A Stickiness Coefficient for Longitudinal Data

Revised Version
November 2011

Andrea Gottlieb
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: 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

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 an underlying smooth stochastic process, the stickiness coefficient summarizes the relationship between the value of the process at one time with the value it assumes at another time via a single measure. In particular, the stickiness coefficient quantifies the extent to which deviations from the mean trajectory tend to co-vary over time. The proposed estimation scheme is shown to be consistent even in situations where longitudinal data are sparsely observed at irregular times and may be corrupted by noise. We also demonstrate asymptotic convergence rates for the proposed estimates of the stickiness coefficient and illustrate stickiness coefficients with some theoretical calculations, as well as several economic and health related data examples.

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 reflecting all the information contained in a data set, they allow for summarization and provide valuable initial insight, as well as a simple method for comparing and contrasting multiple data sets. In regression and correlation analysis, the coefficients of determination and correlation provide simple and interpretable summary measures for the relationship between two sets of random variables, which is especially suited to compare different such relationships.

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 various longitudinal data. The proposed stickiness coefficient is designed to capture a property that is intrinsic to longitudinally observed stochastic processes and its estimation could be one of the initial steps of data exploration for longitudinal data.

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

![Growth curves from the Berkeley Growth Study and smooth estimate of the mean function (dashed). Left: Growth curves of 10 randomly selected girls. Right: Growth curves of 10 randomly selected boys.](image)
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 general 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 cross-sectional mean trajectory is denoted by the thick dashed curve and was estimated from all participants in the study. If one follows 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 almost 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 and adolescence. Likewise those who start out shorter than average tend to remain below average height. This signifies ‘stickiness’ in the growth process. Subjects tend to be ‘stuck’ as either below or above the average height trajectory over time.

This notion of ‘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 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, and this stickiness also extends to growth rates of GDP, albeit with opposite signs. 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 non-sticky. In fact, certain processes tend to self-stabilize over time. Recently, Müller and Yao (2010) 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 can be characterized as reinforcing deviations from the mean trajectory over time, while others do not, is a primary motivation for quantifying stickiness of longitudinal data. Accordingly, we propose a coefficient of stickiness 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 very simple and is designed to distinguish processes 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 also be useful in medical studies to help identify health characteristics that are more or less amenable to intervention. Health characteristics with high stickiness coefficients would be viewed as being resistant to change, and those with lower coefficients as more malleable.

As we are interested in the estimation of this coefficient for a general class of longitudinal data, it is necessary to avoid the unduly restrictive assumption that the entire time course of the underlying stochastic process is observable. Doing so allows to calculate stickiness for sparse, irregularly sampled, noise-contaminated longitudinal measurements, which are common in longitudinal studies of health, social and psychological development and also in many economic studies, for example on longitudinal income inequality. A basic assumption is that the data are generated by an underlying smooth stochastic process, which is square integrable and might be observed intermittently. Functional data analysis for such sparsely observed processes has become a much debated topic (James et al. 2000; Zhao et al. 2004; Hall et al. 2006; Yang et al. 2011).

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 \( T \) with length \( |T| \) and \( N_i \) observations of \( X_i \), where \( N_i \) is a realization of a bounded integer-valued 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

\[
Y_{ij} = Y_i(T_{ij}) = X_i(T_{ij}) + \varepsilon_{ij} \quad T_{ij} \in T.
\]
Here, the $\varepsilon_{ij}$ are zero mean i.i.d. measurement errors, with $\text{var}(\varepsilon_{ij}) = \sigma^2$, independent of all other random components. The stickiness coefficient will be defined for data 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 studied in Section 5, followed by concluding remarks in Section 6. Additional details on estimation procedures, auxiliary results, proofs and additional simulation results can be found in the Appendix.

