Journal of Mathematical Finance, 2015, 5, 433-447
Published Online November 2015 in SciRes. http://www.scirp.org/journal/jmf
http://dx.doi.org/10.4236/jmf.2015.55037

Forecasting Density Function: Application in Finance

Rituparna Sen[1], Changie Ma[2]

[1]Applied Statistics Unit, Indian Statistical Institute, Chennai, India
[2]Capital One, Washington DC, USA
Email: rsen@isichennai.res.in, changjiema@gmail.com

Received 25 September 2015; accepted 23 November 2015; published 26 November 2015

Abstract

With increasing availability of data, in many situations it is now possible to reasonably estimate the probability density function (pdf) of a random variable. This is far more informative than using a few summary statistics like mean or variance. In this paper, we propose a method of forecasting the density function based on a time series of estimated density functions. The proposed method uses kernel estimation to pre-process the raw data followed by dimension reduction using functional principal components analysis (FPCA). Then we fit Vector ARMA models to the reduced data to make a prediction of the principal component scores, which can then be used to obtain the forecast for density function. We need to transform and scale the forecasts to ensure non-negativeness and integration to one. We compared our method to [1] for histogram forecasts, on simulated data as well as real data from S&P 500 and the Bombay Stock Exchange. The results showed that our method performed better on both the datasets and the simulation using uniform and Hilbert distance. The time dependence and complexity of density function are different for the two markets, which is captured by our analysis.

Keywords

Density Estimation, Functional Data Analysis, Principal Components Analysis, Vector ARMA

1. Introduction

Contemporaneous aggregation is often the only way to analyze temporal data, for example, considering the observations of a variable measured through time in a population, e.g. the monthly output of firms in a country. If the individuals considered are not the same through time, then it is not possible to deal with the longitudinal data. However if one is interested in the overall evolution of all firms, histograms or densities can still be studied. [2]

How to cite this paper: Sen, R. and Ma, C. (2015) Forecasting Density Function: Application in Finance. *Journal of Mathematical Finance*, **5**, 433-447. http://dx.doi.org/10.4236/jmf.2015.55037

displayed a time series of the weekly returns of the firms in the S&P 500, summarized by histograms. This is an interesting precedent which shows that in some cases, histograms are preferable to averages or totals. [1] used smoothing and non-parametric method to estimate and predict histogram time series on S&P 500 data. In this paper, we follow-up on this idea, but replace histograms by density estimates. Kernel density estimates are smooth estimates of the probability density function and do not depend on the choice of end-points as opposed to histograms.

Density function estimation has been widely used in many areas such as finance [3], energy market [4] and meteorology [5]. Although there is a rich literature on density estimation, there is very little work on density function forecast. The problem studied in this paper is different from the above references, in that, we consider replications of density estimates observed over time as opposed to just one density. To this end, we utilize the technique of Functional Data Analysis (FDA).

FDA is a popular statistical technique that treats entire curves as units of data (see [6] for an introduction). The theory of estimation of individual functions from discrete observations is quite well developed and these curves are eventually used in various applications like regression, clustering, time series etc.

The difference between a general function and a density function lies in the fact that density functions are nonnegative everywhere and their integral over the whole space is always equal to one. These restrictions pose challenges to using functional data methodology to densities directly. [7] [8] used a differential equation method to fit a smooth and monotone function and with the exponential transformation and post hoc procedure to finally obtain a fitted density function. [9] used FPCA to fit density functions and viewed them as independent and identically distributed to make statistical inference. Other possibilities are to deal with the cumulative density function or quantile function (see [10]) instead of density, which also requires addressing challenges like smoothness, monotonicity and fixed support etc.

The paper is organized as follows. In Section 2 an overview of the methodology used in this paper is given, followed by a detail introduction of the three main statistical methods used: kernel estimation, Functional Principal Component Analysis(FPCA) and Vector Autoregression and Moving Average (VARMA). A simulation analysis is conducted on Section 3 to validate the performance of the proposed method and compare it to the performance of the method of [1] for prediction with a histogram time series using uniform and Hilbert norm distances. Two applications on main stock index of U.S. and India-S&P 500 and Bombay Stock Exchange are presented in Section 4 and 5 respectively. We present our conclusions in Section 6.

2. Methodology

Each dataset consists a set of n_t observations from a distribution F_t and density f_t for $t=1,\cdots,T$. The aim is to obtain a forecast for the distribution f_{T+1}. For each dataset, first we estimate a density function for each time point $\hat{f}_t$ using kernel based estimation method as described in detail in section 2.1. In the next step, we do FPCA on the density functions as explained in section 2.2. In section 2.3, we outline the modeling of the principal component scores ξ_{ti}'s as a multivariate time series and fitting a VARMA (*p*, *q*) model. The asymptotic results for time series of functional data are available in [11]. Since we are dealing with density functions, we need extra steps of transformation and normalization, so that the forecast functions are nonnegative and integrate to one. These are explained in section 2.4. The specification of distance between two functions is handled in section 2.5.

2.1. Kernel Density Estimation

Suppose $X_1, X_2, \cdots, X_n \sim_{iid} F$ and *F* is differentiable cumulative distribution function, that is $F(x)=P(X\le x)$. We are interested in estimating the probability density function (pdf) $f=F'$. In the rest of the paper we use density to signify the pdf.

As a motivation of the following method, observe that:

$$f(x)=\lim_{h\to 0}\frac{F\left(x+\frac{1}{2}h\right)-F\left(x-\frac{1}{2}h\right)}{h}\approx\frac{1}{h}\left[F\left(x+\frac{1}{2}h\right)-F\left(x-\frac{1}{2}h\right)\right],$$

where the approximation holds for “small” *h*. Replacing F by the empirical distribution and *h* by a sequence

$h_n = o(1)$ as $n \to \infty$ we obtain the estimator:

$$\hat{f}_n(x) = \frac{1}{h_n}\left[F_n\left(x+\frac{h_n}{2}\right) - F_n\left(x-\frac{h_n}{2}\right)\right] = \frac{1}{nh_n}\sum_{i=1}^{n}\mathbf{1}_{\left[-\frac{1}{2},\frac{1}{2}\right)}\left(\frac{x-X_i}{h_n}\right). \tag{1}$$

This density estimator can be understood as follows. First observe that the discrete empirical distribution gives mass $\frac{1}{n}$ at X_i. The point mass is replaced by a uniform distribution centered at X_i, with support of length h_n. These uniform distributions are averaged to obtain the proposed estimator. Instead of spreading the mass uniformly, an arbitrary pdf $\frac{1}{h_n}K\left(\frac{x-X_i}{h_n}\right)$ centered at X_i can be used. This leads to a class of estimators of the form:

$$\hat{f}_n(x) = \frac{1}{h_n}\sum_{i=1}^{n}K\left(\frac{x-X_i}{h_n}\right), \tag{2}$$

where K is a pdf, *i.e.* $K(x) > 0$, for all x and $\int_{-\infty}^{\infty}K(u)\,\mathrm{d}u = 1$. Such as estimator is called kernel estimator, with the kernel K and bandwidth h_n, both chosen by the user. Observe that $\hat{f}$ automatically is a pdf, as we have $\hat{f}_n(x) \geq 0$ for all *x*, and the integration of $\hat{f}_n(x)$ over the real line is 1.

