Project Description

1 Introduction

With recent technological advancement, an explosive amount of real-time human activity data can be collected through various venues nowadays. For example, social media platforms such as Twitter and Facebook produce a myriad of user generated content on a daily basis, and on-line retailers such as Amazon and eBay maintain detailed transaction records of all registered users. The complexity and magnitude of these new data call for new innovative statistical modeling tools. We plan to answer this call by proposing a series of new nonparameteric and semi-parametric point process models for real-time temporal point patterns of structured human activities. To illustrate, we first introduce two datasets that the team has access to as motivating examples.

Stock Trading Data The team has access to detailed transaction records (i.e., time stamps of buying/selling stocks at the second level) of 1.2 million stock trading accounts from a national leading brokerage house in China from January 4th, 2007 to September 30th, 2009. The study window covers an important period in the history of the Chinese stock market. On January 4th, 2007, the Shanghai Composite Index was only around 2700. However, by October 16th, 2007, it had quickly raised to peak 6092. Since then it spiraled downward drastically and lost about 70% of its peak value in less than a year. The market recovered after the crash, but the index had been hovering around 3000 since then. The data offer a unique opportunity to study the trading behaviors of individual Chinese stock investors. Some questions of interests are: (a) what contributed to variations in an investor's buying or selling activities? (b) how did the investors adjust their trading behaviors before, during and after a stock bubble? (c) how were one's buying and selling activities related? (d) were there any clusters for the trading behaviors?

Social Media Data Sina Weibo is the largest Twitter-type social media in China. When a user publishes a post, it can be either an *original post* or a *repost* from another account. The team has access to data from 54,773 followers of a public Weibo account during the period of Nov. 30th, 2011 to Nov. 16th, 2013. In addition to posting times, user-related information such as the numbers of followers and followees of the account are also available. Figure 1 shows the time stamps for three users in a window of 30 consecutive days. All three users clearly have strong clustered patterns. However, these posting patterns differ significantly in many ways, such as posting frequencies, percentages of original posts and reposts, and strengths of clustering. It is of interest to characterize these activity patterns and quantify their differences. More specific questions include, for example, (a) at what time of a day would a user start using Weibo? (b) how many posts would a user generate during each use? (c) how user-related characteristics were related to the posting patterns? Answers to these questions can help us understand and predict activity patterns of targeted user groups. They can also provide potentially valuable inputs to other tasks such as optimizing on-line advertisement placements.

As we can see in these two motivating examples, temporal point pattern data contain rich information about human activities and have become increasingly prevalent in many disciplines. However, the development of new statistical tools for handling such data lags much behind the data availability. The proposed research aims to narrow this gap by achieving following specific aims.

- *Aim I: Multilevel functional principal component analysis (MFPCA) for temporal point patterns.* Many factors can affect human activity patterns. For example, in the Stock Trading data, an investor's trading activities may be affected by three factors, i.e., investor-related characteristics, daily market incentives and their interactions. The proposed model treats the trading times as a univariate/bivariate temporal point process with a random latent intensity function that can be decomposed into three different levels. Variations of the intensity functions both between users and across days will be studied using MFPCA.