2 Stickiness Coefficient

2.1 Population Definition of the Stickiness Coefficient

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

$$S_X = \frac{E\left[\{X(T_1) - \mu(T_1)\}\{X(T_2) - \mu(T_2)\}\right]}{\left[\text{Var}\{X(T_1) - \mu(T_1)\}\right]^{1/2}\left[\text{Var}\{X(T_2) - \mu(T_2)\}\right]^{1/2}},$$

(2)

where $\mu(t) = E[X(t)]$ is the mean trajectory and $T_1$ and $T_2$ are two random times that are independently and 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 and only one realization of the time series is observed.

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.

### 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 $T$ with mean function $EX(t) = \mu(t)$ and auto-covariance function $\text{cov}(X(s), X(t)) = G(s, t)$, $s, t \in T$. Here $G(s, t)$ is a smooth, symmetric and non-negative definite surface.

With $G$ as kernel, we define the linear Hilbert-Schmidt operator $(A_G f)(t) = \int_T G(s, t)f(s)ds$.

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 $\phi_k$, 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, $\xi_{ik} = \int_T (X_i(t) - \mu(t))\phi_k(t)dt$, $k = 1, 2, \ldots$, are the functional principal components (FPCs) of the random trajectories $X_i$, for $k = 1, 2, \ldots$, which are uncorrelated random variables with $E(\xi_{ik}) = 0$ and $E\xi_{ik}^2 = \lambda_k$, with $\sum_k \lambda_k < \infty$ (Ash and Gardner 1975).

The eigenfunctions $\phi_k$ 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 shows 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) that

$$S_X = \frac{E[E\{X(T_1) - \mu(T_1)\}\{X(T_2) - \mu(T_2)\}|T_1, T_2]}{E\{\text{Var}(X(T_1) - \mu(T_1))|T\} + \text{Var}\{E(X(T_1) - \mu(T_1))|T\}}$$

$$= \frac{E(\sum_{k=1}^{\infty} \xi_k\phi_k(T_1)\sum_{j=1}^{\infty} \xi_j\phi_j(T_2))}{E\{\text{Var}(X(T_1) - \mu(T_1))|T_1\}}$$
\[ = \sum_{k=1}^{\infty} \{ \lambda_k \int_T \phi_k(t_1) f_{T_1}(t_1) dt_1 \int_T \phi_k(t_2) f_{T_2}(t_2) dt_2 \} \]

whence

\[ S_X = \frac{1}{|T|} \sum_{k=1}^{\infty} \lambda_k \left[ \int_T \phi_k(t) dt \right]^2 \sum_{k=1}^{\infty} \lambda_k, \quad (3) \]

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 term of the smooth covariance operator \( G(s,t) \),

\[ S_X = \frac{1}{|T|} \int_{T \times T} G(s,t) dt ds \int_T G(t,t) dt. \quad (4) \]

3 Theoretical Examples and Estimation

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

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

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, corresponding to the zero line, it is unlikely for the process to cross the line 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 will therefore characterize a process as particularly sticky if its stickiness coefficient exceeds \( \frac{2}{3} \).
3.2 Case of One Eigenfunction

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

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

Thus the magnitude of $S_X$ is entirely determined by the shape of $\phi$. For simplicity, we restrict our attention to centered processes, $\mu(t) = 0$, on $T = [0, 1]$. If, for example, $\phi(t) = \sqrt{12} (t - \frac{1}{2})$, 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 $S_X = 1$. For it holds that $S_X = 1$ if, and only if, $G(s,t)$ is a constant function: Without loss of generality, suppose $T = [0,1]$. A straight-forward calculation shows that if $G$ is constant, then indeed $S_X = 1$. Conversely, suppose $S_X = 1$. Then from (3),

$$\sum_{k=1}^{\infty} \lambda_k \left[ \int_T \phi_k(t) dt \right]^2 = \sum_{k=1}^{\infty} \lambda_k. \quad (7)$$
Since \( \int \phi_k(t)dt \leq \int \phi^2_k(t)dt = 1 \), it follows that \( \left[ \int_T \phi_k(t)dt \right]^2 \leq 1 \), for all \( k \). Thus (7) can only hold when \( \left[ \int \phi_k(t)dt \right]^2 = 1 \) for all \( k \). Now, the two statements \( \left[ \int \phi_k(t)dt \right]^2 = 1, \int \phi^2_k(t)dt = 1 \) can only hold simultaneously if \( \phi_k(t) \equiv 1 \). This means there can only be one eigenfunction, \( \phi(t) \equiv 1 \), and thus \( G(s,t) \) is constant.