The choice of the kernel usually is not that crucial. The estimator in (1) is a special case with $K = \mathbf{1}_{\left[-\frac{1}{2},\frac{1}{2}\right)}$, the uniform kernel. Some common choices for *K* are Uniform, Normal, Logistic and Epanechnikov.

The bandwidth h_n is also called the smoothing parameter, as it determines how far the mass is spread out locally and hence how smooth the resulting estimator becomes. The choice of the bandwidth has a big influence on the performance of the estimator. For instance, for the uniform kernel, if the bandwidth is very large, the estimator would be flat. On the other hand, as the bandwidth becomes smaller, $\hat{f}_n$ consists of larger spikes around the observations.As estimators of a pdf, both of the extreme cases are undesirable. The bias increases with increasing bandwidth, and the variance increases with decreasing bandwidth. This trade off indicates that an optimal choice of the bandwidth balancing between bias and variance, can be found through a minimization problem using standards like mean squared error. Under certain mild regularity condition, if $h_n = o(1)$ as $n \to \infty$, then the kernel estimator as given in (2) satisfies:

$$\mathrm{MSE}\left[\hat{f}_n(x)\right] = h_n^4\left[\frac{\sigma_K^4\left(f''(x)\right)^2}{4} + o(1)\right] + \frac{1}{nh_n}\left[\|K\|_2^2 f(x) + o(1)\right], \tag{3}$$

where $\sigma_K^2 = \int_{-\infty}^{\infty}u^2K(u)\,\mathrm{d}u < \infty$ and $\|K\|^2 = \int_{-\infty}^{\infty}K^2(u)\,\mathrm{d}u < \infty$. Therefore if $f''(x) \neq 0$ then this asymptotic MSE in (3) is minimized for $h_n = n^{-\frac{1}{5}}t_0$ with $t_0 = \left(\frac{\|K\|_2^2 f(x)}{\sigma_K^4\left(f''(x)\right)^2}\right)^{\frac{1}{5}}$.

2.2. Principal Component Analysis of Functional Data

Consider a sample of *T* smooth random trajectories $\left(f_t(x)\right)_{x\in[\alpha_1,\alpha_2]}, t = 1,\cdots,T$ generated from a process *f*. Following [12], throughout we assume that *f* is an element of the Hilbert space $\mathcal{H} := L^2(\mathcal{T})$ endowed with the inner product $\langle f,g\rangle_{\mathcal{H}} = \int_{\mathcal{T}} f(t)g(t)\,\mathrm{d}t$ and the norm $\|f\| = \sqrt{\langle f,f\rangle_{\mathcal{H}}} < \infty$ a.s. The sample trajectories are modeled as realization of a stochastic process $f(x)$ that has mean $\mathrm{E}\left[f(x)\right] = \mu_f(x)$ and covariance function $\mathrm{cov}\left(f(x), f(y)\right) = G(x,y)$. There is an orthogonal expansion of *G* in terms of eigenfunctions ϕ_i and non-increasing eigenvalues λ_i as:

$$G(x,y)=\sum_i \lambda_i \phi_i(x)\phi_i(y), \quad x,y\in[\alpha_1,\alpha_2],$$

where $[\alpha_1,\alpha_2]$ are the domain of each trajectory. The Karhunen Loève theorem then provides a representation of individual random pdfs, given by

$$f_t(x)=\mu_f(x)+\sum_i \xi_{ti}\phi_i(x), x\in[\alpha_1,\alpha_2], t=1,\cdots,T, \tag{4}$$

where the ξ_i are uncorrelated random variables with zero mean and variance $\mathrm{E}\left[\xi_i^2\right]=\lambda_i$, where $\sum_i \lambda_i<\infty$. The deviation of each sample trajectory from the mean is thus a sum of orthogonal curves with uncorrelated random amplitudes.

Often it is realistic to incorporate uncorrelated measurement errors with mean zero and constant variance σ^2 into the model, reflecting additional variation in the measurements, compare [13]. Let $\tilde{f}_t(x_{tj})$ be the observations of the random function $f_t(\cdot)$ at time grids x_{tj} and W_{tj} additional measurement errors are assumed to be i.i.d and independent of the random coefficients $\xi_{ti}, t=1,\cdots,T, j=1,\cdots,m_t$ *i.e.*,

$$\tilde{f}_t(x_{tj})=f_t(x_{tj})+W_{tj}=\mu_f(x_{tj})+\sum_i \xi_{ti}\phi_i(x_{tj})+W_{tj}, \quad x_{tj}\in[\alpha_1,\alpha_2], \tag{5}$$

where $\mathrm{E}\left[W_{tj}\right]=0, \mathrm{var}\left(W_{tj}\right)=\sigma^2$. In special cases, one might assume in addition that the ξ_{ti}, W_{tj} are all jointly normally distributed, but generally we do not make such assumption.

Under Equation (5) and with indicator function $I(.)$, we can get:

$$\mathrm{E}\left(\tilde{f}(x)\right)=\mu_f(x), \mathrm{Cov}\left(\tilde{f}(x),\tilde{f}(y)\right)=G(x,y)+\sigma^2 I(x=y).$$

This implies that the smooth mean function $\mu_f(x)$ and the smooth covariance function $G(x,y)$ can be consistently estimated from pooling the sample of T trajectories and smoothing the resulting scatter plot. Well known procedure exists to infer eigenfunctions and eigenvalues [14].

Processes *f* are then approximated by substituting estimates and using a chosen finite number of principal components. The specific number of principal components to be retained in the model is chosen by some optimization criterion like cross-validation, AIC, BIC or a scree plot.

2.3. VARMA Modeling

A sequence $\left(f_t=f_t(\alpha), \alpha\in\mathcal{T}; 0<t\le T\right)$ of random functions with values in $\mathcal{H}$ is said to follow ARH (1) if it is stationary and such that

$$f_t=\theta(f_{t-1})+\epsilon_t, \tag{6}$$

where $\left(\epsilon_t(\alpha), \alpha\in\mathcal{T}, 0<t\le T\right)$ is an $\mathcal{H}$ white noise and the operator $\theta:\mathcal{H}\to\mathcal{H}$ is linear and compact.