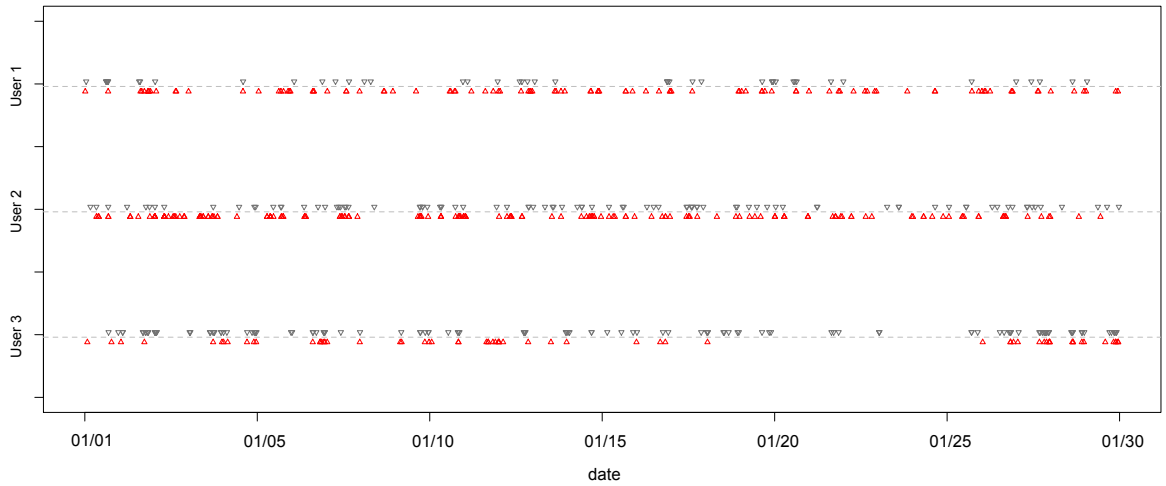

Figure 1: The posting times of original posts and reposts from three selected users. The events above the dashed line are original posts; the events below the dashed line are reposts.

- *Aim II: Soft-clustering of multi-level functional point processes.* An important goal in studying human activity patterns is to identify user groups displaying similar behavioral patterns. One can further look into each cluster to better understand the underlying cause for certain activity patterns and to make some necessary adjustments (e.g. behavioral interventions). The goal of this project is to develop a unified approach to model human activity patterns and simultaneously form user clusters accordingly.
- *Aim III: Semi-parametric modeling of bivariate clustered point processes.* We propose a new class of bivariate point process models that are flexible enough to model the complex behaviors of modern social media users. Compared to the nonparametric models proposed in Aims I and II, the parameters in this model class have straightforward interpretations and can therefore provide meaningful insights into a user's content generating behavior.

2 Aim 1: Multilevel Functional PCA for temporal point patterns

2.1 Introduction

Functional data analysis (FDA; Ramsay, 2006) has been a hot research topic in recent years, see, for example, Yao et al. (2005); Hall et al. (2006); Li and Hsing (2010); Li et al. (2013). To analyze functional data with complicated structures, a particularly useful tool is the so-called multi-level functional data analysis (MFDA; Di et al., 2009; Staicu et al., 2010; Zhou et al., 2010), where functional components can be studied at different levels. Functional time series data (Hormann and Kokoszka, 2012; Hörmann et al., 2015) can also be viewed as multi-level functional data if one views the time of an observed function as a level. While the aforementioned work have been focused on Gaussian-type functional data, Goldsmith et al. (2015) recently considered generalized multi-level functional data from an exponential family, following the idea of Hall et al. (2008).

There has also been some but scarce work on FDA for point processes. Bouzas et al. (2006); Illian et al. (2006); Wu et al. (2013) and Gervini (2015) considered data with independent replicates of point patterns and modeled a summary measure or the intensity function of the point process as functional data. Li and Guan (2014) considered FDA for spatio-temporal point processes where the intensity function is spatially correlated. Other examples include Panaretos et al. (2016).

Our Stock Trading data consist of daily buying/selling time stamps for many accounts over 672 trading days. As alluded in Aim I, the data can be viewed as a multi-level temporal point process with account, day and account-day being the different levels. For each account on a given day,

the trading times are assumed to be generated by a log-Gaussian Cox process with some random latent intensity function, to which multi-level functional PCA (MFPCA) can be applied. To our best knowledge, MFPCA for point processes has never been considered before.

2.2 Model formulation

To best illustrate, we use the Stock Trading data to present our model. Suppose activities of n accounts were monitored over a fixed time window $[0,1]$ for an m-day period. Let $N_{ij} = \{T_{ij,l} : l = 1,\ldots,\#\{N_{ij}\}\}$, where $0 \le T_{ij,1} < T_{ij,2} < \cdots < T_{ij,\#\{N_{ij}\}} \le 1$, be the set of trading times observed for account i on day j, $i = 1,\ldots,n$ and $j = 1,\ldots,m$. Assume that conditioned on a (random) latent intensity function $\lambda_{ij}(\cdot)$, the trading times $T_{ij,l}$'s are generated by an inhomogeneous Poisson process over the time window $[0,1]$. By generalizing the two-way functional ANOVA model (Di et al., 2009) to our setting, we propose the following decomposition:

