Rank Dynamics for Functional Data[☆]

Yaqing Chen[1,], Matthew Dawson[1], Hans-Georg Müller[1]

[a]*Department of Statistics, University of California, Davis*
[b]*Graduate Group in Biostatistics, University of California, Davis*

Abstract

We study the dynamic behavior of cross-sectional ranks over time for functional data and show that the ranks of the observed curves at each time point and their evolution over time can yield valuable insights into the time dynamics of functional data. This approach is of interest in various application areas. To analyze the dynamics of ranks, we obtain estimates of the cross-sectional ranks of functional data and discuss several statistics of interest in ranked functional data. To quantify the evolution of ranks over time, we develop a model for rank derivatives, in which we decompose rank dynamics into two components. One component corresponds to population changes and the other to individual changes that both affect the rank trajectories of individuals. We establish the joint asymptotic normality for suitable estimates of these two components. These approaches are illustrated with simulations and three longitudinal data sets: Growth curves obtained from the Zürich Longitudinal Growth Study, monthly house price data in the US from 1980 to 2015 and Major League Baseball offensive data for the 2017 season.

Keywords: Decomposition of rank derivatives; Functional data analysis; House price dynamics; Major League Baseball; Zürich Longitudinal Growth Study.

1. Introduction

In many statistical applications, practitioners are interested in relative, as opposed to absolute, behavior of random quantities. For example, in growth studies, one is often interested in growth faltering, stunting and more generally determining whether children are tall, normal or small for their age. Such determinations are based on an assessment of how individuals rank relative to others, where an individual's rank will change as the individual ages. In sports, many interested parties aim to track the longitudinal changes in the relative rankings of the best players and teams. For example, the compensation a player receives is tied to relative performance. Related studies have been done on regression models for conditional distribution functions

[☆]Declarations of interest: none.
[*]Corresponding author

and quantiles [? ? ? ? ? ? , for example], while our focus here is on modeling the temporal evolution of longitudinal ranks.

In the case of univariate measurements, ranking data is straightforward and well-studied. However, one cannot rank multivariate data because there is no total ordering in $\mathbb{R}^p$. For the same reason, functional data that correspond to infinite-dimensional objects similarly cannot be ordered [for overviews, see, e.g., ? ? ?]. In related work, the analysis of sports data with functional data analysis techniques has been recently considered in ?], archetypoids of functional trajectories were applied to sports statistics in ?], and ?] studied epigraph and hypograph indices which are the proportions of sample trajectories entirely lying above or below certain curves.

While functional data cannot be ordered, they are time-indexed and a total ordering exists cross-sectionally at each fixed time. This can be utilized to transform functional data into trajectories that consist of ranks, viewed as functions of time. Of interest then is the modeling of the ranks of individuals and their patterns over time. In this paper, we discuss statistical tools to study such rank dynamics. In particular, we introduce a novel decomposition for rank dynamics, where we show that rank derivatives can be naturally decomposed into two components, corresponding to a population and an individual contribution to the rank evolution, respectively. A simple example for the effect of the population on individual ranks occurs when the scores of the population improve overall, but a particular individual stays the same, say a runner maintains a certain level of speed but the population of runners at large is getting faster — then the individual runner's rank will drop within the population, even though the individual's performance is not worse than before.

As rank dynamics depend on the interplay between individual and population changes and make reference to the cross-sectional population at each time t where functional values are obtained, rank dynamics is quite different from common dynamic models in functional data analysis, where only the time dynamics of individuals viewed by themselves are the focus, with the associated notions of derivatives of observed trajectories and empirical dynamics. These previous approaches could be characterized as dynamics learning from functional data, and include derivative principal components, identification of differential equations, and dynamic regression modeling [? ? ? ? ? ? ? ?].

More specifically, to study rank dynamics one first transforms the observed functional data through a probability transform that is implemented at each time point. We assume that the functional data are densely sampled with negligible noise and that there is a stochastic process Y with square integrable trajectories which are in the Hilbert space L^2. The process Y generates the sample of trajectories, which are the observed functional data. If the functional data are measured on a time grid with additive noise, one can implement a pre-smoothing step [? ?].

Our starting point is the cross-sectional distribution

$$P(Y(t) \leq y) = F_t(y), \tag{1}$$

for each $t \in \mathcal{T}$, where the domain $\mathcal{T}$ is a compact interval. Without loss of generality, we consider $\mathcal{T} = [0, 1]$. The process of local probability transforms $R(t)$ associated with Y is then

$$R(t) = F_t(Y(t)), \ t \in \mathcal{T}.$$

Since the subject-specific random process $R(t)$ conveys the information which fraction of individuals has larger and which fraction has lower values at time t compared to a selected individual, we refer to $R(t)$ as the *rank process* associated with the functional process Y.

We note that the range of the rank process is always the interval $[0, 1]$ and multiplying it by the sample size n gives the actual ranks. Indeed, the distribution of $R(t)$ is uniform on $[0, 1]$ for every $t \in \mathcal{T}$, as it corresponds to the local probability transform. In a finite sample situation there are various ways to carry out the probability transform from a sample of data $Y(t)$, depending on how one estimates the cumulative distribution function F. If one uses the empirical distribution function one obtains the actual ranks, but one can also use smooth versions of empirical distribution functions, which often are advantageous [**?**] and yield approximate ranks.

The paper is organized as follows. In Section **??**, we introduce a time-dynamic model for ranked functional data to quantify the temporal evolution of rank processes, which is a key contribution of this paper. In Section **??**, we discuss several measures for the central tendency and variation of the rank trajectories, and in Section **??** the estimation of these population quantities. Asymptotic distributions and finite-sample performance of the proposed estimates are demonstrated in Sections **??** and **??**, respectively. Data illustrations are provided in Section **??**, where we demonstrate rank dynamics for three scenarios including Zürich growth curves, house price trajectories and Major League Baseball data.

2. A Time-Dynamic Model for Ranked Functional Data

Increases or decreases in an individual's rank trajectory depend on both the subject's functional trajectory $Y(t)$ and the functional trajectories of all other individuals in the sample, as the subject's rank at time t depends on these two inputs. This decomposition is exemplified by the *keeping up with the Joneses* paradigm, where subjects' happiness is assessed through an individual's relative standing and its changes, compared to their peers, i.e. critically important are the subject's rank and especially the changes in rank [e.g., **? ?**].

To quantify relative changes in a sample of functional data, it is expedient to utilize derivatives $R'(t)$. Recalling that $F_t(y)$ is the cross-sectional distribution of Y at time t and $R(t) = F_t(Y(t))$ and taking the derivative of R with respect to t leads to

$$\begin{aligned} R'(t) &= C_1(t) + C_2(t) \\ &:= D_1(Y(t), t) + D_2(Y(t), t)Y'(t), \end{aligned} \tag{2}$$

where

$$D_1(y,t) \coloneqq \frac{\partial F_t(y)}{\partial t} \quad \text{and} \quad D_2(y,t) \coloneqq \frac{\partial F_t(y)}{\partial y} = f_t(y). \tag{3}$$

The two terms in (**??**) provide the decomposition of the rank derivative into two components for each subject. The first component $C_1(t)$ reflects the changes in the distribution of the original process Y with respect to time. More specifically, $C_1(t)$ indicates how population changes influence the rank of a given subject, where positive (negative) values of $C_1(t)$ for a specific subject mean that the underlying functional trajectories $Y(t)$ for the other subjects are generally decreasing (increasing) at time t, which leads to an increase (decrease) in rank for the selected subject that is entirely due to a change in the characteristics of the general population. On the other hand, the second component $C_2(t)$ represents the subject's own contribution to the rank dynamics. Since $D_2(y,t) = f_t(y) \geq 0$, positive (negative) values of $Y'(t)$ contribute to an increase (decrease) in rank due to individual change. Note that even if a subject's underlying functional trajectory $Y(t)$ is increasing, the population change $C_1(t)$ could increase even faster and potentially overpower a subject's own contribution, leading to a decrease in rank.

To gain a better understanding of the nature of the model in (**??**), it is helpful to consider the case where $Y(t)$ is a constant function. In this case, we have that $C_2(t) = 0$ for all $t \in \mathcal{T}$, and the change in rank is completely determined by the rest of the population, i.e. the rank only changes when the population changes. Similarly, for a subject that traverses on a constant rank trajectory, it holds that $R'(t) = 0$ for all t, which means that population and subject driven components match each other, $C_1(t) = -C_2(t)$ for all t.

To determine the contributions of population and individual effects, it is then of interest to quantify the overall contributions of C_1 and C_2 to the rank derivative. For this, we define the rank component contributions

$$\Lambda_1 \coloneqq \frac{\int_{\mathcal{T}} E(|C_1(t)|)\ \mathrm{d}t}{\int_{\mathcal{T}} E(|C_1(t)|)\ \mathrm{d}t + \int_{\mathcal{T}} E(|C_2(t)|)\ \mathrm{d}t}, \quad \Lambda_2 \coloneqq 1 - \Lambda_1.$$

When Λ_1 is large, changes in rank are primarily dictated by changes in the population trajectories. In contrast, if Λ_2 dominates Λ_1, the changes in rank are due to changes in individual trajectories.

