A Stickiness Coefficient for Longitudinal Data

June 2011

Andrea Gottlieb[1]
Graduate Group in Biostatistics
University of California, Davis
1 Shields Avenue
Davis, CA 95616 U.S.A.
Phone: 1 (530) 752-2361
Fax: 1 (530) 752-7099
Email: gottlieb@wald.ucdavis.edu

Hans-Georg Müller
Department of Statistics
University of California, Davis
1 Shields Avenue
Davis, CA 95616 U.S.A.
Phone: 1 (530) 752-2361
Fax: 1 (530) 752-7099
Email: mueller@wald.ucdavis.edu

[1]Corresponding Author

ABSTRACT

In this paper, we introduce the stickiness coefficient, a summary statistic for time-course and longitudinal data, which is designed to characterize the time dynamics of such data. The stickiness coefficient provides a simple, intuitive and informative measure that captures key information contained in time-course data. Under the assumption that the data are generated by the trajectories of a smooth underlying stochastic process, the stickiness coefficient illuminates the relationship between the value of the process at one time with the value it assumes at another time via a single numeric measure. In particular, the stickiness coefficient summarizes the extent to which deviations from the mean trajectory tend to co-vary over time. The estimation scheme we propose will allow for estimation even in the case that the longitudinal data are sparsely observed at irregular times and may be corrupted by noise. We demonstrate an estimation procedure for the stickiness coefficient and establish asymptotic consistency as well as asymptotic convergence rates. We illustrate the resulting stickiness coefficient with some theoretical calculations as well as several economic and health related data examples.

KEY WORDS: Functional Data Analysis, Longitudinal Data, Empirical Dynamics.

1 Introduction

Any course in elementary statistics includes a section on basic summary measures for univariate data. Topics typically include measures of central tendency as well as measures of variability. Although none of these statistics are capable of expressing all the information contained in a data set, they allow for summarization and provide valuable initial insight. Furthermore, they provide a clear method for comparing and contrasting multiple data sets. In regression analysis, the coefficient of determination provides a simple and interpretable summary of the goodness of fit of a particular model and a basic methodology for comparing two models. The purpose of this paper is to introduce a similarly salient and intuitive coefficient for longitudinal data that allows for data summarization and comparison between data sets. The proposed stickiness coefficient is designed to capture a property that is intrinsic to longitudinal processes and its estimation could be one of the initial steps of data exploration.

Consider a classic data set in the study of longitudinal and functional data. Figure 1 shows the growth curves of 10 randomly selected girls and boys from the Berkeley Growth Study (Tanner et al. 1966).

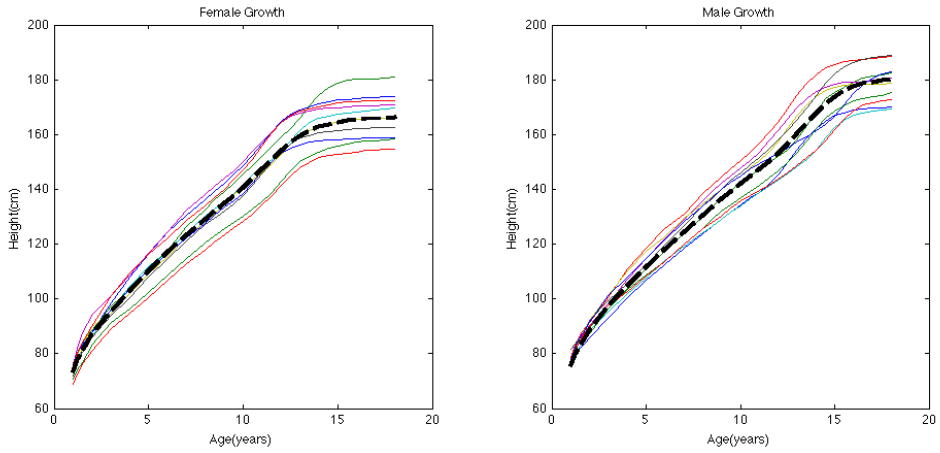

Figure 1: Growth curves from the Berkeley Growth Study and smooth estimate of the mean function (dash). Left: Growth curves of 10 randomly selected girls. Right: Growth curves of 10 randomly selected boys.

There is a considerable body of literature concerning the analysis of growth curves (Boularan et al. 1994; Gasser et al. 1984; Kirkpatrick and Heckman 1989; Rao 1958) and in particular of the Berkeley Growth Study due to the availability of these data (Jones and Bayley 1941). Growth data have played an important role in the development of functional data analysis (Jones and Rice 1992; Ramsay and Silverman 2005; Rice and Wu 2001). In particular, these data demonstrate the importance of curve-alignment (Gervini and Gasser 2004) and derivative estimation (Zhou and Wolfe 2000), two key topics in this field.

The proposed stickiness coefficient highlights another characteristic of these data. In Figure 1, the mean trajectory is denoted by the thick dashed curve and was estimated from all participants in the study. If you follow any of the individual trajectories with respect to their relative location to mean curve over time a noticeable pattern emerges. Curves tend to remain either above or below the mean curve over the entire domain. That is to say, babies who are longer than average at the beginning of life tend to remain taller than average throughout childhood. Likewise, those who start out shorter than average tend to remain below average height. This signifies a certain amount of 'stickiness' in the growth process. Subjects tend to be 'stuck' as either below or above the average height trajectory over time.

This particular 'stickiness' is certainly not unique to the human growth process. Many economic processes will share this quality. An obvious example would be a measure such as Gross Domestic Product, GDP, per capita. Developed countries tend to have a higher than average GDP per capita over time whereas less developed countries tend to have a lower than average GDP per capita over time with little exchange between the two groups over time. An interesting example that we will explore below is income inequality within a nation and its evolution over time.

Not all longitudinal processes display the stickiness feature as there are many processes that are distinctively non-sticky. In fact, certain processes tend to self-stabilize over time. Recently, **?** investigated longitudinal online auction data. The data consisted of price bids on eBay for 156 online auctions of Palm Personal Digital Assistants. They found that prices tend to self-stabilize over time. In other words, it is quite unlikely to observe a price trajectory that runs away at levels much higher than the mean trajectory or plummets way below the mean trajectory since other bidders will be reluctant to make a higher bid on an already high priced item (Liu and Müller 2009).

The notion that some processes tend to reinforce deviations from the mean trajectory over time while others do not is the primary motivation for developing a measure of stickiness for longitudinal data. In this paper, we propose a stickiness coefficient that summarizes the extent to which the deviation of the process from the mean trajectory at one time tends to co-vary with the deviation of the process from the mean trajectory at another time. The proposed coefficient is simple and designed to distinguish between processes that have realizations that tend to remain on one side of the mean function from those that show upwards and downwards mobility across the mean function. The stickiness coefficient could be useful in medical studies to help identify health measures that are more or less amenable to intervention. Measures with high stickiness coefficients, on average, are resistant to change, while those with lower coefficients tend to be more malleable.