$$\lambda_{ij}(t) = \lambda_0(t)\exp\{X_i(t) + Y_j(t) + Z_{ij}(t)\}, \tag{1}$$

where $\lambda_0(t)$ is a fixed function representing the overall pattern during a day and $X_i(t), Y_j(t), Z_{ij}(t)$ are mutually independent random functions from zero-mean Gaussian processes, reflecting deviations from the overall pattern at the account, day and account-day levels, receptively.

Using the Karhunen-Loève expansion (Watanabe, 1965), the random functions $X_i(t), Y_j(t)$ and $Z_{ij}(t)$ can be approximated as follows

$$X_i(t) = \sum\nolimits_{k=1}^{p_X} \xi_{ik}^X \phi_k^X(t), \quad Y_j(t) = \sum\nolimits_{k=1}^{p_Y} \xi_{jk}^X \phi_k^Y(t), \quad Z_{ij}(t) = \sum\nolimits_{k=1}^{p_Z} \xi_{ijk}^Z \phi_k^Z(t), \tag{2}$$

for some positive integers p_X, p_Y, p_Z, where ξ_{ik}^X's, ξ_{ik}^Y's, and ξ_{ik}^Z's are mutually independent normal random variables with mean 0 and variances λ_k^X, λ_k^Y, λ_k^Z, respectively, and $\phi_k^X(t)$'s, $\phi_k^Y(t)$'s and $\phi_k^Z(t)$'s are orthogonal basis functions determined by the covariance kernels of $X_i(t), Y_j(t), Z_{ij}(t)$. We further assume that (a) ξ_{ik}^X and $\xi_{i'k'}^X$ are independent if $k \neq k'$; (b) ξ_{jk}^Y and $\xi_{j'k'}^Y$ are independent if $k \neq k'$ (c) ξ_{ijk}^Z and $\xi_{i'j'k'}^Z$ are independent if $(i,j,k) \neq (i',j',k')$.

It is worth pointing out that assumptions (a)-(c) allow potential correlations among $X_i(t)$'s as well as among $Y_j(t)$'s, which is different from existing literature of multi-level FDA (Di et al., 2009). However, to ensure estimation consistency, it is necessary to impose some constraints on the strength of dependence. In this proposal, we assume the following condition

$$\text{[C1]} \sup_{s,t\in[0,1]} \frac{1}{n}\sum_{i=1}^{n}\sum_{i'=1}^{n} |\mathrm{Cov}\left[X_i(s), X_{i'}(t)\right]| < \infty, \text{ and } \sup_{s,t\in[0,1]} \frac{1}{m}\sum_{j=1}^{m}\sum_{j'=1}^{m} \left|\mathrm{Cov}\left[Y_j(s), Y_{j'}(t)\right]\right| < \infty.$$

Condition C1 is a mild condition that ensures that the overall dependence strength is weak. For $Y_j(\cdot)$'s, it allows short-range longitudinal correlations among daily trading intensities. For $X_i(\cdot)$'s, it allows investors within a certain social circle to affect each other's trading activities. Thus, condition C1 is more realistic for the Stock Trading data than assuming independence. See also Figure 2 for an illustration of correlations among $Y_j(\cdot)$'s.

2.3 Eigenvalue-eigenfunction estimation

The first step of the model estimation is to estimate eigenfunctions and eigenvalues, i.e., (ϕ_k^X, λ_k^X)'s, (ϕ_k^Y, λ_k^Y)'s and $(\phi_k^Z, , \lambda_k^Z)$'s. Standard FDA theory suggests that it suffices to estimate corresponding covariance kernels, denoted as $C_X(\cdot,\cdot)$, $C_Y(\cdot,\cdot)$, $C_Z(\cdot,\cdot)$, consistently on $[0,1]^2$. To do so, we first

introduce the following quantities:

$$\text{(Same account, same day): } A_h(s,t) = \frac{1}{mn}\sum_{i=1}^{n}\sum_{j=1}^{m}\sum_{u,v\in N_{ij}}^{u\neq v} K_h(u-s)K_h(v-t),$$

$$\text{(Same account, different days): } B_h(s,t) = \frac{1}{mn}\sum_{i=1}^{n}\sum_{j=1}^{m}\sum_{j'\neq j}\sum_{u\in N_{ij},v\in N_{ij'}}\sum K_h(u-s)K_h(v-t),$$

$$\text{(Different accounts, same day): } C_h(s,t) = \frac{1}{mn}\sum_{i=1}^{n}\sum_{i'\neq i}\sum_{j=1}^{m}\sum_{u\in N_{ij},v\in N_{i'j}}\sum K_h(u-s)K_h(v-t), \quad (3)$$

$$\text{(Different accounts and days): } D_h(s,t) = \frac{1}{mn}\sum_{i=1}^{n}\sum_{i'\neq i}\sum_{j=1}^{m}\sum_{j'\neq j}\sum_{u\in N_{ij},v\in N_{i'j'}}\sum K_h(u-s)K_h(v-t),$$

where $K_h(u) = h^{-1}K(u/h)$ for some kernel function $K(\cdot)$ (e.g. Epanechnikov kernel) with a $h > 0$.

For ease of illustration, we assume that $X_i(\cdot)$ and $X_{i'}(\cdot)$ are independent if $i \neq i'$ and $Y_j(\cdot)$ and $Y_{j'}(\cdot)$ are independent if $j \neq j'$. Let $\rho(t) = \lambda_0(t)\exp\left[\left\{\sigma_X^2(t) + \sigma_Y^2(t) + \sigma_Z^2(t)\right\}/2\right]$. By Campbell's formula, $\mathbb{E}\left\{B_h(s,t)\right\} = \int_0^1\int_0^1 \rho(u)\rho(v)\exp\left\{C_X(u,v)\right\}K_h(u-s)K_h(v-t)dudv$ and $\mathbb{E}\left\{D_h(s,t)\right\} = \int_0^1\int_0^1 \rho(u)\rho(v)K_h(u-s)K_h(v-t)dudv$. Hence, for a sufficiently small h, $\mathbb{E}\{B_h(s,t)\}/\mathbb{E}\{D_h(s,t)\} \approx \exp\{C_X(s,t)\}$. More generally, under the weak dependence required by condition C1, we expect the following proposition to be true.

Proposition 2.1. *Assuming intensity function* (1)*. For some $h \to 0$, it can be shown that*

$$\mathbb{E}\{B_h(s,t)\}/\mathbb{E}\{D_h(s,t)\} \approx c_1\exp\{C_X(s,t)\}, \quad \mathbb{E}\{C_h(s,t)\}/\mathbb{E}\{D_h(s,t)\} \approx c_2\exp\{C_Y(s,t)\},$$
$$\mathbb{E}\{A_h(s,t)\}\mathbb{E}\{D_h(s,t)\}/\mathbb{E}\{B_h(s,t)\}\mathbb{E}\{C_h(s,t)\} \approx c_3\exp\{C_Z(s,t)\}, \quad (4)$$

for any $s,t \in [0,1]$ and some positive constant c_1, c_2, c_3.

Based on Proposition 2.1, sensible estimators of the covariance kernels are

$$\widehat{C}_{X,h}(s,t) = \log\frac{B_h(s,t)}{c_1 D_h(s,t)},\ \widehat{C}_{Y,h}(s,t) = \log\frac{C_h(s,t)}{c_2 D_h(s,t)},\ \widehat{C}_{Z,h}(s,t) = \log\frac{A_h(s,t)D_h(s,t)}{c_3 B_h(s,t)C_h(s,t)}, \quad (5)$$

for any $s,t \in [0,1]$. An interesting feature of the estimators given in (5) is that none of them depends on the overall mean intensity $\lambda_0(t)$, which is typically not the case for Gaussian-type functional data. To ensure the estimators in (5) are positive definite, we can trim eigenvalue-eigenfunction pairs with negative eigenvalues as suggested by Hall et al. (2008); Yao et al. (2005).