3. Summary Measures for Rank Processes

Suppose we have a sample of trajectories Y_i that are subject-specific independently and identically distributed realizations of a smooth underlying process Y, for $i = 1, \ldots, n$. It is then of interest to have measures that quantify longitudinal central tendency and stability of both subject-specific and population ranks that are functionals of the corresponding rank processes $R_i(t) = F_t(Y_i(t))$ with F_t as per (**??**)

and $i = 1, \ldots, n$. A beneficial feature of the rank process approach is that like other rank-based methods, the analysis does not depend on the scale of the data and allows for direct comparisons of different data sources and measurement scales through comparing the corresponding rank processes.

Subject-specific integrated rank. A natural way to summarize a subject's overall rank is to integrate the subject's rank trajectory over the time domain, i.e. to consider the subject-specific measure

$$\rho_i \coloneqq \int_{\mathcal{T}} R_i(t) \, \mathrm{d}t. \tag{4}$$

Subject-specific rank volatility. It is also of interest to quantify how variable a subject is in terms of rank, which can be quantified by

$$\nu_i \coloneqq \int_{\mathcal{T}} [R_i(t) - \rho_i]^2 \, \mathrm{d}t. \tag{5}$$

Subject-specific rank dynamics. For smooth rank processes, one can define a rank derivative $R'(t)$, $t \in \mathcal{T}$. If it is non-zero, then the subject's rank trajectory crosses the trajectories of other subjects, i.e. the rank of the subject will change over time. Pertinent measures include

$$\zeta_i \coloneqq \int_{\mathcal{T}} R_i'(t) \, \mathrm{d}t = R_i(1) - R_i(0), \quad \text{and} \quad \eta_i \coloneqq \int_{\mathcal{T}} R_i'^2(t) \, \mathrm{d}t, \tag{6}$$

quantifying how variable the rank of a subject is over the time interval.

Population rank stability. Since $E[R(t)] = 1/2$ for all $t \in [0, 1]$, we have that $E[R'(t)] = 0$ under mild assumptions. Although the mean functions are therefore not interesting, the variation of R' on subdomains is of interest, as it can pinpoint temporal regions where ranks tend to change and the intensity of pairwise crossings of the rank trajectories is high. We define time-dependent rank stability as

$$\gamma(t) \coloneqq \mathrm{var}[R'(t)] = E[R'(t)^2]. \tag{7}$$

Integrating this quantity leads to an overall population rank stability coefficient, for which we choose

$$G \coloneqq \exp\left(-\int_{\mathcal{T}} \gamma(t) dt\right). \tag{8}$$

Note that if the underlying functional data never cross paths, then $\gamma(t) = 0$ for all t, and thus the overall rank stability is $G = 1$, while the closer G is to 0, the lower is rank stability, i.e., the trajectories of the functional data exhibit more frequent crossings.

4. Estimation

The starting point is to estimate the rank trajectories $R_i(t)$. Suppose for all subjects, processes Y_i are observed on a regular dense grid $t_{i1} < \cdots < t_{im_i}$ on the time domain, i.e., there exists a design distribution function $\theta : \mathcal{T} \to [0,1]$ such that $t_{ij} = \theta^{-1}((j-1)/(m_i-1))$ for $j = 1, \ldots, m_i$ and $Y_{ij} = Y_i(t_{ij})$. We assume that the underlying surface $F_t(y) = P(Y(t) \leq y)$ is differentiable in both y and t. To obtain smooth estimates of the rank process, we utilize a kernel function K, which is a pdf, and an integrated kernel H, which is a cdf. Furthermore, we assume:

(A1) With probability 1, the process Y has continuously differentiable sample paths and there exists a constant $M > 0$ such that $\sup_{t\in\mathcal{T}} |Y'(t)| \leq M$.

(A2) The kernel K is a symmetric pdf on $\mathbb{R}$ such that,

$$\int x^l K(x)\, \mathrm{d}x < \infty, \text{ for } l = 2, 4.$$

The kernel H is a cdf such that its derivative $H'(\cdot)$ exists almost everywhere, is bounded on $\mathbb{R}$ and is a symmetric pdf such that

$$\int x^l H'(x)\, \mathrm{d}x < \infty, \text{ for } l = 2, 4.$$

(A3) The kernel K has a compact support, assumed to be $[-1, 1]$. On $(-1, 1)$, the first and second derivatives K' and K'' exist and are bounded.

(A4) The design distribution function θ is four times continuously differentiable on $[0, 1]$. There exist $0 < a_1 < a_2$ such that $a_1 \leq \theta'(t) \leq a_2$ for all $t \in [0, 1]$.

We provide two strategies for the estimation of $R(t)$ based on the sample $\{t_{ij}, Y_{ij}\}$ as follows.

Cross-sectional empirical distributions. The most straightforward approach to obtain a ranked sample from a dense functional sample is to estimate the empirical distribution at each time point $t \in \mathcal{T}$. Obtaining cross-sectional empirical distributions in this manner is equivalent to taking cross-sectional ranks and scaling them, i.e.,

$$\widehat{R}_i(t) = \frac{1}{n} \sum_{l \neq i} \mathbf{1}_{\{Y_l(t) \leq Y_i(t)\}}. \tag{9}$$

The empirical ranking defined in (**??**) has several benefits. It is very simple to implement, and its interpretation is very clear. However, since we aim to obtain differentiable rank functions that allow us to study the decomposition of rank dynamics into population and individual components, we need smooth estimates of the rank processes.

Smooth rank functions. Smooth estimation of conditional/cross-sectional distribution functions has been well investigated [e.g., **? ? ? ?**]. Define

$$\widetilde{Q}_{1i}(y,t) = \frac{1}{m_i}\sum_{j=1}^{m_i} h_T^{-1} H\left(\frac{y - Y_{ij}}{h_Y}\right) K\left(\frac{t - t_{ij}}{h_T}\right),$$

$$\widetilde{Q}_{2i}(y,t) = \frac{1}{m_i}\sum_{j=1}^{m_i} h_T^{-1} K\left(\frac{t - t_{ij}}{h_T}\right),$$

and for $l = 1, 2$,

$$\overline{Q}_l(y,t) = \frac{1}{n}\sum_{i=1}^{n} \widetilde{Q}_{li}(y,t),$$

where $h_Y, h_T > 0$ are bandwidths. Here, we utilize a kernel estimate of $F_t(y)$ given by well-established methods described in **?**] and **?**],

$$\widetilde{F}_t(y) = \frac{\overline{Q}_1(y,t)}{\overline{Q}_2(y,t)}. \tag{10}$$

Thus, a smooth estimator for the rank process $R_i(t)$ can be obtained by

$$\widetilde{R}_i(t) = \widetilde{F}_t(Y_i(t)). \tag{11}$$

We will discuss the selection of bandwidths h_Y and h_T in the Supplementary Material.

Using one of the two methods described above, we obtain the estimated rank for level Y_{ij} at time t_{ij}, yielding the surface $\{t_{ij}, Y_{ij}, \widehat{R}_i(t_{ij})\}$ or $\{t_{ij}, Y_{ij}, \widetilde{R}_i(t_{ij})\}$, and hence estimate the measures ρ_i, ν_i and ζ_i given in (**??**)–(**??**), respectively, by plugging in either of the two estimators of $R_i(t)$, applying numerical integration. Estimation of the measures η_i, $\gamma(t)$ and G (**??**)–(**??**) requires the estimation of the rank derivatives $R'(t)$, while identifying the components of the time-dynamic model as per (**??**) requires estimation of $D_1(y,t)$, $D_2(y,t)$, and $Y'(t)$.

For estimating $Y'(t)$, one can make use of local polynomial smoothing, or a similar method. To estimate $D_1(y,t)$ and $D_2(y,t)$ defined in (**??**), we take partial derivatives of (**??**), yielding

$$\widetilde{D}_1(y,t) = \frac{\overline{Q}_3(y,t)}{\overline{Q}_2(y,t)} - \frac{\overline{Q}_1(y,t)\overline{Q}_4(y,t)}{\overline{Q}_2(y,t)^2} \quad \text{and} \quad \widetilde{D}_2(y,t) = \frac{\overline{Q}_5(y,t)}{\overline{Q}_2(y,t)}, \tag{12}$$

where

$$\widetilde{Q}_{3i}(y,t) = \frac{1}{m_i}\sum_{j=1}^{m_i} h_T^{-2} H\left(\frac{y - Y_{ij}}{h_Y}\right) K'\left(\frac{t - t_{ij}}{h_T}\right),$$

$$\widetilde{Q}_{4i}(y,t) = \frac{1}{m_i}\sum_{j=1}^{m_i} h_T^{-2} K'\left(\frac{t - t_{ij}}{h_T}\right),$$

$$\widetilde{Q}_{5i}(y,t) = \frac{1}{m_i}\sum_{j=1}^{m_i} h_Y^{-1} h_T^{-1} K\left(\frac{y - Y_{ij}}{h_Y}\right) K\left(\frac{t - t_{ij}}{h_T}\right),$$

and for $l = 3, 4, 5$,

$$\overline{Q}_l(y,t) = \frac{1}{n}\sum_{i=1}^{n} \widetilde{Q}_{li}(y,t),$$

where $h_Y, h_T > 0$ are bandwidths as in $\widetilde{Q}_{1i}$ and $\widetilde{Q}_{2i}$.