We are interested in the estimation of this coefficient for a general class of longitudinal data. We avoid the rather restrictive assumption that the entire time course of the process is observable. Avoiding this assumption allows for the analysis of sparse, irregularly sampled, noise-contaminated longitudinal measurements, which are common in longitudinal studies of health, social and psychological development. A basic assumption is that the data are generated by an underlying smooth and square integrable stochastic process, which might only be observed only intermittently. Functional data analysis for such sparsely observed processes has become a much debated topic in the recent literature (Hall et al. 2006; Staniswalis and Lee 1998; Zhao et al. 2004).

An appropriate data model for longitudinal measurements, which reflects that the data consist of sparse, irregular and noise corrupted measurements of an underlying smooth random trajectory for each subject or experimental unit, is as follows. Given n realizations X_i of the underlying process X on an interval $\mathcal{T}$ with length $|\mathcal{T}|$ and N_i of an integer-valued bounded random variable N, we assume that N_i measurements Y_{ij}, $i = 1, \ldots, n$, $j = 1, \ldots, N_i$, are obtained at random times T_{ij}, for the i-th subject or unit, according to

$$Y_{ij} = Y_i(T_{ij}) = X_i(T_{ij}) + \varepsilon_{ij} \quad T_{ij} \in \mathcal{T}, \tag{1}$$

where ε_{ij} are zero mean i.i.d. measurement errors, with $\mathrm{var}(\varepsilon_{ij}) = \sigma^2$, independent of all other random components. The stickiness coefficient will be defined for data generated as in (1).

The paper is organized as follows. In Section 2 we introduce the population stickiness coefficient. We also review expansions in eigenfunctions and functional principal components, which we use as a tool for dimension reduction. Next, an expansion of the proposed coefficient in terms of the eigen-decomposition and covariance structure is presented. Section 3 consists of several theoretical examples including Brownian motion, followed by a discussion of estimation procedures. In Section 4 we explore several longitudinal health and economic data examples. Asymptotic properties are considered in Section 5, followed by concluding remarks in Section 6. We defer technical results to an Appendix, with estimation procedures in Appendix A.1, assumptions and auxiliary results in Appendix A.2 and proofs in Appendix A.3.

2 Stickiness Coefficient

2.1 Population Definition for Stickiness Coefficient

The proposed stickiness coefficient aims to provide a standardized measure of the extent to which deviations of the process from the mean trajectory tend to co-vary over the time course of the process. Given a stochastic process X defined on a interval domain $\mathcal{T}$, define

$$S_X = \frac{E[\{X(T_1) - \mu(T_1)\}\{X(T_2) - \mu(T_2)\}]}{[\mathrm{Var}\{X(T_1) - \mu(T_1)\}]^{\frac{1}{2}}[\mathrm{Var}\{X(T_2) - \mu(T_2)\}]^{\frac{1}{2}}} \tag{2}$$

where $\mu(t) = E[X(t)]$ is the mean trajectory and T_1 and T_2 are independent random times, independent of the process X, that are uniformly distributed on $\mathcal{T}$.

For fixed t_1 and t_2 in $\mathcal{T}$, this simply represents the correlation between the random variables $X(t_1)$ and $X(t_2)$. However, since we are interested in a stickiness measure that reflects the entire time course of the data, we take expected values with respect to random sampling times T_1 and T_2. Superficially, the stickiness coefficient appears related to the concept of the auto-covariance function in time series analysis. However, the expression in (2) is a coefficient rather than a bivariate function and is designed for a sample of non-stationary functional data, rather than a time series where stationarity usually is a basic assumption.

The stickiness coefficient S_X has a convenient representation in terms of the eigenvalues

and eigenfunctions of the covariance operator of the process X. In order to explore this connection, we give a brief review of functional principal components below.

2.2 Functional Principal Components

Functional Principal Component Analysis (FPCA) is a crucial methodology for both modeling and dimension reduction of sparse, irregular noise-corrupted longitudinal data. We make the minimal assumptions that the underlying unobserved random trajectories $X(t)$ that generate the available sparse observations are square integrable on the domain $\mathcal{T}$ with mean function $EX(t) = \mu(t)$ and auto-covariance function $\text{cov}(X(s), X(t)) = G(s,t),\ s,t \in \mathcal{T}$. Here $G(s,t)$ is a smooth, symmetric and non-negative definite surface. Using G as the kernel in a linear operator leads to the Hilbert-Schmidt operator $(A_G f)(t) = \int_{\mathcal{T}} G(s,t)f(s)ds$. Denoting the ordered eigenvalues (in declining order) of this operator by $\lambda_1 \geq \lambda_2 \geq \ldots \geq 0$ and the corresponding orthonormal eigenfunctions by ϕ_k, one obtains the well-known representation $G(s,t) = \sum_{k=1}^{\infty} \lambda_k \phi_k(s)\phi_k(t)$ for the covariance surface and the Karhunen-Loève representation $X_i(t) = \mu(t) + \sum_{k=1}^{\infty} \xi_{ik}\phi_k(t)$ for the individual trajectories. Here, $\xi_{ik} = \int_{\mathcal{T}}(X_i(t) - \mu(t))\phi_k(t)dt$, $k = 1, 2, \ldots$, are the functional principal components (FPCs) of the random trajectories X_i which are uncorrelated random variables with $\text{E}(\xi_{ik}) = 0$, $E\xi_{ik}^2 = \lambda_k$, and $\sum_k \lambda_k < \infty$ (Ash and Gardner 1975). The eigenfunctions ϕ_k are the solutions of the eigen-equations $\int G(s,t)\phi_k(s)\,ds = \lambda_k\phi_k(t)$, $k = 1, 2, \ldots$, under the constraint of orthonormality. Estimation of these components will be discussed in Appendix A.1.

