Cluster Analysis of High-Dimensional High-Frequency Financial Time Series

Syed A. Pasha and Philip H.W. Leong

Abstract— **Recently the availability of tick data is driving renewed interest in statistical tools for the analysis of high-dimensional irregularly spaced time series. Since the standard tools require that the data are evenly spaced, the traditional multivariate time series analysis techniques are inadequate for the analysis of tick data. We develop for perhaps the first time a proper procedure that performs cluster analysis of tick data using the joint information of the temporal process and the continuous-valued data at the actual sampling times. A simulation example studies the problem with the standard approach and demonstrates the reliability of our proposed method. Data analyses of major stock market indices and currencies are provided.**

I. Introduction

RECENTLY in a number of application areas it has become possible to record measurements at ultra-high frequency. In computational finance, these data are referred to as tick data and comprise of recordings at irregularly spaced intervals and at ultra-fine time scales typically on the order of $1\,ms$. The availability of simultaneous recordings of tick data from a large number of channels is driving renewed interest in statistical tools for the analysis of high-dimensional tick data.

Statistical methods for multivariate time series analysis generally suffer from the curse of dimensionality problem, i.e., as the number of dimensions increases the number of parameters required increases exponentially. The traditional approach is to perform a dimension reduction such as principal components analysis (PCA) [1] or equivalently K-means clustering [2] which has been shown to be closely related to PCA [3]. Unfortunately, for data with temporal continuity, such standard tools require that the data are evenly spaced.

In high-frequency econometrics, the standard practice has been to aggregate the irregularly spaced time series to regular intervals to which traditional multivariate statistical tools can be applied [4]. However, there are two issues with this approach. Firstly, the transformation of the irregularly spaced time series to a discrete time process loses temporal information which has been well known in neural coding [5] and more recently in econometrics where it has been found to typically lead to blurring of the market microstructure and spurious inference [6]. Secondly, it is not obvious what constitutes a judicious selection of the interval for resampling since a different resampling interval is likely to give a different outcome. This problem is further compounded in the multivariate time series setting. The traditional multivariate time series analysis techniques are therefore inadequate for the analysis of tick data.

Ideally, one would prefer to avoid aggregating the tick data and treat the process directly. A natural means to represent the temporal information is through the point process formalism. The point process approach based on the stochastic intensity [7], [8] has been recognized as a flexible framework for modeling tick data to capture information such as volatility embedded in the inter-arrival times [9], [6]. In the case of additional information observed with the temporal process the marked point process [8] approach may be useful, but the mark space is restricted to a discrete set. At present there is no unified way to model a temporal process and continuous-valued information observed with the temporal process such as price or volume in a satisfactory manner.

Spectral analysis based on a least squares fit of sinusoids to irregularly spaced time series was developed in [10], [11] which encodes the temporal as well as the continuous-valued information observed jointly with the temporal process. The procedure is better known as the Lomb-Scargle method [12], [13] after the authors established the equivalence of periodogram analysis and least squares fitting of sinusoids to irregularly spaced data. Computationally efficient implementation based on the fast Fourier transform (FFT) can construct the Lomb-Scargle periodogram at M frequencies in $\mathcal{O}(M \log M)$ as opposed to a direct implementation requiring $\mathcal{O}(Mn)$ operations for n samples of the time series [14]. The utility of the Lomb-Scargle method has been recognized in a number of application areas [15], [16], [17], [18] and more recently in high-frequency finance [19]. In [20] the Lomb-Scargle method is used to reveal periodicities in biological rhythm. [21] takes a similar approach to reveal periodicity in the foreign exchange tick data. But these works are limited to the univariate time series case and do not address the curse of dimensionality problem indicated above.

The main contributions of this work are:

(i) we develop a reliable procedure for cluster analysis of multivariate tick data for apparently the first time which does not suffer from the problems with the standard approach, i.e., loss of temporal information and using a fixed interval to resample the multivariate time series. Our proposed procedure uses the Lomb-Scargle method to perform cluster analysis based on the joint information of the temporal distribution and the process observed at the irregular sampling times;

(ii) a simulation example is provided to study the effect

S.A. Pasha and P.H.W. Leong are with the Department of Electrical and Information Engineering, The University of Sydney, NSW 2006, Australia (email: {ahmed.pasha, philip.leong}@sydney.edu.au).

This work was supported by the Australian Research Council under the Linkage Project LP110200413

978-1-4673-5921-4/13/$31.00 ©2013 IEEE

of different resampling intervals on the cluster analysis and a comparison of the clustering accuracy of the standard approach and the proposed approach is shown;

(iii) cluster analyses of multivariate tick data of major stock market indices and currencies are presented and a discussion of the results is offered.

The rest of this paper is organized as follows: Section II reviews the Lomb-Scargle method and the computational complexity. Section III discusses a low-dimensional representation of the spectral density estimate. Section IV is a brief discussion on clustering in the K-means cluster centroid subspace. A simulation study and two data analyses are presented in Section V. A modified model is proposed and a least squares estimation procedure is sketched in Section VI. Finally, Section VII summarizes the paper.

Notation. We will use matrix $\Phi_{d\times p}$ of dimensions $d \times p$ and the singular value decomposition (SVD) $\Phi = U_{d\times d}\Sigma_{d\times p}V^T_{p\times p}$. $\Phi = (\phi_1, ..., \phi_p)$ and the matrix vectorization operator vec [22], i.e. $\text{vec}(\Phi) = \vec{\phi} = (\phi_1^T, ..., \phi_p^T)^T$. The d-vector $\vec{y} = (\bar{y}_1, ..., \bar{y}_d)^T$ where $\bar{y}_k \in \mathbb{R}_+$ with $\mathbb{R}_+$ denoting the positive reals. tr is the trace operator.

II. Spectral Analysis of Irregularly Spaced Data

A. The Lomb-Scargle Method

For a regularly sampled time series, the standard tools for analysis in the frequency domain are the Fourier methods based on the fast Fourier transform (FFT). The extension of the Fourier methods to irregularly spaced time series is the Lomb-Scargle periodogram [12], [13].