For subject i, the estimated components are

$$\widetilde{C}_{1i}(t) = \widetilde{D}_1(Y_i(t),t) \quad \text{and} \quad \widetilde{C}_{2i}(t) = \widetilde{D}_2(Y_i(t),t)\widetilde{Y}_i'(t),$$

where $\widetilde{Y}_i'(t)$ is an estimate of the derivative for example by local polynomial smoothing. From these estimators we obtain the estimated decomposition $\widetilde{R}_i'(t) = \widetilde{C}_{1i}(t) + \widetilde{C}_{2i}(t)$. The component contributions Λ_1 and Λ_2 may be estimated by numerically integrating the estimated components $\widetilde{C}_{1i}(t)$ and $\widetilde{C}_{2i}(t)$,

$$\widetilde{\Lambda}_1 = \frac{\int_{\mathcal{T}} n^{-1}\sum_{i=1}^{n} |\widetilde{C}_{1i}(t)| \, \mathrm{d}t}{\int_{\mathcal{T}} n^{-1}\sum_{i=1}^{n} |\widetilde{C}_{1i}(t)| \, \mathrm{d}t + \int_{\mathcal{T}} n^{-1}\sum_{i=1}^{n} |\widetilde{C}_{2i}(t)| \, \mathrm{d}t}, \quad \widetilde{\Lambda}_2 = 1 - \widetilde{\Lambda}_1.$$

The measures η_i in (**??**) can then be estimated by plugging in $\widetilde{R}_i'(t)$ based on trajectory $Y_i(t)$; estimators for $\gamma(t)$ and G in (**??**) and (**??**) are obtained using the sample mean of $\widetilde{R}_i'(t)^2$.

5. Theoretical Justifications

We demonstrate the asymptotic normality of $\widetilde{F}_t(y)$, the joint asymptotic normality of $[\widetilde{D}_1(y(t),t), \widetilde{D}_2(y(t),t)y'(t)]^\top$, given a curve $y(t)$, and the asymptotic normality of $\widetilde{R}'(t) = \widetilde{D}_1(y(t),t) + \widetilde{D}_2(y(t),t)y'(t)$. We denote convergence in distribution by $\xrightarrow{\mathcal{D}}$, and define

$$\sigma^2(K) = \int x^2 K(x) \, \mathrm{d}x, \quad \sigma^2(H') = \int x^2 H'(x) \, \mathrm{d}x.$$

All proofs and auxiliary results are in the Supplementary Material. Throughout, we use the notations $F_{s,s'}(z,z') = P(Y(s) \le z, Y(s') \le z')$ and $f_{s,s'}(z,z')$ for the joint cdf and pdf of $Y(s)$ and $Y(s')$, and also the notation $\sim$, where $h_n \sim n^\alpha$ indicates $\lim_{n\to\infty} h_n n^{-\alpha} = 1$. We further need to assume:

(A5) The partial derivatives $\frac{\partial^{k+l}}{\partial t^k \partial y^l} F_t(y)$ are bounded over $t \in [0,1]$ and $y \in \mathbb{R}$, for $(k,l) \in \{(3,0),(0,3),(2,1),(1,2)\}$.

(A6) The partial derivatives $\frac{\partial^2}{\partial s \partial s'} F_{s,s'}(z,z')$, $\frac{\partial^2}{\partial z \partial z'} F_{s,s'}(z,z')$ and $\frac{\partial^2}{\partial s \partial z'} F_{s,s'}(z,z')$ are bounded over $s, s' \in [0,1]$ and $z, z' \in \mathbb{R}$.

The following proposition is similar to some results in literature, for example **?**]; we omit the proof. Theorem **??** is our main result. The proof and auxiliary lemmas are in the Supplementary Material.

Proposition 1. *Assume* **??**–**??**, *optimal bandwidth sequences* $h_Y \sim n^{-1/4}$ *and* $h_T \sim n^{-1/4}$, *as* $n, m_i \to \infty$ *with* $\lim_{n\to\infty} \max_{1\leq i\leq n} m_i^{-1} n^{1/2} = 0$. *Then the estimate for* $F_t(y)$ *as defined in* (**??**) *satisfies*

$$\sqrt{n}\left[\widetilde{F}_t(y) - F_t(y)\right] \xrightarrow{\mathcal{D}} \mathcal{N}\left(\beta_{\widetilde{F}}, \sigma^2_{\widetilde{F}}\right),$$

where

$$\beta_{\widetilde{F}} = \frac{1}{2}\sigma^2(H')\frac{\partial}{\partial y} f_t(y) + \frac{1}{2}\sigma^2(K)\left[\frac{\partial^2}{\partial t^2} F_t(y) + 2\frac{\theta''(t)}{\theta'(t)}\frac{\partial}{\partial t} F_t(y)\right],$$
$$\sigma^2_{\widetilde{F}} = F_{t,t}(y,y) - F_t(y)^2.$$

Theorem 1. *Assume* **??**–**??**. *Given a curve* $y(t)$, *the estimates* $\widetilde{C}_1(t) = \widetilde{D}_1(y(t), t)$, $\widetilde{C}_2(t) = \widetilde{D}_2(y(t), t)y'(t)$ *with* $\widetilde{D}_1$ *and* $\widetilde{D}_2$ *defined in* (**??**) *for the two components* $C_1(t) = D_1(y(t), t)$ *and* $C_2(t) = D_2(y(t), t)y'(t)$ *with* D_1 *and* D_2 *as per* (**??**) *are jointly asympotically normal. With bandwidths* $h_Y \sim n^{-1/4}$ *and* $h_T \sim n^{-1/4}$, *as* $n, m_i \to \infty$ *such that* $\lim_{n\to\infty} \max_{1\leq i\leq n} m_i^{-1} n^{3/4} = 0$,

$$\sqrt{n}\left[\begin{pmatrix}\widetilde{C}_1(t)\\ \widetilde{C}_2(t)\end{pmatrix} - \begin{pmatrix}C_1(t)\\ C_2(t)\end{pmatrix}\right] \xrightarrow{\mathcal{D}} \mathcal{N}\left(\boldsymbol{\beta}_{\widetilde{C}}, \boldsymbol{\Sigma}_{\widetilde{C}}\right),$$

where

$$\boldsymbol{\beta}_{\widetilde{C}} = \begin{pmatrix}\frac{1}{2}\sigma^2(H')\frac{\partial^2}{\partial t \partial y} f_t(y(t)) + \frac{1}{2}\sigma^2(K)\left[\frac{\partial^3}{\partial t^3} F_t(y(t)) + 2\frac{\partial}{\partial t}\frac{\theta''(t)\frac{\partial}{\partial t}F_t(y(t))}{\theta'(t)}\right]\\ \frac{1}{2}\sigma^2(K)y'(t)\left[\frac{\partial^2}{\partial y^2} f_t(y(t)) + \frac{\partial^2}{\partial t^2} f_t(y(t)) + 2\frac{\theta''(t)\frac{\partial}{\partial t} f_t(y(t))}{\theta'(t)}\right]\end{pmatrix},$$

and

$$\boldsymbol{\Sigma}_{\widetilde{C}} = \begin{pmatrix}\Sigma_{11} & \Sigma_{12}\\ \Sigma_{12} & \Sigma_{22}\end{pmatrix},$$

with

$$\Sigma_{11} = \frac{\partial^2}{\partial s \partial s'} F_{t,t}(y(t), y(t)) - \left[\frac{\partial}{\partial t} F_t(y(t))\right]^2,$$
$$\Sigma_{12} = y'(t)\left[\frac{\partial^2}{\partial s \partial z'} F_{t,t}(y(t), y(t)) - f_t(y(t))\frac{\partial}{\partial t} F_t(y(t))\right],$$
$$\Sigma_{22} = y'(t)^2\left[f_{t,t}(y(t), y(t)) - f_t(y(t))^2\right].$$

By continuous mapping, the asymptotic normality of $\widetilde{R}'(t) = \widetilde{C}_1(t) + \widetilde{C}_2(t)$ follows.