2.3 Alternative Representation of the Stickiness Coefficient

Expressing the proposed coefficient in terms of the eigenvalues and eigenfunctions of the covariance operator easily demonstrates that the coefficient can only take values in the interval $[0, 1]$ and also leads to a potential estimation scheme. Using the joint uniform distribution of (T_1, T_2) and conditioning we obtain from (2),

$$\begin{aligned} S_X &= \frac{E[E[\{X(T_1) - \mu(T_1)\}\{X(T_2) - \mu(T_2)\}|T_1, T_2]]}{E\{\text{Var}(X(T_1) - \mu(T_1))|T_1\} + \text{Var}\{E(X(T_1) - \mu(T_1))|T_1\}} \\ &= \frac{E\{\sum_{k=1}^{\infty} \xi_k\phi_k(T_1) \sum_{j=1}^{\infty} \xi_j\phi_j(T_2)\}}{E\{\text{Var}(X(T_1) - \mu(T_1))|T_1\}} \end{aligned}$$

$$= \frac{\sum_{k=1}^{\infty}\{\lambda_k \int_{\mathcal{T}} \phi_k(t_1) f_{T_1}(t_1) dt_1 \int_{\mathcal{T}} \phi_k(t_2) f_{T_2}(t_2) dt_2\}}{\frac{1}{|\mathcal{T}|}\sum_{k=1}^{\infty}\lambda_k},$$

whence

$$S_X = \frac{1}{|\mathcal{T}|}\frac{\sum_{k=1}^{\infty}\lambda_k[\int_{\mathcal{T}}\phi_k(t)dt]^2}{\sum_{k=1}^{\infty}\lambda_k}. \tag{3}$$

It follows immediately from the Cauchy-Schwarz inequality that $0 \leq S_X \leq 1$. An alternate expression for S_X can also be given directly in terms of the smooth covariance operator $G(s,t)$,

$$S_X = \frac{1}{|\mathcal{T}|}\frac{\int_{\mathcal{T}\times\mathcal{T}} G(s,t)dsdt}{\int_{\mathcal{T}} G(t,t)dt}. \tag{4}$$

3 Theoretical Examples

3.1 Stickiness Coefficient for Brownian Motion

While Brownian motion has a physical interpretation in terms of the diffusion of a particle suspended in a fluid, it is also a key process for modeling time-dynamic phenomena from financial markets to biological growth and development (Karatzas and Shreve 1991). Brownian motion on $[0,1]$ is a Gaussian process $X(t)$, with $X(0) = 0$, and covariance function $\text{cov}(X(s), X(t)) = G(s,t) = \min(s,t)$ for $s,t \in [0,1]$. It has eigenfunctions $\phi_k(t) = \sqrt{2}\sin\{(k-\frac{1}{2})\pi t\}$, eigenvalues $\lambda_k = \frac{4}{(2k-1)^2\pi^2}$ and independent functional principal components $\xi_k \sim N(0,\lambda_k)$, $k = 1,2,\ldots$.

From (3), one finds

$$S_X = \frac{\sum_{k=1}^{\infty}\frac{4}{(2k-1)^2\pi^2}\frac{8}{(2k-1)^2\pi^2}}{\sum_{k=1}^{\infty}\frac{4}{(2k-1)^2\pi^2}} = \frac{8}{\pi^2}\frac{\sum_{k=1}^{\infty}\frac{1}{(2k-1)^4}}{\sum_{k=1}^{\infty}\frac{1}{(2k-1)^2}} = \frac{2}{3}. \tag{5}$$

Figure 2 shows three simulated trajectories of Brownian motion on $[0,1]$, demonstrating a certain degree of inherent stickiness. Once a trajectory deviates significantly from the mean function, the zero line, it is unlikely for the process to cross the line again in the near future. We can view the stickiness of Brownian motion in (5) as providing a natural threshold between processes that are sticky and those that are not. We would therefore characterize a process as particularly sticky if its stickiness coefficient exceeds $\frac{2}{3}$.

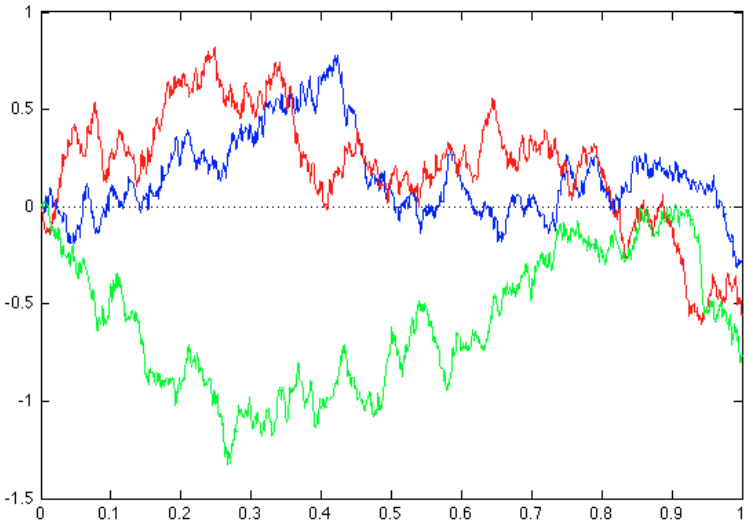

Figure 2: Three simulated Brownian motion trajectories.

3.2 Case of One Eigenfunction

In this section we explore the case where the expansion of the process $X(t)$ consists of a single eigenfunction ϕ. This is an important theoretical example, but is also of practical importance as many processes are generated by one dominant eigenfunction. In this situation, (3) reduces to

$$S_X = \frac{1}{|\mathcal{T}|}\left[\int_{\mathcal{T}} \phi(t)dt\right]^2. \tag{6}$$

Thus the magnitude of S_X is entirely determined by the shape of ϕ. For simplicity, we restrict our attention to centered processes, $\mu(t) = 0$, on $\mathcal{T} = [0, 1]$. If $\phi(t) = \sqrt{12}(t - \frac{1}{2})$, a quick calculation verifies $S_X = 0$. In this situation, in the absence of measurement error, all trajectories will necessarily cross the mean trajectory. On the other hand if $\phi(t) \equiv 1$, $S_X = 1$. In this case, in the absence of measurement error, trajectories will always remain on one side or another of the mean trajectory.

In fact, the preceding case is the only situation in which $S_X = 1$. It holds that, $S_X = 1$ if, and only if, $G(s, t)$ is a constant function. Without loss of generality, suppose $\mathcal{T}$=[0,1]. A straightforward calculation shows that if $G(s, t) = \lambda$, then indeed $S_X = 1$. Conversely suppose $S_X = 1$. Then from (3),

$$\sum_{k=1}^{\infty} \lambda_k \left[\int_{\mathcal{T}} \phi_k(t)dt\right]^2 = \sum_{k=1}^{\infty} \lambda_k. \tag{7}$$

Since $\int \phi(t)dt \leq \int \phi^2(t)dt = 1$, it follows that $\left[\int_{\mathcal{T}} \phi_k(t)dt\right]^2 \leq 1$, for all k. Thus (7) can only hold when $\left[\int \phi_k(t)dt\right]^2 = 1$ for all k such that $\lambda_k > 0$. Now, the two statements $\left[\int \phi_k(t)dt\right]^2 = 1$, $\int \phi_k^2(t)dt = 1$ can only hold simultaneously if $\phi_k(t) \equiv 1$. This means there can only be one eigenfunction with a non-zero eigenvalue, $\phi(t) \equiv 1$ and thus $G(s,t)$ is constant.