### 3.3 Estimation Procedures

Equation (3) suggests a natural method for the estimation of \( S_X \) by using plug-in estimators to estimate the eigenvalues \( \lambda_k \) 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, \( K(n) \rightarrow \infty \) as \( n \rightarrow \infty \) is required. The resulting estimator is

\[
\hat{S}_X^K = \frac{1}{|T|} \frac{\sum_{k=1}^K \hat{\lambda}_k \left[ \int_T \hat{\phi}_k(t)dt \right]^2}{\sum_{k=1}^K \hat{\lambda}_k}.
\]  

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 case of longitudinal measurements. In contrast, the proposed estimation procedure in (8) allows for the estimation of these components even in the case that the longitudinal observations are sparse, irregular and noise corrupted. This situation is handled by borrowing strength from the entire sample. Details can be found in (Yao et al. 2005a,b); 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

\[
\hat{S}_{X,A} = \frac{1}{|T|} \frac{\int_{T \times T} \hat{G}(s,t)dt ds}{\int_T \hat{G}(t,t)dt}.
\]  

This method does not require estimating the eigen-components and instead 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 produce estimates that are negative or greater than 1. For further details, we refer to Appendix A.4.
4 Stickiness in Action for Longitudinal Data

We present several data examples from health and economics to demonstrate the wide variety of design schemes that can arise in longitudinal studies. Our analyses include the point estimates $\hat{S}_X$ of the stickiness coefficients based on eq. (8), where the number of included components $K$ was chosen such that the fraction of variance explained is at least 0.9 (for more details on this selection criterion, see Yao et al. (2005a)). Confidence intervals for the stickiness coefficients were constructed via bootstrap, by resampling from the sample of observed time courses, and obtaining empirical quantiles from the resulting bootstrap distribution of estimates $\hat{S}_X$.

4.1 Berkeley Growth Data

The Berkeley growth data contain 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, two 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 the intuition about the stickiness of human growth based on Figure 1. Bootstrap confidence intervals at the 95% level are given by [0.81, 0.95] and [0.79, 0.90], 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 mm Hg) 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 missed visits. As a result, both number of observations per subject and observation times vary widely (see Yao et al.
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 ages 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 two eigenfunctions were used to summarize SBP, accounting for 93% of the variability. We then obtain coefficient estimates of 0.95 and 0.80 and 95% bootstrap confidence intervals $[0.90, 1.0]$ and $[0.68, 0.87]$ 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.

Since the BMI and SBP measurements are observed on the same collection of men, we can easily obtain a 95% bootstrap confidence interval for the difference between their respective stickiness coefficients. This interval is found to be $[0.06, 0.29]$, providing strong evidence that the BMI trajectory is indeed stickier than the SBP trajectory. This confirms our intuition that on average, body weight is a particularly sticky process over the entire lifetime and requires a great deal of sustained effort to change. In comparison, SBP is also relatively sticky, but much less so than BMI.

### 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 that was repeatedly observed. 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. Our 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 \([0.58, 0.78]\). So the degree of stickiness in this bidding process is not far from that of a Brownian motion.

### 4.4 Stickiness of Economic Indices

In this section we analyze data obtained from the World Bank (http://data.worldbank.org/). We begin by studying 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 truly sparse and irregular data set, as many countries have only a handful of measurements over this time period.

In order to obtain meaningful results, we only considered countries with at least two measurements in the time window \([1979, 2009]\). This led to a sample size of 93 countries entering the analysis. The underlying process is well described by two eigenfunctions that account for 97% of the variability. Not unexpectedly, the point estimate for the stickiness coefficient of this processes was found to be quite high at 0.89, with 95% bootstrap confidence interval \([0.53, 0.97]\).

