Biometrika (201?), **0**, 0, *pp*. 1–18

Printed in Great Britain

Advance Access publication on

High-dimensional volatility matrix estimation via wavelets and thresholding

BY P. FRYZLEWICZ

Department of Statistics, London School of Economics, Houghton Street, London WC2A 2AE, U.K.

p.fryzlewicz@lse.ac.uk

SUMMARY

We propose a locally stationary linear model for the evolution of high-dimensional financial returns, where the time-varying volatility matrix is modelled as a piecewise constant function of time. We introduce a new wavelet-based technique for estimating the volatility matrix, which combines four ingredients: a Haar wavelet decomposition, variance stabilization of the Haar coefficients via the Fisz transform prior to thresholding, a bias correction, and extra time-domain thresholding, soft or hard. Under the assumption of sparsity, we demonstrate the interval-wise consistency of the proposed estimators of the volatility matrix and its inverse in the operator norm, with rates which adapt to the features of the target matrix. We also propose a version of the estimators based on the polarization identity, which permits a more precise derivation of the thresholds. We discuss the practicalities of the algorithm, including parameter selection and how to perform it online. A simulation study shows the benefits of the method, which is illustrated using a stock index portfolio.

Some key words: Financial return; Haar–Fisz transformation; High dimensionality; Local stationarity; Sparsity; Thresholding; Volatility matrix; Wavelet.

1. INTRODUCTION

The estimation of volatility matrices, i.e., covariance matrices of multivariate asset returns, has been a fundamental problem in financial statistics at least since the seminal work of Markowitz (1952, 1959). Allocating a Markowitz-efficient portfolio in practice requires accurate estimation of the associated volatility matrix and its inverse. In another interesting application, an estimate of the volatility matrix is required in the estimation of factors and their loadings in the factor analysis of panels of asset returns, see, e.g., Motta et al. (2011).

Volatility, be it univariate or multivariate, is a model-dependent quantity, and its interpretation and estimation varies between models. For example, considering the univariate situation, in the ARCH model (Engle, 1982) and its many subsequent variants (Lunde & Hansen, 2005), volatility is understood as the variance of the returns process conditional on its own past values; in stochastic volatility modelling (Taylor, 1986; Andersen et al., 2009) it is the variance conditional on a possibly external random process; in the non-stationary deterministic approach of Starica & Granger (2005), Fryzlewicz (2005) and Fryzlewicz et al. (2006), it is the unconditional local variance of the returns process.

The latter approach offers a particularly easy way of introducing non-stationarity into volatility modelling. This is desirable, as some authors point out that the typical stylized facts of financial returns data, i.e., heavy-tailed marginal distribution and significant autocorrelation of absolute values and squares, can be better explained by resorting to non-stationary models, see, e.g., Mikosch & Starica (2004), Starica & Granger (2005) and Fryzlewicz et al. (2008) for arguments in the univariate case. Janeway (2009) goes further and claims that traditional models' stationarity might have been a contributing factor in the recent financial crisis. In this work, we consider multivariate volatility to be the local unconditional covariance matrix of asset returns, which varies over time. Rodriguez-Poo & Linton (2001) and Herzel et al. (2006) both assume a similar model, and use kernel smoothing for estimation.

When a fixed bandwidth is used, kernel volatility estimators are non-adaptive, which means they evolve at the same speed, irrespective of the current market conditions, which many practitioners find undesirable from the point of view of transaction costs. Thus, it makes sense to seek more adaptive estimators of multivariate volatility, which would adjust their speed of evolution as necessary. One such estimator is proposed by Härdle et al. (2003), who search for the longest interval of approximate constancy of volatility via iterative hypothesis testing.

In this work, we model the time-varying multivariate volatility as piecewise-constant, with the number of change-points possibly increasing with the sample size and approaching each other in rescaled time. This ensures that each component of our volatility matrix, viewed as a curve over time, can approximate an arbitrary piecewise-continuous curve in the limit. We propose a new wavelet-based technique for adaptive estimation of the time-evolving correlation matrix and covariance matrix of multivariate returns. The method combines Haar wavelets, nonlinear wavelet thresholding and the variance-stabilizing Fisz transformation. Haar wavelets are natural here as they furnish estimators which automatically adapt, interval-wise, to the piecewise constant volatility in terms of their rates of convergence. These estimators are fast to compute and are also valid at the right-hand end of the data, i.e., at the current time $t = T$.

We also propose an extra thresholding step in the time domain, which ensures that our estimator remains useful also in the high-dimensional setting, where the number of assets considered is perhaps even higher than the effective number of observations for each asset, provided the target volatility matrix is sparse. Bickel & Levina (2008) and El Karoui (2008) proposed thresholding estimators of a sparse stationary covariance matrix, and Wang & Zou (2010) adapted the former technique to the context of large stationary volatility matrix estimation for high-frequency financial data. In this work, we consider low-frequency data, but in a non-stationary setting.

Classical function estimation via wavelet thresholding in the function plus noise setting requires that the standard deviation of the noise should be constant over time. In our setting however, the standard deviation of the sample local cross-covariance is a function of the local cross-covariance itself, and thus variance stabilization is required. We adapt and use the Haar–Fisz technique (Fryzlewicz & Nason, 2004; Fryzlewicz, 2008), in which, roughly speaking, empirical wavelet coefficients are standardized by the local maximum likelihood estimates of their own standard deviations, which ensures variance stabilization. This technique was applied to univariate volatility estimation in Fryzlewicz et al. (2006); however, critical and interesting differences arise in the multivariate setting.

2. THE MULTIVARIATE MODEL

Let $X_{t,T}$ $(t = 1, \ldots, T)$, be a p-dimensional process of daily, or less frequent, log-returns on financial instruments, with components $X_{j,t,T}$ for $j = 1, \ldots, p$, where p can be large and possibly larger than T. Marginally, each $X_{j,t,T}$ is modelled as

$$X_{j,t,T} = \sigma_j(t/T)\,\varepsilon_{j,t}, \tag{1}$$

where $\sigma_j(u)$ is a positive left-continuous piecewise-constant function of $u \in (0,1]$, bounded from above and away from zero, with an unknown number of jumps of unknown locations and magnitudes. The vector random variables $\varepsilon_t = (\varepsilon_{1,t}, \ldots, \varepsilon_{p,t})^{\mathrm{T}}$ are independent, and distributed as $\varepsilon_t \sim N\{0, \Gamma(t/T)\}$, where the elements of the $p \times p$ matrix $\Gamma(t/T) = \{\rho_{i,j}(t/T)\}_{i,j=1}^{p}$ are such that $\rho_{i,i}(u) \equiv 1$, and $\rho_{i,j}(u)\,(i \neq j)$ is a left-continuous piecewise-constant function, with an unknown number of jumps of unknown locations and magnitudes. Let $\Sigma(t/T) = \{c_{i,j}(t/T)\}_{i,j=1}^{p}$ denote the variance matrix of $X_{t,T}$, and let $D(t,T)$ be a diagonal $p \times p$ matrix with $\sigma_i(t/T)\,(i = 1, \ldots, p)$ on the diagonal. We have the decomposition $\Sigma(t/T) = D(t/T)\,\Gamma(t/T)\,D(t/T)$. Marginally, each $X_{j,t,T}$ follows the univariate model of Fryzlewicz et al. (2006).

Here, $\Sigma(u)$, or alternatively the pair $\{D^2(u), \Gamma(u)\}$, can be viewed as the time-dependent parameter of the proposed model. Note that $\Sigma(u)$ is defined over the interval $(0,1]$, which is common practice in nonparametric regression in order to enable meaningful estimation theory. We assume that the jumps in $\Sigma(u)$ can approach each other in rescaled time, and therefore $\Sigma(u)$ can approximate continuous or piecewise continuous volatility matrices. Piecewise-constant modelling of multivariate volatilities was also considered by Härdle et al. (2003).

3. HAAR–FISZ ESTIMATION OF THE VOLATILITY MATRIX $\Sigma(u)$

3·1. *Methodology and theory*

We first consider the estimation of a single time-varying component of the matrix $\Sigma(u)$, i.e., the function $c_{i,l}(u)$, from a single stretch of observations $\{X_{i,t,T}X_{l,t,T}\}_{t=1}^{T}$. Our theoretical results concern the quality of the estimation of the entire matrix $\Sigma(u)$ in the operator norm. The starting point to our estimation procedure is the formulation

$$X_{i,t,T}X_{l,t,T} = c_{i,l}(t/T) + X_{i,t,T}X_{l,t,T} - c_{i,l}(t/T) = c_{i,l}(t/T) + \xi_{i,l,t,T},$$

where the noise $\xi_{i,l,t,T}$ is such that $E(\xi_{i,l,t,T}) = 0$. The Gaussianity of $X_{t,T}$ implies that

$$\mathrm{Var}(\xi_{i,l,t,T}) = c_{i,i}(t/T)c_{l,l}(t/T) + c_{i,l}^2(t/T). \tag{2}$$

Our estimator of $c_{i,l}(u)$ will be based on Haar wavelets, which we briefly introduce below; there are several excellent monographs on wavelets in statistics, including Vidakovic (1999).

The input to our Haar–Fisz estimation algorithm is the vector $\{X_{i,t,T}X_{l,t,T}\}_{t=1}^{T}$: here, we assume that T is an integer power of two and denote $J = \log_2 T$. The algorithm follows.

Step 1. For all of the following combinations of indices: $(\eta, \upsilon) = (i,i), (l,l), (i,l)$, compute the Haar decompositions of $\{X_{\eta,t,T}X_{\upsilon,t,T}\}_{t=1}^{T}$, obtaining the quantities $s_{j,k}^{(\eta,\upsilon)}$, $d_{j,k}^{(\eta,\upsilon)}$ and $\tilde{s}_{j,k}^{(\eta,\upsilon)}$ as follows. Let $s_{J,k}^{(\eta,\upsilon)} = X_{\eta,k,T}X_{\upsilon,k,T}$ $(k = 1, \ldots, 2^J)$. For each $j = J-1, J-2, \ldots, 0$, recursively form vectors $s_j^{(\eta,\upsilon)}$, $d_j^{(\eta,\upsilon)}$, $\tilde{s}_j^{(\eta,\upsilon)}$ with elements:

$$s_{j,k}^{(\eta,\upsilon)} = 2^{-1/2}(s_{j+1,2k-1}^{(\eta,\upsilon)} + s_{j+1,2k}^{(\eta,\upsilon)}), \quad d_{j,k}^{(\eta,\upsilon)} = 2^{-1/2}(s_{j+1,2k-1}^{(\eta,\upsilon)} - s_{j+1,2k}^{(\eta,\upsilon)}),$$
$$\tilde{s}_{j,k}^{(\eta,\upsilon)} = 2^{(j-J)/2}s_{j,k}^{(\eta,\upsilon)} \quad (k = 1, \ldots, 2^j).$$

Step 2. Obtain the variance-stabilized coefficients via the Fisz transformation

$$f_{j,k}^{(i,l)} = d_{j,k}^{(i,l)} \left\{ \tilde{s}_{j,k}^{(i,i)} \tilde{s}_{j,k}^{(l,l)} + \left(\tilde{s}_{j,k}^{(i,l)} \right)^2 \right\}^{-1/2}.$$

Step 3. Denote $\mu_{j,k}^{(i,l)} = E(d_{j,k}^{(i,l)})$. Estimate $\mu_{j,k}^{(i,l)}$ by $\hat{\mu}_{j,k}^{(i,l)} = d_{j,k}^{(i,l)} I(|f_{j,k}^{(i,l)}| > \lambda)$ for scales $j = 0, \dots J^*$ with $2^{J^*} = T^{1-\delta}$ for some $\delta \in (0,1)$, and $\hat{\mu}_{j,k}^{(i,l)} = 0$ otherwise, where $I(\cdot)$ is the indicator function.