Corollary 1. *Under the assumptions of Theorem* **??**, *with* $R'(t) = C_1(t) + C_2(t)$,

$$\sqrt{n}\left[\widetilde{R}'(t) - R'(t)\right] \xrightarrow{\mathcal{D}} \mathcal{N}\left(\beta(t), \sigma^2(t)\right),$$

where

$$\beta(t) = \frac{1}{2}\sigma^2(H')\frac{\partial^2}{\partial t \partial y} f_t(y(t)) + \frac{1}{2}\sigma^2(K)\left[\frac{\partial^3}{\partial t^3}F_t(y(t)) + 2\frac{\partial}{\partial t}\frac{\theta''(t)\frac{\partial}{\partial t}F_t(y(t))}{\theta'(t)}\right] + \frac{1}{2}\sigma^2(K)y'(t)\left[\frac{\partial^2}{\partial y^2}f_t(y(t)) + \frac{\partial^2}{\partial t^2}f_t(y(t)) + 2\frac{\theta''(t)\frac{\partial}{\partial t}f_t(y(t))}{\theta'(t)}\right],$$

and

$$\sigma^2(t) = \frac{\partial^2}{\partial s \partial s'}F_{t,t}(y(t), y(t)) - \left[\frac{\partial}{\partial t}F_t(y(t))\right]^2 + 2y'(t)\left[\frac{\partial^2}{\partial s \partial z'}F_{t,t}(y(t), y(t)) - f_t(y(t))\frac{\partial}{\partial t}F_t(y(t))\right] + y'^2(t)\left[f_{t,t}(y(t), y(t)) - f_t(y(t))^2\right].$$

These results provide rates of convergence and theoretical justifications for the estimated rank dynamics.

6. Simulation

For the implementation of the dynamic model in Section **??** and the summary measures in Section **??**, two important auxiliary parameters h_Y and h_T are involved to obtain the kernel estimators for the rank trajectories $R_i(\cdot)$ and the two components, $C_1(t)$ and $C_2(t)$, of the rank derivatives. In this section, we use simulations to evaluate the finite-sample performance of the bandwidth selection method in the Supplementary Material, and the kernel estimators for $C_1(t)$ and $C_2(t)$ in model (**??**).

Denote ϕ and Φ as the probability density function and cumulative distribution function of the standard Gaussian distribution. Suppose we observe trajectories $Y_i(t) = \sum_{k=1}^{5}\xi_{ik}\psi_k(t)$ for subjects $i = 1, \ldots, n$ on a dense time grid $\{j/m : j = 0, 1, \ldots, m\} \subset \mathcal{T} = [0, 1]$, where $\psi_1(t) = 6(t-0.5)^2\mathbf{1}_{\{t>0.5\}}$, $\psi_2(t) = 0.4 + (70/9)\phi((t-0.5)/0.09)$, $\psi_3(t) = 0.6\cos(8\pi t)$, $\psi_4(t) = \sin(2\pi t) + 1$, $\psi_5(t) = 8\phi((t-0.2)/0.05)$, $\xi_{i1} \sim \mathcal{N}(1.4, 1.7^2)$, $\xi_{i2} \sim \mathcal{N}(1, 0.6^2)$, $\xi_{i3} \sim \mathcal{N}(0, 0.5^2)$, $\xi_{i4} \sim \mathcal{N}(0.8, 0.4^2)$, and $\xi_{i5} \sim \mathcal{N}(0.4, 0.2^2)$, independently across $i = 1, \ldots, n$. Hence, the true values of

$R_i(t)$, $C_{1i}(t)$ and $C_{2i}(t)$ are respectively

$$R_i(t) = \Phi\left(\frac{\sum_{k=1}^5(\xi_{ik}-\mu_k)\psi_k(t)}{\sqrt{\sum_{k=1}^5\sigma_k^2\psi_k(t)^2}}\right),$$

$$C_{1i}(t) = \left[\frac{-\sum_{k=1}^5\mu_k\psi_k'(t)}{\sqrt{\sum_{k=1}^5\sigma_k^2\psi_k(t)^2}} - \frac{\left[\sum_{k=1}^5(\xi_{ik}-\mu_k)\psi_k(t)\right]\left[\sum_{k=1}^5\sigma_k^2\psi_k(t)\psi_k'(t)\right]}{\left[\sum_{k=1}^5\sigma_k^2\psi_k(t)^2\right]^{3/2}}\right]$$
$$\cdot\,\phi\left(\frac{\sum_{k=1}^5(\xi_{ik}-\mu_k)\psi_k(t)}{\sqrt{\sum_{k=1}^5\sigma_k^2\psi_k(t)^2}}\right),$$

$$C_{2i}(t) = \frac{\sum_{k=1}^5\xi_{ik}\psi_k'(t)}{\sqrt{\sum_{k=1}^5\sigma_k^2\psi_k(t)^2}}\cdot\phi\left(\frac{\sum_{k=1}^5(\xi_{ik}-\mu_k)\psi_k(t)}{\sqrt{\sum_{k=1}^5\sigma_k^2\psi_k(t)^2}}\right).$$

To assess the performance of the cross-validation (CV) selected bandwidths $(h_Y^{\mathrm{CV}}, h_T^{\mathrm{CV}})$, we compared the mean integrated squared error (MISE) of $\widetilde{C}_{1i}(t)$ and $\widetilde{C}_{2i}(t)$ obtained with the CV bandwidths as well as with the optimal choice given by

$$(h_Y^{\mathrm{opt}}, h_T^{\mathrm{opt}}) = \underset{(h_Y,h_T)\in\mathcal{H}}{\operatorname{argmin}}\ \mathrm{MISE}(h_Y,h_T;\widetilde{C}_1) + \mathrm{MISE}(h_Y,h_T;\widetilde{C}_2),$$

where $\mathcal{H}\in\mathbb{R}^2$ is the set of bandwidth pairs considered,

$$\mathrm{MISE}(h_Y,h_T;\widetilde{C}_1) = \frac{1}{n}\sum_{i=1}^n\int_{h_{\max}}^{1-h_{\max}}\left[\widetilde{C}_{1i}(t)-C_{1i}(t)\right]^2\,\mathrm{d}t,$$

$$\mathrm{MISE}(h_Y,h_T;\widetilde{C}_2) = \frac{1}{n}\sum_{i=1}^n\int_{h_{\max}}^{1-h_{\max}}\left[\widetilde{C}_{2i}(t)-C_{2i}(t)\right]^2\,\mathrm{d}t,$$

and $h_{\max}$ is the maximum value of h_T considered. The impact of boundary effects is known to distort bandwidth selection and is removed by cutting off $[0, h_{\max})$ and $(1-h_{\max}, 1]$ in the integration.

In the simulations, we used $m = 31$, $\mathcal{H} = \{(h_Y,h_T) = (2.4\times0.6^u, 0.3\times0.6^v) : u,v = 0,1,2,3\}$, and considered three different sample sizes $n = 20$, 50 and 200. The kernels K and H used in Sections **??** and **??** are the pdf and cdf of standard normal distributions truncated on $[-4,4]$, respectively. Specifically,

$$K(x) = \phi(x)\mathbf{1}_{[-1,1]}(x/4)/[\Phi(4)-\Phi(-4)], \quad \text{and}$$
$$H(x) = [\Phi(x)-\Phi(-4)]\mathbf{1}_{[-1,1]}(x/4)/[\Phi(4)-\Phi(-4)] + \mathbf{1}_{(1,\infty)}(x/4),$$

where ϕ and Φ are the pdf and cdf of standard normal distributions, respectively. We use these kernels as due to their smoothness in practical implementations they

yield smooth estimates $\widetilde{F}$, $\widetilde{D}_1$, and $\widetilde{D}_2$. Boxplots of the MISEs corresponding to the optimal bandwidths chosen by MISE and CV in each of the 1000 Monte Carlo runs for $n = 20$, 50 and 200 are shown in Figure **??**. The main message is that CV performs satisfactorily, as it tracks the optimal choice closely, especially for larger sample sizes n.

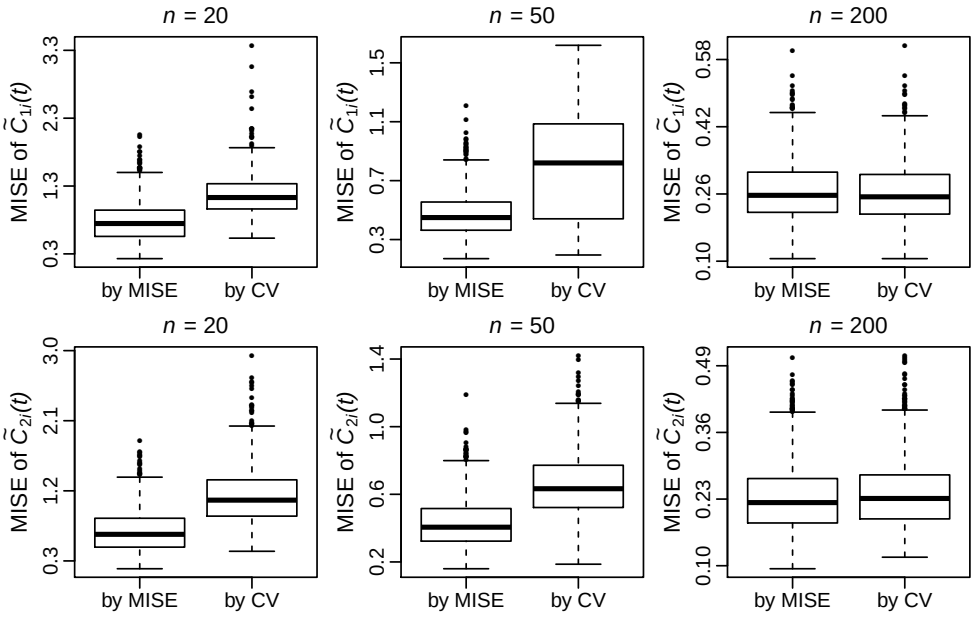

Figure 1: Boxplots of the MISEs of $\widetilde{C}_{1i}(t)$ and $\widetilde{C}_{2i}(t)$ corresponding to the optimal bandwidths chosen by MISE and CV in 1000 runs.