Given a time series $y_i := y(t_i)$ at irregularly spaced intervals $t_i - t_{i-1}$ with $0 < t_i \leq T, i = 1, ..., n$, the Lomb-Scargle (normalized) periodogram is

$$\bar{y}(\omega) = \frac{1}{2\sigma^2}\Big[\frac{(\Sigma_1^n(y_i - \hat{y})\cos\omega(t_i - \tau))^2}{\Sigma_1^n \cos^2\omega(t_i - \tau)} + \frac{(\Sigma_1^n(y_i - \hat{y})\sin\omega(t_i - \tau))^2}{\Sigma_1^n \sin^2\omega(t_i - \tau)}\Big]$$

where

$$\hat{y} = \frac{1}{n}\Sigma_1^n y_i,\ \sigma^2 = \frac{1}{n}\Sigma_1^n(y_i - \hat{y})^2$$

and

$$\tau = \frac{1}{2\omega}\arctan\Big(\frac{\Sigma_1^n \sin 2\omega t_i}{\Sigma_1^n \cos 2\omega t_i}\Big)$$

with $\omega \equiv 2\pi f, f > 0$.

Typically $\bar{y}(\omega)$ is evaluated at $2n$ or $4n$ frequencies $f \in [f_l, f_h]$. The lowest frequency f_l to be examined is $1/T$ and the highest frequency $f_h = n/(2T)$.

B. Computational Details

Given a sequence of data $y_i, i = 1, ..., n$, a direct implementation to construct the Lomb-Scargle periodogram at M frequencies requires $\mathcal{O}(Mn)$ operations. In [14], the authors show that the computations can be done in $\mathcal{O}(M \log M)$ as follows.

Define

$$\begin{aligned} S_h &:= \Sigma_1^n(y_i - \hat{y})\sin\omega t_i, C_h := \Sigma_1^n(y_i - \hat{y})\cos\omega t_i, \\ S_2 &:= \Sigma_1^n \sin^2\omega(t_i - \tau), C_2 := \Sigma_1^n \cos^2\omega(t_i - \tau) \end{aligned}$$

then,

$$\begin{aligned} \Sigma_1^n(y_i - \hat{y})\cos\omega(t_i - \tau) &= C_h\cos\omega\tau + S_h\sin\omega\tau, \\ \Sigma_1^n(y_i - \hat{y})\sin\omega(t_i - \tau) &= S_h\cos\omega\tau - C_h\sin\omega\tau \end{aligned}$$

and

$$\begin{aligned} \Sigma_1^n \cos^2\omega(t_i - \tau) &= \frac{n}{2} + \frac{1}{2}C_2\cos 2\omega\tau + \frac{1}{2}S_2\sin 2\omega\tau, \\ \Sigma_1^n \sin^2\omega(t_i - \tau) &= \frac{n}{2} - \frac{1}{2}C_2\cos 2\omega\tau - \frac{1}{2}S_2\sin 2\omega\tau \end{aligned}$$

Interpolate the values y_i on the M partitions in $[f_l, f_h]$ and take the FFT to obtain S_h, C_h. Interpolate the constant values 1 on the M partitions and take the FFT; after some manipulation yields S_2, C_2.

III. Spectral Estimate in the Low-Dimensional Space

A. Basis Representation

In the high-dimensional multivariate setting the statistical analysis of high-frequency data quickly becomes prohibitive due to the high computational power required to process large volume of data. An efficient approach to overcome this impediment is to use a basis representation which gives a compact form by a weighted sum of a sufficient number of basis functions which is considerably smaller than the number of samples of the univariate time series. A least squares procedure to fit the vector-valued spectral density is the following.

Given a d-dimensional vector of the spectral estimate $\vec{y}_\omega = (\bar{y}_1(\omega), ..., \bar{y}_d(\omega))^T, \bar{y}_k(\omega) \in \mathbb{R}_+, \omega \equiv 2\pi f$, evaluated at M frequencies $f \in [f_l, f_h]$, we suppose that the spectral estimate can be represented by

$$\vec{y}_\omega = \bar{c} + \bar{\Phi}\bar{\psi}_\omega + \epsilon_\omega$$

where $\bar{c}$ is the d-vector of means, $\bar{\psi}_\omega$ is a m-vector of basis functions, $\bar{\Phi}$ is the $d \times m$ matrix of coefficients and ϵ_ω is the d-vector of residuals.

We rewrite the model above more compactly as

$$\vec{y}_\omega = \Phi\psi_\omega + \epsilon_\omega$$

with $\Phi = (\bar{c}, \bar{\Phi}),\ \psi_\omega = (1, \bar{\psi}_\omega^T)^T$. The augmented matrix Φ is determined by minimizing the least squares criterion,

$$\int_{\omega_l}^{\omega_h} \frac{1}{2\delta}||\vec{y}_\omega - \Phi\psi_\omega||^2 d\omega$$

where the limits of integration $\omega_l = 2\pi f_l, \omega_h = 2\pi f_h$ and δ is a constant defined below.

Partitioning the interval $[\omega_l, \omega_h]$ into tiny bins of size δ given by $\omega_h - \omega_l = M\delta$, for $\omega_l < \omega \leq \omega_h$ we have $\omega - \omega_l = i\delta, i = 1, ..., M$. Defining $\vec{y}_i := \vec{y}_{\omega_l + i\delta}, \psi_i := \psi_{\omega_l + i\delta}$ gives the discretized equivalent,

$$\Sigma_1^M \frac{1}{2}||\vec{y}_i - \Phi\psi_i||^2$$

Recognizing $\Phi\psi_i = \text{vec}(\Phi\psi_i)$, we obtain $\Phi\psi_i = (\psi_i^T \otimes I)\vec{\phi}$ with $\vec{\phi} := \text{vec}(\Phi)$, $\otimes$ is the Kronecker product and I is the $d \times d$ identity matrix.

The least squares optimization can then be posed as

$$\vec{\phi} = \arg\min_{\vec{\phi}} \Sigma_1^M \frac{1}{2} ||\vec{y}_i - (\psi_i^T \otimes I)\vec{\phi}||^2 \tag{1}$$

We rapidly find the solution

$$\vec{\phi} = (\Sigma_1^M (\psi_i \otimes I)(\psi_i^T \otimes I))^{-1} \Sigma_1^M (\psi_i \otimes I) y_i$$

Proof: Using the result for the derivative of the two-norm, $\frac{\partial}{\partial x}||x|| = \frac{x}{||x||}$, the derivative with respect to $\vec{\phi}$ of the sum in (1) gives $-\Sigma_1^M (\psi_i \otimes I)(y_i - (\psi_i^T \otimes I)\vec{\phi})$. Setting the derivative to zero and rearranging the terms gives the required result. The proof is complete.

