Toeplitz Approximation to Empirical Correlation Matrix of Asset Returns: A Signal Processing Perspective

Ali N. Akansu, *Fellow, IEEE*, and Mustafa U. Torun

***Abstract*—Empirical correlation matrix of asset returns has its intrinsic noise component. Eigen decomposition, also called Karhunen-Loeve Transform (KLT), is employed for noise filtering where an identified subset of eigenvalues replaced by zero. The filtered correlation matrix is utilized for calculation of portfolio risk and rebalancing. We introduce Toeplitz approximation to symmetric empirical correlation matrix by using auto-regressive order one, AR(1), signal model. It leads us to an analytical framework where the corresponding eigenvalues and eigenvectors are defined in closed forms. Moreover, we show that discrete cosine transform (DCT) with implementation advantages provides comparable performance as a good approximation to KLT for processing the empirical correlation matrix of a portfolio with highly correlated assets. The energy packing of both transforms degrade for lower values of correlation coefficient. The theoretical reasoning for such a performance is presented. It is concluded that the proposed framework has a potential use for quantitative finance applications.**

***Index Terms*—Karhunen-Loeve transform, discrete cosine transform, AR(1) model, empirical correlation matrix, portfolio management, risk management.**

I. Introduction

A PORTFOLIO is comprised of multiple financial assets. The standard deviation of portfolio return is a widely used risk metric in finance [1]. A desirable portfolio delivers maximum return on investment with minimum risk. Therefore, the return of each asset is individually assessed, and also compared against competing assets in the portfolio. Pair-wise correlations of asset returns populate the empirical correlation matrix that reveals significant information on portfolio risk and its variations in time. A portfolio manager monitors these variations and rebalances portfolio in order to keep the risk within allowed range for the desired return.

Severe non-stationarity with high level of intrinsic noise is common in asset returns of a portfolio. Hence, empirical correlation matrix needs to be tamed accordingly. Eigen analysis, also called principal component analysis (PCA) or KLT, has been successfully employed to filter out this undesirable noise component from the measured correlations [2]. The caveat is the computational cost of KLT operations. Therefore, we revisit DCT as an approximation to KLT, and compare their performance for noise cleaning of empirical correlation matrix of asset returns. DCT is quite attractive over KLT due to its computational efficiency. The filtered empirical correlation matrix is represented by its eigenvectors and eigenvalues that lead us to creation of eigenportfolios. Our goal is to compare performances of fixed transform DCT and input dependent KLT for empirical correlation matrices of various portfolios in order to justify the use of the former as an efficient replacement to the latter in practice.

Mathematical preliminaries are given in Section II. We introduce the basics of orthogonal transforms and performance metrics to compare their merit. We briefly discuss about modeling of random signal sources and focus on auto-regressive model of order one, AR(1). Its Toeplitz correlation matrix is defined. We also describe eigendecomposition of a matrix, and present closed form expressions for eigenvalues and eigenvectors of an AR(1) source. Moreover, we give the kernel of DCT in this section. In Section III, symmetric empirical correlation matrix is approximated by employing AR(1) model with Toeplitz correlation matrix in two different ways. Namely, the first one utilizes only one AR(1) model approximation for the entire empirical correlation matrix. In contrast, the second one uses one AR(1) model per row (asset). The merit of these Toeplitz approximations and the use of DCT as a fast KLT implementation in finance applications are highlighted [3], [4]. Then, we introduce portfolio risk and its calculations through KLT and DCT based methods in Section IV. Our conclusions are presented in Section V.

II. Mathematical Preliminaries

In this section, we present theoretical foundations of orthogonal transforms that provide underlying mathematical steps utilized in many applications including financial signal processing.

A. Orthogonal Transforms

A family of linearly independent N orthonormal discrete-time sequences on the interval $0 \leq n \leq N-1$ satisfies [5]

$$\sum_{n=0}^{N-1} \phi_k(n)\phi_l^*(n) = \delta_{k-l} = \begin{cases} 1, & k = l \\ 0, & \text{otherwise} \end{cases} \quad (1)$$

Manuscript received June 30, 2011; revised January 23, 2012; accepted May 02, 2012. Date of publication June 15, 2012; date of current version July 13, 2012. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Sanjeev Kulkarni.

The authors are with the Electrical Engineering Department, New Jersey Institute of Technology, Newark, NJ 07102 USA (e-mail: akansu@njit.edu; mustafa.torun@njit.edu).

Digital Object Identifier 10.1109/JSTSP.2012.2204724

Orthonormality may also be expressed on the unit circle of the complex plane where $z = e^{j\omega}; -\pi \leq \omega \leq \pi$ as follows

$$\sum_{n=0}^{N-1} \phi_k(n)\phi_l^*(n) = \frac{1}{2\pi}\int_{-\pi}^{\pi} \Phi_k(e^{j\omega})\Phi_l^*(e^{j\omega})d\omega = \delta_{k-l} \tag{2}$$

In matrix form, $\{\phi_k(n)\}$ are the rows of the transform matrix, and they are also called as basis functions

$$\Phi = [\phi_k(n)] : k, \quad n = 0, 1, 2, \ldots, N-1 \tag{3}$$

with orthogonality

$$\Phi\Phi^{-1} = \Phi\Phi^{*T} = \mathbf{I} \tag{4}$$

where $*T$ indicates conjugated and transposed version of a matrix. A signal vector $\mathbf{x}$ is mapped into the orthonormal space through forward transform operator

$$\boldsymbol{\theta} = \Phi\mathbf{x} \tag{5}$$

where $\boldsymbol{\theta}$ is transform coefficients vector. Similarly, the inverse transform yields the signal vector

$$\mathbf{x} = \Phi^{-1}\boldsymbol{\theta} \tag{6}$$

The vector $\mathbf{x}$ is populated by a wide-sense stationary (WSS) stochastic process that satisfies the properties

$$\begin{aligned} E\{x(n)\} &= \mu(n) = \mu \\ E\{x(n)x^*(n+m)\} &= R_{xx}(m) \end{aligned} \tag{7}$$

where $E\{\cdot\}$ is the expectation operator. The correlation and covariance matrices of such a random vector process $\mathbf{x}$ are defined, respectively,

$$\begin{aligned} \mathbf{R}_x &= E\{\mathbf{x}\mathbf{x}^{*T}\} \\ \mathbf{R}_x &= \begin{bmatrix} R_{xx}(0) & R_{xx}(1) & \cdots & R_{xx}(N-1) \\ R_{xx}(1) & R_{xx}(0) & \cdots & R_{xx}(N-2) \\ \vdots & \vdots & \ddots & \vdots \\ R_{xx}(N-1) & R_{xx}(N-2) & \cdots & R_{xx}(0) \end{bmatrix} \\ \mathbf{C}_x &= E\{(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^{*T}\} = \mathbf{R}_x - \boldsymbol{\mu}\mu^{*T} \end{aligned} \tag{8}$$

where

$$\boldsymbol{\mu}\boldsymbol{m}u^{*T} = \mu^2 \begin{bmatrix} 1 & 1 & \cdots & 1 \\ 1 & 1 & \cdots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & \cdots & 1 \end{bmatrix}$$