Finally, the eigenvalue-eigenfunction pairs (ϕ_k^X, λ_k^X)'s, (ϕ_k^Y, λ_k^Y)'s and (ϕ_k^Z, λ_k^Z)'s can be estimated using procedures in Yao et al. (2005) and Jacques and Preda (2014). For example, $(\widehat{\phi}_k^X, \widehat{\lambda}_k^X)$ can be obtained by solving the integral equation $\int_0^1 \widehat{C}_{X,h}(s,t)\widehat{\phi}_k^X(s)\,ds = \widehat{\lambda}_k^X\widehat{\phi}_k^X(t)$ for $k = 1,\ldots,p_X$, where $\widehat{\phi}_k^X$'s are subject to constraints $\int_0^1\{\widehat{\phi}_k^X(s)\}^2\,ds = 1$ and $\int_0^1 \widehat{\phi}_m^X(s)\widehat{\phi}_k^X(s)\,ds = 0$ for $m < k$.

Conjecture 2.2. *Under suitable conditions and C1, we have that* $\sup_{s,t\in[0,1]}\max\{|\widehat{C}_{X,h}(s,t) - C_X(s,t)|,\ |\widehat{C}_{Y,h}(s,t) - C_Y(s,t)|, |\widehat{C}_{Z,h}(s,t) - C_Z(s,t)|\} \xrightarrow{p} 0$ *as $h \to 0$ and $n, m \to \infty$.*

Conjecture 2.3. *Under suitable conditions and C1, all eigenvalue-eigenfunction pairs, (ϕ_k^X, λ_k^X)'s, (ϕ_k^Y, λ_k^Y)'s and (ϕ_k^Z, λ_k^Z)'s, can be uniformly consistently estimated as $h \to 0$ and $n, m \to \infty$.*

Conjecture 2.3 follows directly from Conjecture 2.2 using Theorem 2 in Yao et al. (2005). To show Conjecture 2.2, it suffices to show uniform convergence in probability for all quantities defined in (3) following Yao et al. (2005), with similar conditions on the kernel $K(\cdot)$ and the bandwidth h. We need an additional condition to bound all denominators in (5) away from 0 in probability. Let $\rho_{ij,i'j',2}(s,t) = \mathbb{E}[\lambda_{ij}(s)\lambda_{i'j'}(t)]$, $i, i' = 1, \ldots, n$, we anticipate following condition will be sufficient:

[C2] There exist constants $0 < c_0 < C_0$ such that $c_0 \leq \rho_{ij,i'j',2}(s,t) \leq C_0$ for any s, t, i, i', j, j'.

2.4 Principal component score estimation

Having estimated the eigenvalues and eigenfunctions, the next step is to estimate the functional principal component (FPC) scores. Traditional approaches in FDA do not apply here since our latent functions $\lambda_{ij}(\cdot)$'s are not observable. To overcome this challenge, we propose a new procedure based on conditional likelihood. We will focus on estimation of the FPC scores at the account level, i.e., $\boldsymbol{\xi}_i^X$'s. Similar results can be obtained for the FPC scores at the day level, i.e., $\boldsymbol{\xi}_j^Y$'s. There are typically very few data at the account-day level and $\boldsymbol{\xi}_{ij}^Z$'s will therefore be difficult to estimate.

To estimate the FPC scores for $X_i(t)$, define $N_i^X = \bigcup_{j=1}^m N_{ij}$ and $N_{-i}^X = \bigcup_{i' \neq i} \bigcup_{j=1}^m N_{i'j}$. In the Stock Trading data, N_i^X and N_{-i}^X are the aggregated trading times over m days for account i and for all accounts but account i. For ease of presentation, assuming again that $X_i(\cdot)$'s are mutually independent, then conditioned on $X_i(\cdot)$, the intensity functions of N_i^X and N_{-i}^X are $\tilde{\lambda}_i^X[t|X_i(\cdot)] = m\rho(t)\exp\left[X_i(t) - \sigma_X^2(t)/2\right]$ and $\tilde{\lambda}_{-i}^X[t|X_i(\cdot)] = (n-1)m\rho(t)$. With $X_i(t)$ approximated by (2) with estimated $\widehat{\phi}_k^X(t)$'s, given presence of an event at t, the probability for it to be from account i is