B. Computational Details

The solution involves inversion of a $pd \times pd$ symmetric matrix where $p = m + 1$. The matrix is a lattice of p^2 elements where each element is a scaled $d \times d$ identity matrix. Partitioning the matrix, the matrix inversion lemma [23] delivers, $\begin{pmatrix} A & B \\ B^T & C \end{pmatrix}^{-1} = \begin{pmatrix} P & Q \\ Q^T & R \end{pmatrix}$ where $P = S^{-1}, Q = -S^{-1}BC^{-1}, R = C^{-1}B^T S^{-1} BC^{-1} + C^{-1}$ and $S = A - BC^{-1}B^T$ is the Schur complement. With $A_{d \times d}$, C^{-1} can be computed by recursively partitioning and using the matrix inversion lemma until the partition comprises of $d \times d$ scaled identity matrices where C^{-1}, S^{-1} only require taking the reciprocal of the scalings.
By undoing the vec operation we obtain the augmented matrix Φ.

The profile of the spectral estimate suggests that a basis representation such the cosines $\psi_\omega = (\psi_\omega^{(1)}, ..., \psi_\omega^{(m)})^T, \psi_\omega^{(u)} = \cos(\frac{u\pi(\omega - \omega_l)}{\omega_h - \omega_l})$ should give a reasonably good approximation with the number of basis functions $m << M$. This leads to a finite and a more compact dimensional representation of the spectral estimate within $\Phi_{d \times p}$.

IV. Clustering in the Centroid Subspace

Clustering refers to the partitioning of data into disjoint groups whereby data within a cluster are similar based on some criterion while data in different clusters are dissimilar. The K-means algorithm [2], [24] for example minimizes the sum of the squared errors, i.e.,

$$\Sigma_1^K \Sigma_{i \in C_k} ||y_i - \tilde{y}_k||^2$$

for some data $\{y_1, ..., y_n\}$ where K is the number of clusters, C_k is the k-th cluster, $\tilde{y}_k = \frac{1}{|C_k|}\Sigma_{i \in C_k} y_i$ is the centroid of C_k and $|C_k|$ is the cardinality of C_k. The solution of the standard implementation has been found to often converge to a local minima (see [3] and references therein).

Clustering in the cluster subspace has been shown to be more robust to clustering in the original space [3, Prop. 3.4]. This is because in the cluster subspace while the between-cluster distances remain more or less the same as in the original space, the within-cluster distances shrink. Given K clusters, the transformation $P^T y$ of any vector y yields a vector in the subspace spanned by the K cluster centroids where [3], $P = \Sigma_1^K |C_k| \tilde{y}_k \tilde{y}_k^T$.
The subspace can be determined by a singular value decomposition (SVD) as follows.
Given $\Phi_{d \times p}$, the decomposition of $\Phi = U_{d \times d} \Sigma_{d \times p} V_{p \times p}^T$ where Σ is the diagonal matrix of singular values and the columns of U, V form orthonormal bases in $\mathbb{R}^d, \mathbb{R}^p$ respectively, i.e., $U^T U = I_{d \times d}$ and $V^T V = I_{p \times p}$. For r principal components with $d \geq p \geq r$, the subspace spanned by $K = r + 1$ cluster centroids is $U_{d \times r} \Sigma_{r \times r}^2 U_{r \times d}^T$.

V. Simulation and Data Analyses

A simulation example is first provided which demonstrates the problem in aggregating an irregularly spaced time series in the multivariate setting. Cluster analysis using the proposed approach which does not lose the temporal information in the irregularly spaced time series is then presented. Then, data analyses of tick data for major stock market indices and currencies is presented and a discussion on the results is offered.

A. Simulation Example

We consider a $d = 60$-dimensional multivariate time series which comprises of independent univariate time series. The irregularly spaced intervals of the univariate time series are independent and identically distributed with an exponential distribution with parameter λ. The multivariate time series is partitioned into 3 groups based on the values $\lambda = 1, 5, 10\, Hz$ but the lack of distinct periodicities will make cluster analysis difficult.

The intervals of the univariate time series are simulated on $0 < t \leq T$ with $T = 1000\, s$ by first sampling the counts n from a Poisson distribution for a given λ and then drawing n samples $u_1, ..., u_n$ from a uniform distribution so that the sampling time $t_i = -\lambda^{-1} \ln u_i, i = 1, ..., n$. The state of the time series at the sampling time t_i is modeled by Brownian motion that has quadratic variation on $[0, t]$ of t [25]. In [19] the authors use the autoregressive AR(1) to model the dynamics but this model does not have a natural extension in the irregularly sampled data setting. Fig. 1 shows a raster plot of the uneven sampling times of the multivariate time series for $100\, s$ where the 3 groups are clearly visible.

1) Cluster Analysis of Aggregated Time Series: The traditional approach is to aggregate the multivariate time series using a fixed interval but as already mentioned it is not obvious how to select an interval to resample the data particularly in the multivariate setting. Here we present cluster analysis for a range of values of the discretization step δ.

The univariate time series are discretized on $0 < t \leq T$ at intervals of $\delta = 2, 1, 0.1, 0.01\, s$ using cubic spline interpolation. For coarse δ steps cluster analysis can be applied directly to the aggregated time series but for fine δ values the time series is very high dimensional and a direct application of the cluster analysis of the aggregated time series becomes prohibitive. We have found the basis representation useful since a cluster analysis on the much lower dimensional space of the coefficients matrix gives almost identical results at considerably less computational time. The aggregated time series is fit using B-splines which are suitable since the time series is aperiodic. We use $m = 40$ cubic spline basis elements for the fitting. For $\delta = 1\,s$ the true, aggregated and fitted time series and the error in the basis representation are shown for $100\,s$ in Fig. 2 which suggests that the basis function candidate gives a reasonably good representation of the aggregated time series.