Note that $\mathbf{R}_x = \mathbf{C}_x$ for a zero mean WSS process where $\mu = 0$. Hence, one can derive the covariance matrix of transform coefficients as follows

$$\begin{aligned} \mathbf{R}_{\boldsymbol{\theta}} &= E\{\boldsymbol{\theta}\boldsymbol{\theta}^{*T}\} \\ &= E\{\Phi\mathbf{x}\mathbf{x}^{*T}\Phi^{*T}\} \\ &= \Phi E\{\mathbf{x}\mathbf{x}^{*T}\}\Phi^{*T} \\ &= \Phi\mathbf{R}_x\Phi^{*T} \end{aligned} \tag{9}$$

Energy preserving property of an orthonormal transform yields the relationship between signal variance and variances of corresponding coefficients

$$\begin{aligned} E\{\boldsymbol{\theta}\boldsymbol{\theta}^{*T}\} &= E\{\mathbf{x}\mathbf{x}^{*T}\} \\ E\{\boldsymbol{\theta}\boldsymbol{\theta}^{*T}\} &= \sum_{k=0}^{N-1} E\left\{\theta_k^2\right\} = \sum_{k=0}^{N-1} \sigma_k^2 \\ E\{\mathbf{x}\mathbf{x}^{*T}\} &= \sum_{n=0}^{N-1} \sigma_x^2(n) = N\sigma_x^2 \\ \sigma_x^2 &= \frac{1}{N}\sum_{k=0}^{N-1} \sigma_k^2 \end{aligned} \tag{10}$$

It is noted that the linear transformation of a stationary random vector process results in a non-stationary random vector process. In practice, the variance values of transform coefficients $\{\sigma_k^2\}$ are desired to have smaller geometric mean in order to justify the transform domain processing of random signals. KLT provides optimal geometric mean of coefficient variances with a diagonal correlation matrix $\mathbf{R}_\theta$ with best possible repacking of signal vector energy into as few transform coefficients as possible. The compaction efficiency of a transform is defined as

$$\eta_E(L) = \frac{\sum_{k=0}^{L-1}\sigma_k^2}{N\sigma_x^2}; \quad L = 1, 2, \ldots, N-1 \tag{11}$$

This is an important metric to assess the efficiency of a transform for the given signal type. The gain of transform coding over pulse code modulation (PCM) performance of an $N \times N$ unitary transform for a given input correlation is particularly significant and widely utilized in transform coding applications as defined [5]

$$G_{\mathrm{TC}}^N = \frac{\frac{1}{N}\sum_{k=0}^{N-1}\sigma_k^2}{\left(\prod_{k=0}^{N-1}\sigma_k^2\right)^{1/N}} \tag{12}$$

Similarly, decorrelation efficiency of a transform is defined as

$$\eta_c = 1 - \frac{\sum_{k=0}^{N-1}\sum_{l=1;l\neq k}^{N-1}|\mathbf{R}_\theta(k,l)|}{\sum_{k=0}^{N-1}\sum_{l=1;l\neq k}^{N-1}|\mathbf{R}_x(k,l)|} \tag{13}$$

Note that $\eta_c = 1$ for KLT where transform coefficients are perfectly decorrelated (pairwise), and signal energy is optimally packed as measured in (12) and (13) for the given $\mathbf{R}_x$ and transform size N. Therefore, KLT is the optimal block transform for a given input statistics offering the best possible performance with high computational cost. Its basis set needs to be recalculated whenever signal statistics changes. In contrast, fixed transform DCT with efficient implementation is an attractive alternative to KLT particularly for highly correlated processes. We highlight this point in the following section.

B. AR(1) Signal Model

Random signal sources are mathematically described by a variety of models including auto-regressive (AR), moving average

(MA), and auto-regressive moving average (ARMA) types. AR source models, also called all-pole models, have been successfully used in speech compression and recognition applications for decades. AR source model with order one, AR(1), is a first approximation to many natural signals, and it has been widely employed in various engineering applications. AR(1) signal is generated through the regression formula as written [5]

$$x(n) = \rho x(n-1) + w(n) \tag{14}$$

where $w(n)$ is a white noise with zero mean. The first order correlation coefficient is in the range of $-1 < \rho < 1$, and the noise variance is related to the variance of $x(n)$ as follows

$$\sigma_w^2 = E\{w(n)w(n+k)\} = (1-\rho^2)\sigma_x^2\delta(k) \tag{15}$$

Auto-correlation sequence of $x(n)$ is expressed as

$$R_{xx}(k) = \sigma_x^2\rho^{|k|}; k = 0, \pm 1, \pm 2, \ldots \tag{16}$$

The resulting Toeplitz correlation matrix of size $N \times N$ is defined as

$$\mathbf{R}_x = \sigma_x^2 \begin{bmatrix} 1 & \rho & \rho^2 & \cdots & \rho^{N-1} \\ \rho & 1 & \rho & \cdots & \rho^{N-2} \\ \rho^2 & \rho & 1 & \cdots & \rho^{N-3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho^{N-1} & \rho^{N-2} & \rho^{N-3} & \cdots & 1 \end{bmatrix} \tag{17}$$

AR(1) model is utilized to approximate empirical correlations of asset returns in the following sections of the paper with the purpose of developing an analytical framework, and also a fast KLT approximation by using DCT as a replacement.