Step 4. Take the inverse Haar transform of $\hat{\mu}_{j,k}^{(i,l)}$ to obtain an initial estimate $\hat{c}_{i,l}(t/T)$ of the covariance function $c_{i,l}(t/T)$.

Step 5. Correct the estimate by replacing its value on each interval of constancy by the local average of the sequence $\{X_{i,t,T} X_{l,t,T}\}_{t=1}^T$ over the same interval. Denote this bias-corrected estimate by $\tilde{c}_{i,l}(t/T)$.

Step 6. If $i \neq l$, apply additional thresholding in the time domain, i.e., construct the final estimate by either of the two operations

$$\bar{c}_{i,l}^{(h)}(t/T) = \tilde{c}_{i,l}(t/T) I\{|\tilde{c}_{i,l}(t/T)| > \lambda_1 \tilde{c}_{i,i}^{1/2}(t/T)\tilde{c}_{l,l}^{1/2}(t/T)\} \quad \text{(hard thresholding)},$$

$$\bar{c}_{i,l}^{(s)}(t/T) = \text{sign}\{\tilde{c}_{i,l}(t/T)\} \max\{|\tilde{c}_{i,l}(t/T)| - \lambda_1 \tilde{c}_{i,i}^{1/2}(t/T)\tilde{c}_{l,l}^{1/2}(t/T), 0\} \quad \text{(soft thresholding)},$$

denoting $\bar{\Sigma}^{(h)}(t/T) = \{\bar{c}_{i,l}^{(h)}(t/T)\}_{i,l=1}^p$ and $\bar{\Sigma}^{(s)}(t/T) = \{\bar{c}_{i,l}^{(s)}(t/T)\}_{i,l=1}^p$.

In view of (2) and the fact that $\tilde{s}_{j,k}^{(i,l)}$ is the local sample mean of the sequence $\{X_{i,t,T} X_{l,t,T}\}_{t=1}^T$ over the interval $t \in [(k-1)2^{J-j}+1, \dots, k\,2^{J-j}]$, the coefficient $f_{j,k}^{(i,l)}$ can be viewed as a variance-stabilized, or studentized, version of the Haar coefficient $d_{j,k}^{(i,l)}$. This variance-stabilization step permits the use of a threshold λ independent of scale j or location k. This is in the spirit of the Haar–Fisz transform; see, e.g., Fryzlewicz & Nason (2004) and Fryzlewicz (2008), both inspired by Fisz (1955). We refer to this variance stabilization as the Fisz transformation of $d_{j,k}^{(i,l)}$ to $f_{j,k}^{(i,l)}$.

The bias correction in step 5 is non-standard in a wavelet estimation context, but essential for the time-domain thresholding in step 6 to ensure that zero covariances are estimated as exactly zero with high probability, which helps reduce the overall estimation error in the operator norm under the assumption of sparsity.

As with the hard- and soft-thresholding covariance estimators proposed for independent and identically distributed data (Bickel & Levina, 2008; Rothman et al., 2009), our estimators are also not guaranteed to be positive-definite in finite samples for an arbitrary true covariance structure and arbitrary λ_1. Even outside the estimation context, hard- or soft-thresholded covariance matrices are not automatically positive-definite, as argued in Bickel & Levina (2008). However, as our results later demonstrate, our estimators converge to a positive definite limit with probability tending to one. Also, $\bar{\Sigma}^{(s)}(t/T)$ is guaranteed to be positive-definite for arbitrary finite samples, provided that λ_1 is large enough. This is because unlike hard thresholding, soft thresholding is a continuous operation and hence as λ_1 increases, $\bar{\Sigma}^{(s)}(t/T)$ converges continuously to the matrix containing $\tilde{c}_{i,i}(t/T)$ on the diagonal and zeros elsewhere, which is positive-definite. Therefore, $\bar{\Sigma}^{(s)}(t/T)$ will necessarily be positive-definite from a certain λ_1 onwards.

Shrinkage-type estimators for stationary covariance matrices have also been considered, e.g., in Haff (1980), Dey & Srinivasan (1985) and Ledoit & Wolf (2003). In some nonparametric models, one route to obtaining nonparametric function estimators which are exactly zero on parts of their domain is through the fused lasso approach of Tibshirani et al. (2005), and our time domain thresholding could in some cases serve as an alternative to this technique.

In order to analyse the behaviour of our estimator, we first introduce some notation. For any $p \times p$ matrix $M = (m_{i,l})_{i,l=1}^{p}$, we denote its ordered eigenvalues by $\lambda_{\max}(M) = \lambda_1(M) \geq \cdots \geq \lambda_p(M) = \lambda_{\min}(M)$. With $\|v\|_2$ denoting the l_2 norm of a vector v, the operator norm of M is defined as $\|M\| = \sup\{\|Mv\|_2 : \|v\|_2 = 1\}$, and for symmetric matrices, e.g., covariance matrices, is given by $\|M\| = \max_{1\leq i\leq p} |\lambda_i(M)|$. It is well known (Golub & Van Loan, 1989, Section 2.3.3) that for symmetric matrices, we have $\|M\| \leq \max_l \sum_i |m_{i,l}|$. Further, let $\Sigma = (c_{i,l})_{i,l=1}^{p}$ be any constant volatility matrix. We define a class of sparse constant volatility matrices as $\mathcal{U}\{c_0(p)\} = \{\Sigma : c_{i,i} = 1, \max_i \sum_{l=1}^{p} I(c_{i,l} \neq 0) \leq c_0(p)\}$, and a class of invertible sparse constant volatility matrices as $\mathcal{U}\{c_0(p), \epsilon_0\} = \{\Sigma : \Sigma \in \mathcal{U}\{c_0(p)\}, \lambda_{\min}(\Sigma) \geq \epsilon_0 > 0\}$. A dyadic interval is defined as any interval of the form $(\{k-1\}2^{-j}, k2^{-j}]$ $(j = 0, \ldots, J-1; k = 1, \ldots, 2^j)$. Our main result follows.

THEOREM 1. *Assume that the true volatility matrix $\Sigma(u)$ satisfies the following:*

(i) There exists a dyadic interval $\mathcal{I}$ of length at least 2^{-J^*}, such that for for each i and l, the function $c_{i,l}(u)$ is constant for all $u \in \mathcal{I}$.

(ii) For each i and l, if there are change-points in $c_{i,l}(u)$ to the left or right of $\mathcal{I}$, then the nearest one on either side is covered by a dyadic interval denoted $\mathcal{J}_{i,l}^{1}$ on the left-hand side, and $\mathcal{J}_{i,l}^{2}$ on the right-hand side, of length at least 2^{-J^*}, containing no other change-point, not intersecting with $\mathcal{I}$ and such that

$$\min_{i,l,m} \int_{\mathcal{J}_{i,l}^{m}} \left\{ c_{i,l}(u) - |\mathcal{J}_{i,l}^{m}| \int_{\mathcal{J}_{i,l}^{m}} c_{i,l}(z)dz \right\}^2 du \geq C_3 T^{-\beta}, \quad \beta \in [0, 1-\delta). \tag{3}$$

Further, assume that diag$^{-1/2}\{\Sigma(\mathcal{I})\}\Sigma(\mathcal{I})$diag$^{-1/2}\{\Sigma(\mathcal{I})\} \in \mathcal{U}\{c_0(p)\}$, and that its size p is at most of order $O(T^{\zeta})$ for some fixed $\zeta > 0$. Assume also that the elements of diag$\{(\Sigma(\mathcal{I})\}$ are uniformly bounded from below and above by constants. Recall the notation $\rho_{i,l}(u) = c_{i,l}(u)c_{l,l}^{-1/2}(u)c_{i,i}^{-1/2}(u)$ and let $\underline{c}(A) = \inf_{i,u\in A} c_{i,i}(u)$, $\bar{c}(A) = \sup_{i,u\in A} c_{i,i}(u)$. Let the thresholds λ and λ_1 satisfy

$$C \log^{1/2} T \geq \lambda \geq \left[2 \left\{ 2 \log\, p + (1-\delta) \log\, T + \log \frac{1}{a_{p,T}} \right\} \right]^{1/2},$$

$$\lambda_1 \geq \left\{ 2T^{-1}|\mathcal{I}|^{-1} \left(2 \log\, p + \log \frac{1}{a_{p,T}} \right) \right\}^{1/2},$$

$$\lambda_1 \leq \{1 + \varpi/\underline{c}(\mathcal{I})\}^{-1} \left\{ \min_{\rho_{i,l}(\mathcal{I})\neq 0} |\rho_{i,l}(\mathcal{I})| - \varpi/\bar{c}(\mathcal{I}) \right\},$$

for some $C > 0$, where $a_{p,T}$ tends to zero as $T \to \infty$ but no faster than $O(T^{-\zeta})$, and

$$\varpi = 2\bar{c}(\mathcal{I}) \left\{ \frac{\log\, p + \log\, c_0(p) + \log a_{p,T}^{-1}}{T|\mathcal{I}|} \right\}^{1/2}.$$

The following holds with probability of at least $1 - C_1\, a_{p,T}$ for some positive C_1:

(a) The estimator $\bar{\Sigma}^{(h)}(u)$ is constant for $u \in \mathcal{I}$ and such that $\bar{c}_{i,l}^{(h)}(\mathcal{I}) = 0$ if $c_{i,l}(\mathcal{I}) = 0$ and $\bar{c}_{i,l}^{(h)}(\mathcal{I})$ is a local sample mean of the sequence $\{X_{i,t,T}X_{l,t,T}\}_t$ over a subinterval $t/T \in \mathcal{K}_{i,l}$ where $\mathcal{I} \subseteq \mathcal{K}_{i,l}$ and $c_{i,l}(\mathcal{I}) = c_{i,l}(\mathcal{K}_{i,l})$, if $c_{i,l}(\mathcal{I}) \neq 0$.