A higher order of autoregression process-ARH (p) (see [15]), could now be defined as:

$$f_t=\theta_1(f_{t-1})+\theta_2(f_{t-2})+\cdots+\theta_p(f_{i-p})+\epsilon_t.$$

A natural extension would be to consider the series of functions follows the ARMAH (p, q) model with mean $\mu\in\mathcal{H}$:

$$f_t(.)-\mu=\theta_1\left(f_{t-1}(.)-\mu\right)+\cdots+\theta_p\left(f_{t-p}(.)-\mu\right)+\epsilon_t(.), \tag{7}$$

where

$$\epsilon_t(.)=\eta_t(.)+\varphi_1\eta_{t-1}(.)+\cdots+\varphi_q\eta_{t-q}(i),$$

$\eta_t(.)$ is $\mathcal{H}$ white noise and $\theta_1,\cdots,\theta_p$ are linear functions. The expansion in Equation (4) is still valid, as long as the process is second-order stationary. Combining (4) and (7) we have:

$$\mu_f+\sum_i \xi_{ti}\phi_i(.)-\mu=\theta_1\left(\mu_f+\sum_i \xi_{(t-1)i}\phi_i(.)-\mu\right)+\cdots+\theta_p\left(\mu_f+\sum_i \xi_{(t-p)i}\phi_i(.)-\mu\right)+\epsilon_t(.).$$

Using linearity of $\theta_1,\cdots,\theta_p$, this implies:

$$\mu_f+\sum_i \xi_{ti}\phi_i(.)-\mu=\sum_i \xi_{(t-1)i}\theta_1\phi_i(.)+\theta_1(\mu_f-\mu)+\cdots+\sum_i \xi_{(t-p)i}\theta_p\phi_i(.)+\theta_p(\mu_f-\mu)+\epsilon_t(.).$$

Combining all the terms involving μ and μ_f into $\tilde{\mu}$ and using vector representation, we have the following result:

$$\Phi(.)\Xi_t=\tilde{\mu}+\theta_1(\Phi(.))\Xi_{t-1}+\cdots+\theta_p(\Phi(.))\Xi_{t-p}+\epsilon(.), \tag{8}$$

where $\Phi=(\phi_1,\phi_2,\cdots)$ and $\Xi=(\xi_{1t},\xi_{2t},\cdots)^{\mathrm{T}}$. Since the columns of are orthonormal, we can left-multiply Equation (8) by Φ^{T} to get:

$$\Xi_t=\Phi^{\mathrm{T}}\tilde{\mu}+\Phi^{\mathrm{T}}\theta_1(\Phi(.))\Xi_{t-1}+\cdots+\Phi^{\mathrm{T}}\theta_p(\Phi(.))\Xi_{t-p}+\Phi^{\mathrm{T}}\epsilon(.),$$

which implies a VARMA (*p*, *q*) structure on the vector of principal component scores Ξ. We can use this theory to model the time series structure of functions as time series structure of the first few principal component scores, which largely reduces the complexity of the problem.

2.4. Rescaling and Non-Negativeness

As mentioned before, the difference between density function estimation and general function estimation lies in that density function are required to be non-negative everywhere and integrate to one.

However, the fitted function after FPCA estimation is not guaranteed to be positive everywhere. To address this, we took logarithm transformation of the fitted kernel density function before the FPCA estimation and used exponential transformation after the FPCA to guarantee the non-negativeness. In order to ensure that the fitted function integrates to one, we referred to [7] and imposed post hoc to rescale the fitted function.

2.5. Distance Measure between Two Functions

To compare the performance of FDA method and Arroyo's method, we used two different distance measures between predicted functions and actual functions. These are the uniform distance D_U and the Hilbert distance D_H, defined as follows:

$$D_U(f,\hat{f})=\frac{\int\left[f(x)-\hat{f}(x)\right]^2\mathrm{d}x}{\int f(x)^2\mathrm{d}x+\int\hat{f}(x)^2\mathrm{d}x}$$

and

$$D_H(f,\hat{f})=\frac{\sup_x\left|f(x)-\hat{f}(x)\right|}{\sup_x f(x)},$$

where *f* is the actual function and $\hat{f}$ is the predicted function.

3. Comparison to Existing Methods

In [1], the authors implemented exponential smoothing and *k* nearest neighbor methods to address the histogram time series (HTS) prediction problem. The authors defined the histogram data as:

$$h_{X_i}=\left\{\left([x]_{i1},\pi_{i1}\right),\cdots,\left([x]_{in_i},\pi_{in_i}\right)\right\},\ \text{for } i=1,2,\cdots,m,$$

where π_{ij} is the frequency for interval $[x]_{ij}$. The distance between two histogram is defined as

$$MDE^q\left(\{h_{X_t}\},\{\hat{h}_{X_t}\}\right)=\left(\frac{\sum_{t=1}^{T}D_X^q\left(h_{X_t},\hat{h}_{X_t}\right)}{T}\right)^{\frac{1}{q}},$$

where $q=1$ or 2, is the order and $D\left(h_{X_t},\hat{h}_{X_t}\right)$ is a distance measure such as Wasserstein or the Mallows distance. It is assumed that the data points are uniformly distributed within each bin of the histogram. Under this assumption, the CDF $H_X(x)$ of a histogram $h_X=\left\{\left([x]_{i_X},\pi_{i_X}\right)\right\}$, with $i_X=1,2,\cdots,n_X$ is defined as:

$$H_X(x)=\int_{-\infty}^{x}h_X(x)\,\mathrm{d}x=\begin{cases}0, & \text{if } x<\underline{x}_1;\\ w_{i_X-1}+\dfrac{x-\underline{x}_{i_X}}{\overline{x}_{i_X}-\underline{x}_{i_X}}, & \text{if } x\in\left[\underline{x}_{i_X},\overline{x}_{i_X}\right);\\ 1, & \text{if } x\geq\overline{x}_{n_X},\end{cases}$$

where $w_i=\sum_{j=1}^{i}\pi_j$ is the cumulative weight associated with the interval i.

By using this definition of the CDF of a histogram, the Wasserstein and Mallows distances formula can be written as functions of the centers and radii of the histogram bins:

$$D_W\left(h_X,h_Y\right)=\sum_{j=1}^{n}\pi_j\left|x_{C_j}-y_{C_j}\right|$$

$$D_M^2\left(h_X,h_Y\right)=\sum_{j=1}^{n}\pi_j\left[\left(x_{C_j}-y_{C_j}\right)^2+\frac{1}{3}\left(x_{R_j}-y_{R_j}\right)^2\right].$$

3.1. Exponential Smoothing

The idea of exponential smoothing is to predict the next observation by a weighted average of previous observation and its estimate. Let h_{X_t} $t=1,2,\cdots,T$ be a histogram time series, the exponential smoothing forecast is given by:

$$\hat{h}_{X_{t+1}}=\alpha h_{X_t}+(1-\alpha)\hat{h}_{X_t}.$$

The authors show that the forecast is also the solution to the following optimization problem:

$$\hat{h}_{X_{t+1}}\equiv\arg\min_{\hat{h}_{X_{t+1}}}\left(\alpha D^2\left(\hat{h}_{X_{t+1}},h_{X_t}\right)+(1-\alpha)D^2\left(\hat{h}_{X_{t+1}},\hat{h}_{X_t}\right)\right),$$

where $D(\cdot,\cdot)$ is the Mallows distance. The use of the Wasserstein distance is not suitable in this case because of the properties of the median, which will ignore the weighting scheme. For t large, the exponential smoothing formula can be approximated by:

$$\hat{h}_{X_{t+1}}\approx\sum_{j=1}^{t}\alpha(1-\alpha)^{j-1}h_{X_{t-(j-1)}}. \tag{9}$$

In the analysis below, we let $\hat{h}_{X_1}=0$ and used the training data to estimate the α. Subsequently we plug in the estimated α in the approximation rule in Equation (9) to get the prediction.