C. Eigen Decomposition of AR(1) Process

An eigenvalue λ and an eigenvector $\boldsymbol{\phi}$ of a matrix Φ with size $N \times N$ must satisfy the eigenvalue [5], [6]

$$\Phi\boldsymbol{\phi} = \lambda\boldsymbol{\phi} \tag{18}$$

It is rewritten as

$$\Phi\boldsymbol{\phi} - \lambda\mathbf{I}\boldsymbol{\phi} = (\Phi - \lambda\mathbf{I})\boldsymbol{\phi} = 0 \tag{19}$$

such that $(\Phi - \lambda\mathbf{I})$ is an invertible matrix. Namely,

$$\det(\Phi - \lambda\mathbf{I}) = 0 \tag{20}$$

This determinant requirement leads us, in general, to the characteristic polynomial

$$P(\lambda) = (\lambda - \lambda_1)^{n_1}(\lambda - \lambda_2)^{n_2}\ldots(\lambda - \lambda_l)^{n_{m_l}} \tag{21}$$

where

$$\sum_{i=1}^{m_l} n_i = N; 1 \le m_l \le n_l \tag{22}$$

This results in the eigenpair set $\{\lambda_l, \boldsymbol{\phi}\}$ where $1 \le l \le m_l$ for a single eigenvector. It is emphasized that if more than one eigenvector share the same eigenvalue, those eigenvectors along with the zero vector form a linear subspace of the vector space, and it is called an eigenspace. Moreover, the eigenvectors with different eigenvalues are linearly independent, and matrix Φ with size $N \times N$ is called defective if it does not have linearly independent eigenvectors. For the case of diagonal matrix, its eigenvectors are basis vectors and eigenvalues are its component values on the diagonal. Now, for a non-defective Φ matrix with distinct eigenvectors one can write the following eigenmatrix equation

$$\Phi\mathbf{A}_{\mathrm{KLT}}^{*T} = \mathbf{A}_{\mathrm{KLT}}^{*T}\boldsymbol{\Lambda}$$
$$\Phi = \mathbf{A}_{\mathrm{KLT}}^{*T}\Lambda\mathbf{A}_{\mathrm{KLT}} = \sum_{k=0}^{N-1}\lambda_k\boldsymbol{\phi}_k\boldsymbol{\phi}_k^{*T} \tag{23}$$

where $\boldsymbol{\Lambda} = \mathrm{diag}(\lambda_k); k = 0, 1, \ldots, N-1$, and kth column of $\mathbf{A}_{\mathrm{KLT}}^{*T}$ matrix is the kth eigenvector $\boldsymbol{\phi}_k$ of Φ with the corresponding eigenvalue λ_k. Note that $\lambda_k = \sigma_k^2$ for the given signal statistics as $\mathbf{R}_x = \Phi$ in (8). The eigenvalues for an AR(1) source model of (14) is calculated in closed form as follows [3]

$$\sigma_k^2 = \lambda_k = \frac{1-\rho^2}{1 - 2\rho\cos(\omega_k) + \rho^2}; \quad 0 \le k \le N-1 \tag{24}$$

where $\{\omega_k\}$ are the positive roots of the polynomial

$$\tan(N\omega_k) = -\frac{(1-\rho^2)\sin(\omega_k)}{\cos(\omega_k) - 2\rho + \rho^2\cos(\omega_k)} \tag{25}$$

and the resulting matrix $\mathbf{A}_{\mathrm{KLT}} = [A(k,n)]$ of size $N \times N$ is calculated as

$$[A(k,n)] = \frac{2}{N+\lambda_k}\sin\left[\omega_k\left(n - \frac{N-1}{2}\right) + \frac{(k+1)\pi}{2}\right] \tag{26}$$

where $0 \le k, n \le N-1$. KLT is the signal dependent unique transform with jointly optimal energy packing and perfect decorrelation features for a given $\mathbf{R}_x$. The computation of the KLT transform is difficult in practice. Therefore, fixed transforms are preferred in many applications that are concerned with the implementation cost of KLT. In contrast to input dependent KLT, Discrete Cosine Transform (DCT) is a fixed transform and offers efficient implementation algorithms. DCT matrix of size N is defined as [7]

$$\mathbf{A}_{\mathrm{DCT}} = [A(k,n)] = \frac{1}{c_k}\cos\left[\frac{(2n+1)k\pi}{2N}\right] \tag{27}$$

where $0 \le k, n \le N-1$ and

$$c_k = \begin{cases} \sqrt{N} & k = 0 \\ \sqrt{N/2} & k \ne 0 \end{cases}$$

It has been shown in the literature that the basis vectors of DCT approach to the eigenvectors of AR(1) process as the correlation coefficient goes to one [4]. It was reported that the DCT basis

functions are eigenvectors of a symmetric tridiagonal matrix as defined

$$\mathbf{R}_{\mathrm{DCT}} = \begin{bmatrix} (1-\alpha) & -\alpha & 0 & \cdots & 0 \\ -\alpha & 1 & -\alpha & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \vdots & \vdots & -\alpha & 1 & -\alpha \\ 0 & 0 & 0 & -\alpha & (1-\alpha) \end{bmatrix} \tag{28}$$

Similarly, the covariance matrix of an AR(1) process with the correlation coefficient ρ has the form

$$\mathbf{R}_{\mathrm{AR(1)}} = \frac{1}{\beta^2} \begin{bmatrix} (1-\rho\alpha) & -\alpha & 0 & \cdots & 0 \\ -\alpha & 1 & -\alpha & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \vdots & \vdots & -\alpha & 1 & -\alpha \\ 0 & 0 & 0 & -\alpha & (1-\rho\alpha) \end{bmatrix} \tag{29}$$

where

$$\beta^2 = \frac{1-\rho^2}{1+\rho^2}, \quad \alpha = \frac{\rho}{1+\rho^2}$$

Therefore, it is shown that

$$\beta^2 \mathbf{R}^{-1}_{\mathrm{AR(1)}}|_{\rho \to 1} \cong \mathbf{R}_{\mathrm{DCT}}$$

This very nature of DCT has made it a popular transform that is successfully employed for decomposition of highly correlated signal sources. In particular, image and video compression standards like JPEG and MPEG use DCT based 2-D transform coding. In Fig. 1(a) $\eta_E(L)$ of KLT and DCT as defined in (11) are displayed for various values of correlation coefficient ρ and transform size $N = 31$. Similarly, Fig. 1(b) depicts relative G^N_{TC} performance of (12) for KLT and DCT as a function of ρ for $N = 31$. This figure verifies the use of DCT as a replacement to KLT in image and video processing applications where signals are highly correlated. Moreover, it is noted that the energy packing performance of both transforms degrade for lower values of correlation coefficient. The readers of more interest on the theory of signals and transforms are referred to [5].

III. Toeplitz Approximation to Empirical Correlation Matrix

In this section, we attempt to approximate empirical correlations of asset returns by utilizing AR(1) signal source model as discussed earlier. The main motivation here is to incorporate the closed form expressions of eigenvalues and eigenvectors of AR(1) sources as expressed in (24)–(26) that are utilized for eigenfiltering of empirical correlation matrix accordingly. This procedure will be explained later in Section IV. Moreover, we highlight potential use of DCT for noise filtering of empirical correlation matrix as a replacement to KLT where the former is very efficient to implement.