Boxplots of the MSEs, ISE or SE for the estimation of the rank summary measures (**??**)–(**??**) based on the kernel estimators $\widetilde{R}_i(t)$ and $\widetilde{R}'_i(t)$ obtained with the optimal bandwidths chosen by CV are shown in Figure **??**. Overall the proposed estimators are seen to converge fast to the true values as n increases.

7. Applications

We demonstrate our methods with three functional datasets which are very different in nature. The first is the Zürich longitudinal growth data; the second is US median house price data at the county level; the third is based on the 2017 Major League Baseball (MLB) season, where our interest lies in offensive or hitting performances. We find that by transforming the original processes into rank processes we are able to find new and interesting characterizations for the individuals in each dataset.

7.1. Zürich Longitudinal Growth Data

The Zürich longitudinal growth data consist of dense longitudinal height measurements for 112 girls and 120 boys from birth to age 20 and the measurements

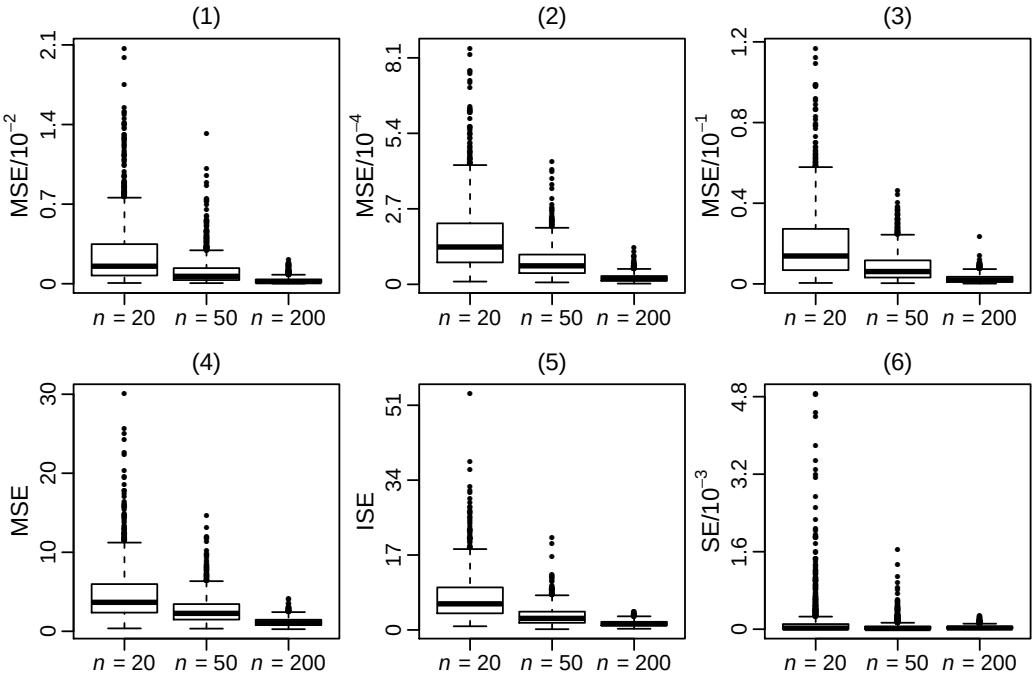

Figure 2: Boxplots of the MSEs, ISE or SE of various rank summary statistics based on 1000 runs and CV bandwidths. Panels (1)–(6) display the estimated values of ρ_i, ν_i, ζ_i, η_i, $\gamma(t)$, and G, respectively, indicating also the equation number where the corresponding quantity is defined.

are known to contain very little noise [**?**]. It is helpful to compare the ranking for individuals; we highlight the same six girls and six boys throughout, with their height trajectories shown in Figure **??**.

We find that the two ranking methods yield similar results, with the smooth rank functions resembling the empirical ranks. Visually, it is clear that taking a ranked perspective with functional data is appealing. For example, from Figure **??**, Girl 1 and Boy 1 are seen to be generally tall throughout, and Girl 2 and Boy 2 are seen to have volatile ranks as they age. Ranks are fairly stable from ages 5 until 10 and 12 for girls and boys, respectively; subsequently, the ranks are more dynamic, with higher volatility.

We also obtained the estimates of the rank summary statistics (**??**)–(**??**) for the Zürich longitudinal growth data, based on the smooth ranks defined in (**??**). In Figure **??** we see that Girl 1 and Boy 1 have very high ranks and that the ranks are almost constant throughout. On the other hand, we find that Girl 2 and Boy 2 have overall middle ranks that are quite volatile. The rank volatility plots are bell-shaped, as subjects with integrated ranks near 0 and 1 cannot have high volatility. On the other hand, subjects with moderate integrated ranks have less restricted volatility. We also highlight the subjects with the highest and lowest values of the subject-specific rank increases from start to end ζ_i as in (**??**) in Figure **??**, where ζ_i captures the overall ranking trend for a subject, i.e., subjects with large values of ζ_i have large

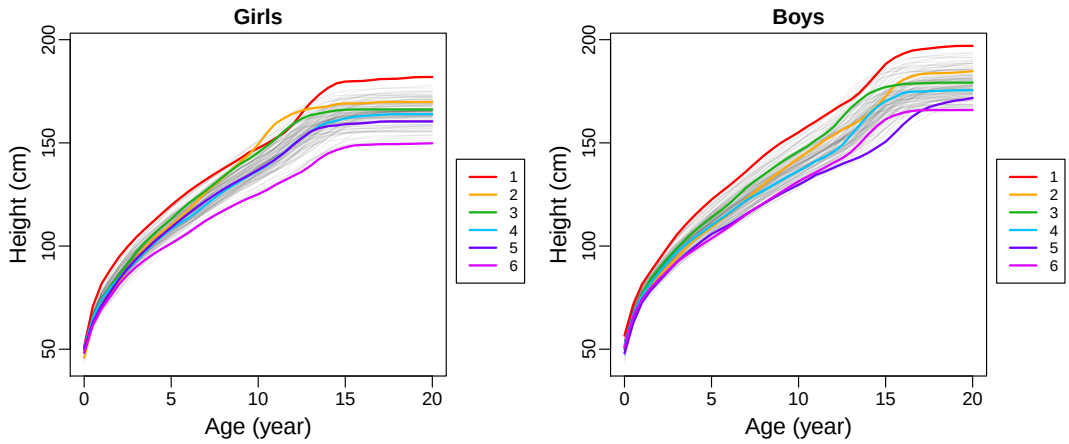

Figure 3: Pre-smoothed Zürich growth curves with six subjects highlighted for boys and girls.

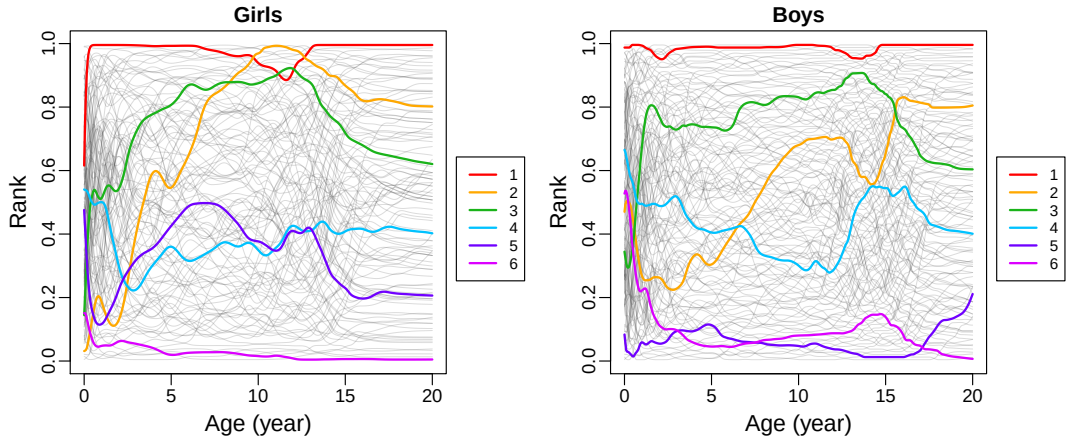

Figure 4: Smoothly ranked Zürich growth trajectories.

increases or decreases in ranks from the beginning to the end of the time domain.

