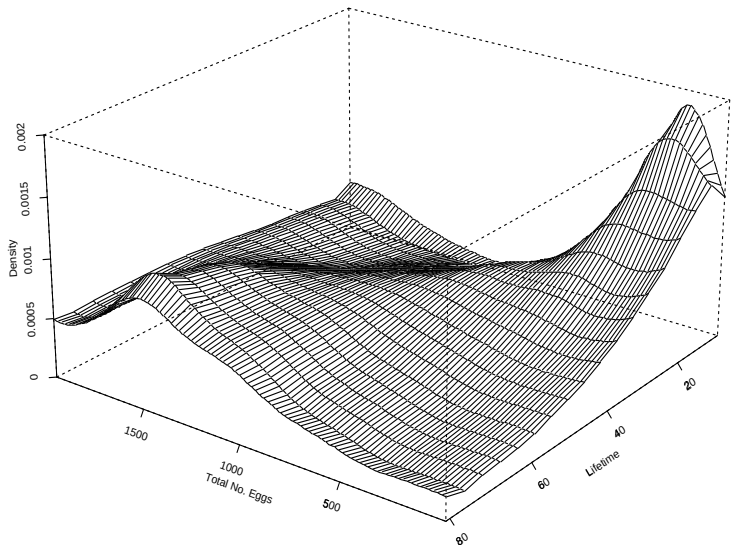

Figure 5: Surface generated by predicted density functions for total numbers of eggs, with lifetime as predictor ($L = 2$).

ADDITIONAL REFERENCES

Capra, W.B. and Müller, H.G. (1997) An Accelerated Time Model for Response Curves. *Journal of the American Statistical Association*, 92, 72-83.

Carey, J., Liedo, P., Müller, H.G., Wang, J.L. and Chiou, J.M. (1998) Relationship of Age Patterns of Fecundity to Mortality, Longevity and Lifetime Reproduction in a Large Cohort of Mediterranean Fruit Fly Females. *Journal of Gerontology*, 53, B245-B251.

Chiou, J.M. and Müller, H.G. (2000) Semiparametric Quasi-likelihood Modeling for

Curve Data. *Manuscript*.

Chiou, J.M. and Müller, H.G. (1998) Quasi-likelihood Regression With Unknown Link and Variance Functions. *Journal of the American Statistical Association*, 93, 1376-1387.

Fan, J.Q. 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.

He, X.M. and Shao, Q.M. (2000) On Parameters of Increasing Dimensions. *Journal of Multivariate Analysis*, 73, 120-135.

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.

Kirkpatrick, M. and Heckman, N. (1989) A Quantitative Genetic Model for Growth, Shape, Reaction Norms and Other Infinite-dimensional Characters. *Journal of Mathematical Biology*, 27, 429-450.

Müller, H.G. and Wang, J.L. (1998) Statistical Tools for the Analysis of Nutrition Effects on the Survival of Cohorts. *Advances in Experimental Medicine and Biology*, 445, 191-206.

Staniswalis, J.G. and Lee, J.J. (1998) Nonparametric Regression Analysis of Longitudinal Data. *Journal of the American Statistical Association*, 93, 1403-1418.