We consider 30 stocks of the index Dow Jones Industrial Average (DJIA) along with the exchange traded fund (ETF) called Dow Jones Industrial Average (DIA) that mimics DJIA (total of N = 31 assets) for most experiments reported in this section.

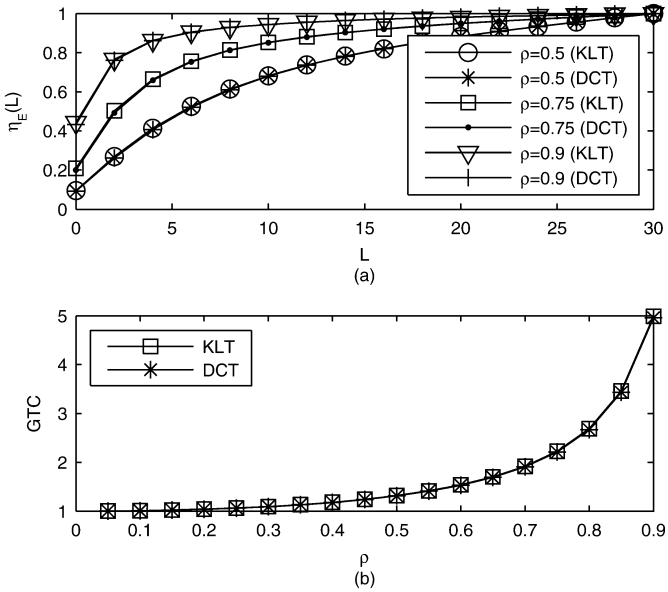

Fig. 1. (a) $\eta_E(L)$ performance of KLT and DCT for various values of ρ and $N = 31$, (b) G^N_{TC} performance of KLT and DCT as a function of ρ for $N = 31$.

Return of the kth asset of the portfolio at discrete-time n is defined as follows

$$r_k(n) = \frac{p_k(n)}{p_k(n-1)} - 1; \quad k = 1, 2, \ldots, 31 \tag{30}$$

where $p_k(n)$ is its price. The mean and variance of $r_k(n)$ are calculated for a measurement window size of W with the ergodicity assumption as written

$$\eta_k(n) = E\{r_k(n)\} = \frac{1}{W}\sum_{m=0}^{W-1} r_k(n-m)$$

$$\sigma_k^2(n) = E\left\{r_k^2(n)\right\} - \eta_k^2(n) = \left[\frac{1}{W}\sum_{m=0}^{W-1} r_k^2(n-m)\right] - \eta_k^2(n) \tag{31}$$

where $k = 1, 2, \ldots, 31$ and $\sigma_k(n)$ is called the volatility of the kth asset at time n that is a widely used risk metric in finance. The return vector of all assets at time n is written as

$$\mathbf{r}(n) = [r_k(n)]; \quad k = 1, 2, \ldots 31 \tag{32}$$

The empirical correlation matrix of returns at time n is defined as

$$\mathbf{R}_E(n) \triangleq [E\{\mathbf{r}(n)\mathbf{r}^T(n)\}] = [R_{k,l}(n)] = \begin{bmatrix} R_{1,1}(n) & R_{1,2}(n) & \cdots & R_{1,31}(n) \\ R_{2,1}(n) & R_{2,2}(n) & \cdots & R_{2,30}(n) \\ \vdots & \vdots & \ddots & \vdots \\ R_{31,1}(n) & R_{31,2}(n) & \cdots & R_{31,31}(n) \end{bmatrix} \tag{33}$$

where

$$R_{k,l}(n) = E\{r_k(n) r_l(n)\} = \frac{1}{W}\sum_{m=0}^{W-1} r_k(n-m) r_l(n-m)$$

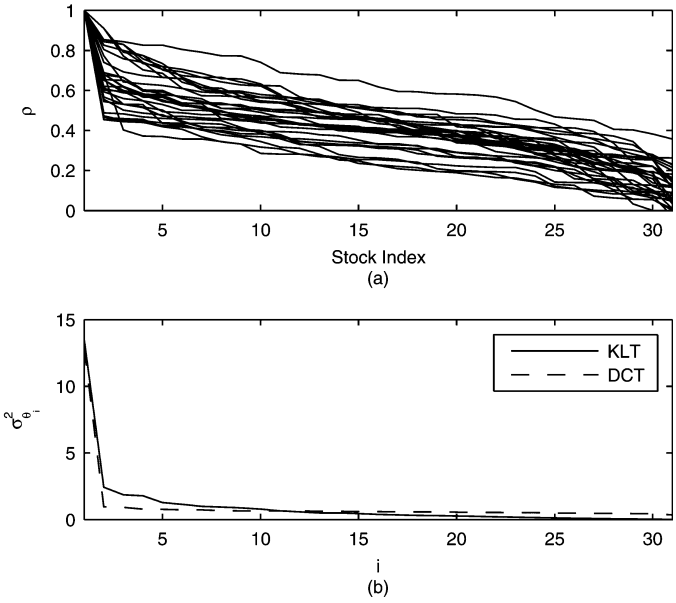

Fig. 2. (a) Rows of $\mathbf{R}_E$ matrix displayed in descending order, (b) KLT and DCT coefficient variances for $\mathbf{R}_E$ of DJIA & DIA EOD returns calculated at time $n = 16:00$ with $N = 31$.

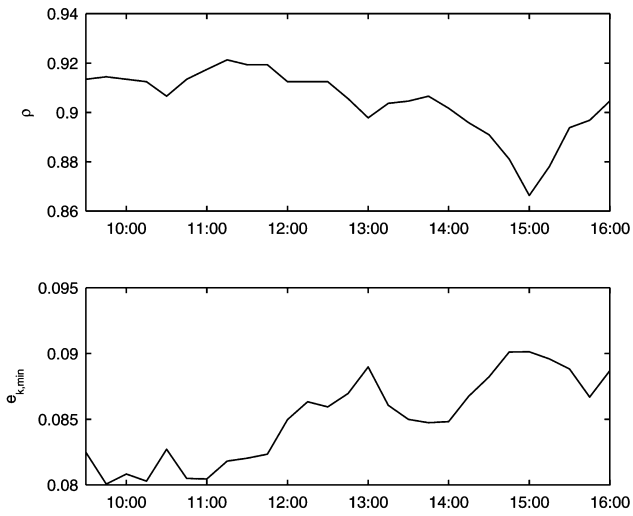

Fig. 3. Variations of optimal correlation coefficient and the resulting error of AR(1) approximation, (35), as a function of time with 15 minute sliding intervals with $W = 60$ days for 24 hour returns of 31-asset portfolio (DJIA & DIA) in the interval $n = 9:30 - 16:00$.