Next we discuss Gross Domestic Products (GDP), focusing on GDP growth rates. The
data consist of the annual percentage growth rates of GDP at market prices, based on the local currency for 203 countries over the last 50 years. We might expect growth in GDP 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 [0.14, 0.29]. This indicates that this process is unsticky.

5 Asymptotic Properties

In this section we study rates of convergence and establish asymptotic consistency for \( \hat{S}_X \), the estimator of \( S_X \). These results require a collection of regularity conditions regarding 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 \( h_\mu \) 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 \( N_i \) \text{i.i.d.} \( N \), where \( N \) is a bounded positive discrete random variable with \( P\{N \geq 2\} > 0 \), and \( \{T_{ij}, j \in J_i\}, \{Y_{ij}, j \in J_i\} \) are independent of \( N_i \), for \( J_i \subseteq \{1, \ldots, N_i\} \).

Writing \( T_i = (T_{i1}, \ldots, T_{iN_i})^T \) and \( Y_i = (Y_{i1}, \ldots, Y_{iN_i})^T \), the triples \( \{T_i, Y_i, N_i\} \) are assumed to be i.i.d. 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 \( \phi_k \) and eigenvalues \( \lambda_k \) as their order \( k \) increases. For this we define,

\[
\alpha(K) = \sum_{k=K+1}^{\infty} \lambda_k, \quad \delta_1 = \lambda_1 - \lambda_2, \quad \delta_k = \min_{j \leq k}(\lambda_{j-1} - \lambda_j, \lambda_j - \lambda_{j+1}), \quad k \geq 2.
\]

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),
\[
\left| \hat{S}_K X - S X \right| = O(\alpha(K)) + \left( \sum_{k=1}^{K} \delta_k^{-1} + K \right) O_p \left( \frac{1}{\sqrt{n h_G}} + h_G^2 \right).
\]

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

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.

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 \( 0 < \alpha < \frac{1}{4} \). Then the number of included eigen-components \( K(n) = -\frac{1}{2\rho} \log \left( n^{\alpha - \frac{1}{2}} + n^{-2\alpha} \right) \) minimizes the leading terms of the right hand side of (11).

6 Discussion

The purpose of the proposed stickiness coefficient is to provide the data analyst with a summary statistic for longitudinal data, which is also useful for initial data exploration. 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. It proves useful for applications that the entire underlying process need not be observed for the estimation of the coefficient. Rather, the observations can be very sparse, in fact, two observations of the process at random locations suffice (Hall et al. 2006; Yao et al. 2005b). The reason for this surprising result is that one gains strength for statistical inference by pooling all the data together for estimation of covariances as well as eigenvalues and eigenfunctions.
While human growth data provide a motivating example, the stickiness coefficient is applicable to any field where trajectory data are collected. In economics, the stickiness coefficient is of interest to discriminate between sticky phenomena such as income inequality and those which are non-sticky such as GDP growth rate. In health and human development, human growth is found to be a sticky process and body mass index is very sticky. In contrast, systolic blood pressure turns out to be considerably less sticky. In such applications, the stickiness coefficient reflects crucial information about the structure of the underlying process that generates the data.

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 are encountered 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 obtain 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 $\kappa_1$ and bandwidth $h_\mu$, one would minimize

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

for each $t$ with respect to $\alpha_0$ and $\alpha_1$ from which one obtains $\hat{\mu}(t) = \hat{\alpha}_0$ (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 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

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

(13)
for a fixed \((s, t)\) with respect to \(\alpha_0, \alpha_1,\) and \(\alpha_2\). 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, \(\text{cov}(Y_{ij}, Y_{ij}|T_{ij}) = \text{cov}(X(T_{ij}, X(T_{ij})) + \sigma^2\). 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 \(\sigma\) 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.