3.3 Estimation Procedures

Equation (3) suggests a natural method for the estimation of S_X in which plug-in estimators are used to estimate the eigenvalues λ_k and eigenfunctions $\phi_k(t)$. For estimation, one must truncate the expression at a finite number $K = K(n)$ of included eigen-components, which needs to be determined from the data. For asymptotic consistency, $K(n) \to \infty$ as $n \to \infty$ is required. The resulting estimator is

$$\hat{S}_X^K = \frac{1}{|\mathcal{T}|} \frac{\sum_{k=1}^K \hat{\lambda}_k [\int_{\mathcal{T}} \hat{\phi}_k(t)dt]^2}{\sum_{k=1}^K \hat{\lambda}_k}. \tag{8}$$

An alternative estimator could be based on directly targeting expected values in (2), conditioning on pairs of times T_1 and T_2 and using an empirical covariance estimator. However, this would require a dense, balanced design across subjects, which often is not available in the case of longitudinal measurements. In contrast, the proposed estimation procedure in (8) allows for the estimation of the required components even in the case that the longitudinal observations are sparse, irregular and noise corrupted. This strategy requires us to borrow strength from the entire sample. Details about this procedure can be found in (Yao et al. 2005a,b) and a brief summary is provided in Appendix A.1.

A second alternative estimator would involve directly estimating the quantities in expression (4) yielding the estimate,

$$\hat{S}_{X,A} = \frac{1}{|\mathcal{T}|} \frac{\int_{\mathcal{T}\times\mathcal{T}} \hat{G}(s,t)dtds}{\int_{\mathcal{T}} \hat{G}(t,t)dt}. \tag{9}$$

This method does not require estimating the eigen-components and rather relies on an estimate of $G(s,t)$. One can obtain such an estimate from two-dimensional surface smoothing, as described in Appendix A.1. However, there is no guarantee that the outcome of such an estimation procedure yields a positive definite surface and therefore this method can yield estimates that are negative or greater than 1. For further details, we refer to Appendix B.1.

4 Stickiness in Action for Longitudinal Data

We present several health and economic related data examples to demonstrate the wide variety of design schemes that can arise in longitudinal studies. In addition to providing a point estimate for the stickiness coefficient in each case based on Equation (8), confidence intervals for the estimates can be obtained via bootstrap.

4.1 Berkeley Growth Data

The Berkeley growth data contains height measurements for 54 girls and 38 boys from 1 to 18 years of age. The children were measured 4 times between the ages of 1 and 2, yearly from ages 2 to 8 and twice yearly from ages 8 to 18. This design results in 31 measurements over years 1 to 18. For both girls and boys 2 eigen-components were used to summarize the process which accounts for 95% and 94% of the variability, respectively. We obtain an estimated stickiness coefficient of 0.87 and 0.83, for girls and boys, respectively, confirming our intuition about the stickiness of human growth based on Figure 1. Bootstrap confidence intervals at the 95% level are given by [0.813, 0.945] and [0.787, 0.898], respectively. These results indicate that there is a strong tendency in the growth process for deviations from the mean curve to co-vary over time. As anticipated, this implies that a taller than average one-year old is more likely to be a taller-than average child over his or her entire childhood.

4.2 Aging Related Health Measures

In this section we analyze data from a longitudinal study on aging (Pearson et al. 1997; Shock et al. 1984). Systolic Blood Pressure (SBP, measured in mmHg) and Body Mass Index (BMI, measured in kg/m^2) were recorded on each visit of 1590 male volunteers bi-annually. This data set is truly sparse and noisy as many study participants frequently missed visits. As a result, both number of observations per subject and observation times vary widely (see Yao et al. 2005b). Despite the large noise in the measurements and the highly irregular sampling times, these data can still be reasonably viewed as being generated by underlying smooth random trajectories of blood pressure and body mass index. For analysis we select subjects for whom at least two measurements are available between age 40 and 70, which is the minimum number of repeated measurements needed for meaningful analysis (see Section

(4) assumption (A1)). Furthermore, we only consider subjects who survived beyond 70, to avoid problems of selection bias due to non-survival. The resulting sample size is $n = 266$ subjects.

Stickiness coefficients were estimated for the BMI and SBP processes separately. The BMI process is almost completely described by the first eigenfunction which accounts for 98% of the variability, while 2 eigenfunctions were used to summarize the SBP process accounting for 93% of the variability. We obtain stickiness coefficient estimates of 0.95 and 0.80 and 95% bootstrap confidence intervals [0.901, 0.998] and [0.678, 0.872] for BMI and SBP, respectively. These confidence intervals indicate that there is a relatively large amount of sampling variability in the estimates, but nevertheless these processes are clearly quite sticky. Furthermore, since the two measurements are observed on the same collection of men, we can easily obtain a 95% bootstrap confidence interval for the difference between the two stickiness coefficients. We obtain [0.063, 0.285], providing strong evidence that the Body Mass Index trajectory is indeed stickier than the Systolic Blood Pressure trajectory. This confirms our intuition that on average, weight is a particularly sticky process over one's lifetime and requires a great deal of sustained effort to change.

4.3 Online Auction Data

There has been a fair amount of recent interest in the statistical analysis of online auction data (Bapna et al. 2008; Reddy and Dass 2006; Wang et al. 2008). In this section we explore the stickiness of a particular eBay auction. The data consist of 156 auctions of Palm Personal Digital Assistants in 2003 (courtesy of Wolfgang Jank). The data correspond to 'live bids' that are entered by bidders at irregular times over a seven day period and correspond to the actual price a winning bidder would pay for the item. More details on the eBay bidding process can be found in Jank and Shmueli (2006) and Liu and Müller (2009). The time unit of these 7-day auctions is hours and thus the domain is [0, 168]. In order to estimate the stickiness coefficient, we log transform the data and restrict our analysis to the last four days of the seven day bidding cycle because there is a considerable amount of initial variability which is of less interest. This process is well described by two eigenfunctions accounting for 95% of the variability.

As discussed in the Introduction, we would expect that auction prices tend to self-stabilize

over time. Price trajectories rarely remain significantly below or above the average price trajectory for long periods of time. This is quite reasonable given the mechanisms of online auctions. For example, a particularly cheap item is unlikely to remain under-priced over time as bidders notice the deal and eventually bid the price up. The estimated stickiness coefficient is consistent with this intuition. A point estimate for the coefficient was found to be 0.72 which is smaller than any of the health related data examples and the 95% bootstrap confidence interval was found to be [0.575, 0.777].