3.2. *k*-Nearest Neighbor

The *k*-Nearest Neighbor (*k*-NN) method is a classic pattern recognition procedure that can be used for time series forecasting. The *k*-NN forecasting method in classic time series consists of two steps: identification of the *k* sequences in the time series that are more similar to the current one, and computation of the forecast as the weighted average of the sequences determined in the previous step.

The adaptation of the *k*-NN method to forecast HTS can be described in the following steps:

1) The HTS, $\left\{h_{X_t}\right\}$ with $t=1,\cdots,T$, is organized as a series of *d*-dimensional histogram valued vectors $\left\{h_{X_t}^d\right\}$, where

$$h_{X_t}^d = \left(h_{X_t}, h_{X_{t-1}}, \cdots, h_{X_{t-(d-1)}}\right)',$$

where $d \in \mathbb{N}$ is the number of lags and $t = d, \cdots, T$.

2) The dissimilarity between the most recent histogram valued vector $h_{X_T}^d$ and the rest of the vectors $h_{X_t}^d$ is computed by implementing the following distance measure

$$D_t\left(h_{X_T}^d, h_{X_t}^d\right) = \left(\frac{\sum_{i=1}^{d}\left(D^q\left(h_{X_{T-i+1}}, h_{X_{t-i+1}}\right)\right)}{d}\right)^{\frac{1}{q}},$$

where $D^q\left(h_{X_{T-i+1}}, h_{X_{t-i+1}}\right)$ is the Mallows or the Wasserstein distance of order q.

3) Once the dissimilarity measures are computed for each $h_{X_t}^d, t = T-1, T-2, \cdots, d$, we select the k-closest vectors to $h_{X_T}^d$. These vectors are denoted by $h_{X_{T_1}}^d, h_{X_{T_2}}^d, \cdots, h_{X_{T_k}}^d$.

4) Given the k-closest vectors, their subsequent values, $h_{X_{T_1+1}}, h_{X_{T_2+1}}, \cdots, h_{X_{T_k+1}}$ are averaged by means of the barycenter approach to obtain the final forecast $\hat{h}_{X_{T+1}}$ in the following minimization problem:

$$\hat{h}_{X_{T+1}} \equiv \arg\min_{\hat{h}_{X_{T+1}}}\left[\sum_{p=1}^{k} w_p D^q\left(\hat{h}_{X_{T+1}}, h_{X_{T_p+1}}\right)\right]^{\frac{1}{q}},$$

where w_p is the weight assigned to the neighbor *p*, with $w_p \geq 0$ and $\sum_{p=1}^{k} w_p = 1$. For example, the weights may be assumed to be equal for all the neighbors or inversely proportional to the distance between the last sequence $h_{X_T}^d$ and the considered sequence $h_{X_{T_p}}^d$.

In the analysis, we used equal weights when performing the minimization. The optimal parameter $\hat{k}$ and $\hat{d}$, which minimize the mean distance error defined in the previous section in the estimation period, are obtained by conducting a two-dimensional grid search. Then the estimated parameter are plugged in the whole procedure again to get the prediction.

3.3. Simulation Results

Simulation was carried out to compare the performance of the proposed FDA method to the method of [1] for prediction with a histogram time series using uniform and Hilbert norm distances.

The data was simulated following Autoregressive Hilbertian (ARH) process as described in Equation (6).

Suppose $\beta : \mathcal{T}^2 \to \mathcal{R}$ satisfies $\int_{\mathcal{T}}\int_{\mathcal{T}} \beta^2(s,t)\,\mathrm{d}s\mathrm{d}t < \infty$. We can define the operator $\theta : \mathcal{H} \to \mathcal{H}$ by the kernel β in the following way:

$$\theta(x) = \int_{\mathcal{T}} \beta(\alpha, t)\, x(t)\,\mathrm{d}t, \quad \text{for } \alpha \in \mathcal{T}.$$

Specifically, in our simulation, we used $\mathcal{T} = [0,1]$ and the following $\beta(s,t)$ function with *h* as bandwidth:

$$\beta(s,t) = \begin{cases} \left(-\log\left(\left|\frac{t-s}{h}\right|\right) - 2\log\left(1 - \left|\frac{t-s}{h}\right|\right)\right)\Big/6h, & \text{if } 0 < \left|\frac{t-s}{h}\right| < 1, \\ 0, & \text{if } \left|\frac{t-s}{h}\right| \geq 1, \\ 100, & \text{if } t-s=0 \end{cases} \tag{10}$$

Then our simulation consisted of the following steps:

- Considered 16 different initial density functions ξ_1: all beta distributed $\left(\text{Beta}(\alpha,\beta)\right)$ with integer α

values from 2 to 5 and integer β values from 2 to 5. Also, considered 3 different bandwidths of h in Equation (10) as 0.05. 0.08 and 0.1. Therefore, there are totally 48 combinations.

- Used Equation (6) with $\epsilon_i(t)$'s being i.i.d normally distributed with standard deviation equals to 0.05 to simulate 48 ARH (1) process each with length T equals to 250.
- Used FDA method and Arroyo's method to fit models on the first 200 density functions to predict the next 50 density functions.
- Evaluated the performance of FDA method and compare the performance of FDA method and Arroyo's method.

The performance evaluation of FDA method and comparison with Arroyo's method are shown in **Table 1**. We observe that:

- Most of the time (40 out of 48, 83%), FDA method chose the correct underlying process (ARH (1)).
- The choice of number of principal components varied.
- The FDA method outperformed Arroyo's to a great extent in all metrics and both in uniform measure and Hilbert's measure. Specifically, using uniform distance, FDA method is 90% less in average mean distance and 24% less in average standard deviation of distance than Arroyo's method; using Hilbert's distance, FDA method is 92% less in average mean distance and 89% less in average standard deviation of distance than Arroyo's method.