Assuming the true number of clusters is known, we proceed by computing the cluster centroid subspace spanned by the $K-1$ components of the B-splines coefficients matrix. For $\delta = 1\,s$, K-means clustering in the cluster centroid subspace assigns only 3 members to one of the clusters. The clustering result is summarized in the confusion matrix,

$$C = \begin{pmatrix} 13 & 13 & 10 \\ 6 & 6 & 9 \\ 1 & 1 & 1 \end{pmatrix}$$

where $[C_{ij}]$ is the number of times a data point of cluster j is assigned to cluster i. The clustering accuracy is $\frac{1}{d}\text{tr}(C) = 33.33\%$. For $\delta = 2\,s$, K-means clustering assigns only 2 members to one of the clusters and yields identical clustering result to the case $\delta = 1\,s$.
For $\delta = 0.1\,s$, only 1 member is assigned to one of the clusters. The clustering result is

$$C = \begin{pmatrix} 14 & 14 & 10 \\ 6 & 6 & 9 \\ 0 & 0 & 1 \end{pmatrix}$$

and the clustering accuracy is 35%. For $\delta = 0.01\,s$, the clustering result is identical to the case $\delta = 0.1\,s$.
Note that for finer δ values, $T/\delta >> n_k$, the number of samples of the univariate time series, and will not yield meaningful result which is also suggested by the cluster analysis of the aggregated time series with $\delta = 1\,ms$,

$$C = \begin{pmatrix} 8 & 7 & 7 \\ 7 & 5 & 5 \\ 5 & 8 & 8 \end{pmatrix}$$

We have found that $m > 40$ gives similar results for the cluster analysis.

Note that without the information of the true number of clusters, the SVD of the B-splines coefficients matrix gives $r = 6$ principal components (the singular values are shown in Fig. 3) which yields an incorrect number of of clusters $K = r + 1 = 7$.

2) Cluster Analysis of Irregularly Spaced Time Series: The Lomb-Scargle periodogram for the multivariate time series is constructed by computing the spectral density estimate for the univariate time series independently on a fine grid of $M = 4\max(N) = 42400$ partitions in $[f_l, f_h]$ with $f_l = 1/T, f_h = 0.25$ and $N = (n_1, ..., n_d)^T$ is the vector of number of samples of the multivariate time series. The spectral density for one member of each of the 3 groups is shown in Fig. 4.

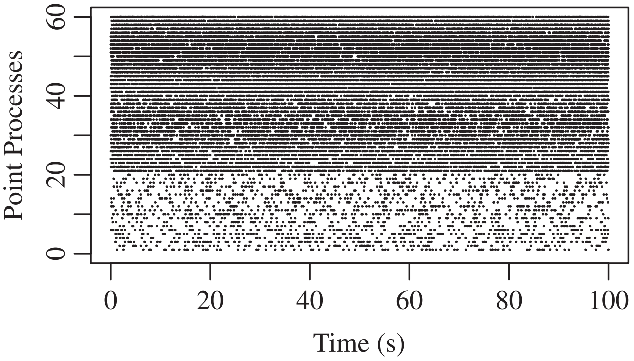

Fig. 1. Raster Plot of Multivariate Poisson Processes with $\lambda = 1, 5, 10\,Hz$

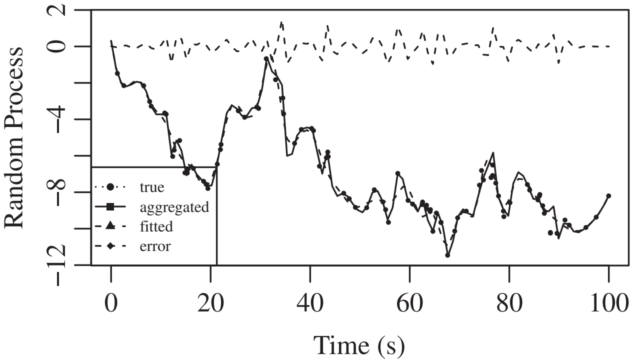

Fig. 2. Actual and Aggregated Times Series and B-Splines Representation and Error in Basis Representation

For reasons already mentioned we perform cluster analysis in the low dimensional space derived from a basis representation of the spectral estimate. We have found the cosine basis function a good candidate for the spectral estimate. Using $m = 40$ cosine basis elements to represent $\vec{y}_\omega$, the spectral density, cosine basis representation and error in the representation are shown in Fig. 5 where the high dimensional multivariate time series is represented in a much more compact dimensional space within Φ.

We assume no prior knowledge of the true number of clusters. The number of clusters $K = r + 1 = 3$ is taken where r is the number of principal components given by the SVD of Φ which coincides with the true number of clusters. Fig. 6 shows the singular values of Φ. K-means clustering is performed in the cluster centroid subspace. The clustering result is summarized in the confusion matrix,

$$C = \begin{pmatrix} 14 & 0 & 10 \\ 3 & 20 & 0 \\ 3 & 0 & 10 \end{pmatrix}$$

The clustering accuracy is 73.33% which is much higher than that obtained using the traditional approach. As already mentioned the lack of clear periodicites in the point process makes the exact recovery of the actual clusters extremely difficult.

To summarize, we can infer that aggregating the multivariate

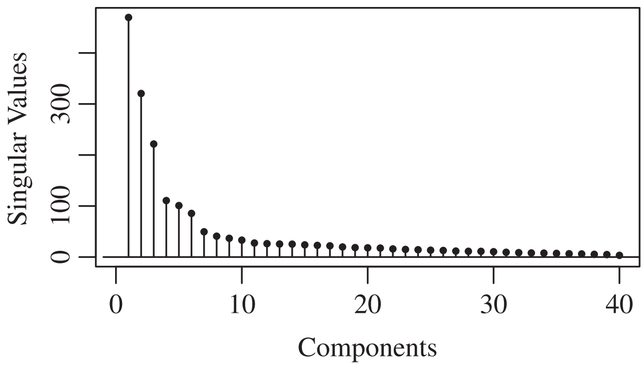

Fig. 3. Singular Values of the B-splines Coefficients Matrix

time series loses the temporal information in the uneven sampling times which cannot subsequently be recovered despite using a very fine interval for resampling. Furthermore, cluster analysis of the aggregated multivariate time series is likely to be highly unreliable. The Lomb-Scargle method which encodes both the temporal as well as continuous-valued information observed at the actual sampling times captures the underlying structure which is subsequently recovered with reasonable accuracy by performing cluster analysis of the spectral estimate.