represents measured correlation between returns of kth and lth assets with a measurement window of W samples. Note that the returns are normalized to be zero mean and unit variance. The empirical correlation matrix $\mathbf{R}_E(n)$ is real, symmetric and positive definite. Fig. 2(a) displays its elements for each row in a descending order for 31 assets being considered in this section and calculated at the end of day (EOD) for $W = 60$ days. The kth sequence represents pairwise correlations of the kth asset with all assets of the portfolio. For simplicity, we drop the time variable in (33) and rewrite it as

$$\mathbf{R}_E = \begin{bmatrix} R_E(1,1) & R_E(1,2) & \cdots & R_E(1,N) \\ R_E(2,1) & R_E(2,2) & \cdots & R_E(2,N) \\ \vdots & \vdots & \ddots & \vdots \\ R_E(N,1) & R_E(N,2) & \cdots & R_E(N,N) \end{bmatrix} \tag{34}$$

$\mathbf{R}_E$ is normalized such that $R_E(k,k) = 1\ \forall_k$. Fig. 2(b) shows variances of corresponding KLT and DCT coefficients for the empirical correlation matrix $\mathbf{R}_E$ of DJIA & DIA EOD returns for $W = 60$ days as displayed graphically in Fig. 2(a). This figure confirms their similar behavior for financial signals of this type.

We consider two cases where AR(1) signal model with Toeplitz correlation matrix is employed to approximate symmetric $\mathbf{R}_E$ matrix as follows.

A. AR(1) Approximation to Empirical Correlation Matrix $\mathbf{R}_E$

We find the optimal correlation coefficient ρ_{opt} of AR(1) source as stated in (16) that minimizes the approximation error

$$e = \frac{1}{N^2}\sum_{k=1}^{N}\sum_{l=1}^{N}\left[R_E(k,l) - \rho_{\text{opt}}^{|k-l|}\right]^2 \tag{35}$$

Accordingly, symmetric empirical correlation matrix is approximated by a Toeplitz matrix as

$$\tilde{\mathbf{R}}_E = \begin{bmatrix} 1 & \rho_{\text{opt}} & \cdots & \rho_{\text{opt}}^{N-1} \\ \rho_{\text{opt}} & 1 & \cdots & \rho_{\text{opt}}^{N-2} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{\text{opt}}^{N-1} & \rho_{\text{opt}}^{N-2} & \cdots & 1 \end{bmatrix} \tag{36}$$

Therefore, one can calculate the resulting eigenvalues and eigenvectors of AR(1) model according to (24) and (26) as approximations to their measured values, respectively. Fig. 3 displays variations of correlation coefficient ρ_{opt} for 31 assets of DJIA & DIA under consideration along with approximation errors of (35). The returns are measured for 24 hour intervals with sliding time intervals of 15 minutes and measurement window of $W = 60$ trading days for a trading day of 6.5 hours in this figure. The last sample corresponds to EOD return of an asset. This figure shows highly correlated nature of EOD and 24 hour returns along with the merit of Toeplitz approximation to empirical correlation matrix.

B. AR(1) Approximation to Each Row of Empirical Correlation Matrix $\mathbf{R}_E$

In this method we approximate each row of empirical correlation matrix by the optimal correlation sequence of AR(1) signal model with the correlation coefficients $\{\rho_{k,\text{opt}}\}$. Hence, the rows are approximated as

$$\hat{\mathbf{R}}_E = \begin{bmatrix} 1 & \rho_{1,\text{opt}} & \cdots & \rho_{1,\text{opt}}^{N-1} \\ \rho_{2,\text{opt}} & 1 & \cdots & \rho_{2,\text{opt}}^{N-2} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{N,\text{opt}}^{N-1} & \rho_{N,\text{opt}}^{N-2} & \cdots & 1 \end{bmatrix} \tag{37}$$

where the optimum $\rho_{k,\mathrm{opt}}$ for the kth row of $\mathbf{R}_E$ (kth asset of a portfolio) is obtained by minimizing the approximation error as defined

$$e_k = \frac{1}{N}\sum_{l=1}^{N}[R_E(k,l) - \hat{R}_E(k,l)]^2 \tag{38}$$

and, $\hat{R}_E(k,l)$ is the element of matrix $\hat{\mathbf{R}}_E$ located at the kth row and lth column. Then, each row is AR(1) approximated independently and we rewrite (37) as

$$\hat{\mathbf{R}}_E = \sum_{k=1}^{N}\mathbf{S}_k\mathbf{R}_{E,k} \tag{39}$$

where the selection matrix $\mathbf{S}_k$ is defined as

$$\mathbf{S}_k \triangleq \begin{cases} s_{k,k}=1 & \text{for } k \\ 0 & \text{otherwise} \end{cases};\ k=1,2,\ldots,N \tag{40}$$

and, the resulting $\mathbf{R}_{E,k}$ matrix is a Toeplitz matrix as expressed

$$\mathbf{R}_{E,k} = \begin{bmatrix} 1 & \rho_{k,\mathrm{opt}} & \cdots & \rho_{k,\mathrm{opt}}^{N-1} \\ \rho_{k,\mathrm{opt}} & 1 & \cdots & \rho_{k,\mathrm{opt}}^{N-2} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{k,\mathrm{opt}}^{N-1} & \rho_{k,\mathrm{opt}}^{N-2} & \cdots & 1 \end{bmatrix} \tag{41}$$

for $k=1,2,\ldots,N$. Similarly, we decompose each $\mathbf{R}_{E,k}$ to its eigenvectors individually as shown

$$\mathbf{R}_{E,k} = \mathbf{A}_{\mathrm{KLT},k}^T\mathbf{\Lambda}_k\mathbf{A}_{\mathrm{KLT},k};\ k=1,2,\ldots,N \tag{42}$$

Therefore, we can rewrite the Toeplitz approximation of (39) as

$$\hat{\mathbf{R}}_E = \sum_{k=1}^{N}\mathbf{S}_k\mathbf{A}_{\mathrm{KLT},k}^T\mathbf{\Lambda}_k\mathbf{A}_{\mathrm{KLT},k} \tag{43}$$

where $\mathbf{A}_{\mathrm{KLT},k}$ and $\mathbf{\Lambda}_k$ are comprised of the kth set of eigenvectors and eigenvalues, respectively. Then, we calculate the resulting eigenvalues and eigenvectors according to the closed form expressions of (24) and (26) for the given set of AR(1) correlation coefficients $\{\rho_{k,\mathrm{opt}}\}$. Fig. 4 displays variations of correlation coefficients and resulting approximation errors of this method for 31 assets under consideration. Similarly, the returns are for 24-hour intervals with 15 minute sliding windows for a trading day of 6.5 hours in these figures. It is noted that the approximation error of this method is lower than the first one. The trade-off is the increased computational cost of the multiple Toeplitz approximations. The histogram for correlation coefficients of Fig. 4 is shown in Fig. 5. The resulting mean and variance values are 0.8756 and 0.0125, respectively. These correlation coefficient values coupled with the KLT and DCT performance comparisons displayed in Fig. 1 suggest the use of DCT as fast KLT approximation for this type of signals.