4.4 Stickiness of Economic Indices

In this section we analyze data obtained from the World Bank (http://data.worldbank.org/). We begin by analyzing income inequality data alluded to in the Introduction. The Gini index measures the extent of the deviation of household incomes within a nation from a perfectly equal distribution. Therefore a nation with a Gini index of 100 would have complete inequality whereas a nation with a Gini index of 0 would have perfect equality. The World Bank recorded Gini indices for 204 countries sporadically over the last 30 years. This is a sparse and irregular data set as many countries have only a handful of measurements over this time period. In order to obtain a meaningful analysis we only considered countries with at least two measurements in [1979, 2009]. This led to a sample size of 93 countries entering the analysis. This process was well described by two eigenfunctions accounting for 97% of the variability. The point estimate for the stickiness coefficient of this processes was high at 0.89. A 95% bootstrap confidence interval was [0.526, 0.973]. This indicates a strong tendency for the distribution of wealth within a country to persist over time.

Next we discuss the economic index Gross Domestic Product (GDP), focusing on GDP growth rates. The data consist of the annual percentage growth rate of GDP at market prices, based on the local currency for 203 countries over the last 50 years. We might expect GDP growth to be a very volatile process, sensitive to all types of global and local economic pressures. In fact, 9 eigenfunctions were required to adequately describe this process, accounting for 93% of the variability. Based on our sample, we obtain an estimated stickiness coefficient of 0.19 with 95% bootstrap confidence interval of [0.137, 0.286]. This indicates that this process is decidedly unsticky.

5 Asymptotic Properties

In this section we obtain rates of convergence and establish asymptotic consistency for $\hat{S}_X$, the estimator of S_X. These results require a collection of regularity conditions concerning the distribution of the design points and the behavior of the eigenfunctions and eigenvalues as their order increases. Also required are assumptions about the large sample behavior of the smoothing bandwidths h_μ for the estimation of the mean function $\mu(t)$, and h_G for the estimation of the covariance surface $G(s,t)$. Specifically, for the observations (T_{ij}, Y_{ij}), $i = 1, \ldots, n$, $j = 1, \ldots, N_i$, corresponding to the i-th trajectory, we require that

(A1) N_i are random variables with $N_i \overset{\text{i.i.d.}}{\sim} N$, where N is a bounded positive discrete random variable with $P\{N \geq 2\} > 0$, and measurement times T_{ij} as well as responses Y_{ij}, $1 \leq \cdots \leq N_i$, are independent of N_i, $1 \leq i \leq n$.

Writing $\boldsymbol{T}_i = (T_{i1}, \ldots, T_{iN_i})^T$ and $\boldsymbol{Y}_i = (Y_{i1}, \ldots, Y_{iN_i})^T$, the triples $\{\boldsymbol{T}_i, \boldsymbol{Y}_i, N_i\}$ are assumed to be i.i.d. and the pairs $\{\boldsymbol{T}_i, N_i\}$ are assumed to be independent of the process X. For the bandwidths used in the smoothing steps for $\mu(t)$ and $G(s,t)$, we require that as $n \to \infty$,

(A2) $\max(h_\mu, h_G) \to 0$, $nh_\mu \to \infty$, $nh_G^2 \to \infty$.

For asymptotic consistency, we will require assumptions about the behavior of the eigenfunctions ϕ_k and eigenvalues λ_k as their order k increases. For this we define

$$\alpha(K) = \sum_{k=K+1}^{\infty} \lambda_k, \qquad \delta_k = \min_{1 \leq j \leq k} (\lambda_j - \lambda_{j+1}). \tag{10}$$

In addition to assumptions (A1)-(A2) given above, additional assumptions are needed about the kernels used in the local linear smoothing steps and the underlying densities. These more technical conditions, (B1)-(B3), are deferred to the Appendix.

Theorem 1. *Under (A1)-(A2) and (B1)-(B3), if $\lambda_k > 0$ for all $k \geq 1$,*

$$\left| \hat{S}_X^K - S_X \right| = O\left(\alpha(K)\right) + \left(\sum_{k=1}^{K} \delta_k^{-1} + K \right) O_p\left(\frac{1}{\sqrt{n} h_G} + h_G^2 \right). \tag{11}$$

Theorem 1 provides a convergence rate for the proposed estimate of the stickiness coefficient. Asymptotic consistency follows immediately if we require the additional assumption

(A3) $\sum_{k=1}^{K} \delta_k^{-1} = o(\min\{\sqrt{n} h_G^2, h_G^{-2}\}), K = o(\min\{\sqrt{n} h_G^2, h_G^{-2}\})$, as $n \to \infty$.

Note that if the process is well approximated by the first few leading terms of its eigen-expansion, then condition (A3) is easily satisfied. The proof of Theorem 1 and additional details about the required regularity conditions are provided in Appendix A.2. Note that if the assumption made in Theorem 1 that $\lambda_k > 0$ for all $k \geq 1$ does not hold, then there exists a K_0 such that for all $k < K_0$, $\lambda_k > 0$ and for all $k \geq K_0$, $\lambda_k = 0$. In this case, it is easy to see that $\left|\hat{S}_X^K - S_X\right| = O_p\left(\frac{1}{\sqrt{n}h_G} + h_G^2\right)$.

Theorem 1 can be used to determine the best choice $K = K(n)$ given a particular form of the eigenvalues and a choice of the smoothing parameter h_G. For example, in the special case that the eigenvalues are decaying exponentially, that is $\lambda_k = e^{-\rho k}$ for some $\rho > 0$, we can explicitly solve for the best choice of included eigen-components $K(n)$. In this case, $\alpha(K) = \frac{e^{-\rho(K+1)}}{1-e^{-\rho}}$ and $\sum_{k=1}^{K} \frac{1}{\delta_k} = \frac{e^{\rho}(e^{\rho K}-1)}{(1-e^{-\rho})(e^{\rho}-1)}$. Suppose that the smoothing parameter h_G decays at a rate $h_G = n^{-\alpha}$ for some $0 < \alpha < \frac{1}{2}$. Then the number of included eigen-components $K(n) = -\frac{1}{2\rho}\ln\left(n^{\alpha-\frac{1}{2}} + n^{-2\alpha}\right)$ minimizes the leading terms of the right hand side of (11).