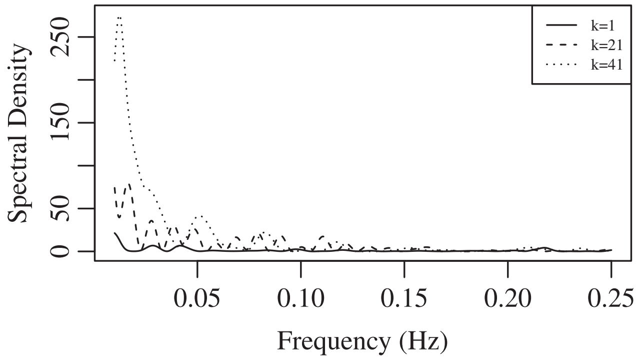

Fig. 4. Spectral Density of 3 Time Series

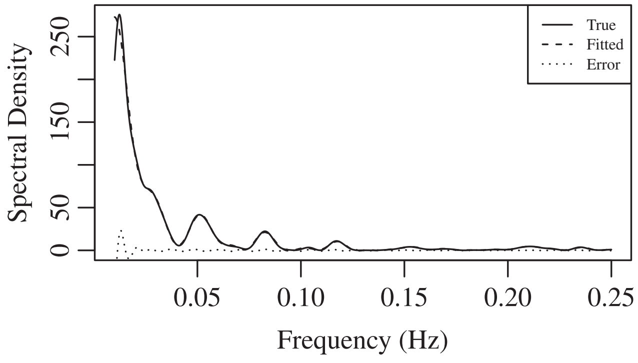

Fig. 5. Spectral Density, Cosine Basis Representation and Residual

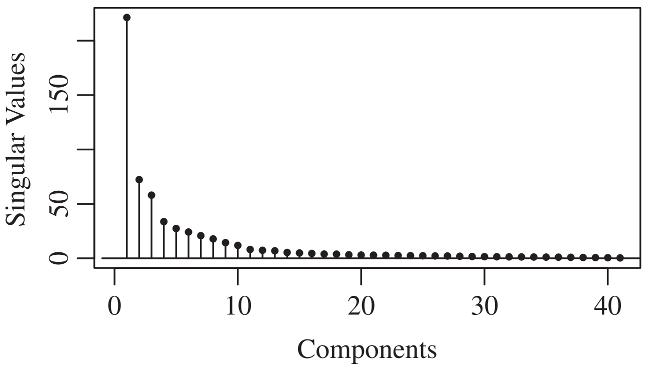

Fig. 6. Singular Values of Φ

B. Stock Market Indices

The tick history of 26 of the major world stock market indices is analyzed. The data comprise of the price of the indices on September 25, 2012 and the timestamp with an average of around $11,000$ recordings per index. A raster plot of the timestamp for the multivariate tick history is shown in Fig. 7 which shows the highly irregular behaviour of price movements with significant lapses in recordings. A pattern emerges from Fig. 7 but the interest here is not to cluster the indices based on the temporal distribution of the tick history alone but rather to determine homogenous groups based on the joint information of the temporal distribution and the price movements. Clearly, there is no way to represent such time series using the standard discrete time or continuous time models without losing temporal information.

The spectral density estimate of the tick history of each index was computed using the Lomb-Scargle method for $11.58 \times 10^{-5} \leq f \leq 2 \times 10^{-3}\, Hz$ with $M = 125,332$ partitions. The spectral estimate of the All Ordinaries (.AORD), FTSE Italia All-Share (.FTITLMS) and Euro STOXX 50 (.STOXX) is shown in Fig. 8. The high dimensional multivariate spectral estimate is subsequently expressed in considerably low dimension with $m = 200$ cosine basis functions. The spectral estimate, basis representation and the residual for the .AORD are shown in Fig. 9 which suggests that the representation is reasonably good.

By taking the SVD of Φ we determine the number of principal components as $r = 1$ and the subspace spanned by the $K = r+1$ cluster centroids. The result of clustering in the subspace is summarized in Fig. 10 which shows two groups of 7 and 19 members. The group on the left comprises of the FTSE All-Share (.FTAS), FTSE MID 250 (.FTMC), FTSE MIB (.FTMIB), .FTSE, German DAX (.GDAXI), NASDAQ Composite (.IXIC) and NASDAQ-100 (.NDX) while the group on the right comprises of indices such as the .AORD, Bats 1000 (.BATSK), Dow Jones Industrial Average (.DJI), CAC 40 (.FCHI), S&P/TSX Composite (.GSPTSE), Hang Seng (.HSI), S&P 500 (.INX) etc.

The recursive application of K-means clustering on the groups identified reveals homogeneity/heterogeneity within a group which is not obvious in Fig. 10. The result is summarized in Fig. 11 which suggests for example that within the group on the left in Fig. 10, .GDAXI, .IXIC and .NDX form a group and .FTAS, .FTMC, .FTMIB and .FTSE form a group. It is interesting to note that the cluster analysis groups the FTSE indices in one cluster. Similarly, within the group on the right in Fig. 10, .DJI and the Swiss Market Index (.SSMI) form a group and .AORD, .HSI and Nikkei 225 (.N225) also form a group.

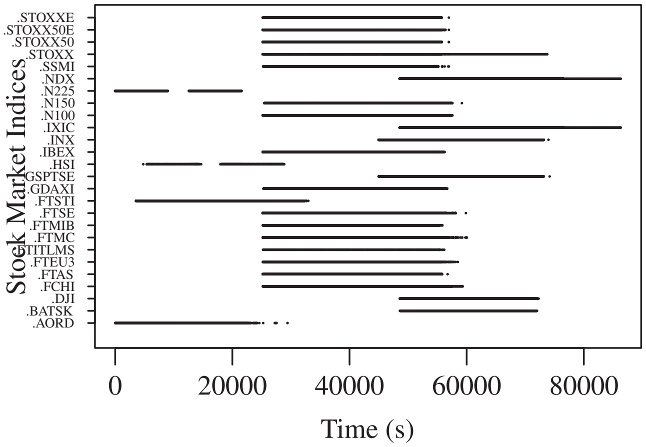

Fig. 7. Raster Plot of Timestamp for Multivariate Tick Data

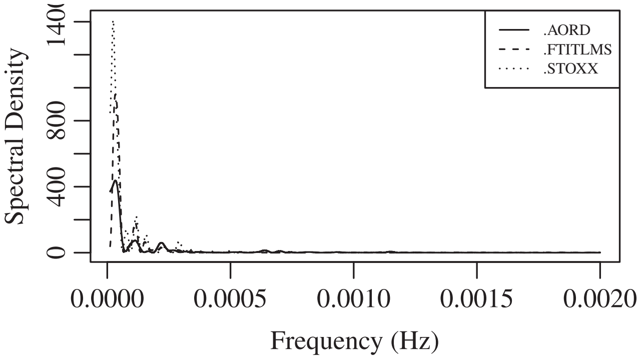

Fig. 8. The Spectral Density Estimate for the All Ordinaries (.AORD), FTSE Italia All-Share (.FTITLMS) and Euro STOXX 50 (.STOXX)