IV. Portfolio Risk and Eigenfiltering of Empirical Correlation Matrix

Eigen decomposition of empirical correlation matrix of asset returns in a portfolio has been widely employed for risk analysis and rebalancing [2], [8]. The return of an *N*-asset portfolio at time n is calculated as

$$R_p(n) = \sum_{k=1}^{N} d_k(n)r_k(n) = \mathbf{d}^T\mathbf{r} \tag{44}$$

where $\mathbf{r}$ is the return vector, and $\mathbf{d}$ is the normalized investment vector of the portfolio with

$$\mathbf{d}^T\mathbf{1} = 1;\ \mathbf{1} \triangleq [1 \quad 1 \quad \cdots \quad 1]^T \tag{45}$$

Therefore, we can calculate mean of portfolio return at time n as

$$\eta_p(n) = \sum_{k=1}^{N} d_k(n)\eta_k(n) = \mathbf{d}^T\boldsymbol{\eta} \tag{46}$$

Elements of vector $\boldsymbol{\eta}$ are expected returns of assets in the portfolio. We can also calculate the variance of portfolio return at time n as

$$\begin{aligned}\sigma_p^2(n) &= E\left\{r_k^2(n)\right\} - \eta_p^2(n) \\ &= \mathbf{d}^T(n)\mathbf{V}^T(n)\mathbf{R}_E(n)\mathbf{V}(n)\mathbf{d}(n) \\ &= \sum_{k=1}^{N}\sum_{k=1}^{N} d_k(n)d_l(n)R_E(k,l)\sigma_k(n)\sigma_l(n)\end{aligned} \tag{47}$$

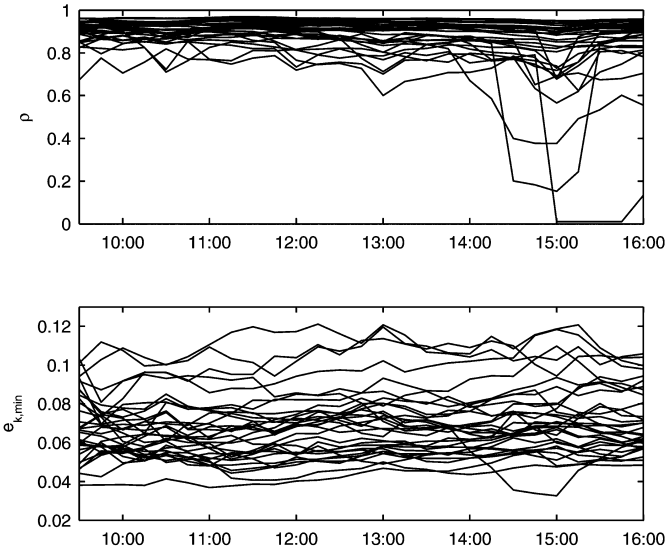

Fig. 4. Variations of correlation coefficients and the resulting errors of AR(1) approximations as a function of time with 15 minute sliding intervals for 24 hour returns of 31-asset portfolio (DJIA & DIA) with $W = 60$ days in the interval $n = 9:30-16:00$.

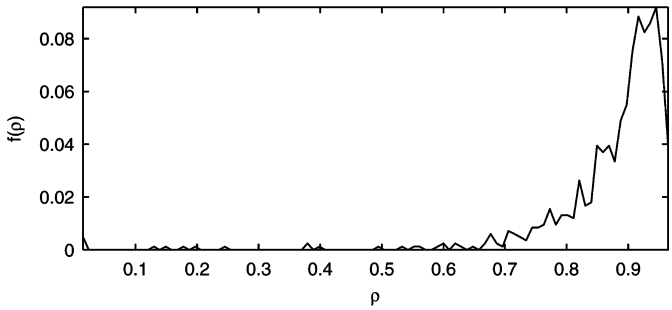

Fig. 5. Histogram of correlation coefficients displayed in Fig. 4.

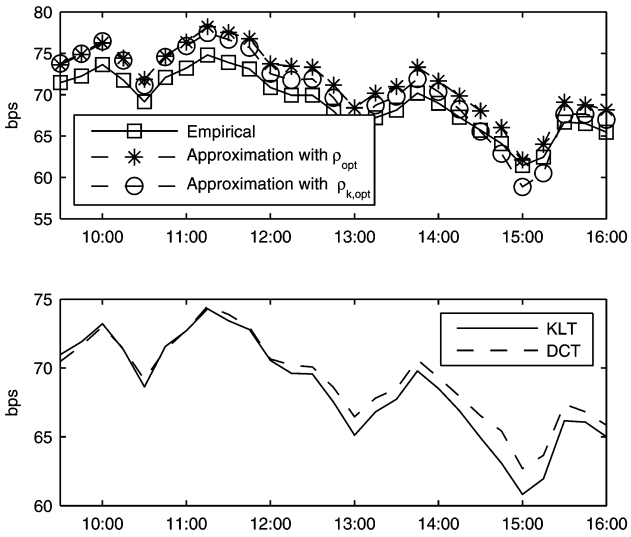

Fig. 6. (a) Portfolio risk for symmetric empirical correlation matrix $\mathbf{R}_E(\mathrm{n})$, and its Toeplitz approximations $\bar{\mathbf{R}}_E$ of (36), and $\hat{\mathbf{R}}_E$ of (37) as a function of time with 15 minute sliding intervals for 24 hour returns and $W = 60$ trading days of 31-asset portfolio (DJIA & DIA) in the interval $n = 9:30 - 16:00$, (b) Portfolio risk calculated from (48) using two versions of filtered empirical correlation matrix $\mathbf{R}_E(\mathrm{n})$ for 5 factors as a function of time with 15 minute sliding intervals and $W = 60$ trading days for 24 hour returns of DJIA & DIA in the interval $n = 9:30 - 16:00$.