We also applied the rank decomposition (**??**) to the Zürich growth data. Figure **??** shows the rank derivative decomposition for all subjects in the study. The population trends quantified by the negative terms C_1 tend to lower an individual's rank as the population of children at large is growing, while individuals are also growing as reflected by the positive terms C_2. For the growth data, this decomposition indicates that the population and individual components of the rank derivative are roughly equal in size. Indeed, the estimated contributions from the first component $\widetilde{\Lambda}_1$ for girls and boys are 0.487 and 0.486, with $\widetilde{\Lambda}_2 = 0.513$ and 0.514, respectively, for the second component. We conclude that in human growth an individual's change in rank is the result of a fine balance of individual growth which is counterbalanced by population trends in growth when considering individual rank trajectories. Rank volatility is seen to increase during times of growth spurts, where the population tends to grow relatively fast while individuals may have accelerated or delayed growth, with

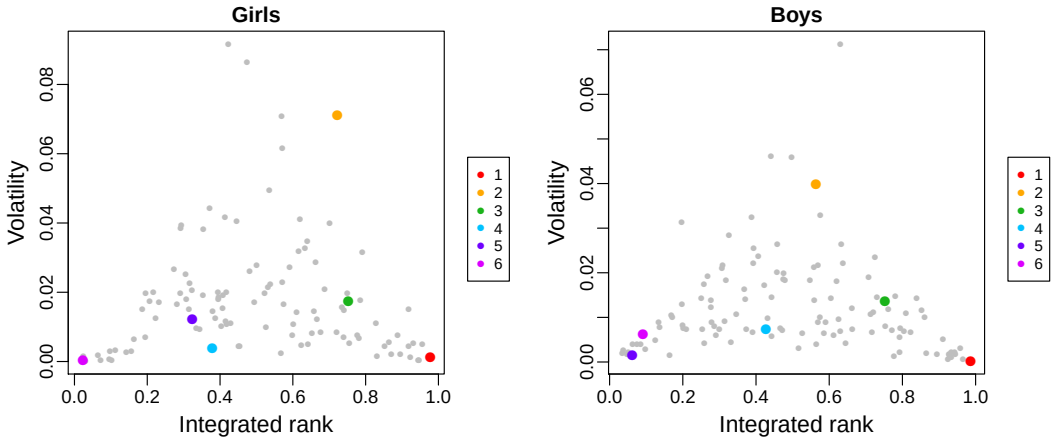

Figure 5: Rank volatility versus integrated rank in the Zürich growth data, with the same six subjects highlighted as in Figures 3 and 4.

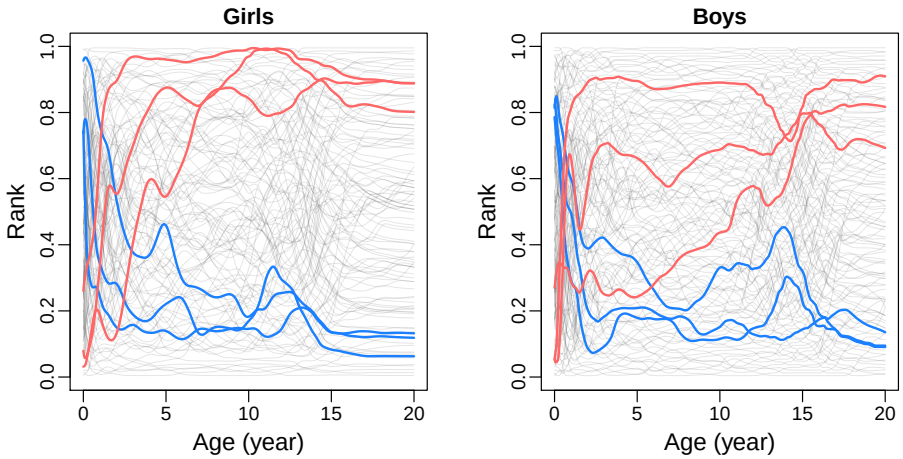

Figure 6: Smoothly ranked Zürich growth data. Here we highlight the subjects with the highest (light red) and lowest (blue) subject-specific rank stability measures ζ_i.

resulting rank changes.

7.2. House Price Data

House price data are available from Zillow. We consider here monthly longitudinal median house prices after inflation adjustment for house transactions in 306 counties in the US from May 1996 to August 2015. To compare the ranking for individual markets, we highlight the same six counties throughout, as in Figure **??**. Adopting the smooth rank function version defined in (**??**), in Figure **??** house prices in Contra Costa and Fayette are seen to be generally high and low throughout, respectively, and those in Fresno are seen to have significant rank variation. We find that ranks were fairly stable before 2002 and became more dynamic afterward.

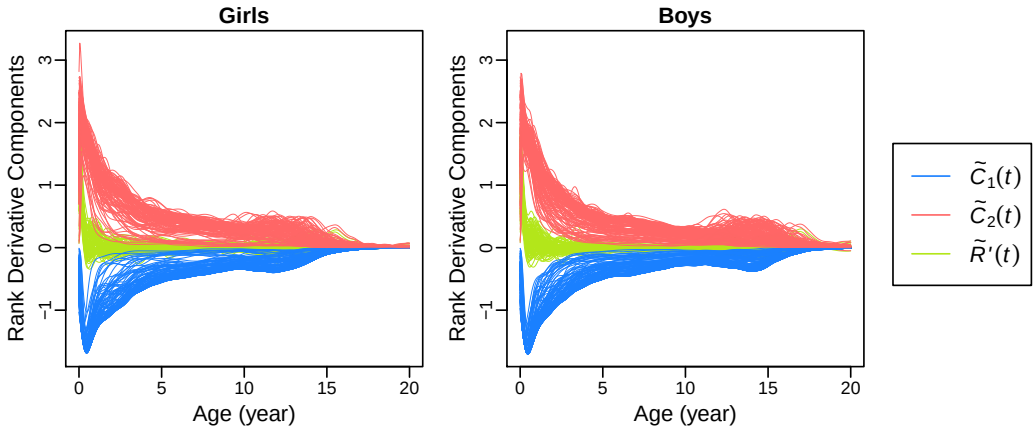

Figure 7: Rank derivative components for girls (left) and boys (right) in the Zürich growth data.

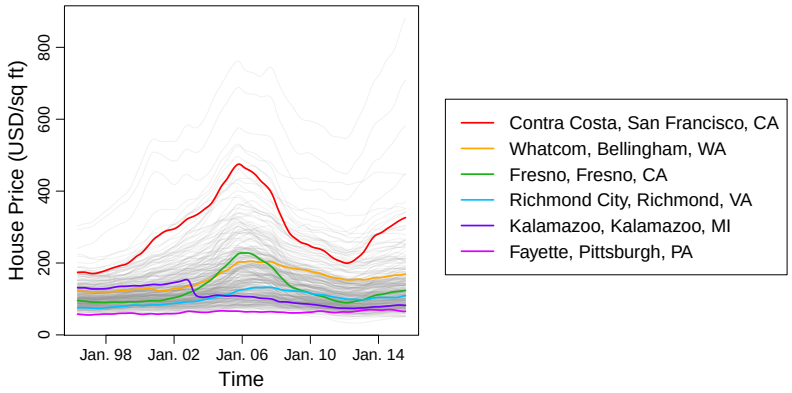

Figure 8: Pre-smoothed inflation-adjusted median house price curves for U.S. counties with six counties highlighted.

We also estimated the rank summary statistics for the house price data. In Figure **??** we see that Contra Costa county and Fayette county have very high and low ranks respectively and that their ranks were almost constant throughout the time period considered. On the other hand, we find that Kalamazoo has moderate ranks that are very volatile. These findings are in agreement with Figure **??** and the rank volatility plot has a similar shape to that in Figure **??**, as expected. Highlighting the counties with the highest and lowest gains in rank ζ_i as in (**??**) in Figure **??**, we find that the magnitudes of difference in ranks between the beginning and the ending for the house price data are not as large as those for the Zürich growth curves.

We also applied the dynamic rank decomposition (**??**) to the house price curves. Figure **??** shows the rank derivative decomposition for all counties in the study. The house price ranks were more volatile a few years before and after the 2008 financial crisis. The population components also reveal that county median house prices were increasing in general before 2006, turned to drop from 2007, and then gradu-

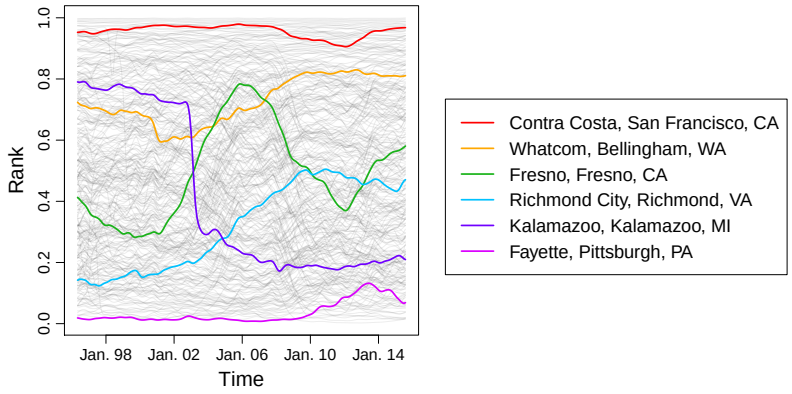

Figure 9: Smoothly ranked house price trajectories.