6 Discussion

There is a substantial body of literature on the interface of longitudinal and functional data analysis that covers diverse and complicated statistical issues. In this paper, we took a more elementary approach to the topic and introduced the stickiness coefficient, a summary statistic for longitudinal data. The purpose is to provide the data analyst with a simple and intuitive calculation in the initial data exploration of a set of sparsely sampled curves. Formally, this coefficient summarizes the extent to which deviations in individual curves from the mean trajectory tend to co-vary over time. Intuitively, populations of curves with high stickiness coefficients are those for which individual curves tend to get stuck above or below the mean trajectory over time. We choose the word 'stickiness' to highlight this property of the coefficient.

The estimation of this coefficient was discussed for a general class of longitudinal data. The only assumption is that the observed data are generated by a smooth process, although the entire process need not be observed. Rather, the observations can be sparsely observed and corrupted by noise. In fact, if the locations are random, two observations of the process

suffice (Hall et al. 2006; Yao et al. 2005b). The reason for this surprising results is that one gains strength for statistical inference by pooling all the data together for estimation of eigenvalues and eigenfunctions. In this context, we established a rate of convergence that depends on smoothing parameters and number of included eigen-components. Using this rate, we were able to determine an asymptotically ideal choice of included eigen-components for a particular eigenvalue structure.

Although human growth data provided a motivating example, the stickiness coefficient was not developed solely for a particular application. Rather, this coefficient is applicable to any field where trajectory data are collected. In the context of financial markets, the stickiness coefficient could be used to evaluate potential winners in the stock market. In a market with a particularly low stickiness index, one should pick a stock with returns below the mean because it is likely to cross above the mean in the near future. Conversely in a very sticky market choosing a stock with returns below the mean is a poor idea as it is likely to remain below average.

In our applications section we examined data sets from the health, social and economic sciences. For the health related examples, the coefficient estimates confirmed our intuition that human growth is a sticky process and that weight (measured as BMI) is a particularly sticky process. Concerning the socio-economic data, we observed that income inequality tends to persist over time in a country while GDP growth can be much more variable. In each case, we were able to ascertain some crucial information about the structure of the process generating the data by examining the stickiness coefficient.

Appendix

A.1 Estimation Procedures

For estimation, we require a method that can handle not only entirely observed functional trajectories but also the type of sparse, irregular data that arises in many longitudinal studies. Although we assume such data are generated by an underlying smooth random process, the observed measurements can be sparse, irregular and corrupted by noise. We follow the procedures introduced in Yao et al. (2005a) and extended in Yao et al. (2005b). In order to overcome the limitations of sparse designs, we borrow strength across subjects by pooling the data to achieve estimates of $\mu(t)$ and $G(s,t)$.

The first step involves aggregating measurements across subjects into one scatterplot and applying a local linear smoother to obtain an estimate for the mean function $\mu(t)$. For a given univariate density function κ_1 and bandwidth h_μ, one would minimize

$$\sum_{i=1}^{n}\sum_{j=1}^{N_i}\kappa_1\left(\frac{T_{ij}-t}{h_\mu}\right)\Big\{Y_{ij}-\{\alpha_0+\alpha_1(T_{ij}-t)\}\Big\}^2 \tag{12}$$

for each t with respect to α_0 and α_1 from which one obtains $\hat{\mu}(t)=\hat{\alpha}_0(t)$ (Fan and Gijbels 1996).

Also required is an estimate of the covariance surface $G(s,t)$. In this step, one forms a pooled scatterplot of pairwise raw covariances $G_i(T_{ij},T_{il})=(Y_{ij}-\hat{\mu}(T_{ij}))(Y_{il}-\hat{\mu}(T_{il}))$, $j\neq l$, and minimizes the objective function

$$\sum_{i=1}^{n}\sum_{1\leq j\neq l\leq N_i}\kappa_2\left(\frac{T_{ij}-t}{h_G},\frac{T_{il}-s}{h_G}\right)\Big\{G_i(T_{ij},T_{il})-(\alpha_0+\alpha_{11}(T_{ij}-t)+\alpha_{21}(T_{il}-s))\Big\}^2 \tag{13}$$

for a fixed (s,t) with respect to α_0,α_{11} and α_{21}. One must choose a bivariate density function as kernel $\kappa_2(s,t)$ and a bandwidth h_G. This leads to the estimate $\hat{G}(s,t)=\hat{\alpha}_0(s,t)$. Notice here that elements along the diagonal are excluded because for these points, $\mathrm{cov}(Y_{ij},Y_{ij}|T_{ij})=\mathrm{cov}(X(T_{ij}),X(T_{ij}))+\sigma^2$. Therefore the diagonal of the raw covariances should be excluded and only $G_i(T_{ij},T_{il})$ for $j\neq l$ should be included in the smoothing step. One can obtain a consistent estimator for σ by taking the difference between a smoother that uses only the diagonal elements and the diagonal estimate obtained from smoothing step (13). We refer to Yao et al. (2005a) for more details.

The last step is to obtain estimates of λ_k and $\phi_k(t)$. This is achieved by numerically solving the eigen-equations,

$$\int_{\mathcal{T}} \hat{G}(s,t)\hat{\phi}_k(s)ds = \hat{\lambda}_k\hat{\phi}_k(t) \tag{14}$$

where the $\hat{\phi}_k(t)$ are subject to $\int_{\mathcal{T}} \hat{\phi}_k(t)^2dt = 1$ and $\int_{\mathcal{T}} \hat{\phi}_k(t)\hat{\phi}_m(t)dt = 0$ for $m < k$.

For the theoretical analysis developed in the next sections one requires that $K = K(n) \to \infty$ as $n \to \infty$. However, for any data analysis, the number of included eigen-terms K must be chosen by the practitioner. There are several methods for doing so, including AIC/BIC criterion based on marginal/conditional pseudo-likelihood or thresholding of the total variation explained by the included components (Yao et al. 2005b). For the data analysis of Section 3, the latter method was employed with threshold 0.9.

A.2 Additional Assumptions and Auxiliary Results

Assumptions about the distribution of the design points and behavior of the bandwidths h_μ, for estimation of the mean function $\mu(t)$ and h_G, for the estimation of the covariance surface $G(s,t)$ were stated as assumptions (A1)-(A2).

In addition to (A1)-(A2), assumptions about the kernels used in the local linear smoothing steps and underlying densities and moment functions are required. We denote the densities of T and (T_1, T_2) by $f_1(t)$ and $f_2(s,t)$ respectively, and embed the interval $\mathcal{T} = [a, b]$, within $\mathcal{T}_\delta = [a - \delta, b + \delta]$ for some $\delta > 0$. Regularity conditions for the densities and the targeted moment functions, where ℓ_1, ℓ_2 are non-negative integers, are given by (B1)-(B3).