C. Currency Exchange Rates

The exchange rate tick history of major currencies is analyzed. The data comprise of exchange rates of 156 currencies against the USD on September 25, 2012. The timestamp is recorded up to $1\,ms$ accuracy. Currencies with at least 300 recordings were analyzed which meant that 55 of the 156 currencies with an average of around $18,000$ recordings per currency are studied. This was done merely for presentation of results and is otherwise not a limitation of the proposed approach. A raster plot of the timestamp for the multivariate tick history is shown in Fig. 12. The sporadic behaviour is characteristic of the exchange rate and the significant lapses in recording illustrate why discrete or continuous time modeling is inadequate for the multivariate tick data.

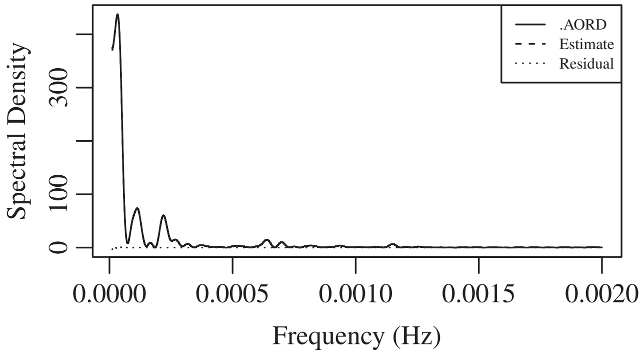

Fig. 9. Spectral Density Estimate, Basis Representation with $m = 200$ Cosine Elements and Residual for the All Ordinaries (.AORD)

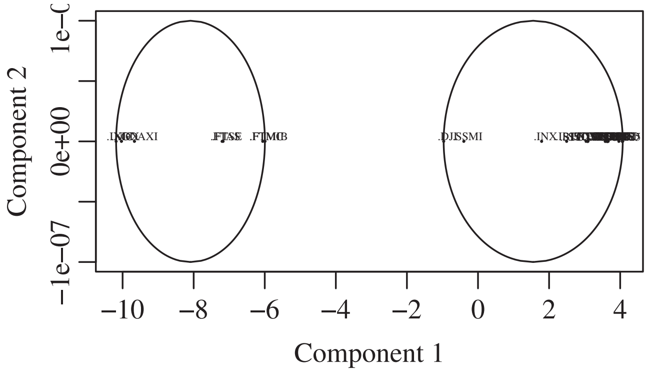

Fig. 10. Clustering in the Subspace of $K = 2$ Cluster Centroids

The spectral density estimate of the tick history of each currency was computed using the Lomb-Scargle method for $11.57 \times 10^{-5} < f \leq 2 \times 10^{-3}\,Hz$ with $M = 239,300$ partitions. The spectral estimates of the Argentine Peso (ARS), Singapore Dollar (SGD) and Chilean Peso (CLP) are shown in Fig. 13. The basis representation with $m = 200$ cosine elements and the residual for the Chinese Yuan Renminbi (CNY) are shown in Fig. 14 which demonstrates the suitability of the cosine basis functions.

The cluster centroid subspace was computed using the SVD of Φ. The result of clustering is summarized in Fig. 15 which shows 4 groups of $5, 14, 12$ and 24 members. The first group comprises of the Swiss Franc (CHF), Danish Krone (DKK), Euro (EUR), Moroccan Dirham (MAD) and Swedish Krona (SEK), mainly the Western European currencies. The second group comprises of currencies such as the Bulgarian Lev (BGN), Czech Koruna (CZK), Croatian Kuna (HRK), Polish Zloty (PLN) and Serbian Dinar (RSD) which are predominantly Central European currencies. In addition it includes currencies such as the Australian Dollar (AUD), Canadian Dollar (CAD), British Pound Sterling (GBP) etc. The third comprises of the Eastern European currencies such as the Hungarian Forint (HUF) and the Romanian New Leu (RON) and South American currencies such as the Brazilian Real (BRL) and the Columbian Peso (COP) etc., while the remaining which is the largest group is dominated by Middle Eastern currencies such as the United Arab Emirates

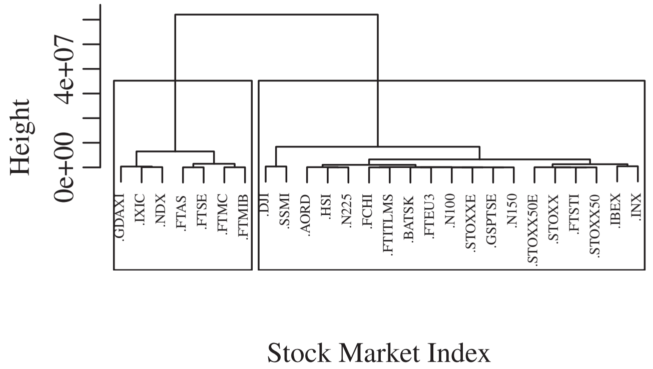

Fig. 11. Recursive K-means Clustering

Dirham (AED), Bahraini Dinar (BHD), Omani Riyal (OMR), Qatari Riyal (QAR) and the Saudi Riyal (SAR) and African currencies such as the Algerian Dinar (DZD), South African Rand (ZAR), Lesotho Maloti (LSL), Swazi Lilangeni (SZL) and the West African CFA Franc (XOF).

Fig. 16 shows the result of recursive K-means clustering which uncovers homogeneity/heterogeneity within each group not apparent in Fig. 15. As expected the Middle Eastern currencies in the largest cluster form a homogeneous group within the cluster and so on.