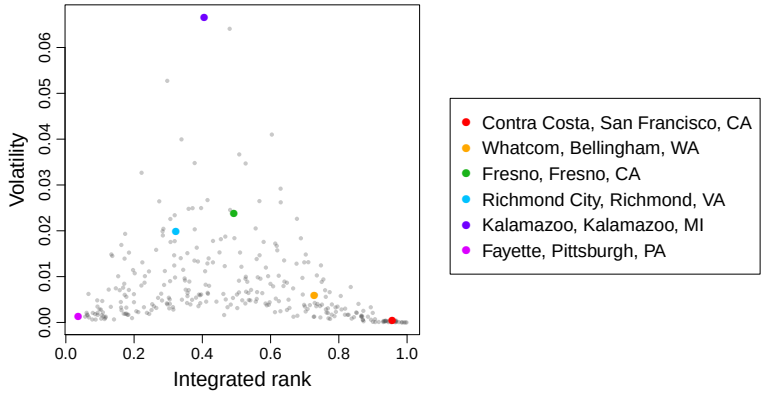

Figure 10: Rank volatility versus integrated rank for the house price data, with the same six counties highlighted for clarity.

ally recovered and increased again since 2012. The individual component is seen to contribute more to the rank derivative than the population component. This is also reflected by the estimated contributions from the two components, $\widetilde{\Lambda}_1 = 0.458$ and $\widetilde{\Lambda}_2 = 0.542$. As shown in Figure **??**, a general trend cana be discerned from the house price trajectories: Prices initially increased until 2005, decreased from 2005 to 2012, and then increased again. The house price population dynamics points predominantly downwards until 2008, with individual markets exercising strong counterforces; this means a county where price growth was sluggish fell back in rank; the opposite happened between 2008 and 2012 — a county where house prices were stable was gaining against the population and its rank increased.

7.3. Major League Baseball Offensive Data

Another area where relative rank is important is in sports. Major League Baseball (MLB) teams routinely spend over $100 million on player salaries every year. It is therefore of paramount interest to rank players in terms of ability so that teams

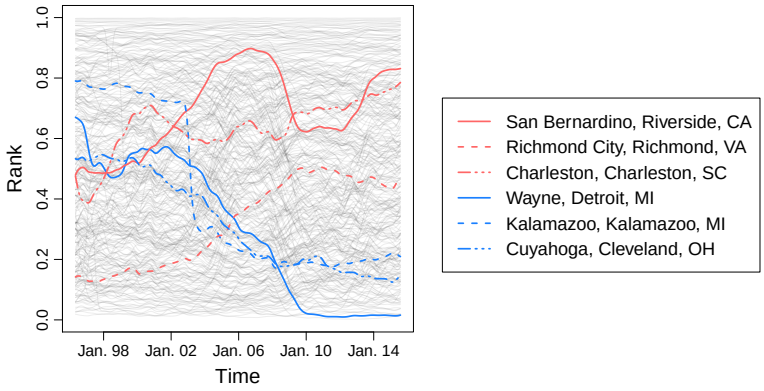

Figure 11: Smoothly ranked house price data, highlighting the counties with the highest (light red) and lowest (blue) county-specific rank gains ζ_i.

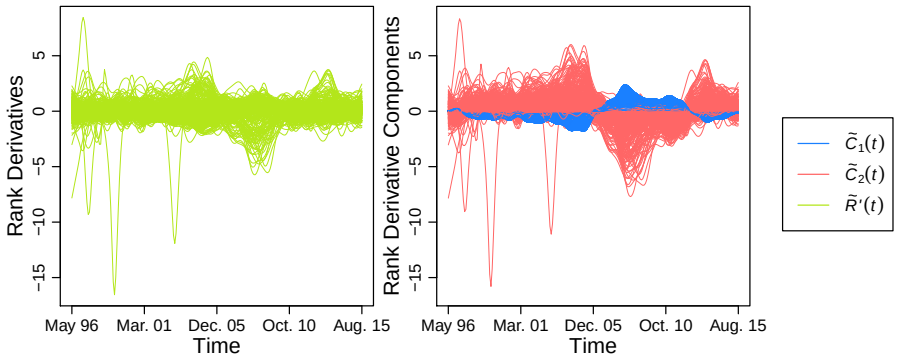

Figure 12: Rank derivative components for the house price data.

can invest efficiently in individual players. Although there are many factors which contribute to the overall value of a player, one of the most important is offensive performance, and accordingly we focus on ranking MLB players in terms of offense.

Baseball has recently become a game dominated by statistics [see **? ?** , and the movie *Moneyball* for instance]. As such, statisticians and sabermetricians look for simple yet informative measures for assessing player performance. By far, the most widely used statistic to quantify offensive performance is the batting average (BA), which is the number of hits a player has divided by the number of attempts. While the batting average is simple to understand, it has several shortcomings; for example, late in the season, when the number of attempts or at-bats is high, the average will not easily reveal changes in performance.

In light of the drawbacks of using batting average as a response, we tracked the number of hits a player accrued for each day in the 2017 MLB season (http://www.baseballmusings.com/), and then took the derivative of this trajectory, which we used as our functional response. This derivative can be viewed as a local batting

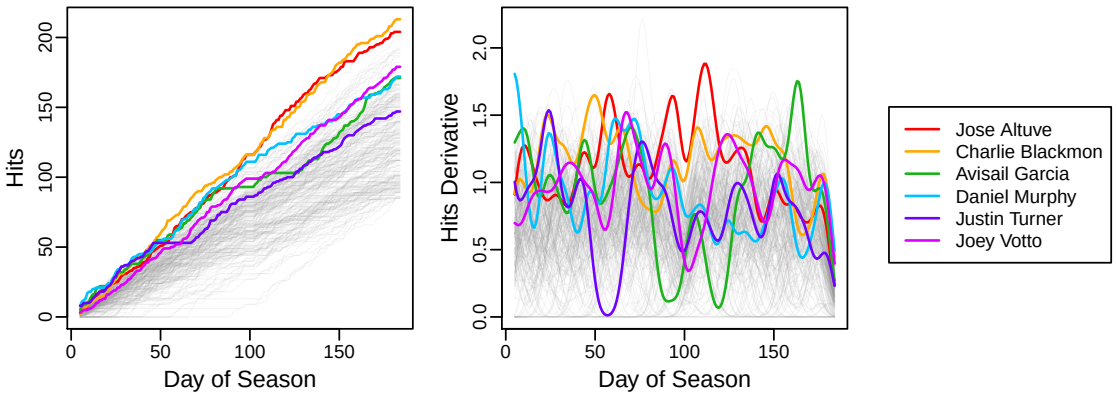

Figure 13: Cumulative hits (left) and hits derivatives (right) for each day in the 2017 Major League Baseball season for 237 players. Six curves are highlighted which correspond to the players with the highest batting averages.

average, or the change in hits divided by the change in days. It is thus less affected by long-term history because it is an instantaneous measure. This response therefore characterizes the *heat* of a player, which is the level of their current performance. The original hits trajectories and corresponding hits derivatives trajectories in Figure **??**, obtained by local polynomial smoothing, are our starting point for the rank analysis. The objective is to quantify the player's ranks and changes in ranks in this dataset, aiming to identify top players. We first transform the hit derivative trajectories into rank trajectories using the smooth representation in (**??**), visualized in Figure **??**, where the differences in rank for the six highlighted players are highlighted. For example, this visualization makes it clear that Joey Votto improved drastically throughout the season, moving from a rank near 0.25 at the beginning of the season, to finishing with a rank of nearly 1.

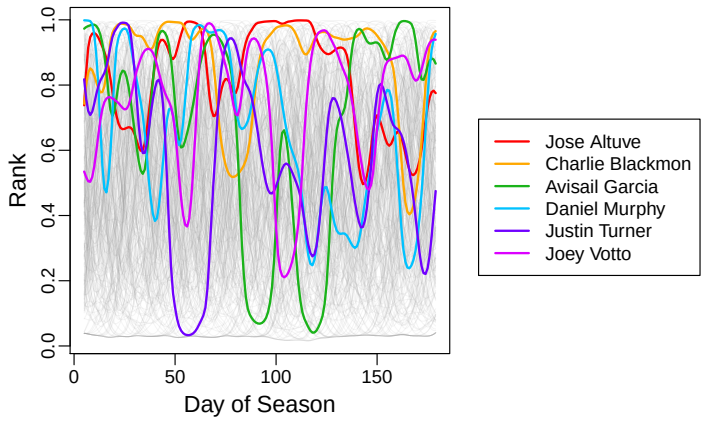

Figure 14: Rank transformed baseball data, with the same six players highlighted.