4. S&P 500 Data Analysis

The Standard & Poor's 500 (S&P 500) is a free-float capitalization-weighted (movements in the prices of stocks with higher market capitalizations have a greater effect on the index than companies with smaller market caps) index of the prices of 500 large-cap common stocks actively traded in the United States. It has been widely regarded as the best single gauge of the large cap U.S. equities market since the index was first published in 1957. The stocks included in the S&P 500 are those of large publicly held companies that trade on either of the two largest American stock market exchanges: the New York Stock Exchange and the NASDAQ. These 500 large-cap American companies included in S&P 500 capture about 75% coverage of the American equity market by capitalization. It covers various leading industries in United States, including energy (e.g. including companies like Exxon Mobil Corp.), materials (e.g. Dow Chemical), industrials (e.g. General Electric Co.), consumer discretionary (e.g. McDonald's Corp.), consumer staples (e.g. Procter & Gamble), health care (e.g. Johnson & Johnson), financials (e.g. JPMorgan Chase & Co.), information technology (e.g. Apple Inc.), telecommunication services (e.g. AT&T Inc.), and utilities (e.g. PG&E Corp.). Though the list of the 500 companies is fairly stable, Standard & Poors does update the components of the S&P 500 periodically, typically in response to acquisitions, or to keep the index up to date as various companies grow or shrink in value. For example, TRIP (TripAdvisor Inc.) was added to replace TLAB (Tellabs Inc.) on Dec 20, 2011 due to the fact that Expedia Inc. spun off TripAdvisor Inc and WPX (WPX Energy Inc.) was added to replace CPWR (Compuware) on Dec 31, 2011 due to market cap changes.

The dataset we have is daily returns of all the constituents of the S&P 500 for 245 days from August 21, 2009 to August 20, 2010. This is the same data used by [1] and can be downloaded at http://pages.swcp.com/stocks/. **Figure 1** shows the histogram of the first 4 days returns of all constituents. The first 3 days' histograms look like a bell shape, indicating possibly normal distribution while the fourth day's histogram is very centralized.

4.1. Kernel Density Estimation

After using the ksdensity function of Matlab on the S&P 500 data, we found out that over 40% of the fitted density function contains many extremely small (less than 0.0001) probability points, no matter how big bandwidth is, mainly due to some extreme returns each day. Example of fitted density functions that contain

Table 1. Summary of performance comparison between the FDA method and arroyo's method on simulation results using both uniform distance and Hilbert's distance. U-Uniform distance; H-Hilbert's distance.

Method	Avg. Mean (U)	Avg. SD. (U)	Avg. Mean (H)	Avg. SD. (H)
FDA	0.0874	0.0076	0.3662	0.0222
Arroyo	0.8405	0.0103	4.552	0.2109

many extremely small probability points can be seen on **Figure 2**. Retaining these points is a problem since we need to take logarithm of the extremely small numbers, which will make the fitting procedure later non-applicable.

Therefore, we drop the top 5% and bottom 5% fitted density points and keep the other 90% of it. After using this procedure, we get rid of the extremely small probability points problem completely. However, one thing we need to keep in mind is that the method in this section cannot be used for problems in extremes like value-at-risk and expected shortfall.

4.2. Principal Component Analysis of Functional Data

We use the PACE program in MATLAB ([16]) for this step. The program has AIC, BIC, and FVE (fraction of variance explained) method to determine to number of principal component functions. We observe that the AIC or BIC based method is not conservative enough and includes some extra functions which are very generic. Therefore, we decided to use FVE method and use scree plot to select the optimal number of principal component functions. See **Figure 3** for the scree plot of the fitting of the S&P 500 data. Based on the scree plot, we chose 2 components for the fitting.

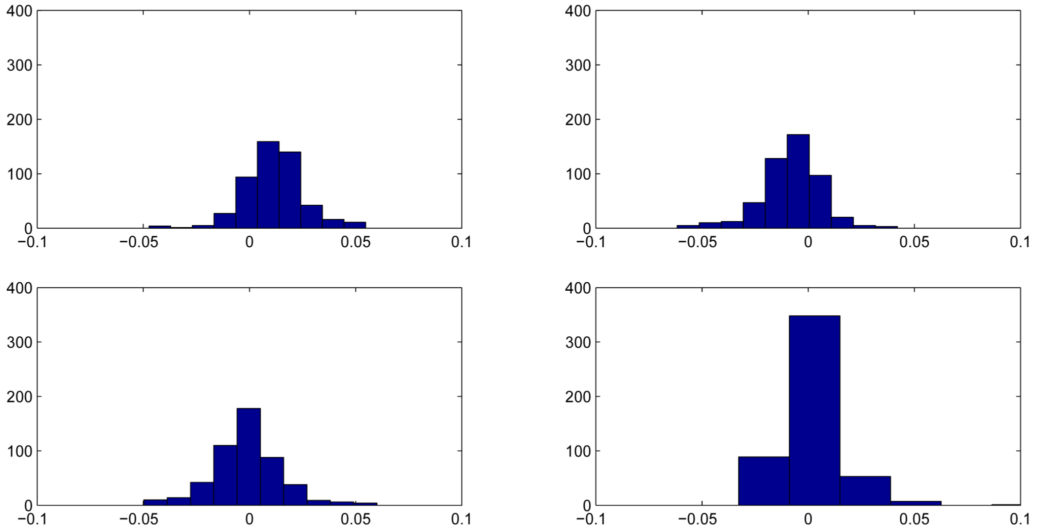

Figure 1. Histogram of daily returns of all constituents of S&P 500 between 21-Aug-2009 and 26-Aug-2009.

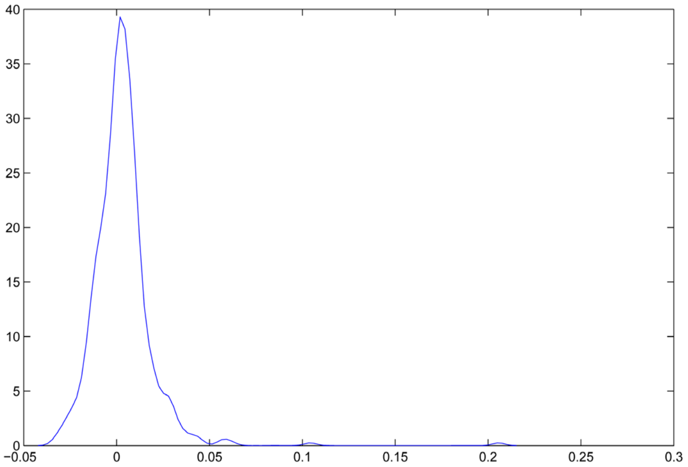

Figure 2. Fitted density function that contains many extremely small probability points of S&P 500-1.

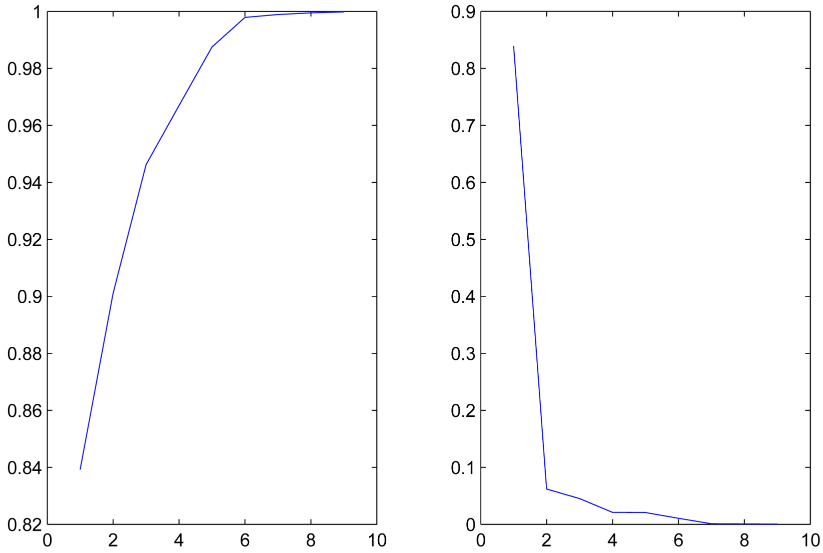

Figure 3. Scree plot of the fitting of the S&P 500 data. Left: Cumulative variance; Right: Variance.