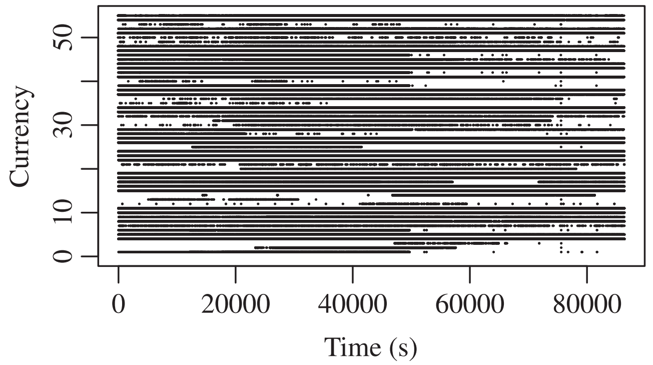

Fig. 12. Raster Plot of Timestamp for Multivariate Currency Tick History

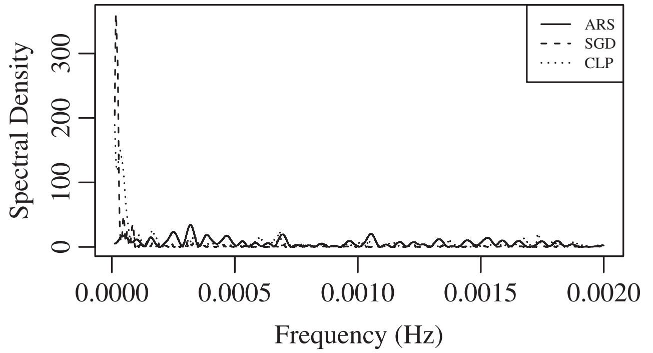

Fig. 13. The Spectral Density Estimate for the Argentine Peso (ARS), Singapore Dollar (SGD) and Chilean Peso (CLP)

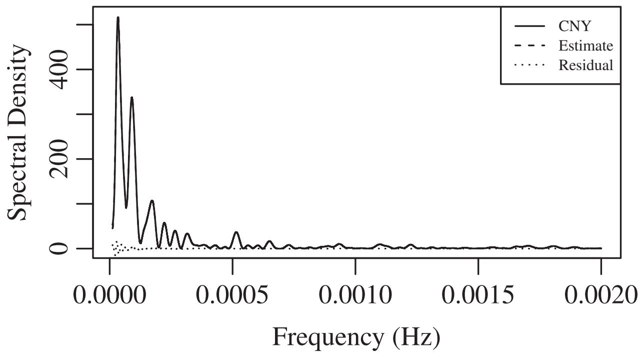

Fig. 14. Spectral Density Estimate, Basis Representation with $m = 200$ Cosine Elements and Residual for the Chinese Yuan Renminbi (CNY)

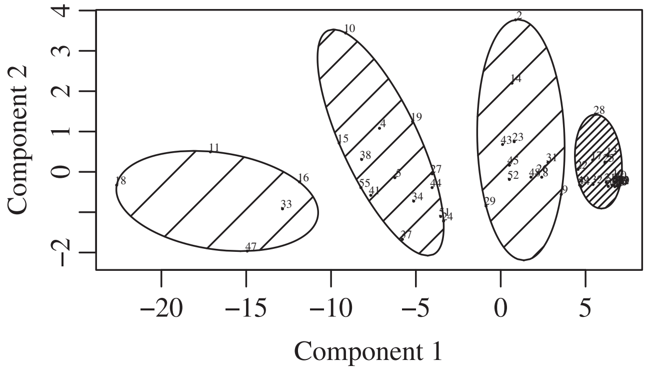

Fig. 15. Clustering in the Subspace of $K = 4$ Cluster Centroids

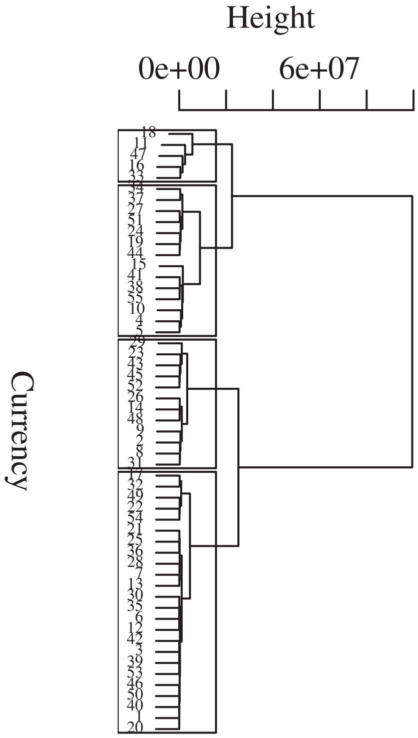

Fig. 16. Recursive K-means Clustering for Exchange Rate Tick History

VI. Non-negativity Constraint

The model in Section III ignores the non-negativity of the spectral density. Here we show how the model can be modified to ensure non-negativity of the spectral estimate and only sketch the procedure for fitting the non-negative data due to lack of space. The empirical analysis will be pursued elsewhere.

Given the d-vector of spectral density $\vec{y}_\omega = (\bar{y}_1(w), ..., \bar{y}_d(w))^T$, suppose $\bar{y}_k(w)$ is given by

$$\bar{y}_k(w) = e^{\bar{c}_k + \bar{\phi}_k^T \bar{\psi}_\omega} + \epsilon_\omega$$

where $\bar{c}_k$ is the mean, $\bar{\psi}_\omega$ is a m-vector of basis functions, $\bar{\phi}_k$ is the m-vector of coefficients and ϵ_ω is the residual. Note that the model ensures non-negativity of the data without explicit constraints in the optimization. The natural approach is to fit $\log(\vec{y}_\omega)$ using the procedure in Section III but $\log(\vec{y}_\omega)$ is non-smooth as the spectral data approaches zero. A procedure to fit $\vec{y}_\omega$ with the non-negativity constraint is outlined below.