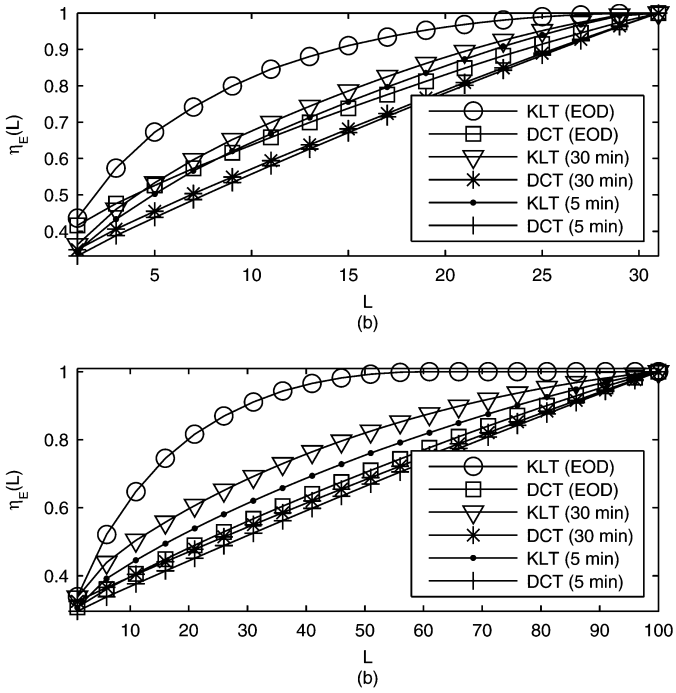

Fig. 7. (a) Compaction efficiencies of KLT and DCT for multiple intervals (frequencies) of 31-asset portfolio DJIA & DIA, (b) Compaction efficiencies of KLT and DCT for multiple intervals of 100-asset portfolio NASDAQ100.

where $\mathbf{V}(n) = \mathrm{diag}[\sigma_k(n)]$. Now, we introduce eigendecomposition of $\mathbf{R}_E$ as defined in (23), and rewrite the portfolio variance as follows

$$\begin{aligned}\sigma_p^2(n) &= \mathbf{d}^T(n)\mathbf{V}^T(n)\mathbf{R}_E(n)\mathbf{V}(n)\mathbf{d}(n) \\ &= \mathbf{d}^T(n)\mathbf{V}^T(n)\mathbf{A}_{\mathrm{KLT}}^T\mathbf{\Lambda}\mathbf{A}_{\mathrm{KLT}}\mathbf{V}(n)\mathbf{d}(n) \quad (48)\end{aligned}$$

where

$$\mathbf{R}_E(n) = \mathbf{A}_{\mathrm{KLT}}^T\mathbf{\Lambda}\mathbf{A}_{\mathrm{KLT}} = \sum_{k=1}^{N}\lambda_k\boldsymbol{\phi}_k\boldsymbol{\phi}_k^T \quad (49)$$

One may utilize Toeplitz approximation to symmetric matrix $\mathbf{R}_E(n)$ as described in Section III. This provides an analytical framework for further studies. Moreover, it suggests to reduce computational cost of eigen decomposition, particularly for large values of N, by employing DCT as a fast KLT approximation.

Fig. 6(a) displays portfolio risk $\sigma_p(n)$ calculated from (48) for empirical correlation matrix $\mathbf{R}_E(n)$, and its Toeplitz approximations $\bar{\mathbf{R}}_E(n)$ of (36), and $\hat{\mathbf{R}}_E(n)$ of (43) as a function of time with 15 minute sliding intervals and a measurement window of $W = 60$ trading days for 24 hour returns of 31-asset portfolio DJIA & DIA.

Eigenfiltering utilizes eigen decomposition where a subset of L most significant eigenvalues, $L \ll N$, along with eigenvectors are kept to represent filtered empirical correlation matrix as expressed

$$\mathbf{R}_E^S(n) = \sum_{k=1}^{L}\lambda_k\boldsymbol{\phi}_k\boldsymbol{\phi}_k^T \quad (50)$$

and with the remaining noise matrix

$$\mathbf{R}_E^\epsilon(n) = \sum_{k=L+1}^{N}\lambda_k\boldsymbol{\phi}_k\boldsymbol{\phi}_k^T \quad (51)$$

where one can write the equation

$$\mathbf{R}_E(n) = \mathbf{R}_E^S(n) + \mathbf{R}_E^\epsilon(n) \quad (52)$$

In order to preserve the total energy in transform domain where $\sum_{k=1}^{N}\lambda_k = N$ we introduce the diagonal matrix $\hat{\mathbf{R}}_E^\epsilon(n) = \mathrm{diag}[\mathbf{R}_E^\epsilon(n)]$ as included in the modified version of $\mathbf{R}_E(n)$ as

$$\bar{\mathbf{R}}_E(n) = \mathbf{R}_E^S(n) + \hat{\mathbf{R}}_E^\epsilon(n) \quad (53)$$

Fig. 6(b) compares portfolio risks of 24 hour returns of 31-asset portfolio DJIA & DIA calculated from (48) employing KLT and DCT filtering methods with five factors $(L = 5)$ as a function of time for 15 minute sliding intervals in a given 6.5 hour long trading day. It is observed from the figure that their performances are very similar. Fig. 7 displays compaction efficiencies, (11), of KLT and DCT for three different empirical correlation matrices calculated for sampling (rebalancing) periods of 24 hours (EOD), 30 minutes, and 5 minutes, and for two different portfolios. Namely, they are of 31-asset DJIA & DIA and 100-asset NASDAQ100 portfolios. It is observed from these figures that KLT and DCT perform similarly for the filtering of empirical correlation matrices of asset returns experimented in various frequencies [2], [8]–[11] and two portfolios. It is noted that their energy compaction performance degrades when the sampling interval is reduced where the value of correlation coefficient drops due to the Epps Effect [12], [13].

V. Conclusion

In this paper, we introduced two methods for Toeplitz approximation to symmetric empirical correlation matrix of asset returns in a portfolio with corresponding closed form expressions for eigen decomposition. We showed the merit of the proposed framework through two portfolios, DJIA & DIA and NASDAQ100. Furthermore, we forwarded and verified the use of DCT as a fast and effective KLT approximation for the analysis of an empirical correlation matrix. It is concluded that the proposed analytical framework and easy implementation may be used in data-intensive finance applications.