4.3. VARMA Modeling

We fitted multiple VARMA models of different order using Maximum Likelihood, Yule-Walker estimation methods as well as state-space models. These are not presented here, but are available from the authors on request. We observe that:

- VARMA (1,1) is the best model when considering either AIC or BIC.
- The AIC/BIC performance are usually better when using the Maximum Likelihood Estimation approach than the Yule-Walker approach, except when the model considered is VAR (1) model. In that case, the AIC/BIC performance are the same for both approaches.
- The most promising procedure of state space model fitting in this data set is the brute force technique.

The daily S&P 500 Data has been reduced to a 2 dimensional time series in the previous procedure. Therefore, a VARMA (1,1) model only needs to estimate $4 \times 2 = 8$ parameters, which has good power and accuracy. Based on this and the corresponding AIC/BIC performance, we decided to use VARMA (1,1) model for this data.

4.4. Comparison between FDA Method and Arroyo's Method on S&P 500 Data

To compare the prediction result of FDA method and Arroyo's method, we divided the S&P 500 sample into 185 days (around 75% of all data) as training period and 60 days (around 25% of all data) as prediction period. In the *k*-NN procedure, we also kept away the first 50 days' data from the training period, since the estimation needs to begin with more data when *k* and *d* are large.

We used the training data to fit the FPCA model and obtained the corresponding 2 estimated principal component functions, the mean function, and the estimated principal component scores. Then we used VARMA (1,1) model of the principal component scores for next-day prediction. After getting the next-day prediction of principal component scores for 60 days, we combined those with the principal component functions and mean function obtained in previous training steps to get the predicted densities for each of the 60 days. Finally, we used Uniform Norm and Hilbert Norm to measure the distance between the predicted densities and the original densities. The distance between the predicted densities and the original densities (in both histogram and kernel form) using Arroyo's method are also computed for comparison.

The time series plot of the Hilbert Norm distance of the 60 Days' prediction period is shown in **Figure 4**. From the plot, we can clearly see that the one using FDA method outperforms the one using Arroyo's method, using distance between the original density and the predicted density. It not only has a small value of distance on almost every day, but also has more stable performance.

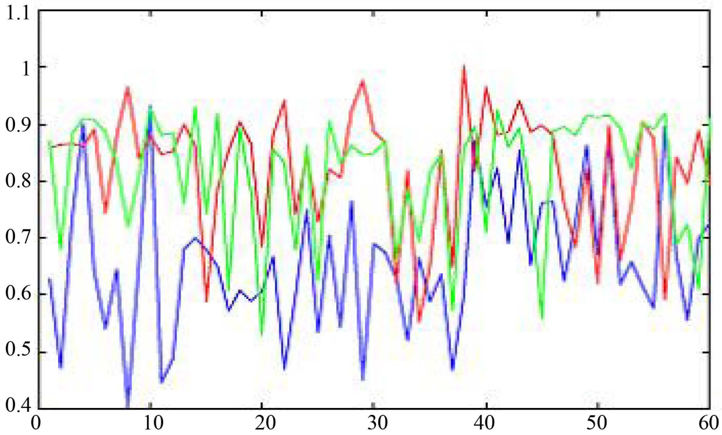

Figure 4. Time series plot of hilbert norm distance of the 60 days' prediction period on S&P 500 data. Blue: FDA method; Red: Arroyo's with exponential smoothing; Green: Arroyo's with *k*-NN.

The descriptive statistic of the Uniform Norm Distance and Hilbert Norm distance of the 60 days' prediction period is given in **Table 2** and **Table 3**. For Uniform Norm Distance measure, the distance using FDA method has smallest mean, second smallest median, and smallest standard deviation which indicates better and more stable performance. The fact that the minimum distance of the sixty days using FDA method is largest does not show any disadvantage of FDA method since we are looking at the overall performance across the 60 days. When it comes to Hilbert Norm Distance measure, the distance using FDA method has much smaller mean and median when compared with distance using Arroyo's method. The standard deviation using FDA method is the largest, which is due to the fact that FDA method has result in a couple of much smaller distance values which actually indicates good performance.

In all, from the time series plot and descriptive statistic, the overall performance of FDA method is better than Arroyo's method, in both Uniform Norm Distance measure and Hilbert Norm Distance measure.

5. BSE Data Analysis

The Bombay Stock Exchange (BSE) is a stock exchange located in Mumbai, India and is the oldest stock exchange in Asia. The equity market capitalization of the companies listed on the BSE was US$1.7 trillion as of January 2015, making it the 4th largest stock exchange in Asia and the 11th largest in the world. The BSE has the largest number of listed companies in the world with over 5500 listed companies. The dataset we had was weekly returns of 507 stocks of the BSE from from January 1997 to December 2004, totally 365 weeks. **Figure 5** shows the histogram of the first 4 weeks' returns of all the stocks. The first and third week's histograms look like a bell shape, indicating possibly normal distribution while the second and forth week's histogram is very centralized. There is also more skewness than in the S&P data.

5.1. Kernel Density Estimation

We used the same procedure as discussed in Section 1 on the BSE data. The weekly BSE data also suffers from the small probability points problem after applying the ksdensity function in Matlab (over 43% of the fitted density function contains many extremely small, less than 0.0001, probability points) and similar procedure was used to bypass this problem.

Examples of fitted density functions that contain many extremely small probability points can be seen in **Figure 6**.

5.2. Principal Component Analysis of Functional Data

FVE method of PACE package and scree plot is used again to select the optimal number of principal component functions. See **Figure 7** for the scree plot of the fitting of the BSE data. Based on the scree plot and FVE procedure, we chose 4 components for the fitting.

Table 2. Descriptive statistic of uniform norm distance of the 60 days' prediction period on S&P 500 data. (1) and (2) denote exponential smoothing and *k*-NN respectively.

Method	Mean	Median	Std. Dev.	Maximum	Minimum
FDA	0.3200	0.2756	0.1614	0.8288	0.1233
Arroyo (1)	0.3551	0.3725	0.2076	0.7272	0.0284
Arroyo (2)	0.3728	0.2676	0.2312	0.8366	0.1060

Table 3. Descriptive statistic of uniform norm distance of the 60 days' prediction period on S&P 500 data. (1) and (2) denote exponential smoothing and *k*-NN respectively.