To obtain estimates of \(\lambda_k\) and \(\phi_k(t)\), one numerically solves the eigen-equations

\[
\int_T \hat{G}(s, t) \hat{\phi}_k(s) ds = \hat{\lambda}_k \hat{\phi}_k(t),
\]

where the \(\hat{\phi}_k(t)\) are subject to \(\int_T \hat{\phi}_k(t)^2 dt = 1\) and \(\int_T \hat{\phi}_k(t) \hat{\phi}_m(t) dt = 0\) for \(m \neq k\). For the theoretical analysis we require that \(K = K(n) \to \infty\) as \(n \to \infty\). In 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 type criteria based or selecting a fraction of variance explained by the included components (Yao et al. 2005b).

A.2 Additional Assumptions and Auxiliary Results

In addition to assumptions (A1)-(A2) in Section 5, requirements for the kernels used in the local linear smoothing steps and for the 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 \(T = [a, b]\), within \(T_\delta = [a - \delta, b + \delta]\) for some \(\delta > 0\), which makes it possible to ignore boundary effects. For non-negative integers \(\ell_1, \ell_2\), the required additional assumptions are as follows:

(B1) \(f^{(4)}_1(t)\) exists and is continuous on \(T_\delta\) 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 \(T_\delta^2\) for \(\ell_1 + \ell_2 = 4\).

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

(B3) The kernel \(\kappa_1\) is a symmetric, continuous, non-negative density with compact support.

The kernel \(\kappa_2\) 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 the $L_2$ convergence rate for the eigenfunction estimates $\hat{\phi}_k$.

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

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

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). An important feature is that the constant in the $L_2$ convergence for the eigenfunction estimates $\hat{\phi}_k$ is uniform across $k$.

### 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 apply the inequalities $\int_T \phi_k(t) dt^2 \leq |T|$ and $\int_T \hat{\phi}_k(t) dt^2 \leq |T|$, where the first inequality is implied by the Cauchy-Schwarz 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 $\int_T \hat{\phi}_k^2(t) dt = 1$. Noting that

$$\Delta := |T| \sum_{k=1}^{\infty} \lambda_k \left| \hat{S}_X^K - S_X \right|$$

we find by Lemma 1,

$$I \leq \frac{|T|} {\sum_{k=1}^{K} \hat{\lambda}_k} \sum_{k=1}^{\infty} \lambda_k - \sum_{k=1}^{K} \hat{\lambda}_k \leq |T| \left( \sum_{k=1}^{K} \hat{\lambda}_k - \lambda_k \right) + \alpha(K) = KO_p \left( \frac{1}{\sqrt{nh_G}} + h^2 G \right) + |T| \alpha(K)$$

and, again by Lemma 1,

$$II \leq \sum_{k=1}^{K} \lambda_k \left( \left[ \int_T \hat{\phi}_k(t) dt \right]^2 - \left[ \int_T \phi_k(t) dt \right]^2 \right) + \sum_{k=1}^{K} \left[ \int_T \hat{\phi}_k(t) dt \right]^2 (\lambda_k - \hat{\lambda}_k) - \sum_{k=K+1}^{\infty} \lambda_k \left[ \int_T \phi_k(t) dt \right]^2$$

$$\leq \sum_{k=1}^{K} \lambda_k O_p \left( \frac{1}{\lambda_k \delta_k} \left( \frac{1}{\sqrt{nh_G}} + h^2 G \right) \right) + KO_p \left( \frac{1}{\sqrt{nh_G}} + h^2 G \right) + |T| \alpha(K).$$
Therefore, $\Delta \leq I + II = O(\alpha(K)) + \left(\sum_{k=1}^{K} \delta_k^{-1} + K\right) O_p\left(\frac{1}{\sqrt{nh}} + h^2\right)$, completing the proof.

### A.4 Comparisons of Estimation Methods $\hat{S}_{X,A}$ and $\hat{S}_X^K$

The data analysis in Section 4 was completed using both estimators $\hat{S}_X^K$, given in equation (8), and $\hat{S}_{X,A}$, given 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 the stickiness coefficient when using various estimators, including those for $\hat{S}_X^K = \hat{S}_X^{K(p)}$ for different $K = K(p)$, where $K(p)$ is selected as the minimum number of eigen-components such that 100$p\%$ of the variation is explained. We find that there is good general agreement between the two types of estimators, however $\hat{S}_{X,A}$, due to its unconstrained nature (see discussion after eq. (9)), produces an estimate of the stickiness of BMI that exceeds 1.