(b) We have

$$\|\bar{\Sigma}^{(h)}(\mathcal{I}) - \Sigma(\mathcal{I})\| \leq 2c_0(p)\bar{c}(\mathcal{I}) \left\{ \frac{\log\, p + \log\, c_0(p) + \log\, a_{p,T}^{-1}}{T|\mathcal{I}|} \right\}^{1/2}.$$

(c) If, in addition, $diag^{-1/2}\{\Sigma(\mathcal{I})\}\Sigma(\mathcal{I})diag^{-1/2}\{\Sigma(\mathcal{I})\} \in \mathcal{U}\{c_0(p), \epsilon_0\}$*, then*

$$\|(\bar{\Sigma}^{(h)}(\mathcal{I}))^{-1} - \Sigma(\mathcal{I})^{-1}\| \leq C_2 c_0(p) \left\{ \frac{\log\, p + \log\, c_0(p) + \log\, a_{p,T}^{-1}}{T|\mathcal{I}|} \right\}^{1/2}$$

for some positive C_2.

The parameter δ, which appears in step 3 of the estimation algorithm and impacts the magnitudes of J^* and β, is required to be less than 1 for various technical reasons, including guaranteeing uniform strong asymptotic normality of $d_{j,k}^{(i,l)}$ and $s_{j,k}^{(i,l)}$. The lower its value, the less strict the assumptions (i) and (ii) of Theorem 1, i.e. the larger the class of volatilities $\Sigma(u)$ for which our method is applicable, but, potentially, the worse the error bounds in statements (b) and (c). The variance-type condition (3) specifies how large, or how isolated, the nearest change-point needs to be before our estimator reacts to it.

The reason why p is not allowed to grow exponentially with T is that $\log(p)/(T|\mathcal{I}|)$ needs to tend to zero to lead to consistency; however, the only assumption about $\mathcal{I}$ is that $|\mathcal{I}| \geq T^{\delta-1}$ with δ being possibly arbitrarily close to zero.

The application of the variance-stabilizing Fisz transformation in the computation of $f_{j,k}^{(i,l)}$ allows the threshold λ to be independent of $c_{i,l}(u)$. The lower bound for the threshold λ_1 is also independent of $c_{i,l}(u)$ as it is calibrated under the hypothesis that, locally, $c_{i,l}(u) = 0$.

Parameter $a_{p,T}$ determines the probability with which the results of Theorem 1 hold; the higher the desired probability, the worse the error bounds. As in the stationary set-up of Bickel & Levina (2008), the magnitude of the error bounds specifies how fast the sparsity parameter $c_0(p)$ is permitted to grow with p before consistency fails.

The convergence rates in Theorem 1 depend on $|\mathcal{I}|$, so the estimator exhibits interval-wise adaptation to the features of the target matrix. In practice, heuristically speaking, this means that our estimator, whose explicit form appears in statement (a) of Theorem 1, will tend to be based on longer samples of data, thereby leading to more slowly-changing estimated volatility, in periods when the true volatility is changing slowly or not at all, and on shorter samples in periods of rapid changes in volatility. To the practitioner, the first potential benefit of this property is reduction of unnecessary transaction costs, incurred as a result of changes in estimated volatility, in periods of slowly changing volatility, compared to non-adaptive estimators such as those based on GARCH-type models or exponential smoothing. The second potential benefit is faster reaction to significant changes in volatility in comparison with non-adaptive estimators.

The extra thresholding in the time domain ensures stable invertibility of our estimator and hence accurate estimation of the precision matrix as evidenced in statement (c) of Theorem 1. The latter is of importance in tasks such as optimal portfolio allocation in Markowitz's mean-variance paradigm. A similar consistency result can be formulated for the $\bar{\Sigma}^{(s)}(u)$ estimator, but we omit it for lack of space.

3·2. *Alternative approach via polarization identity*

In this section, we propose an alternative to the initial estimator $\hat{c}_{i,l}(t/T)$, based on the polarization identity $X_{i,t,T}X_{l,t,T} = 1/4\left\{(X_{i,t,T} + X_{l,t,T})^2 - (X_{i,t,T} - X_{l,t,T})^2\right\}$. Define an operator $\mathcal{F}$ by $\hat{c}_{i,l}(t/T) = \mathcal{F}(\{X_{i,t,T}X_{l,t,T}\}_{t=1}^T)$. Note that $\mathcal{F}$ is a nonlinear smoothing operator, since it involves the nonlinear operation of thresholding by λ. Thus, in general, by the polarization identity,

$$\hat{c}_{i,l}(t/T) \neq \frac{1}{4}\left\{\mathcal{F}(\{(X_{i,t,T}+X_{l,t,T})^2\}_{t=1}^T) - \mathcal{F}(\{(X_{i,t,T}-X_{l,t,T})^2\}_{t=1}^T)\right\}.$$

In this section, we propose and motivate the following alternative to $\hat{c}_{i,l}(t/T)$:

$$\begin{aligned}\hat{c}^P_{i,l}(t/T) &= \frac{1}{4}\left\{\mathcal{F}(\{(X_{i,t,T}+X_{l,t,T})^2\}_{t=1}^T) - \mathcal{F}(\{(X_{i,t,T}-X_{l,t,T})^2\}_{t=1}^T)\right\}\\ &= \frac{1}{4}\left\{\hat{\sigma}^{+\,2}_{i,l}(t/T) - \hat{\sigma}^{-\,2}_{i,l}(t/T)\right\}.\end{aligned}$$

Both $X_{i,t,T}+X_{l,t,T}$ and $X_{i,t,T}-X_{l,t,T}$ follow the multiplicative models $X_{i,t,T}+X_{l,t,T} = \sigma^+_{i,l}(t/T)\varepsilon^+_{i,l,t}$ and $X_{i,t,T}-X_{l,t,T} = \sigma^-_{i,l}(t/T)\varepsilon^-_{i,l,t}$, where the functions $\sigma^\pm_{i,l}(t/T)$ are piecewise constant, $\varepsilon^\pm_{i,l,t}$ are independent $N(0,1)$, and $\sigma^{\pm\,2}_{i,l}(t/T) = \sigma^2_i(t/T) + \sigma^2_l(t/T) \pm 2c_{i,l}(t/T)$. Thus, to estimate $\sigma^{\pm\,2}_{i,l}(t/T)$, and therefore compute $\hat{c}^P_{i,l}(t/T)$, we can use the algorithm of §3·1 with $\{(X_{i,t,T} \pm X_{l,t,T})^2\}_{t=1}^T$ as input. It is possible to derive the exact distribution of the corresponding Haar–Fisz coefficients of $(X_{i,t,T} \pm X_{l,t,T})^2$, denoted $f^{(i,l,\pm)}_{j,k}$ here to differentiate them from $f^{(i,l)}_{j,k}$, under the null hypothesis of the local constancy of $\sigma^{\pm\,2}_{i,l}(t/T)$ over the corresponding sub-interval, which leads to a more accurate, non-asymptotic selection of the threshold λ. To see this, first note that

$$\begin{aligned}f^{(i,l,\pm)}_{j,k} &= \frac{d^{(i,l,\pm)}_{j,k}}{2^{(1+j-J)/2}s^{(i,l,\pm)}_{j,k}}\\ &= 2^{(J-j-1)/2}\frac{\sum_{t=(k-1)2^{J-j}+1}^{(k-1/2)2^{J-j}}\sigma^{\pm\,2}_{i,l}(t/T)\varepsilon^{\pm\,2}_{i,l,t} - \sum_{t=(k-1/2)2^{J-j}+1}^{k\,2^{J-j}}\sigma^{\pm\,2}_{i,l}(t/T)\varepsilon^{\pm\,2}_{i,l,t}}{\sum_{t=(k-1)2^{J-j}+1}^{(k-1/2)2^{J-j}}\sigma^{\pm\,2}_{i,l}(t/T)\varepsilon^{\pm\,2}_{i,l,t} + \sum_{t=(k-1/2)2^{J-j}+1}^{k\,2^{J-j}}\sigma^{\pm\,2}_{i,l}(t/T)\varepsilon^{\pm\,2}_{i,l,t}},\end{aligned}$$

which, under the local hypothesis of constancy of $\sigma^{\pm\,2}_{i,l}(t/T)$, with $\sigma^{\pm\,2}_{i,l}(t/T) \neq 0$, leads to

$$2^{(j+1-J)/2}f^{(i,l,\pm)}_{j,k} = \frac{\sum_{t=(k-1)2^{J-j}+1}^{(k-1/2)2^{J-j}}\varepsilon^{\pm\,2}_{i,l,t} - \sum_{t=(k-1/2)2^{J-j}+1}^{k\,2^{J-j}}\varepsilon^{\pm\,2}_{i,l,t}}{\sum_{t=(k-1)2^{J-j}+1}^{(k-1/2)2^{J-j}}\varepsilon^{\pm\,2}_{i,l,t} + \sum_{t=(k-1/2)2^{J-j}+1}^{k\,2^{J-j}}\varepsilon^{\pm\,2}_{i,l,t}}. \tag{4}$$

However, by Lemma 1 of Fryzlewicz et al. (2006), $2^{(j+1-J)/2}f^{(i,l,\pm)}_{j,k}$ in (4) is distributed as $2Y-1$, where $Y \sim \beta(2^{J-j-2}, 2^{J-j-2})$. Knowledge of this distribution can lead to the choice of λ based on the exact quantiles of the beta distribution; this contrasts with the results of Theorem 1 where the choice of λ is based on strong asymptotic normality arguments. The distribution of the Haar–Fisz coefficients is only readily available in the case of the polarized estimator $\hat{c}^P_{i,l}(t/T)$; indeed, it is not clear how to obtain the exact distribution of $f^{(i,l)}_{j,k}$, i.e., the Haar–Fisz coefficients in the computation of the non-polarized estimator $\hat{c}_{i,l}(t/T)$, when $i \neq l$.

As an example of how the knowledge of the distribution of $f^{(i,l,\pm)}_{j,k}$ can help in selecting the threshold λ, which can possibly depend on the scale j and will therefore be denoted by $\tilde{\lambda}_j$, consider the case where the true volatility is constant, $\Sigma(u) = \Sigma$. To ensure that our initial polarized estimator $\hat{\Sigma}^P(u) = \{\hat{c}^P_{i,l}(t/T)\}_{i,l=1}^p$ is also constant with probability no less than $1 - a_{p,T}$, it is sufficient to require that $\mathrm{pr}(\bigcup_{i,l}\bigcup_{j,k}\bigcup_{s\in\{+,-\}}|f^{(i,l,s)}_{j,k}| \geq \tilde{\lambda}_j) \leq a_{p,T}$. Setting

$\mathrm{pr}(|f_{j,k}^{(i,l,s)}| \geq \tilde{\lambda}_j)$ to be independent of j and using the Bonferroni inequality, the above is implied by $2p^2T^{1-\delta}\mathrm{pr}(|2^{(j+1-J)/2}\, f_{j,k}^{(i,l,s)}| \geq 2^{(j+1-J)/2}\,\tilde{\lambda}_j) = a_{p,T}$, which can easily be solved numerically for each j separately using the quantiles of the relevant beta distribution.