Method	Mean	Median	Std. Dev.	Maximum	Minimum
FDA	0.6578	0.6554	0.1215	0.9327	0.4028
Arroyo (1)	0.8236	0.8564	0.1044	1.0031	0.5521
Arroyo (2)	0.8167	0.8587	0.1069	0.9301	0.5294

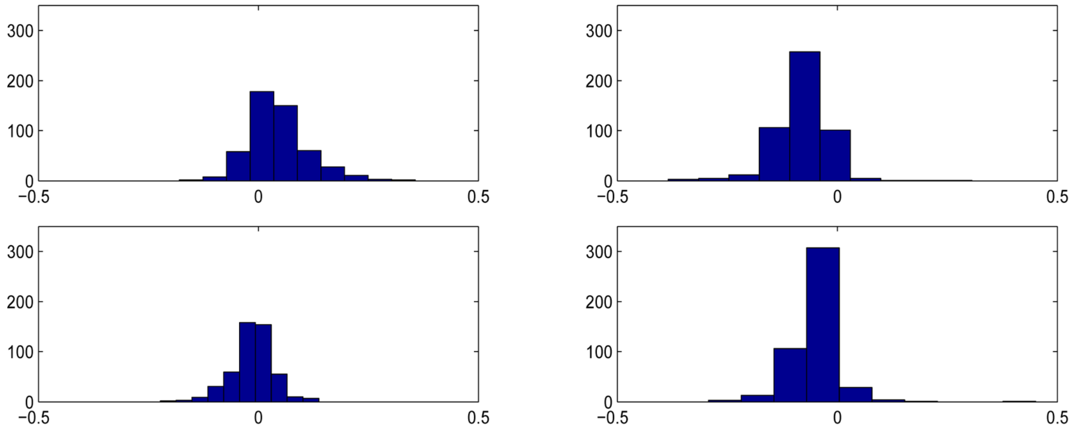

Figure 5. Histogram of weekly returns of selected 507 stocks of BSE of January 1997.

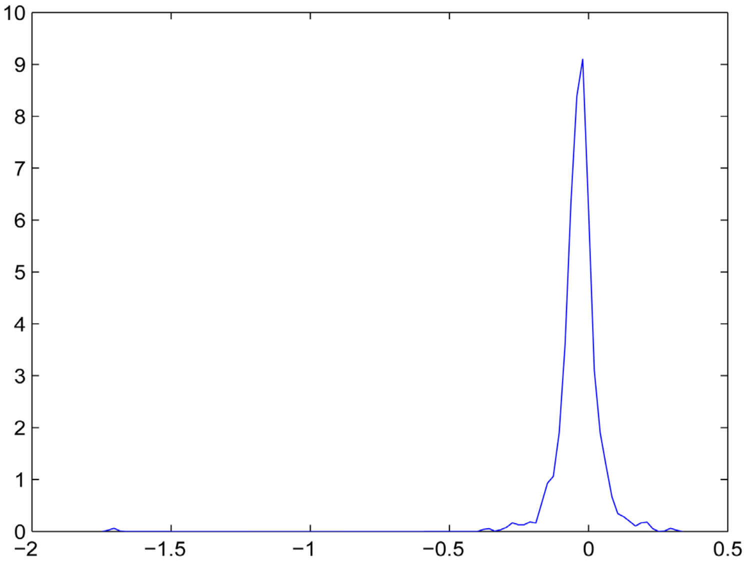

Figure 6. Fitted density function that contains many extremely small probability points of BSE.

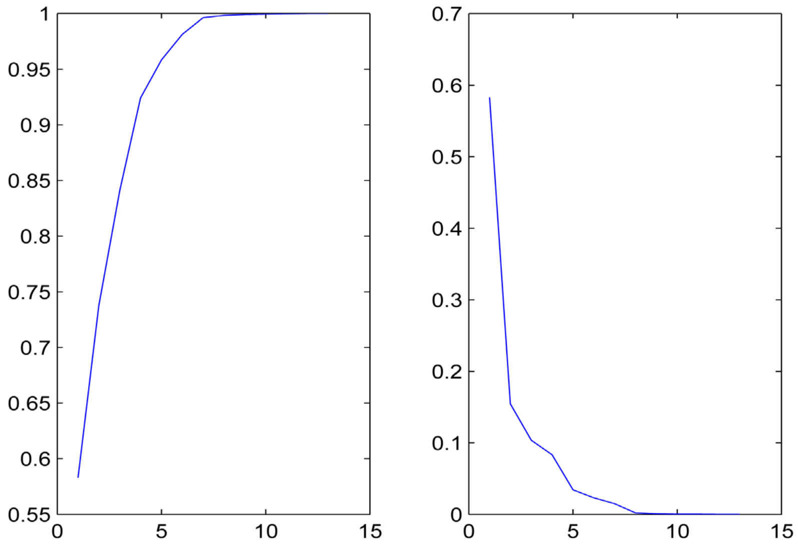

Figure 7. Scree plot of the fitting of the BSE data. Left: Cumulative variance; Right: Variance.

5.3. VARMA Modeling

We did the similar VARMA modeling analysis on the BSE Data, namely fitted multiple models using different estimation methods and compared their AIC/BIC score. We observe that:

- VAR (6) is the best model when considering AIC only.
- VAR (1) is the best model when considering BIC only.
- The AIC/BIC performance are usually better when using the Maximum Likelihood Estimation approach than the Yule-Walker approach, except when the model considered is VAR (1) model. In that case, the AIC/BIC performance are the same for both approaches.
- The most promising procedure of state space model fitting in this data set is also the brute force technique.
- Model chosen by AIC or BIC criteria has MA degree zero.

The daily BSE Data has been reduced to a 4 dimensional time series in the previous procedure. Therefore, a VAR (6) model needs to estimate $16 \times 6 = 96$ parameters while a VAR (1) model only needs to estimate 16 parameters, which has far more power and accuracy. Based on this and the corresponding AIC/BIC performance, we decided to use VAR (1) model for this data.

5.4. Comparison between FDA Method and Arroyo's Method on BSE Data

For BSE data, from time series plots (**Figure 8**) using uniform distance, we can clearly see that the one using FDA method outperforms the one using Arroyo's method, using distance between the original density and the predicted density.

The descriptive statistic of the Uniform Norm Distance and Hilbert Norm distance of the 50 days' prediction period is given in **Table 4** to **Table 5**. Under Uniform Norm Distance measure, the distance using FDA method has better performance in all metrics. When it comes to Hilbert Norm Distance measure, FDA method has best performance in terms of mean and median although it suffers from the maximum being considerably large (4.2083) and has much larger standard deviation.

In all, from the time series plot and descriptive statistic, the overall performance of FDA method is better than Arroyo's method, in both Uniform Norm Distance measure and Hilbert Norm Distance measure.