(B1) $f_1^{(4)}(t)$ exists and is continuous on $\mathcal{T}_\delta$ with $f_1(t) > 0$, $\frac{\partial^4}{\partial t^{\ell_1}\partial s^{\ell_2}} f_2(s,t)$ exists and is continuous on $\mathcal{T}_\delta^2$ for $\ell_1 + \ell_2 = 4$.

(B2) $\mu^{(4)}(t)$ exists and is continuous on $\mathcal{T}_\delta$, $\frac{\partial^4}{\partial t^{\ell_1}\partial s^{\ell_2}} G(s,t)$ exists and is continuous on $\mathcal{T}_\delta^2$ for $\ell_1 + \ell_2 = 4$.

(B3) Kernel function κ_1 is a symmetric, continuous, non-negative density with compact support. Kernel κ_2 is a continuous bivariate density with compact support such that $\kappa_2(u,v)$ is symmetric in u for each fixed v and is also symmetric in v for each fixed u.

The following lemma provides an L_2 convergence rate for the eigenfunction estimate $\hat{\phi}_k$.

Lemma 1. *Under (A1)-(A2) and (B1)-(B3), for $\phi_k(t)$ corresponding to λ_k of multiplicity 1,*

$$\|\hat{\phi}_k(t) - \phi_k(t)\| = O_p\left(\frac{1}{\lambda_k \delta_k}(\frac{1}{\sqrt{n}h_G} + h_G^2)\right), \ |\hat{\lambda} - \lambda| = O_p\left(\frac{1}{\sqrt{n}h_G} + h_G^2\right), \tag{15}$$

where the $O_p(\cdot)$ term in (15) is uniform in $k \geq 1$.

A proof of this lemma along with a few additional auxiliary assumptions can be found in Müller and Yao (2010). This lemma establishes a L_2 rate of convergence for the eigenfunction estimate $\hat{\phi}_k$ that is uniform in the order k. Our asymptotic results given in Theorem 1 depends crucially on this uniformity.

A.3 Proof of Theorem 1

Throughout the proof of Theorem 1 we will make repeated use of the Cauchy-Schwarz inequality $|\langle f, g\rangle| \leq \|f\|\|g\|$. Specifically, we will need to make use of both inequalities, $[\int_{\mathcal{T}} \phi_k(t)dt]^2 \leq |\mathcal{T}|$ and $[\int_{\mathcal{T}} \hat{\phi}_k(t)dt]^2 \leq |\mathcal{T}|$. We note here that the first inequality is a consequence of the C-S inequality as well as the required orthonormality of the eigenfunctions of the auto-covariance operation $G(s,t)$. The second inequality is a consequence of our estimation scheme which produces estimates $\hat{\phi}_k(t)$ which satisfy $\int_{\mathcal{T}} \hat{\phi}_k^2(t)dt = 1$. Noting that,

$$\begin{aligned}
\Delta &=: |\mathcal{T}| \sum_{k=1}^{\infty} \lambda_k \left| \hat{S}_X^K - S_X \right| \\
&\leq \frac{\sum_{k=1}^K \hat{\lambda}_k \left[\int_{\mathcal{T}} \hat{\phi}_k(t)dt\right]^2}{\sum_{k=1}^K \hat{\lambda}_k} \left| \sum_{k=1}^{\infty} \lambda_k - \sum_{k=1}^{K} \hat{\lambda}_k \right| + \left| \sum_{k=1}^{K} \hat{\lambda}_k \left[\int_{\mathcal{T}} \hat{\phi}_k(t)dt\right]^2 - \sum_{k=1}^{\infty} \lambda_k \left[\int_{\mathcal{T}} \phi_k(t)dt\right]^2 \right| \\
&= I + II,
\end{aligned}$$

we find by Lemma 1,

$$\begin{aligned}
I &\leq \frac{|\mathcal{T}| \sum_{k=1}^K \hat{\lambda}_k}{\sum_{k=1}^K \hat{\lambda}_k} \left| \sum_{k=1}^{\infty} \lambda_k - \sum_{k=1}^{K} \hat{\lambda}_k \right| \\
&\leq |\mathcal{T}| \left(\sum_{k=1}^{K} \left| \hat{\lambda}_k - \lambda_k \right| + \alpha(K) \right) = K O_p\left(\frac{1}{\sqrt{n}h_G} + h_G^2\right) + |\mathcal{T}|\alpha(K) \text{ and, again by Lemma 1,}
\end{aligned}$$

$$\begin{aligned}
II &\leq \left| \sum_{k=1}^{K} \lambda_k \left(\left[\int_{\mathcal{T}} \hat{\phi}_k(t)dt\right]^2 - \left[\int_{\mathcal{T}} \phi_k(t)dt\right]^2 \right) \right| + \left| \sum_{k=1}^{K} \left[\int_{\mathcal{T}} \hat{\phi}_k(t)dt\right]^2 (\lambda_k - \hat{\lambda}_k) - \sum_{k=K+1}^{\infty} \lambda_k \left[\int_{\mathcal{T}} \phi_k(t)dt\right]^2 \right| \\
&\leq \sum_{k=1}^{K} \lambda_k O_p\left(\frac{1}{\lambda_k \delta_k}\left(\frac{1}{\sqrt{n}h_G} + h_G^2\right)\right) + K O_p\left(\frac{1}{\sqrt{n}h_G} + h_G^2\right) + |\mathcal{T}|\alpha(K).
\end{aligned}$$

Therefore, $\Delta \leq I + II = O\left(\alpha(K)\right) + \left(\sum_{k=1}^{K} \delta_k^{-1} + K\right) O_p\left(\frac{1}{\sqrt{n}h_G} + h_G^2\right)$ and thus we have established Theorem 1.

B.1 Comparison of Estimation Methods $\hat{S}_{X,A}$ and $\hat{S}_X^K$ with Simulation

The data analysis of Section 4 was completed using both estimators $\hat{S}_X^K$ in equation (8) and $\hat{S}_{X,A}$ in equation (9). Estimator $\hat{S}_X^K$ requires a choice of K, the number of included eigenfunctions. In Table 1, we show the results for various estimators including $\hat{S}_X^K$ for different $K = K(p)$ where $K(p)$ is selected as the minimum number of eigen-components such that $100p\%$ of the variation was explained. We find that there is a general agreement between the two types of estimators, however $\hat{S}_{X,A}$ produces an estimate of the stickiness of BMI that exceeds 1.