4. Practicalities, online algorithm and simulation study

4·1. *Current interval of stationarity*

In the following, we take $\bar{\Sigma}(u)$ to denote $\bar{\Sigma}^{(h)}(u)$ or $\bar{\Sigma}^{(s)}(u)$, and $\bar{c}_{i,l}(u)$ to denote the entries of $\bar{\Sigma}(u)$. Having observed $X_{s,T}, (s = 1, \ldots, t)$, the practitioner will be particularly interested in $\bar{\Sigma}(t/T)$, the value of the estimator at the current time t. In the algorithm of §3·1, each estimate $\bar{c}_{i,l}(t/T)$ is a possibly thresholded average of $\{X_{i,s,T}X_{l,s,T}\}_s$ over a certain interval $T\mathcal{K}_{i,l}$ ending at $s = t$. Empirically, it has been found that $\bar{\Sigma}(t/T)$ is more stably invertible if all of its entries are, possibly thresholded, averages of $\{X_{i,s,T}X_{l,s,T}\}_s$ over an interval $T\mathcal{K}$ ending at $s = t$ whose length is constant over i and l. In practice, we choose $\mathcal{K}$ to be the shortest out of the intervals $\mathcal{K}_{i,l}$ over all i and l.

In this and the following paragraph, we use the notation $T\mathcal{K}_t$ to emphasize the dependence of the common interval $T\mathcal{K}$, selected as above, on the current time t. In an online setting, $\bar{\Sigma}(t/T)$ will be recalculated with the arrival of each new observation $X_{t,T}$, leading to a certain sequence of intervals of stationarity $\{T\mathcal{K}_t\}_t$. Let their lengths be denoted by $|T\mathcal{K}_t|$. As an example, if the sequence $|T\mathcal{K}_t|$ progresses over time t as $\ldots, 64, 64, 64, 16, 64, 64, \ldots$, then the 16 is likely to be the result of a type-I error, i.e. detection of a change-point when there are none, and will lead to the estimator $\bar{\Sigma}(t/T)$ having an unnecessary blip for the corresponding t. To rectify this, we propose to use a smoothed version of $T\mathcal{K}_t$, denoted by $T\tilde{\mathcal{K}}_t$ and constructed such that $T\tilde{\mathcal{K}}_t$ ends at t and $|T\tilde{\mathcal{K}}_t| = \mathrm{Mode}(|T\mathcal{K}_{t-m+1}|, \ldots, |T\mathcal{K}_t|)$. We use $m = 10$ in the remainder of the paper. This ensures elimination of blips such as those in the above example.

Due to the dyadic structure of the Haar transform, the intervals $T\tilde{\mathcal{K}}_t$ are likely to be of dyadic length, as in the example from the previous paragraph. However, in an online context, as a new observation arrives, the interval of stationarity should ideally have the property that its length either increases by one if no new change-point is detected, or drops to the smallest permitted length if a new change-point is detected. To enforce this property, we define intervals $T\bar{\mathcal{K}}_t$, ending at t and satisfying $|T\bar{\mathcal{K}}_t| = |T\bar{\mathcal{K}}_{t-1}| + 1$ if $|T\tilde{\mathcal{K}}_t| = |T\tilde{\mathcal{K}}_{t-1}|$, and $|T\bar{\mathcal{K}}_t| = |T\tilde{\mathcal{K}}_t|$ otherwise, so that, e.g., $|T\tilde{\mathcal{K}}_t| = (64, 64, 64, 64, 16, 16)$ results in $|T\bar{\mathcal{K}}_t| = (64, 65, 66, 67, 16, 17)$. The intervals $T\bar{\mathcal{K}}_t$ are used in the computation of $\bar{\Sigma}(u)$ in the remainder of the paper.

4·2. *Selection of* λ_1

Intuitively, the time-domain threshold λ_1 should be as small as possible while enabling stable invertibility of $\bar{\Sigma}(t/T)$. A natural candidate for λ_1 is the lower bound of its permitted theoretical range from Theorem 1, that is, $\{2T^{-1}|\mathcal{I}|^{-1}(2\log\, p + \log\, a_{p,T}^{-1})\}^{1/2}$. The length $T|\mathcal{I}|$ is obviously unknown, but its nearest observable proxy is $|T\bar{\mathcal{K}}_t|$, which leads to our first proposed choice of λ_1, termed universal and defined by $\lambda_1^u = \{2|T\bar{\mathcal{K}}_t|^{-1}(2\log\, p + \log\, a_{p,T}^{-1})\}^{1/2}$. Selection of $a_{p,T}$ is briefly discussed in §4·3. From our technical results in Appendix A, it can be seen that the particular form of λ_1^u is the effect of the Bonferroni inequality, and thus λ_1^u is likely to overestimate the amount of thresholding required.

Again from the technical results, it is apparent that λ_1 represents a bound, with high probability, on the entries of the sample correlation matrix of size $p \times p$, for a sample of length $T|\mathcal{I}|$, under the assumption that the true correlation is I_p, the identity matrix of size $p \times p$. A more precise bound than that furnished by the Bonferroni inequality can be obtained, e.g., by using the distributional results for the maximum entry of the sample correlation matrix, also called its coherence, by Jiang (2004), based on the Chen–Stein Poisson approximation method; see also the refinements of this result in Li et al. (2012), Cai & Jiang (2011) and Cai & Jiang (2012).

In §5, we use the following generic method for selecting λ_1. To guard against λ_1^u being possibly too high, we start with $\lambda_1 = \lambda_1^u$ and gradually decrease it as long as a certain stability condition is satisfied: for example, the condition number of $\bar{\Sigma}(t/T)$ is above a pre-specified positive constant, or the portfolio weights resulting from $\bar{\Sigma}(t/T)$ satisfy a certain desired constraint, e.g. are not too unbalanced, which is a type of exposure constraint, a different form of which was also discussed, e.g., in Fan et al. (2012). If λ_1^u itself does not yield $\bar{\Sigma}(t/T)$ satisfying the desired stability condition, λ_1 should be increased until the condition is satisfied: note that as λ_1 increases, $\bar{\Sigma}(t/T)$ converges to a diagonal, and thus stably invertible, matrix.

It is also possible to select λ_1 by applying the cross-validation technique of Bickel & Levina (2008) to the sample correlation matrix computed over the interval $T\bar{\mathcal{K}}_t$.

4·3. *Simulation study*

We investigate the performance of our method, in an online context, in a set-up where $\Sigma(u)$ changes abruptly at a certain point u_0. We consider the case where the change is caused by the introduction of one common factor to $m = p/2$ of the components of $X_{t,T}$, which can be viewed as a caricature of a situation where some of the markets suddenly become more highly correlated.

We simulate p-variate Gaussian returns $X_{t,T}$ of length $T = 2048$. We are particularly interested in the more challenging problem of estimating the cross-covariance, rather than the marginal volatility, so we use $\sigma_{i,i}(u) \equiv 1$ throughout. The returns are mutually uncorrelated for $t \leq 1024$. For $t \geq 1025$, we have $\mathrm{cov}(X_{i,t,T}, X_{j,t,T}) = \rho^2$ if $1 \leq i \neq j \leq m$, and 0 otherwise.

Over the entire time horizon $T = 2048$, we apply a moving window $[k, k+255]$ for $k = 1, \ldots, 1793$, and for each k, compute our estimator $\bar{\Sigma}_k^{(s)} = \bar{\Sigma}^{(s)}(1)$, that is, compute the estimator at the right edge, indexed $k + 255$, of the currently available data. The competitor is the estimator $\breve{\Sigma}_k^{(s)}$, the sample covariance estimator over the interval $[k, k+255]$, thresholded using the same λ_1 as $\bar{\Sigma}^{(s)}$, for a fair comparison. We do not use any of the selection rules for λ_1 from §4·2 as all of them rely on the interval $T\bar{\mathcal{K}}_t$, which does not feature at all in the sample covariance estimator.

In computing $\bar{\Sigma}_k^{(s)}$, we use $\delta = 0.5$, which leads to the 3 finest scales of the Haar transform being disregarded and to the minimum length of the intervals of constancy of $\bar{\Sigma}_k^{(s)}$ being 8, $a_{p,256} = \log^{-1/2}(256)$, the same rate as that furnished by classical universal thresholding in one-dimensional wavelet setting, and $\lambda = \{2\,(2 \log\, p + (1-\delta) \log\, T - \log\, a_{p,T})\}^{1/2}$, which is the lower end of the permitted range of λ from Theorem 1. Soft thresholding has been found to perform better than hard, and hence we use the former. For completeness, we note than an alternative to this choice of λ would be to use the polarization identity approach, or simply fine-tune λ so that, e.g., empirical residuals in model (1) for each j pass a certain test for independence and identical distribution, as described in the univariate case in Fryzlewicz et al. (2006).

To quantify the estimation accuracy, we use the quantities $\mathrm{MSE}(\bar{\Sigma}^{(s)}) = p^{-2} \sum_{k=1025}^{1274} \|\bar{\Sigma}_k^{(s)} - \Sigma_k\|_F^2$, $\mathrm{MSE}(\breve{\Sigma}^{(s)}) = p^{-2} \sum_{k=1025}^{1274} \|\breve{\Sigma}_k^{(s)} - \Sigma_k\|_F^2$, $\mathrm{MSE}\{(\bar{\Sigma}^{(s)})^{-1}\} = p^{-2} \sum_{k=1025}^{1274} \|(\bar{\Sigma}_k^{(s)})^{-1} - (\Sigma_k)^{-1}\|_F^2$, $\mathrm{MSE}\{(\breve{\Sigma}^{(s)})^{-1}\} = p^{-2} \sum_{k=1025}^{1274} \|(\breve{\Sigma}_k^{(s)})^{-1} - (\Sigma_k)^{-1}\|_F^2$, averaged over 100 simulations, where $\|\cdot\|_F$ denotes the Frobenius norm and Σ_k is the true volatility matrix at time $k + 255$. The range of k in the summations corresponds exactly to the first 250 trading days after the change in the volatility matrix at $t = 1024$, and therefore these error measures are designed to capture how our adaptive and the non-adaptive sample covariance estimator react to the change.

From Table 1, it is clear that our adaptive estimator $\bar{\Sigma}^{(s)}$ outperforms the non-adaptive estimator $\breve{\Sigma}^{(s)}$ for the higher values of ρ, and is not much worse for $\rho = 0.3$. This is unsurprising as being able to react quickly to a change in the correlation structure matters more if the change is more significant. The results seem to be relatively robust with respect to λ_1 for the estimation of Σ_k. For the estimation of Σ_k^{-1}, the differences between the two estimators are more striking for the lower values of λ_1, as these lead to better control of the bias, although not so for $p = 50$, where too low a value of λ_1 is likely to lead to instabilities in terms of invertibility.