Rewriting the model above more compactly as

$$\bar{y}_k(w) = e^{\phi_k^T \psi_\omega} + \epsilon_\omega, \ \phi_k = (\bar{c}_k, \bar{\phi}_k^T)^T, \ \psi_\omega = (1, \bar{\psi}_\omega^T)^T$$

and $\bar{y}_k^{(i)} := \bar{y}_k(\omega_l + i\delta), \psi_i := \psi_{\omega_l + i\delta}$, for a tiny increment δ, the augmented vector ϕ_k is given by the nonlinear optimization $\phi_k = \arg\min_{\phi_k} J$ with

$$J = \Sigma_1^M \frac{1}{2} ||\bar{y}_k^{(i)} - e^{\phi_k^T \psi_i}||^2$$

The gradient and Hessian terms are

$$\begin{aligned}\frac{\partial J}{\partial \phi_k} &= \Sigma_1^M (-\bar{y}_k^{(i)} + e^{\phi_k^T \psi_i}) e^{\phi_k^T \psi_i} \psi_i \\ \frac{\partial^2 J}{\partial \phi_k \partial \phi_l^T} &= \Sigma_1^M (-\bar{y}_k^{(i)} + 2e^{\phi_k^T \psi_i}) e^{\phi_l^T \psi_i} \psi_i \psi_i^T \delta_{k,l}\end{aligned}$$

Then, the Newton-Raphson ϕ_k update step is

$$\phi_k^{(1)} = \phi_k^{(0)} - \left(\frac{\partial^2 J}{\partial \phi_k \partial \phi_k^T} \right)^{-1} \frac{\partial J}{\partial \phi_k}$$

$\phi_k, k = 1, ..., d$ form the columns of Φ^T. K-means clustering can be performed in the cluster centroid subspace of Φ as outlined in Section IV.

VII. Conclusions

In this paper we have discussed for apparently the first time a proper procedure for cluster analysis of multivariate tick data based on the joint information of the irregular sampling times and the continuous-valued process observed at the actual sampling times. The procedure is based on the Lomb-Scargle method which encodes the information in the spectral density. The high dimensional spectral density estimate is given a compact representation using a basis expansion and clustering is performed in the K-means cluster centroid subspace of the coefficients matrix which has been shown to be more robust than clustering in the original space [3]. A simulation study underscores the problem with the standard approach which aggregates the multivariate time series to fixed intervals which not only loses the temporal information but also blurs any structure in the data rendering the analysis unreliable. Our proposed approach is free from such shortcomings and much more reliable in comparison to the standard approach. The cluster analyses of major world stock market indices and currencies is discussed in detail. Future work will consider the non-negativity constraint of the spectral density and cluster analysis of tick data using the modified model in Section VI.

References

[1] I. Joliffe, *Principal Components Analysis*. Heidelberg: Springer Verlag, 1986.

[2] J. MacQueen, "Some methods for classification and analysis of multivariate observations," in *Proc. 5th Berkeley Symp. Mathematical Statistics and Probability*, vol. 1, 1967, pp. 281–297.

[3] C. Ding and X. He, "K-means clustering via principal component analysis," in *ICML*, 2004, pp. 225–232.

[4] M. Dacorogna, R. Genay, U. Müller, R. Olsen, and O. Pictet, *An Introduction to High-Frequency Finance*. Academic Press, 2001.

[5] F. Rieke, D. Warland, R. de Ruyter van Stvenink, and W. Bialek, *Spikes: Exploring the Neural Code*. Boston, MA: MIT Press, 1997.

[6] L. Bauwens and N. Hautsch, "Modelling financial high frequency data using point processes," in *Handbook of Financial Time Series*. Springer, 2009, pp. 953–979.

[7] D. Snyder, *Random Point Processes*. NY: J. Wiley, 1975.

[8] D. Daley and D. Vere-Jones, *An Introduction to the Theory of Point Processes*, 1st ed. NY: Springer-Verlag, 1988.

[9] N. Hautsch, "Modelling irregularly spaced financial data," in *Lecture Notes in Economics and Mathematical Systems*. Berlin: Springer, 2004, vol. 539.

[10] F. Barning, "The numerical analysis of the light-curve of 12 Lacertae," *Bull. Astron. Inst. Netherlands*, vol. 17:22, 1963.

[11] P. Vaniček, "Further development and properties of the spectral analysis by least-squares," *Astrophysics and Space Science,*, vol. 12(1), pp. 10–33, 1971.

[12] N. Lomb, "Least-squares frequency analysis of unequally spaced data," *Astrophysics and Space Science*, vol. 39, pp. 447–462, 1976.

[13] J. Scargle, "Studies in astronomical time series analysis II. Statistical aspects of spectral analysis of unevenly spaced data," *Astrophysical Journal*, vol. 263, pp. 835–853, 1982.

[14] W. Press and G. Rybicki, "Fast algorithm for spectral analysis of unevenly sampled data," *Astrophysical Journal*, vol. 338, 1989.

[15] T. Ruf, "The Lomb-Scargle periodogram in biological rhythm research: Analysis of incomplete and unequally spaced time-series," *Biol. Rhythm Res.*, vol. 30(2), pp. 178–201, 1999.

[16] M. Schulza and K. Statteggerb, "Spectrum: spectral analysis of unevenly spaced paleoclimatic timeseries," *Comput. Geosci.*, vol. 23(9), pp. 929–945, 1997.

[17] S. Baisch and G. Bokelmann, "Spectral analysis with incomplete timeseries: an example from seismology," *Comput. Geosci.*, vol. 25(7), pp. 739–750, 1999.

[18] R. Oliver and J. Ballester, "The north-south asymmetry of sunspot areas during solar cycle 22," *Solar Physics*, vol. 152, 1994.

[19] I. Giampaoli, W. Ng, and N. Constantinou, "Analysis of ultra-high-frequency financial data using advanced Fourier transforms," *Finance Research Letters*, vol. 6, pp. 47–53, 2009.

[20] H. V. D. et. al., "A procedure of multiple period searching in unequally spaced time-series with the LombScargle method," *Biol. Rhythm Res.*, vol. 30(2), pp. 149–177, 1999.

[21] W. Ng, I. Giampaoli, and N. Constantinou, "Periodicities of fx markets in intrinsic time," 2010. [Online]. Available: http://www.essex.ac.uk/ebs/research/working_papers/

[22] J. Magnus and H. Neudecker, *Matrix Differential Calculus with Applications in Statistics and Econometrics*. NY: J. Wiley, 1999.

[23] R. Horn and C. Johnson, *Matrix Analysis*. Cambridge: Cambridge University Press, 1985.

[24] J. Hartigan and M. Wang, "A K-means clustering algorithm," vol. 28, pp. 100–108, 1979.

[25] F. Klebaner, *Introduction to Stochastic Calculus with Applications*. Imperial College Press, 1998.