$$\pi_i^X(t; \boldsymbol{\xi}_i^X) \equiv \frac{\tilde{\lambda}_i^X[t|X_i(\cdot)]}{\tilde{\lambda}_i^X[t|X_i(\cdot)] + \tilde{\lambda}_{-i}^X[t|X_i(\cdot)]} = \frac{\exp\left\{\sum_{k=1}^{p_X} \xi_{ik}^X \widehat{\phi}_k^X(t)\right\}}{\exp\left\{\sum_{k=1}^{p_X} \xi_{ik}^X \widehat{\phi}_k^X(t)\right\} + (n-1)\exp\left\{\widehat{\sigma}_X^2(t)/2\right\}},$$

where $\widehat{\sigma}_X^2(t) = \widehat{C}_{X,h}(t,t)$ is as in (5). We estimate $\boldsymbol{\xi}_i^X$ by maximizing the conditional likelihood

$$L_i^X(\boldsymbol{\xi}_i^X) = \prod\nolimits_{s \in N_i^X} \pi_i^X(s; \boldsymbol{\xi}_i^X) \times \prod\nolimits_{s \in N_{-i}^X} \left\{1 - \pi_i^X(s; \boldsymbol{\xi}_i^X)\right\}. \tag{6}$$

Note that $\rho(t)$, and consequently $\lambda_0(t)$, do not need to be estimated. Denote $\widehat{\boldsymbol{\xi}}_i^X$ as the maximizer of (6), we conjecture that

Conjecture 2.4. *Under suitable conditions and C1,* $\frac{1}{n}\sum_{i=1}^n \|\widehat{\boldsymbol{\xi}}_i^X - \boldsymbol{\xi}_i^X\|_{\max} \xrightarrow{p} 0$ *as* $n, m \to \infty$.

Although we assumed independence among $X_i(\cdot)$'s when motivating (6), we anticipate Conjecture 2.4 to hold under weak dependence required by condition C1. The proof should follow similarly as those in Li and Guan (2014) where the intensity function was approximated using B-splines while in our case it was approximated by (2). Three challenges remain in proving Conjecture 2.4: (1) unlike in the classical FDA setting where PCA scores can be estimated separately, all components in $\boldsymbol{\xi}_i^X$ or $\boldsymbol{\xi}_j^Y$ need to be estimated simultaneously, see, e.g., Yao et al. (2005); (2) we need to quantify the errors introduced by the plug-in $\widehat{\phi}_k^X(t)$'s and $\widehat{\sigma}_X^2(t)$; (3) we need to handle correlations among the account-level as well as the day-level intensities under condition C1.

2.5 Open questions

In the Stock Trading data, we observe two types of events from each account, i.e., buying and selling events. By modeling the underlying processes as bivariate point processes, we may gain insight on the temporal covariation of investors' buying and selling activities. To do so, assume the

decomposition (1) for both buy and sell processes. The Karhunen-Loève expansions of the random functions $X_i^b(t)$ and $X_i^s(t)$ (version of $X_i(\cdot)$ for buy and sell process, respectively) give that $X_i^b(t) = \sum_{k=1}^{p_{X^b}} \xi_{ik}^{X^b} \phi_k^{X^b}(t)$ and $X_i^s(t) = \sum_{k=1}^{p_{X^s}} \xi_{ik}^{X^s} \phi_k^{X^s}(t)$, where the random vectors $\boldsymbol{\xi}_i^{X^b} = (\xi_{i1}^{X^b}, \ldots, \xi_{ip_{x_b}}^{X^b})^T$ and $\boldsymbol{\xi}_i^{X^s} = (\xi_{i1}^{X^s}, \ldots, \xi_{ip_{x_s}}^{X^s})^T$ are jointly multivariate normal, i.e.,