Table 1. *Left column: ratios of MSE*$(\bar{\Sigma}^{(s)})$ *to MSE*$(\breve{\Sigma}^{(s)})$*, the mean-square error of our estimator to the mean-square error of the sample covariance estimator, for dimensionality* $p = 6, 10, 20, 50$*, respectively. Right column: ratios of MSE*$\{(\bar{\Sigma}^{(s)})^{-1}\}$ *to MSE*$\{(\breve{\Sigma}^{(s)})^{-1}\}$*, the mean-square error of the inverse of our estimator to the mean-square error of the inverse of the sample covariance estimator, for the same* p

	Estimation of Σ_k				Estimation of Σ_k^{-1}		
	$\lambda_1 = 0.1$	$\lambda_1 = 0.2$	$\lambda_1 = 0.3$		$\lambda_1 = 0.1$	$\lambda_1 = 0.2$	$\lambda_1 = 0.3$
$\rho = 0.3$	1.07	1.04	1.04	$\rho = 0.3$	1.07	1.03	1.04
$\rho = 0.6$	0.90	0.89	0.93	$\rho = 0.6$	0.90	0.90	0.94
$\rho = 0.9$	0.52	0.49	0.52	$\rho = 0.9$	0.58	0.69	0.78
	$\lambda_1 = 0.1$	$\lambda_1 = 0.2$	$\lambda_1 = 0.3$		$\lambda_1 = 0.1$	$\lambda_1 = 0.2$	$\lambda_1 = 0.3$
$\rho = 0.3$	1.01	1.01	1.01	$\rho = 0.3$	1.04	1.01	1.02
$\rho = 0.6$	0.84	0.87	0.90	$\rho = 0.6$	0.89	0.89	0.92
$\rho = 0.9$	0.45	0.46	0.50	$\rho = 0.9$	0.53	0.68	0.78
	$\lambda_1 = 0.1$	$\lambda_1 = 0.2$	$\lambda_1 = 0.3$		$\lambda_1 = 0.1$	$\lambda_1 = 0.2$	$\lambda_1 = 0.3$
$\rho = 0.3$	1.01	1.01	1.01	$\rho = 0.3$	1.04	1.01	1.02
$\rho = 0.6$	0.80	0.81	0.91	$\rho = 0.6$	0.95	0.84	0.92
$\rho = 0.9$	0.41	0.41	0.48	$\rho = 0.9$	0.48	0.65	0.77
	$\lambda_1 = 0.1$	$\lambda_1 = 0.2$	$\lambda_1 = 0.3$		$\lambda_1 = 0.1$	$\lambda_1 = 0.2$	$\lambda_1 = 0.3$
$\rho = 0.3$	1	1	1	$\rho = 0.3$	1	1	1.01
$\rho = 0.6$	0.71	0.79	0.83	$\rho = 0.6$	0.98	0.96	0.92
$\rho = 0.9$	0.40	0.41	0.41	$\rho = 0.9$	0.75	0.63	0.72

5. EXAMPLE

We consider the multivariate series of log-returns on the daily closing values of 12 stock indices: All Ordinaries, AMEX Major Market Index, Bovespa, BUX, CAC 40, DAX, Dow Jones Industrial Average, FTSE 100, Hang Seng, NASDAQ Composite, Nikkei and S&P 500, on $T = 4097$ trading days ending on 26 October 2012. Marginally, all log-return series have been normalized so that their sample variance over T days equals one.

As in §4·3, we apply a moving window $[k, k + 255]$ for $k = 1, \ldots, 3841$, and for each k, compute our estimator $\bar{\Sigma}_k^{(s)} = \bar{\Sigma}^{(s)}(1)$, that is, compute the estimator at the right edge, indexed $k + 255$, of the currently available data. Except for λ_1, we use the same parameter values as in §4·3. Let $\bar{\Gamma}_k^{(s)}$ be the associated correlation estimator, i.e., $\bar{\Gamma}_k^{(s)} = \mathrm{diag}(\bar{\Sigma}_k^{(s)})^{-1/2} \bar{\Sigma}_k^{(s)} \mathrm{diag}(\bar{\Sigma}_k^{(s)})^{-1/2}$.

To select λ_1, we follow the advice from §4·2 and aim to select the lowest value of λ_1 that still guarantees stable invertibility of $\bar{\Sigma}_k^{(s)}$, or alternatively $\bar{\Gamma}_k^{(s)}$. For λ_1 taking values $1/10, 2/10, \ldots, 9/10$, we compute the condition number c_k, defined as the ratio of the largest and the smallest eigenvalues, of $\bar{\Gamma}_k^{(s)}$, and the ratio b_k of the maximum and minimum components of the vector $(\bar{\Gamma}_k^{(s)})^{-1}1$, where 1 is the column vector of ones of length p. The quantity b_k is a measure of the balancedness of the associated Markowitz portfolio. For $\lambda_1 \leq 4/10$, some large condition numbers c_k lead to instabilities in the inversion of $\bar{\Gamma}_k^{(s)}$, which in turn lead to some extremely large values of b_k. These numerical instabilities do not appear to be present for $\lambda_1 \geq 5/10$, so our recommendation would be to set λ_1 to $5/10$ or $6/10$ for this portfolio.

In Figure 1, the largest peak in the marginal volatility of FTSE 100 corresponds to the most severe phase of the recent financial crisis; this is also when FTSE 100 and S&P 500 become more correlated. There is a drop in the proportion of zeros in $\bar{\Sigma}_k^{(s)}$ around the same time, which serves as yet another piece of evidence for the common wisdom that markets tend to become more correlated in times of crises. The adaptive character of the estimators is apparent, with some smooth sections but also some sharp jumps.

ACKNOWLEDGEMENT

I am grateful to the Editor, Associate Editor and Referees for their very helpful comments.

A. PROOFS

Lemmas 1 and 2 concern the case where the marginal volatilities $\sigma_i(u)$ are known and equal to one, i.e., where the matrix $D(u)$ is the identity matrix. In that case, the difference is that we have $f_{j,k}^{(i,l)} = d_{j,k}^{(i,l)}\{1 + (\tilde{s}_{j,k}^{(i,l)})^2\}^{-1/2}$ and that the time-domain thresholding uses a constant threshold λ_1. Lemma 3 reverts to the setting of §3·1.

LEMMA 1. *Assume that the true volatility matrix $\Sigma(u)$ is constant and such that $\Sigma(u) = \Sigma \in \mathcal{U}\{c_0(p)\}$, and that its size p is at most of order $O(T^\zeta)$ for some fixed $\zeta > 0$. Further, let the thresholds λ and λ_1 satisfy $\lambda \geq \{2(2\log p + (1-\delta)\log T + \log a_{p,T}^{-1})\}^{1/2}$,*

$$\min_{c_{i,l}\neq 0} |c_{i,l}| - 2\left\{\frac{\log p + \log c_0(p) + \log a_{p,T}^{-1}}{T}\right\}^{1/2} \geq \lambda_1 \geq \left\{2T^{-1}\left(2\log p + \log a_{p,T}^{-1}\right)\right\}^{1/2},$$

where $a_{p,T}$ tends to zero as $T \to \infty$ but no faster than $O(T^{-\zeta})$. The following holds with probability of at least $1 - C_1 a_{p,T}$ for some positive C_1:

(a) Our estimator $\bar{\Sigma}^{(h)}(u) = \bar{\Sigma}^{(h)}$ is constant and such that $\bar{c}_{i,l}^{(h)}(t/T) \equiv 0$ if $c_{i,l} = 0$ and $\bar{c}_{i,l}^{(h)}(t/T) \equiv T^{-1}\sum_{t=1}^T X_{i,t,T}X_{l,t,T}$ if $c_{i,l} \neq 0$.
(b) We have $\|\bar{\Sigma}^{(h)} - \Sigma\| \leq 2c_0(p)\{\log p + \log c_0(p) + \log a_{p,T}^{-1}\}^{1/2}T^{-1/2}$.
(c) If, in addition, $\Sigma \in \mathcal{U}\{c_0(p), \epsilon_0\}$, then $\|(\bar{\Sigma}^{(h)})^{-1} - \Sigma^{-1}\| \leq C_2 c_0(p)\{\log p + \log c_0(p) + \log a_{p,T}^{-1}\}^{1/2}T^{-1/2}$ for some positive C_2.

Proof. Note that $\bar{c}_{i,l}^{(h)}(t/T)$ will be constant if and only if all $|f_{j,k}^{(i,l)}|$ fall under the threshold λ. Using the Bonferroni inequality, we have

$$\mathrm{pr}\left\{\bigcup_{j=0}^{J^*}\bigcup_{k=1}^{2^j}(|f_{j,k}^{(i,l)}| > \lambda)\right\} \leq \sum_{j=0}^{J^*} 2^j \mathrm{pr}\left(|f_{j,k}^{(i,l)}| > \lambda\right) \leq \max_j \mathrm{pr}\left(|f_{j,k}^{(i,l)}| > \lambda\right) CT^{1-\delta}, \quad \text{(A1)}$$