<table>
<thead>
<tr>
<th>Data Set</th>
<th>$\hat{S}_{X,A}$</th>
<th>$\hat{S}_X^{0.85}$</th>
<th>$\hat{S}_X^{0.90}$</th>
<th>$\hat{S}_X^{0.95}$</th>
<th>$\hat{S}_X^{0.99}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Female Growth</td>
<td>0.84</td>
<td>0.93</td>
<td>0.87</td>
<td>0.87</td>
<td>0.84</td>
</tr>
<tr>
<td>Male Growth</td>
<td>0.79</td>
<td>0.87</td>
<td>0.83</td>
<td>0.80</td>
<td>0.78</td>
</tr>
<tr>
<td>BMI</td>
<td>1.06</td>
<td>0.95</td>
<td>0.95</td>
<td>0.95</td>
<td>0.94</td>
</tr>
<tr>
<td>SBP</td>
<td>0.76</td>
<td>0.80</td>
<td>0.80</td>
<td>0.75</td>
<td>0.75</td>
</tr>
<tr>
<td>GINI</td>
<td>0.89</td>
<td>0.97</td>
<td>0.89</td>
<td>0.89</td>
<td>0.87</td>
</tr>
<tr>
<td>GDP Growth</td>
<td>0.19</td>
<td>0.20</td>
<td>0.19</td>
<td>0.18</td>
<td>0.17</td>
</tr>
<tr>
<td>Auction Data</td>
<td>0.71</td>
<td>0.72</td>
<td>0.72</td>
<td>0.69</td>
<td>0.68</td>
</tr>
</tbody>
</table>

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

Another small simulation study was carried out to compare the proposed estimators and to verify that both low and high values of $S_X$ can be reliably estimated. In each simulation 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 $[0,1]$ and observations were additively perturbed by a random measurement error $\epsilon_{ij} \sim N(0,0.01)$. 

18
We considered three simulation settings. In simulation 1, we choose the mean function \( \mu(t) = 0 \) and a single eigenfunction \( \phi(t) = \sqrt{12}(t - \frac{1}{2}) \) with eigenvalue \( \lambda = 2 \). Here \( S_X = 0 \), see the discussion in Section 3.2. For simulation 2, we use the mean function \( \mu(t) = 0 \) and a single eigenfunction \( \phi(t) = 1 \) with eigenvalue \( \lambda = 2 \). In this setting \( S_X = 1 \). For simulation 3, we select \( \mu(t) = 0 \) and 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 \( S_X = 0.57 \). Trajectories were generated with functional principal components \( \xi_{ik} \sim N(0, \lambda_k) \). The results are in Table 2.

<table>
<thead>
<tr>
<th>Simulation</th>
<th>( S_X )</th>
<th>( \hat{S}_{X,A} ) (SD)</th>
<th>( \hat{S}^{K(0.9)}_X ) (SD)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.0</td>
<td>-0.004 (0.04)</td>
<td>0.01 (0.02)</td>
</tr>
<tr>
<td>2</td>
<td>1.0</td>
<td>1.004 (0.04)</td>
<td>0.99 (0.02)</td>
</tr>
<tr>
<td>3</td>
<td>0.57</td>
<td>0.59 (0.06)</td>
<td>0.58 (0.06)</td>
</tr>
</tbody>
</table>

Table 2: Simulation study where \( \hat{S}^{K(p)}_X \) means that \( K(p) \) is selected such that \( K \) is the minimum number of eigen-components so that 100\( p \)% 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 \( \hat{S}^{K}_X \) provide similar results except that \( \hat{S}_{X,A} \) can be negative or greater than 1.

**Acknowledgments**

We are grateful to five referees and an Associate Editor, whose comments led to substantial improvements. This research was supported in part by NSF grant DMS-1104426.

**References**