$$\begin{pmatrix} \boldsymbol{\xi}_i^{X^b} \\ \boldsymbol{\xi}_i^{X^s} \end{pmatrix} \sim N\left[\begin{pmatrix} \mathbf{0} \\ \mathbf{0} \end{pmatrix}, \begin{pmatrix} \boldsymbol{\Sigma}_X^b & \boldsymbol{\Sigma}_X^{bs} \\ \boldsymbol{\Sigma}_X^{bsT} & \boldsymbol{\Sigma}_X^s \end{pmatrix} \right], \tag{7}$$

where $\boldsymbol{\Sigma}_X^b$ and $\boldsymbol{\Sigma}_X^s$ are two diagonal matrices of the respective eigenvalues, and $\boldsymbol{\Sigma}_X^{bs}$ is a $p_{X^b} \times p_{X^s}$ matrix. Similar structures can be assumed for $Y_j^b(t)$'s and $Y_j^s(t)$'s, and $Z_{ij}^b(t)$'s and $Z_{ij}^s(t)$'s.

Eigenfunctions and eigenvalues can be estimated following Subsection 2.3 for the buy and sell process separately. Parameters remain to be estimated are in the cross-covariance matrix $\boldsymbol{\Sigma}_X^{bs}$ in (7). Note that the cross-covariance kernel $Q_X(s,t) = \text{Cov}\left[X_i^b(s), X_i^s(t)\right] = [\boldsymbol{\phi}_{X^b}(s)]^T \Sigma_X^{bs} \boldsymbol{\phi}_{X^s}(t)$, where $\boldsymbol{\phi}_{X^b}(t) = (\phi_1^{X^b}(t), \cdots, \phi_{P_{X^b}}^{X^b}(t))^T$, $\boldsymbol{\phi}_{X^s}(t) = (\phi_1^{X^s}(t), \cdots, \phi_{P_{X^s}}^{X^s}(t))^T$. Therefore, to estimate Σ_X^{bs}, it suffices to find a consistent estimator of $Q_X(s,t)$, which can be done in a similar fashion as in Proposition 2.1 with some straightforward changes.

Other issues that need to be carefully addressed are (a) how do we select the bandwidth h used in (3)? (b) How many FPCs, i.e. p_X, p_Y, p_Z, do we need at each level? For the former, one approach would be to develop a new version of "least squares cross validation" in Guan (2007). For the latter, following Di et al. (2009), we can choose p_X, p_Y, p_Z based on the estimated explained variance by examining the eigenvalues of the estimated covariance kernels.

2.6 Preliminary analysis of the Stocking Trading data

We have conducted a preliminary analysis of the Stock Trading data using $10,000$ randomly chosen accounts. Figure 2(a)-(b) show the first three eigenfunctions at the account (X) and day (Y) levels for the buy process, which explain approximately 94.46% and 76.34% of the total variations, respectively. Note that the first eigenfunctions at both levels are flat, suggesting that the overall trading frequency is the most important factor in explaining the variability at each level. At the account level, the second eigenfunction decreases from the start of the day to the end with a small jump at noon. This characterizes trading accounts that have increasing/decreasing activities throughout the day. The third eigenfunction at the account level is approximately flat at zero until one abrupt jump around noon and then decreases from noon to the end of the day. This characterizes trading accounts that are more active in the afternoon and whose afternoon trading activities follow an increasing/decreasing pattern. At the day level, the second eigenfunction shows a roughly decreasing trend in the morning, characterizing days that have increasing/decreasing trading activity in the morning. The third eigenfunction rises until noon and then decreases. This characterizes trading behavior that is concentrated around noon.

The bottom panel of Figure 2 shows the PC scores in the first direction for the buy process at the day level. A larger score indicates a higher trading frequency during that day. The trading frequency first increased from Jan. 1st to May 30th, 2007, on which day the government tripled the transaction tax. Since then, the trading frequency had a slight decrease but remained high until Oct. 16th, 2007, when the SSE market index reached a historic high. After Oct. 16th, 2007, the trading frequency had a noticeable drop especially in the first month. This downward trend continued until Sept. 18th, 2008, when the Chinese government announced the abolition of the buy-side transaction tax; after this tax change trading frequency had a relatively steady recovery. This describes some interesting changes of the individual buying activity patterns responding to governmental interventions.