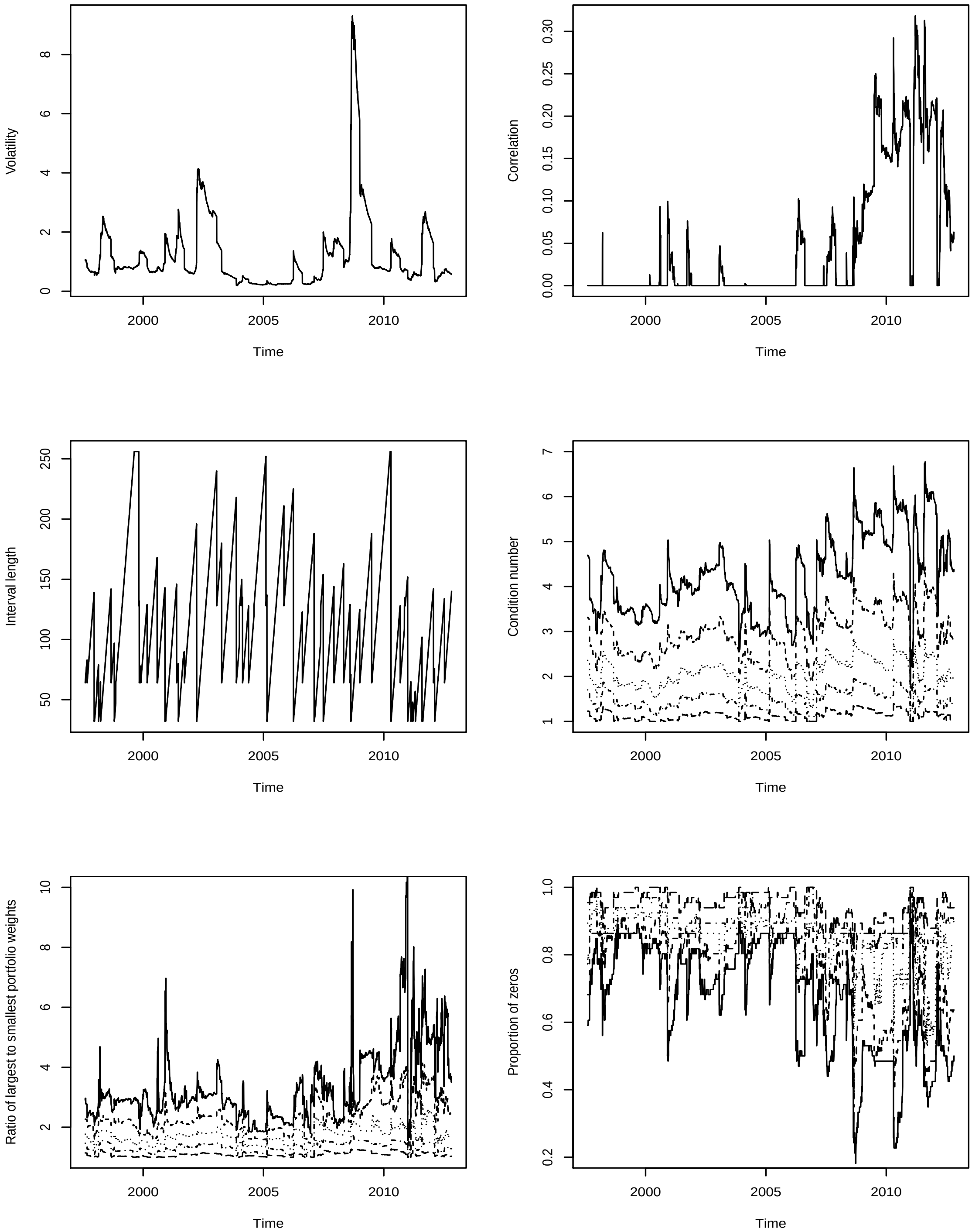

Fig. 1. Various aspects of volatility estimates from §5. Top left: the 8th diagonal component of $\bar{\Sigma}_k^{(s)}$, i.e. estimated marginal volatility of FTSE 100; top right: the (8,12)th component of $\bar{\Gamma}_k^{(s)}$, i.e. estimated correlation between FTSE 100 and S&P 500, with $\lambda_1 = 0.5$; middle left: length of the interval of stationarity in estimating $\bar{\Sigma}_k^{(s)}$; middle right: c_k for $\lambda_1 = 0.5$ (solid), 0.6 (dashed), 0.7 (dotted), 0.8 (dotdash), 0.9 (longdash); bottom left: b_k for the same range of λ_1; bottom right: proportion of zeros in $\bar{\Sigma}_k^{(s)}$ for the same range of λ_1.

where $C, C_1, C_2, \ldots$ are generic fixed positive constants throughout the proof. We now find a bound for the right-hand side term under the assumption that $\lambda \leq C(\log T)^{1/2}$. Of course the same bound will be also valid for higher values of λ. Assessing first the probability term, we have

$$\begin{aligned} \mathrm{pr}\left(|f_{j,k}^{(i,l)}| > \lambda\right) &= \mathrm{pr}\left[|d_{j,k}^{(i,l)}| > \lambda\left\{1 + (\tilde{s}_{j,k}^{(i,l)})^2\right\}^{1/2} \quad | \quad |\tilde{s}_{j,k}^{(i,l)} - c_{i,l}| < \tilde{\delta}_j\right] \mathrm{pr}(|\tilde{s}_{j,k}^{(i,l)} - c_{i,l}| < \tilde{\delta}_j) \\ &+ \mathrm{pr}\left[|d_{j,k}^{(i,l)}| > \lambda\left\{1 + (\tilde{s}_{j,k}^{(i,l)})^2\right\}^{1/2} \quad | \quad |\tilde{s}_{j,k}^{(i,l)} - c_{i,l}| \geq \tilde{\delta}_j\right] \\ &\times \mathrm{pr}(|\tilde{s}_{j,k}^{(i,l)} - c_{i,l}| \geq \tilde{\delta}_j). \end{aligned} \tag{A2}$$

By the convexity of $u(x) = 1 + x^2$,

$$1 + (\tilde{s}_{j,k}^{(i,l)})^2 \geq 1 + c_{i,l}^2 + 2c_{i,l}(\tilde{s}_{j,k}^{(i,l)} - c_{i,l}) \geq 1 + c_{i,l}^2 - 2|c_{i,l}||\tilde{s}_{j,k}^{(i,l)} - c_{i,l}|. \tag{A3}$$

Using this, we bound (A2) by

$$\mathrm{pr}\left\{|d_{j,k}^{(i,l)}| > \lambda(1 + c_{i,l}^2 - 2|c_{i,l}|\tilde{\delta}_j)^{1/2}\right\} + \mathrm{pr}(|\tilde{s}_{j,k}^{(i,l)} - c_{i,l}| \geq \tilde{\delta}_j) = A + B.$$

Starting with A, we have

$$A = \mathrm{pr}\left\{\frac{|d_{j,k}^{(i,l)}|}{\left(1 + c_{i,l}^2\right)^{1/2}} > \lambda\left(\frac{1 + c_{i,l}^2 - 2|c_{i,l}|\tilde{\delta}_j}{1 + c_{i,l}^2}\right)^{1/2}\right\}. \tag{A4}$$

Since $\varepsilon_{i,t}$ are Gaussian, there exist $K > 0$, $\gamma \geq 0$ such that $E(|(\varepsilon_{i,t}\varepsilon_{l,t} - c_{i,l})(1 + c_{i,l}^2)^{-1/2}|^n) \leq K^{n-2}(n!)^{1+\gamma}$, for all $n \geq 3$, uniformly over $c_{i,l} \in [-1, 1]$. Because of this, we are able to apply Theorem 1 and the Corollary underneath it from Rudzkis et al. (1978). In the notation of that paper, computing first the quantity Δ, we get $\Delta = 2^{(J-j)/2}/\{2\max(K, 1)\}$. Since λ is logarithmic in T, the parameter $\tilde{\delta}_j \to 0$ uniformly over j as is detailed below, and $2^{(J-j)/2} \geq T^{\delta/2}$, we have that $\lambda(1 + c_{i,l}^2 - 2|c_{i,l}|\tilde{\delta}_j)^{1/2}(1 + c_{i,l}^2)^{-1/2} = o[\{2^{(J-j)/2-1}\max(K, 1)^{-1}\}^a]$, as $T \to \infty$, uniformly over j, for all $a > 0$. By Theorem 1 from Rudzkis et al. (1978), we uniformly bound (A4) from above by the Gaussian tail probability $C\,\exp\{-\lambda^2/2\,(1 + c_{i,l}^2 - 2|c_{i,l}|\tilde{\delta}_j)/(1 + c_{i,l}^2)\}$. Turning now to B, we have

$$B = \mathrm{pr}\left\{\frac{2^{(J-j)/2}|\tilde{s}_{j,k}^{(i,l)} - c_{i,l}|}{(1 + c_{i,l}^2)^{1/2}} \geq \frac{2^{(J-j)/2}\tilde{\delta}_j}{(1 + c_{i,l}^2)^{1/2}}\right\}. \tag{A5}$$

The random variable on the left-hand side of the argument of the probability function in (A5) is almost the same as that in (A4), except some different signs in the sum, which have no impact on our bounds. So, it is boundable by the corresponding Gaussian tail probability under the same conditions as A. In fact, we can choose $\tilde{\delta}_j$ to be such that the thresholds in (A4) and (A5) are equal, so that there is an exact match between the convergence rates. Equating the thresholds, we get $\lambda^2(1 + c_{i,l}^2 - 2|c_{i,l}|\tilde{\delta}_j) = 2^{J-j}\tilde{\delta}_j^2$, which gives $\tilde{\delta}_j = 2^{j-J}\lambda[\{\lambda^2 c_{i,l}^2 + 2^{J-j}(1 + c_{i,l}^2)\}^{1/2} - \lambda|c_{i,l}|]$. Since λ is logarithmic in T, $\tilde{\delta}_j$ is of order $O(T^{-\alpha})$ for $\alpha > 0$ uniformly over j. Thus, we bound $A + B$ from above by $A + B = 2\,A \leq 2C\,\exp\{-\lambda^2/2\,(1 + c_{i,l}^2 - 2|c_{i,l}|\tilde{\delta}_j)/(1 + c_{i,l}^2)\} = 2C\exp\{-\lambda^2/2\}\exp\{(\lambda^2|c_{i,l}|\tilde{\delta}_j)/(1 + c_{i,l}^2)\} \leq C_1\exp\{-\lambda^2/2\}$. Substituting this in (A1), we bound it by $C_2 T^{1-\delta}\exp(-\lambda^2/2)$. Thus, using the Bonferroni inequality again, the probability of $f_{j,k}^{(i,l)}$ not exceeding λ uniformly over all j, k, i, l can be bounded from above by $C_3 p^2 T^{1-\delta}\exp(-\lambda^2/2)$. Bounding this by the sequence $C_3 a_{p,T}$, we have $\lambda \geq \{2(2\log p + (1 - \delta)\log T + \log a_{p,T}^{-1})\}^{1/2}$, which proves the constancy of our estimator $\bar{\Sigma}^{(h)}$ with the required probability, for the range of λ's as in the statement of Lemma 1.

We now show that $\bar{c}_{i,l}^{(h)}$ is zero if the true covariance $c_{i,l}$ is zero, uniformly over i, l, with the required probability. Under the scenario that all $|f_{j,k}^{(i,l)}| \leq \lambda$, this is equivalent to showing that $|\tilde{s}_{0,1}^{(i,l)}| > \lambda_1$ for any i, l with probability not exceeding a multiple of $a_{p,T}$. Using the same technique as above, for a fixed (i,l) we bound $\mathrm{pr}(|\tilde{s}_{0,1}^{(i,l)}| > \lambda_1) \leq C_4 \exp(-\lambda_1^2 T/2)$. Thus, using the Bonferroni inequality again, we have $\mathrm{pr}(\max_{i,l} |\tilde{s}_{0,1}^{(i,l)}| > \lambda_1) \leq C_5 p^2 \exp(-\lambda_1^2 T/2)$. Bounding this by $C_5 a_{p,T}$, we obtain $\lambda_1 \geq \{2T^{-1}(2\log\, p + \log\, a_{p,T}^{-1})\}^{1/2}$.