References

[1] H. M. Markowitz, *Portfolio Selection: Efficient Diversification of Investments*. New York: Wiley, 1959.

[2] V. Plerou, P. Gopikrishnan, B. Rosenow, L. A. N. Amaral, T. Guhr, and H. E. Stanley, "Random matrix approach to cross correlations in financial data," *Phys. Rev. E*, vol. 65, pp. 066 126-1–066 126-18, Jun. 2002.

[3] W. Ray and R. Driver, "Further decomposition of the Karhunen-Loeve series representation of a stationary random process," *IEEE Trans. Inf. Theory*, vol. IT-16, no. 6, pp. 663–668, Nov. 1970.

[4] A. Jain, "A fast Karhunen-Loeve transform for a class of random processes," *IEEE Trans. Communications*, vol. COM-24, no. 9, pp. 1023–1029, Sep. 1976.

[5] A. N. Akansu and R. A. Haddad, *Multiresolution Signal Decomposition: Transforms, Subbands, and Wavelets*. New York: Academic, 1992.

[6] K. Pearson, "On lines and planes of closest fit to systems of points in space," *Philosoph. Mag.* no. 2, pp. 559–572, 1901 [Online]. Available: http://stat.smmu.edu.cn/history/pearson1901.pdf

[7] N. Ahmed, T. Natarajan, and K. Rao, "Discrete cosine transfom," *IEEE Trans. Comput.*, vol. C-23, no. 1, pp. 90–93, Jan. 1974.

[8] M. U. Torun, A. N. Akansu, and M. Avellaneda, "Portfolio risk in multiple frequencies," *IEEE Signal Process. Mag., Spec. Iss. Signal Process. for Finan. Applicat.*, vol. 28, no. 5, pp. 61–71, Sep. 2011.

[9] J. Bouchaud and M. Potters, *Theory of Financial Risk and Derivative Pricing*. Cambridge, U.K.: Cambridge Univ. Press, 2003.

[10] T. Conlon, H. Ruskin, and M. Crane, "Cross-correlation dynamics in financial time series," *Physica A: Statist. Mech. and its Applicat.*, vol. 388, no. 5, pp. 705–714, 2009.

[11] N. El Karoui, "Concentration of measure and spectra of random matrices: Applications to correlation matrices, elliptical distributions and beyond," *Ann. Appl. Probab.*, vol. 19, no. 6, pp. 2362–2405, 2009.

[12] T. W. Epps, "Comovements in stock prices in the very short run," *J. Amer. Statist. Assoc.*, vol. 74, no. 366, pp. 291–298, 1979.

[13] M. U. Torun and A. N. Akansu, "On Epps effect and rebalancing of hedged portfolio in multiple frequencies," in *Proc. 4th IEEE Int. Workshop Comput. Adv. in Multi-Sensor Adapt. Process.*, Dec. 2011.

Ali N. Akansu (F'08) received the B.S. degree from the Technical University of Istanbul, Turkey, and the M.S. and Ph.D. degrees from the Polytechnic University, Brooklyn, New York, all in electrical engineering. Since 1987, he has been with the New Jersey Institute of Technology, where he is a Professor of Electrical and Computer Engineering. Dr. Akansu has administered and managed a number of research programs and product development projects in academia and private sector, funded by the State & Federal Government agencies, and industry. He was a Founding Director of the New Jersey Center for Multimedia Research (NJCMR) between 1996–2000, and NSF Industry-University Cooperative Research Center (IUCRC) for Digital Video between 1998–2000. Dr. Akansu was the Vice President for Research and Development of IDT Corporation [NYSE: IDT]. He was the founding President and CEO of PixWave, Inc., and Senior Vice President for Technology Development of TV.TV, IDT subsidiaries. He was an academic visitor at David Sarnoff Research Center, IBM T. J. Watson Research Center and at GEC-Marconi Electronic Systems Corp. He was also a Visiting Professor at the Courant Institute of Mathematical Sciences of NYU. He regularly consults to the legal sector and industry, and has sat on the boards of several companies.

Dr. Akansu has published numerous articles and books, gave invited talks, guided theses on the theory of signals and transforms, and their applications in image/video coding, digital communications, Internet multimedia and information security, and quantitative finance. He is a co-author (with R. A. Haddad) of the book *Multiresolution Signal Decomposition: Transforms, Subbands and Wavelets*, Academic Press, 1992 and 2001 (2nd Ed.), and a co-editor (with M. J. T. Smith) of a book entitled *Subband and Wavelet Transforms: Design and Applications*, Kluwer, 1996. He is a co-editor of the book (with M. J. Medley) *Wavelet, Subband and Block Transforms in Communications and Multimedia*, Kluwer, 1999. He is also a co-author of a research monograph (with H. T. Sencar and M. Ramkumar) *Data Hiding Fundamentals and Applications: Content Security in Digital Multimedia*, Elsevier-Academic Press, 2004.

Dr. Akansu is a Fellow of the IEEE. He served as an associate editor of IEEE Transactions on Signal Processing, and IEEE Transactions on Multimedia, as a member of the Signal Processing Theory & Methods, and Multimedia Signal Processing technical committees of the IEEE Signal Processing Society. He organized the first Wavelets Conference in the United States in April 1990. He was the technical program chairman of IEEE Digital Signal Processing Workshop 1996, Loen, Norway. He served as a member of the Steering Committee and the Publications Chair of IEEE ICASSP 2000, Istanbul, Turkey. He was the Lead Guest Editor of the two special issues of IEEE Transactions on Signal Processing on Theory and Application of Filter Banks and Wavelet Transforms (April 1998), and on Signal Processing for Data Hiding in Digital Media and Secure Content Delivery (June 2003). He is also the Lead Guest Editor of the special issue of the IEEE Journal of Special Topics on Signal Processing on Signal Processing Methods in Finance and Electronic Trading (August 2012).

Mustafa U. Torun (mustafa.torun@njit.edu) received his B.S. and M.S. degrees from the Dokuz Eylul University (D.E.U.), Izmir, Turkey, in 2005 and 2007 respectively, both in electrical and electronics engineering. He was a research assistant in the Department of Electrical and Electronics Engineering at D.E.U. from 2005 to 2008. Since 2008, he has been a Ph.D. candidate in the Department of Electrical and Computer Engineering at the New Jersey Institute of Technology, Newark, NJ. His research interests include high-performance computing, data-intensive research in signal processing, multi-resolution signal processing, statistical signal processing, pattern classification, neural networks, genetic algorithms; and their applications in quantitative finance, electronic trading, digital communications, digital imaging, and biomedical engineering.