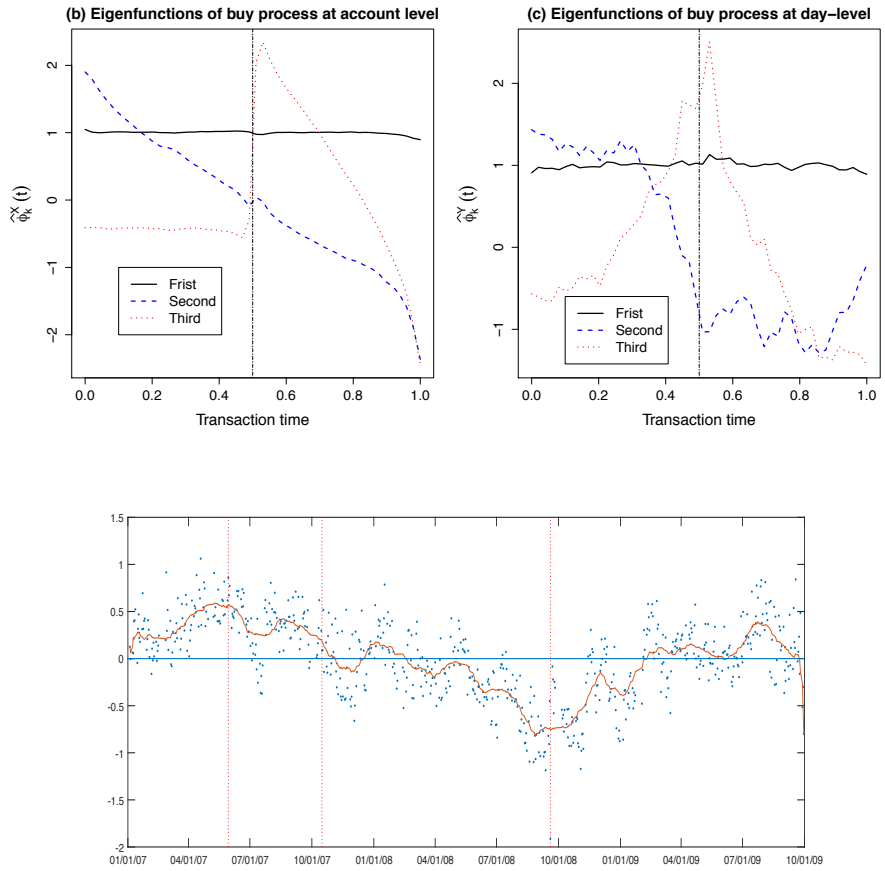

Figure 2: Top panels: Estimated eigenfunctions at account- and day-level. Bottom panel: Estimated first PC scores ξ_{j1}^{Y} at the day-level. The line (in orange) is a fitted smoothed line.

Finally, Table 1 shows the empirical correlation matrices of the estimated FPC scores from the buy and sell processes. We observe strong positive correlations between scores in the same direction and at the same level. For example, the positive correlation (0.9844) between the first buy and sell directions at the account level suggests that accounts with higher buying frequency also have higher selling frequency. The off-diagonal elements in the correlation matrix can also offer insights into the correlation between buying and selling activities. For example, the positive correlation (0.2296) between the first buying and the second selling directions implies accounts with higher overall buying frequencies are more likely to sell earlier in the day than later; the negative correlation (-0.2442) between the second buying and the first selling directions implies accounts with higher overall selling frequencies are more likely to buy later in the day than earlier. These empirically findings clearly suggest applicability of the proposed approach in Section 2.5.

3 Aim 2: Soft-clustering of multi-level functional point processes

3.1 Introduction

Clustering functional data has attracted great attention recently. Existing techniques for clustering functional data can be categorized as: (1) hard-clustering; (2) soft-clustering (Serban and Jiang, 2012). The former refers to a two-step procedure: the functional data are first summarized as a numeric vector of fixed dimension using some dimension reduction tool (e.g., FPCA) and then use some classical tools for finite dimensional data (e.g., k-means algorithm) to form clusters; see, e.g.,