Finally, we show that applying such a threshold λ_1 does not ruin the estimation of $c_{i,l}$ in the case $c_{i,l} \neq 0$. Under the scenario that all $|f_{j,k}^{(i,l)}| \leq \lambda$, this is equivalent to showing that $|\tilde{s}_{0,1}^{(i,l)}| < \lambda_1$ for any i, l with probability not exceeding a multiple of $a_{p,T}$. For a fixed (i,l), we have $\mathrm{pr}(|\tilde{s}_{0,1}^{(i,l)}| < \lambda_1) \leq \mathrm{pr}(|\tilde{s}_{0,1}^{(i,l)} - c_{i,l}| > |c_{i,l}| - \lambda_1) \leq \mathrm{pr}\{T^{1/2}|\tilde{s}_{0,1}^{(i,l)} - c_{i,l}|(1+c_{i,l}^2)^{-1/2} > T^{1/2}(|c_{i,l}| - \lambda_1)(1+c_{i,l}^2)^{-1/2}\}$. Assuming that the threshold on the right-hand side is so low that the normal approximation still works, which is sufficient to consider as the worst-case scenario, we bound the above by $C_6 \exp\{-T/2\,(|c_{i,l}| - \lambda_1)^2(1+c_{i,l}^2)^{-1}\} \leq C_6 \exp\{-T/4\,(|c_{i,l}| - \lambda_1)^2\}$. To obtain a uniform bound across the entire matrix, we first find the number of non-zero $c_{i,l}$'s. Recalling that $\Sigma \in \mathcal{U}\{c_0(p)\}$, we have $\sum_i \sum_l I(c_{i,l} \neq 0) \leq \sum_i \max_i \sum_l I(c_{i,l} \neq 0) \leq pc_0(p)$. Thus, by the Bonferroni inequality, we have $\mathrm{pr}(\min_{i,l} |\tilde{s}_{0,1}^{(i,l)}| < \lambda_1) \leq C_6 pc_0(p) \exp\{-T/4\, \min_{c_{i,l}\neq 0}(|c_{i,l}| - \lambda_1)^2\}$. Bounding the above by $C_6 a_{p,T}$, we get $2T^{-1/2}\{\log\, p + \log\, c_0(p) + \log a_{p,T}^{-1}\}^{1/2} + \lambda_1 \leq \min_{c_{i,l}\neq 0} |c_{i,l}|$, which is satisfied as the left-hand side has a lower order of magnitude than the right-hand side by the assumptions of Lemma 1. This completes the proof of statement (a) of Lemma 1.

For the proof of statement (b), we first calculate the error in estimating the non-zero entries. Proceeding as above, we have

$$\begin{aligned}
\mathrm{pr}(\max_{i,l:c_{i,l}\neq 0} |\bar{c}_{i,l}^{(h)} - c_{i,l}| > \lambda_3) &\leq pc_0(p) \max_{i,l} \mathrm{pr}(|\tilde{s}_{0,1}^{(i,l)} - c_{i,l}| > \lambda_3) \\
&= pc_0(p) \max_{i,l} \mathrm{pr}(T^{1/2}|\tilde{s}_{0,1}^{(i,l)} - c_{i,l}|(1+c_{i,l}^2)^{-1/2} > T^{1/2}\lambda_3(1+c_{i,l}^2)^{-1/2}) \\
&\leq C_7 pc_0(p) \max_{i,l} \exp[-T\lambda_3^2/\{2(1+c_{i,l}^2)\}] \\
&\leq C_7 pc_0(p) \exp(-T\lambda_3^2/4).
\end{aligned}$$

Equating this to $C_7 a_{p,T}$, we get $\lambda_3 = 2T^{-1/2}\{\log\, p + \log\, c_0(p) + \log a_{p,T}^{-1}\}^{1/2}$, which shows that the maximum error is λ_3 with the required large probability. On the other hand, we have shown above that our estimator has a zero error for $c_{i,l} = 0$, uniformly over the entire matrix with probability at least $1 - C_1 a_{p,T}$. Putting together these two facts, we bound $\|\bar{\Sigma}^{(h)} - \Sigma\| \leq \max_l \sum_i |\bar{c}_{i,l}^{(h)} - c_{i,l}| = \max_l \sum_i |\bar{c}_{i,l}^{(h)} - c_{i,l}| I(c_{i,l} \neq 0) \leq \lambda_3 c_0(p) = 2c_0(p)T^{-1/2}\{\log\, p + \log\, c_0(p) + \log a_{p,T}^{-1}\}^{1/2}$, which completes the proof of statement (b) of Lemma 1.

Finally, statement (c) follows since $\|(\bar{\Sigma}^{(h)})^{-1} - \Sigma^{-1}\|$ is of the same order as $\|\bar{\Sigma}^{(h)} - \Sigma\|$ uniformly over the class $\mathcal{U}\{c_0(p), \epsilon_0\}$, as in the proof of Theorem 1 in Bickel & Levina (2008). □

LEMMA 2. *Assume that the true volatility matrix $\Sigma(u)$ satisfies the following:*

(i) There exists a dyadic interval $\mathcal{I}$ of length at least 2^{-J^*}, such that for for each i and l, the function $c_{i,l}(u)$ is constant for all $u \in \mathcal{I}$.

(ii) For each i and l, if there are change-points in $c_{i,l}(u)$ to the left or right of $\mathcal{I}$, then the nearest one on either side is covered by a dyadic interval denoted $\mathcal{J}_{i,l}^1$ on the left-hand side, and $\mathcal{J}_{i,l}^2$ on the right-hand side, of length at least 2^{-J^*}, containing no other change-point, not intersecting with $\mathcal{I}$ and such that

$$\min_{i,l,m} \int_{\mathcal{J}_{i,l}^m} \left\{ c_{i,l}(u) - |\mathcal{J}_{i,l}^m| \int_{\mathcal{J}_{i,l}^m} c_{i,l}(z)dz \right\}^2 du \geq C_3 T^{-\beta}, \quad \beta \in [0, 1-\delta). \tag{A6}$$

Further, assume that $\Sigma(\mathcal{I}) \in \mathcal{U}\{c_0(p)\}$, and that its size p is at most of order $O(T^\zeta)$ for some fixed $\zeta > 0$. Let the thresholds λ and λ_1 satisfy

$$C(\log\, T)^{1/2} \geq \lambda \geq \left[2\left\{2\log\, p + (1-\delta)\log\, T + \log a_{p,T}^{-1}\right\}\right]^{1/2},$$

$$\min_{c_{i,l}(\mathcal{I})\neq 0} |c_{i,l}(\mathcal{I})| - 2\left\{\frac{\log\, p + \log\, c_0(p) + \log a_{p,T}^{-1}}{T|\mathcal{I}|}\right\}^{1/2} \geq \lambda_1 \geq \left\{\frac{2(2\log\, p + \log a_{p,T}^{-1})}{T|\mathcal{I}|}\right\}^{1/2},$$

for some $C > 0$, where $a_{p,T}$ tends to zero as $T \to \infty$ but no faster than $O(T^{-\zeta})$. The following holds with probability of at least $1 - C_1\, a_{p,T}$ for some positive C_1:

(a) Our estimator $\bar{\Sigma}^{(h)}(u)$ is constant for $u \in \mathcal{I}$ and such that $\bar{c}_{i,l}^{(h)}(\mathcal{I}) = 0$ if $c_{i,l}(\mathcal{I}) = 0$ and $\bar{c}_{i,l}^{(h)}(\mathcal{I})$ is a local sample mean of the sequence $\{X_{i,t,T}X_{l,t,T}\}_t$ over a subinterval $t/T \in \mathcal{K}_{i,l}$ where $\mathcal{I} \subseteq \mathcal{K}_{i,l}$ and $c_{i,l}(\mathcal{I}) = c_{i,l}(\mathcal{K}_{i,l})$, if $c_{i,l}(\mathcal{I}) \neq 0$.
(b) We have $\|\bar{\Sigma}^{(h)}(\mathcal{I}) - \Sigma(\mathcal{I})\| \leq 2c_0(p)(T|\mathcal{I}|)^{-1/2}(\log\, p + \log\, c_0(p) + \log a_{p,T}^{-1})^{1/2}$.
(c) If, in addition, $\Sigma(\mathcal{I}) \in \mathcal{U}\{c_0(p), \epsilon_0\}$, then

$$\|\{\bar{\Sigma}^{(h)}(\mathcal{I})\}^{-1} - \Sigma(\mathcal{I})^{-1}\| \leq C_2 c_0(p)\left\{\frac{\log\, p + \log\, c_0(p) + \log a_{p,T}^{-1}}{T|\mathcal{I}|}\right\}^{1/2}$$

for some positive C_2.

Proof. If there is a change-point in $c_{i,l}(u)$ to the left of u_1, then, denoting $2^{-j_0} = |\mathcal{J}_{i,l}^1|$ and decomposing the sampled version of $c_{i,l}(u)$ via a discrete Haar wavelet decomposition over the interval $T\mathcal{J}_{i,l}^1$ at scales $j \geq j_0$, we obtain that only up to one coefficient at each scale j is non-zero. By (A6) and due to the orthonormality of the discrete Haar transform, the sum of the squared Haar coefficients from this decomposition is at least $C_3\, T^{1-\beta}$. At each scale j, the only possibly non-zero squared Haar coefficient is at most of order 2^{J-j}, where the constants of proportionality are uniform over the entire matrix since $|c_{i,l}(u)| \leq 1$. Thus the sum of squared coefficients over the ignored scales $J^* + 1, \ldots, J - 1$ is of order $O(2^{J-J^*}) = O(T^\delta) \leq C_4\, T^\delta$. Thus, the sum of squared Haar coefficients over the non-ignored scales $j_0, \ldots, J^*$ must be at least $C_3\, T^{1-\beta} - C_4\, T^\delta \geq C_5\, T^{1-\beta}$. Therefore, the largest non-squared Haar coefficient must be of magnitude of at least $C_6 T^{1/2-\beta/2} \log^{-1/2} T$, since there are at most $\log_2 T$ decomposition scales. Denote by $j_1(i,l)$ the scale at which the largest coefficient occurs, and note that $j_0 \leq j_1(i,l) \leq J^*$. Similarly denote its location by $k_1(i,l)$.

We wish to investigate if the coefficient $f_{j_1(i,l),k_1(i,l)}^{(i,l)}$ survives thresholding. If it does, then with probability one, there will be a change-point in $\tilde{c}_{i,l}(u)$ at $u = u_0$ where u_0 is the right endpoint of $\mathcal{J}_{i,l}^1$; thus, there will be a change-point in $\tilde{c}_{i,l}(u)$ located between the interval $\mathcal{I}$ and its nearest change-point to the left. But, using the same technique as in the proof of Lemma 1, we can show that $\mathrm{pr}(\min_{i,l} |f_{j_1(i,l),k_1(i,l)}^{(i,l)}| < \lambda) \leq C_7 a_{p,T}$.