We also applied the rank summary statistics, which prove to be informative. The rank volatility versus integrated rank plot, shown in Figure **??** has direct applications in assessing offensive performance from the 2017 MLB season. Naturally, all six of the highlighted players have relatively high ranks. In addition to average performance, we can see that two of the players, Jose Altuve and Charles Blackmon, had high integrated rank and low volatility, which are two features of the most valuable players. These players are consistently performing at a high level with respect to the rest of the sample. As shown in Figure **??**, the player with the highest integrated rank and fairly low volatility is Charlie Blackmon. Taking the viewpoint of a team deciding on which players to acquire, this plot also allows one to select players which have modest average ranks but have low volatility. Players of this type are desirable when looking for consistent backup players, for example. Finally, the player-specific change in ranks ζ_i quantifies whether players are generally improving or deteriorating over the season.

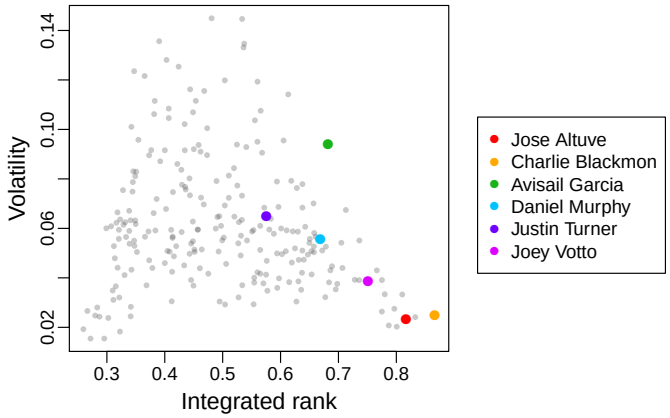

Figure 15: Rank volatility versus integrated rank for the baseball data, with the same six players highlighted.

When fitting the rank derivative decomposition model (**??**) to these baseball data, we find that the subject specific component $C_2(t)$ contributes much more than the population component $C_1(t)$. This is not surprising as the population of hits derivative curves $Y_i(t)$, $i = 1, \ldots, n$ does not have a very clear pattern. Thus rank is determined to a large extent by individual effort alone, with estimated contributions $\widetilde{\Lambda}_1 = 0.165$ for the population component and $\widetilde{\Lambda}_2 = 0.835$ for the individual component. This is visualized in Figures **??** and **??**, where the second component is seen to dominate the first. In addition, an ascent followed by a descent period can be seen in the population component curves around Day 100. This is due to the "All Star Break", which is a break for all the players except the All Stars. i.e., the best players from each team, who play in an exhibition game. Thus, the hits derivatives decrease toward zero for almost all players during the break and then recover after the games are resumed. Hence the population components first ascend and then descend accordingly. The

ascending phase of the population component near the end of the season is due to the same reason, i.e., fewer games are available at that time.

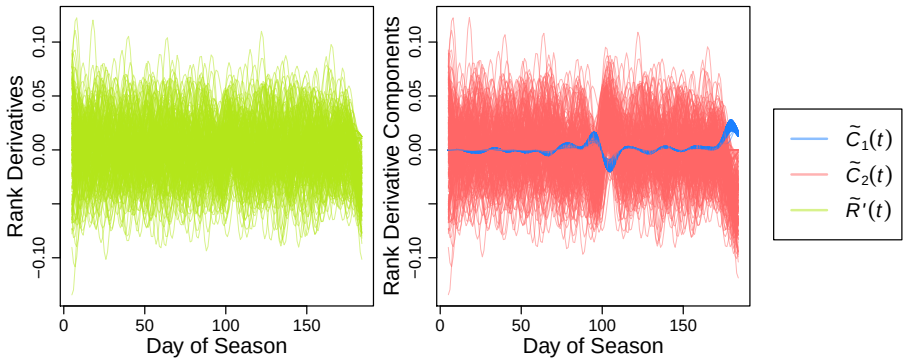

Figure 16: Rank derivative components for 2017 Major League Baseball data.

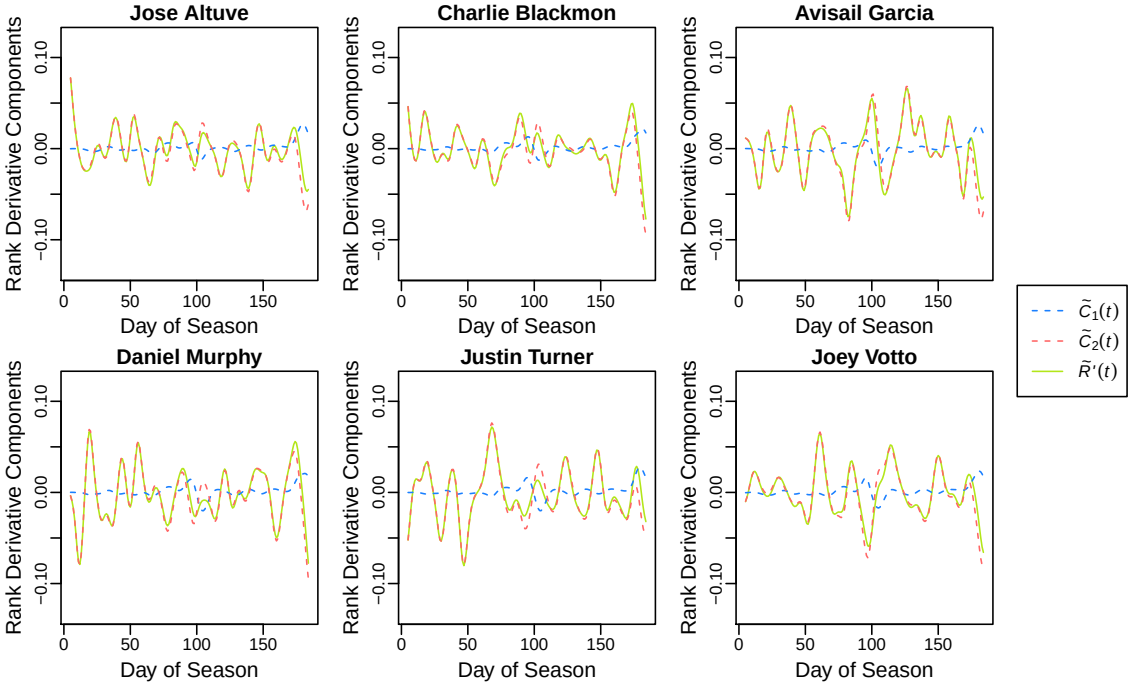

Figure 17: Rank derivative components for the six players with the highest batting averages of 2017.

Finally, the overall rank stability coefficient G in (**??**), which is an overall scaled measure of how variable the rank trajectories are, can be used to compare all three functional data set that we have considered, i.e., the Zürich growth data, the housing price data and the baseball player data. The estimates of G based on the smooth rank estimation are shown in Table **??**. The baseball players' rank curves have the lowest stability, with the most volatility of ranks and a much higher degree of crossing

trajectories. Moreover, the rank trajectories are not much influenced by population trends. In the Zürich growth and house price data, we observe much higher degrees of stability, with the highest level of rank stability and associated lowest rank volatility for the growth data. Especially for the growth data, crossings of rank trajectories are not common. Rank trajectories for the housing data and even more so for the growth data are driven to a large extent by population trends, where population distributions uniformly move to higher levels for the growth data with increasing age, while they have increasing and decreasing phases for the house price data. Notably, for the growth data, the trajectory dynamics are driven in equal parts by population trends and individual growth patterns, while for the housing price data population trends play a slightly smaller role.

Table 1: Estimates of G based on the smooth rank estimation for all the three datasets

Zürich growth		House price	Baseball
Girls	Boys		
0.9866	0.9883	0.4500	3.409×10^{-21}

8. Discussion

Cross-sectional ranking of functional data is a powerful tool for exploratory functional data analysis. To the best of our knowledge, the proposed perspectives in this paper are new to the field of functional data analysis and allow for quantification of the rank dynamics of a stochastic process. These methods are simple to understand and straightforward to implement. The decomposition of rank dynamics into population and individual components allows to better understand the forces that shape observed rank trajectories, and the summary measures of rank volatility, rank stability and rank gain are useful.

For the estimation of the two components $D_1(y,t)$ and $D_2(y,t)$ in (**??**), we could alternatively use local quadratic regression. This would be asymptotically equivalent to the kernel estimator in (**??**) under regularity assumptions on the smoothness of weight functions and the shape of kernels [**?**]. However, the kernel method we employ here has an explicit form which facilitates theoretical derivations, and makes implementation straightforward, while the local quadratic regression involves the inverse of a matrix of dimension at least 5×5. This provides strong motivation for the proposed method.

Our estimation methods and theory are geared towards densely observed functional data. One possible approach for the case of sparsely observed functional data is to divide the time domain into bins in a preprocessing step, followed by estimating the cross-sectional distribution at time t by using local Fréchet regression [**?**] based on the preliminary distributions observed at the midpoints of the bins – these are the empirical distributions derived from the observations falling into each bin.

The two components can then be obtained, e.g., by taking difference quotients of the cross-sectional distribution estimates. To work out the details and full theoretical justification of such a method will be a future research project.

Acknowledgements

Funding: This work was supported by NSF Grant DMS-1712864.

Appendix A. Supplementary Material

A data-driven approach for bandwidth selection for the kernel estimator in (**??**) and details about the theoretical results and proofs are available in the online Supplementary Material.

References