Data Set	$\hat{S}_{X,A}$	$\hat{S}_X^{K(0.85)}$	$\hat{S}_X^{K(0.90)}$	$\hat{S}_X^{K(0.95)}$	$\hat{S}_X^{K(0.99)}$
Female Growth	0.84	0.93	0.87	0.87	0.84
Male Growth	0.79	0.87	0.83	0.80	0.78
BMI	1.06	0.95	0.95	0.95	0.94
SBP	0.76	0.80	0.80	0.75	0.75
GINI	0.89	0.97	0.89	0.89	0.87
GDP Growth	0.19	0.20	0.19	0.18	0.17
Auction Data	0.71	0.72	0.72	0.69	0.68

Table 1: Comparison of various estimators for S_X for the data sets discussed in Section 4 where $\hat{S}_X^{K(p)}$ means that $K(p)$ is selected such that K is the minimum number of eigen-components so that $100p\%$ of the variation was explained

A small simulation study was conducted to compare the proposed estimators and also to verify that both low and high values of S_X can be reliably estimated. In each simulation setting we constructed 100 sparsely observed trajectories where the number of observations from each process was uniformly selected from 2 to 20 measurements. Once the number of

observations was determined, the locations of the measurements were generated uniformly on [0,1]. Additionally each observation was additively perturbed by a random measurement error $\epsilon_{ij} \sim N(0, 0.01)$.

We considered 3 simulation settings. In simulation 1, we used the mean function $\mu(t) = 0$ and a single eigenfunction $\phi(t) = \sqrt{12}(t - \frac{1}{2})$ with eigenvalue $\lambda = 2$. In this setting S_X should be 0. For simulation 2, we used the mean function $\mu(t) = 0$ and a single eigenfunction $\phi(t) = 1$ with eigenvalue $\lambda = 2$. In this setting $S_X = 1$. Finally for simulation 3, we set $\mu(t) = 0$ and used two eigenfunction $\phi_1(t) = \sqrt{2}\sin(\frac{\pi}{2}t)$, $\phi_2(t) = \sqrt{2}\sin(\frac{3\pi}{2}t)$ with eigenvalues $\lambda_1 = 2$, and $\lambda_2 = 1$. In this case S_X= 0.57. For all three simulation settings, each trajectory was generated with FPCs $\xi_{ik} \sim N(0, \lambda_j)$. The simulation results are in Table 2.

Simulation Number	S_X	$\hat{S}_{X,A}$ (SD)	$\hat{S}_X^{K(0.9)}$ (SD)
1	0.0	-0.004 (0.04)	0.01 (0.02)
2	1.0	1.004 (0.04)	0.99 (0.02)
3	0.57	0.59 (0.06)	0.58 (0.06)

Table 2: Simulation study where $\hat{S}_X^{K(p)}$ means that $K(p)$ is selected such that K is the minimum number of eigen-components so that $100p\%$ of the variation was explained.

We find that we are able to reliably estimate small, large and intermediate values of S_X. Estimators $\hat{S}_{X,A}$ and $\hat{S}_X^K$ provide similar results except that $\hat{S}_{X,A}$ can be negative or greater than 1.

References

Ash, R., Gardner, M., 1975. Topics in Stochastic Processes. Academic Press, New York.

Bapna, R., Jank, W., Shmueli, G., 2008. Price formation and its dynamics in online auctions. Decision Support Systems 44, 641–656.

Boularan, J., Ferré, L., Vieu, P., 1994. Growth curves: a two-stage nonparametric approach. Journal of Statistical Planning and Inference 38, 327–350.

Fan, J., Gijbels, I., 1996. Local Polynomial Modelling and its Applications. Chapman & Hall, London.

Gasser, T., Müller, H., Köhler, W., Molinari, L., Prader, A., 1984. Nonparametric regression analysis of growth curves. Annals of Statistics 12, 210–229.

Gervini, D., Gasser, T., 2004. Self-modelling warping functions. Journal of the Royal Statistical Society, Series B 66, 959–971.

Hall, P., Müller, H., Wang, J., 2006. Properties of principal component methods for functional and longitudinal data analysis. Annals of Statistics 34, 1493–1517.

Jank, W., Shmueli, G., 2006. Functional data analysis in electronic commerce research. Statistical Science 21, 155–166.

Jones, H., Bayley, N., 1941. The Berkeley Growth Study. Child Development 12, 167–173.

Jones, M., Rice, J., 1992. Displaying the important features of large collections of similar curves. American Statistician 46, 140–145.

Karatzas, I., Shreve, S., 1991. Brownian Motion and Stochastic Calculus. Springer-Verlag, New York.

Kirkpatrick, M., Heckman, N., 1989. A quantitative genetic model for growth, shape, reaction norms, and other infinite-dimensional characters. Journal of Mathematical Biology 27, 429–450.

Liu, B., Müller, H., 2009. Estimating derivatives for samples of sparsely observed functions, with application to online auction dynamics. Journal of the American Statistical Association 104, 704–717.

Pearson, J., Morrell, C., Brant, L., Landis, P., Fleg, J., 1997. Age-associated changes in blood pressure in a longitudinal study of healthy men and women. The Journals of Gerontology, Series A 52, M177.

Ramsay, J., Silverman, B., 2005. Functional Data Analysis. Springer Series in Statistics, Springer-Verlag, New York. 2 edition.

Rao, C., 1958. Some statistical methods for comparison of growth curves. Biometrics 14, 1–17.

Reddy, S., Dass, M., 2006. Modeling on-line art auction dynamics using functional data analysis. Statistical Science 21, 179–193.

Rice, J., Wu, C., 2001. Nonparametric mixed effects models for unequally sampled noisy curves. Biometrics 57, 253–259.

Shock, N., Greulich, R., Andres, R., Arenberg, D., Costa Jr, P., Lakatta, E., Tobin, J., 1984. Normal Human Aging: The Baltimore Longitudinal Study of Aging. National Institutes of Health .

Staniswalis, J., Lee, J., 1998. Nonparametric regression analysis of longitudinal data. Journal of the American Statistical Association 93, 1403–1404.

Tanner, J., Whitehouse, R., Takaishi, M., 1966. Standards from birth to maturity for height, weight, height velocity, and weight velocity: British children, 1965. Archives of Disease in Childhood 41, 454–471.

Wang, S., Jank, W., Shmueli, G., 2008. Explaining and forecasting online auction prices and their dynamics using functional data analysis. Journal of Business and Economic Statistics 26, 144–160.

Yao, F., Müller, H., Wang, J., 2005a. Functional data analysis for sparse longitudinal data. Journal of the American Statistical Association 100, 577–590.

Yao, F., Müller, H., Wang, J., 2005b. Functional linear regression analysis for longitudinal data. Annals of Statistics 33, 2873–2903.

Zhao, X., Marron, J., Wells, M., 2004. The functional data analysis view of longitudinal data. Statistica Sinica 14, 789–808.

Zhou, S., Wolfe, D., 2000. On derivative estimation in spline regression. Statistica Sinica 10, 93–108.