6. Conclusion

The paper proposes tools from Functional Data Analysis to forecast the probability density function. The technique is found to perform better than the method of [1] to forecast histograms in simulation and real data examples. For both real datasets, the density estimates have long tails. For the components of S&P 500, 2

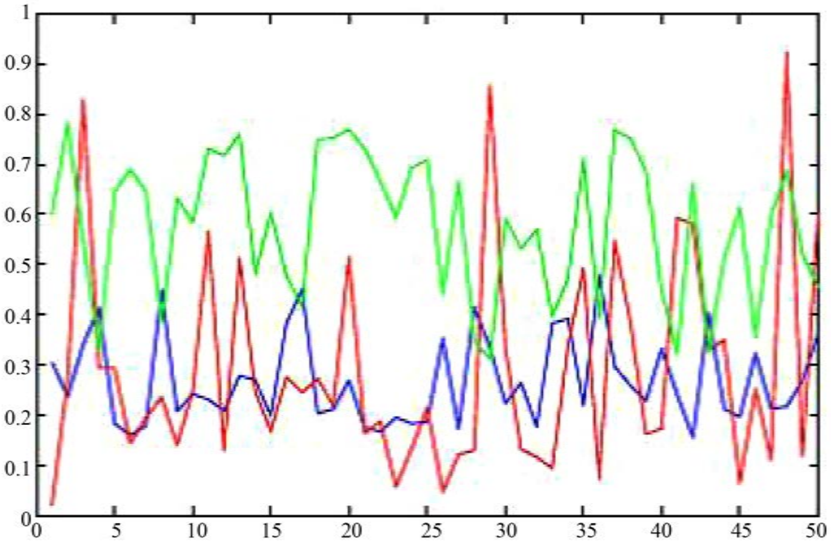

Figure 8. Time series plot of uniform norm distance of the 50 days' prediction period on BSE data. Blue: FDA method; Red: Arroyo's with exponential smoothing; Green: Arroyo's with *k*-NN.

Table 4. Descriptive statistic of uniform norm distance of the 50 days' prediction period on BSE data. (1) and (2) denote exponential smoothing and *k*-NN respectively.

Method	Mean	Median	Std. Dev.	Maximum	Minimum
FDA	0.2136	0.2263	0.0803	0.3301	0.0200
Arroyo (1)	0.2984	0.2496	0.1967	0.7199	0.0313
Arroyo (2)	0.4033	0.3458	0.2148	0.8623	0.1238

Table 5. Descriptive statistic of hilbert norm distance of the 50 days' prediction period on BSE data. (1) and (2) denote exponential smoothing and *k*-NN respectively.

Method	Mean	Median	Std. Dev.	Maximum	Minimum
FDA	0.6485	0.3715	0.7318	4.2083	0.0927
Arroyo (1)	0.8089	0.8266	0.1726	1.2892	0.4414
Arroyo (2)	0.8469	0.8750	0.1141	1.0395	0.4808

principal components are enough to explain most of the variation in the shapes of the kernel densities. For the stocks traded on the Bombay Stock Exchange, 4 principal components are required. Also, the time dependence in the first dataset is ARMA (1,1), whereas for the second it is AR (1). This reflects the variation across markets (mature vs emerging), nature of stocks (large cap vs all) and frequency of observation (daily vs weekly). The method is flexible enough to accommodate these variations. In all the real data examples, forecasts using the FDA method are more efficient than the existing method.

References

[1] Arroyo, J. and Maté, C. (2009) Forecasting Histogram Time Series with k-Nearest Neighbours Methods. *International Journal of Forecasting*, **25**, 192-207. http://dx.doi.org/10.1016/j.ijforecast.2008.07.003

[2] Gonzlez-Rivera, G., Lee, T.H. and Mishra, S. (2008) Jumps in Cross-Sectional Rank and Expected Returns: A Mixture Model. *Journal of Applied Econometrics*, **23**, 585-606. http://dx.doi.org/10.1002/jae.1015

[3] Ait-Sahalia, Y. and Lo, A., (1998) Nonparametric Estimation of State-Price Densities Implicit in Financial Asset Prices. *Journal of Finance*, **53**, 499-547. http://dx.doi.org/10.1111/0022-1082.215228

[4] Taylor, J. and Jeon, J. (2012) Using Conditional Kernel Density Estimation for Wind Power Forecasting. *Journal of the American Statistical Association*, **107**, 66-79. http://dx.doi.org/10.1080/01621459.2011.643745

[5] Carney, M., Cunningham, P., Dowling, J. and Lee, C. (2005) Predicting Probability Distributions for Surf Height Using an Ensemble of Mixture Density Networks. *Proceedings of the* 22nd *International Conference on Machine Learning*, Bonn, 7-11 August 2005, 113-120. http://dx.doi.org/10.1145/1102351.1102366

[6] Ramsay, J.O. and Silverman, B.W. (2005) Functional Data Analysis. Springer, New York. http://dx.doi.org/10.1002/0470013192.bsa239

[7] Ramsay, J. (1998) Estimating Smooth Monotone Functions. *Journal of the Royal Statistical Society*, **60**, 365-375. http://dx.doi.org/10.1111/1467-9868.00130

[8] Ramsay, J. (2000) Differential Equation Models for Statistical Functions. *The Canadian Journal of Statistics*, **28**, 225-240. http://dx.doi.org/10.2307/3315975

[9] Kneip, A. and Utikal, K. (2001) Inference for Density Families Using Functional Principal Component Analysis. *Journal of the American Statistical Association*, **96**, 519-532. http://dx.doi.org/10.1198/016214501753168235

[10] Bernhardt, C., Klüppelberg, C. and Meyer-Brandis, T. (2008) Estimating High Quantiles for Electricity Prices by Stable Linear Models. *The Journal of Energy Markets*, **1**, 3-19.

[11] Sen, R. and Klüppelberg, C. (2015) Time Series of Functional Data. Technical Report, ISI Chennai. http://www.isichennai.res.in/tr/asu/2015/1/ASU-2015-1.pdf

[12] Laukaitis, A. (2007) An Empirical Study for the Estimation of Autoregressive Hilbertian Processes by Wavelet Packet Method. *Nonlinear Analysis*: *Modeling and Control*, **12**, 65-75.

[13] Rice, J. and Wu, C. (2000) Nonparametric Mixed Effects Models for Unequally Sampled Noisy Curves. *Biometrics*, **57**, 253-259. http://dx.doi.org/10.1111/j.0006-341X.2001.00253.x

[14] Müller, H.G., Stadtmüller, U. and Yao, F. (2006) Functional Variance Processes. *Journal of the American Statistical Association*, **101**, 1007-1018. http://dx.doi.org/10.1198/016214506000000186

[15] Damon, J. and Guillas, S. (2005) Estimation and Simulation of Autoregressive Hilbertian Processes with Exogenous Variables. *Statistical Inference for Stochastic Processes*, **8**, 185-204. http://dx.doi.org/10.1007/s11203-004-1031-6

[16] PACE Package for Functional Data Analysis and Empirical Dynamics (Written in Matlab). http://www.stat.ucdavis.edu/PACE