Moreover, since all coefficients $f_{j,k}^{(i,l)}$ computed over the interval of constancy $T\mathcal{I}$ fall under the threshold λ with probability at least $1 - C_8 a_{p,T}$ by Lemma 1, we have that for all i and l, $\tilde{c}_{i,l}(\mathcal{I}) = (|T\mathcal{K}_{i,l}|)^{-1} \sum_{t\in T\mathcal{K}_{i,l}} X_{i,t,T}X_{l,t,T}$, for a certain $\mathcal{K}_{i,l} \supseteq \mathcal{I}$ where $c_{i,l}(\mathcal{I}) = c_{i,l}(\mathcal{K}_{i,l})$, holds with probability at least $1 - C_8 a_{p,T}$.

Therefore, we have a similar situation to the framework of Lemma 1, where all $\tilde{c}_{i,l}(u)$ were, with probability at least $1 - C_8 a_{p,T}$ constant with u and equal to the sample means of $\{X_{i,t,T}X_{l,t,T}\}_{t=1}^T$. Here, the same kind of constancy holds but locally: all $\tilde{c}_{i,l}(u)$ are constant for $u \in \mathcal{I}$ and each equals the sample mean of $\{X_{i,t,T}X_{l,t,T}\}_{t\in\mathcal{K}_{i,l}}$ where $\mathcal{K}_{i,l} \supseteq \mathcal{I}$. Thus, reproducing the argument of Lemma 1, we can show that with probability at least $1 - C_9 a_{p,T}$, we have that $\bar{c}_{i,l}^{(h)}(\mathcal{I}) = 0$ for all those i and l for which $c_{i,l}(\mathcal{I}) = 0$ if $\lambda_1 \geq \{2T^{-1}|\mathcal{I}|^{-1}(2\log\, p + \log a_{p,T}^{-1})\}^{1/2}$. Similarly, with probability at least $1 - C_{10} a_{p,T}$, we have that $\bar{c}_{i,l}^{(h)}(\mathcal{I}) = \tilde{c}_{i,l}(\mathcal{I})$ for all those i, l for which $c_{i,l}(\mathcal{I}) \neq 0$ if $\lambda_1 \leq \min_{c_{i,l}(\mathcal{I})\neq 0} |c_{i,l}(\mathcal{I})| - 2(T|\mathcal{I}|)^{-1/2}\{\log\, p + \log\, c_0(p) + \log a_{p,T}^{-1}\}^{1/2}$. This completes the proof of statement (a). The proofs of statements (b) and (c) proceed analogously to those of the corresponding statements in Lemma 1 by recalling that $|\mathcal{K}_{i,l}| \geq |\mathcal{I}|$ and replacing T with $T|\mathcal{I}|$ where appropriate. □

LEMMA 3. *Assume that the true volatility matrix* $\Sigma(u) = \Sigma$ *is constant and such that* $diag^{-1/2}(\Sigma)\Sigma diag^{-1/2}(\Sigma) \in \mathcal{U}\{c_0(p)\}$, *and that its size* p *is at most of order* $O(T^{\zeta})$ *for some fixed* $\zeta > 0$. *Assume also that the elements of diag*(Σ) *are uniformly bounded from below and above by constants. Recall the notation* $\rho_{i,l} = c_{i,l} c_{l,l}^{-1/2} c_{i,i}^{-1/2}$ *and let* $\underline{c} = \inf_i c_{i,i}$ *and* $\bar{c} = \sup_i c_{i,i}$. *Let the thresholds* λ *and* λ_1 *satisfy* $\lambda \geq [2\{2\log p + (1-\delta)\log T + \log a_{p,T}^{-1}\}]^{1/2}$,

$$(1 + \varpi/\underline{c})^{-1}(\min_{\rho_{i,l}\neq 0} |\rho_{i,l}| - \varpi/\bar{c}) \geq \lambda_1 \geq \left\{\frac{2(2\log p + \log a_{p,T}^{-1})}{T}\right\}^{1/2},$$

where $a_{p,T}$ *tends to zero as* $T \to \infty$ *but no faster than* $O(T^{-\zeta})$, *and*

$$\varpi = 2\bar{c}\left\{\frac{\log p + \log c_0(p) + \log a_{p,T}^{-1}}{T}\right\}^{1/2}.$$

The following holds with probability of at least $1 - C_1 a_{p,T}$ *for some positive* C_1*:*

(a) Our estimator $\bar{\Sigma}^{(h)}(u) = \bar{\Sigma}^{(h)}$ *is constant and such that* $\bar{c}_{i,l}^{(h)}(t/T) \equiv 0$ *if* $c_{i,l} = 0$ *and* $\bar{c}_{i,l}^{(h)}(t/T) \equiv \frac{1}{T}\sum_{t=1}^{T} X_{i,t,T} X_{l,t,T}$ *if* $c_{i,l} \neq 0$.
(b) We have $\|\bar{\Sigma}^{(h)} - \Sigma\| \leq 2c_0(p)\bar{c}T^{-1/2}\{\log p + \log c_0(p) + \log a_{p,T}^{-1}\}^{1/2}$.
(c) If, in addition, diag$^{-1/2}(\Sigma)\Sigma$*diag*$^{-1/2}(\Sigma) \in \mathcal{U}\{c_0(p), \epsilon_0\}$, *then* $\|(\bar{\Sigma}^{(h)})^{-1} - \Sigma^{-1}\| \leq C_2 c_0(p) T^{-1/2}\{\log p + \log c_0(p) + \log a_{p,T}^{-1}\}^{1/2}$ *for some positive* C_2.

Proof. Proceeding as in the proof of Lemma 1, we have

$$\begin{aligned}\mathrm{pr}\left(|f_{j,k}^{(i,l)}| > \lambda\right) &= \mathrm{pr}\left[|d_{j,k}^{(i,l)}| > \lambda\left\{\tilde{s}_{j,k}^{(i,i)}\tilde{s}_{j,k}^{(l,l)} + (\tilde{s}_{j,k}^{(i,l)})^2\right\}^{1/2} \mid A\right]\mathrm{pr}(A)\\ &\quad + \mathrm{pr}\left[|d_{j,k}^{(i,l)}| > \lambda\left\{\tilde{s}_{j,k}^{(i,i)}\tilde{s}_{j,k}^{(l,l)} + (\tilde{s}_{j,k}^{(i,l)})^2\right\}^{1/2} \mid A^c\right]\mathrm{pr}(A^c), \qquad \text{(A7)}\end{aligned}$$

where $A = \{|\tilde{s}_{j,k}^{(i,l)} - c_{i,l}| < \tilde{\delta}_j, |\tilde{s}_{j,k}^{(i,i)} - c_{i,i}| < \tilde{\delta}_j, |\tilde{s}_{j,k}^{(l,l)} - c_{l,l}| < \tilde{\delta}_j\}$. Using (A3), we bound (A7) by

$$\begin{aligned}&\mathrm{pr}\left(|d_{j,k}^{(i,l)}| > \lambda\left\{(c_{i,i} - \tilde{\delta}_j)(c_{l,l} - \tilde{\delta}_j) + c_{i,l}^2 - 2|c_{i,l}|\tilde{\delta}_j\right\}^{1/2}\right) + \mathrm{pr}(A^c)\\ &\quad \leq \mathrm{pr}\left[\frac{|d_{j,k}^{(i,l)}|}{(c_{i,i}c_{l,l} + c_{i,l}^2)^{1/2}} > \lambda\left\{\frac{c_{i,i}c_{l,l} + c_{i,l}^2 - \tilde{\delta}_j(c_{i,i} + c_{l,l} - \tilde{\delta}_j + 2|c_{i,l}|)}{c_{i,i}c_{l,l} + c_{i,l}^2}\right\}^{1/2}\right]\\ &\quad + \mathrm{pr}(|\tilde{s}_{j,k}^{(i,l)} - c_{i,l}| > \tilde{\delta}_j) + \mathrm{pr}(|\tilde{s}_{j,k}^{(i,i)} - c_{i,i}| > \tilde{\delta}_j) + \mathrm{pr}(|\tilde{s}_{j,k}^{(l,l)} - c_{l,l}| > \tilde{\delta}_j).\end{aligned}$$

This leads to practically the same situation as in the proof of Lemma 1, and proceeding analogously, we are able to bound the above by

$$\begin{aligned}&4C\exp\left[-\frac{\lambda^2}{2}\left\{\frac{c_{i,i}c_{l,l} + c_{i,l}^2 - \tilde{\delta}_j(c_{i,i} + c_{l,l} - \tilde{\delta}_j + 2|c_{i,l}|)}{c_{i,i}c_{l,l} + c_{i,l}^2}\right\}\right]\\ &\quad = 4C\exp\left(-\frac{\lambda^2}{2}\right)\exp\left\{\frac{\lambda^2\tilde{\delta}_j(c_{i,i} + c_{l,l} - \tilde{\delta}_j + 2|c_{i,l}|)}{2(c_{i,i}c_{l,l} + c_{i,l}^2)}\right\} \leq C_1\exp\left(-\frac{\lambda^2}{2}\right),\end{aligned}$$

which leads to the same lower bound for λ as in the proof of Lemma 1.

We now show that $\bar{c}_{i,l}^{(h)}$ is zero if the true covariance $c_{i,l}$ is zero, uniformly over i, l, with the required probability. For a fixed (i,l), we use the same technique as above with conditioning on the set $A = \{|\tilde{s}_{0,1}^{(i,i)} - c_{i,i}| < \tilde{\delta}_0, |\tilde{s}_{0,1}^{(l,l)} - c_{l,l}| < \tilde{\delta}_0\}$ to bound $\mathrm{pr}\{|\tilde{s}_{0,1}^{(i,l)}| > \lambda_1(\tilde{s}_{0,1}^{(i,i)}\tilde{s}_{0,1}^{(l,l)})^{1/2}\} \leq C_4\exp(-\lambda_1^2 T/2)$, which leads to the same lower bound for λ_1 as in the proof of Lemma 1.

Finally, we show that applying such a threshold λ_1 does not ruin the estimation of $c_{i,l}$ in the case $c_{i,l} \neq 0$. For a fixed (i,l), using conditioning as above, we have

$$\mathrm{pr}\{|\tilde{s}_{0,1}^{(i,l)}| < \lambda_1 (\tilde{s}_{0,1}^{(i,i)} \tilde{s}_{0,1}^{(l,l)})^{1/2}\} \leq \mathrm{pr}\left[|\tilde{s}_{0,1}^{(i,l)}| < \lambda_1 \left\{(c_{i,i} + \tilde{\delta}_0)(c_{l,l} + \tilde{\delta}_0)\right\}^{1/2}\right] + \mathrm{pr}(A^c). \qquad \text{(A8)}$$

Arguing like in the proof of Lemma 1, there are no more than $pc_0(p)$ non-zero terms $c_{i,l}$ and hence the condition

$$\tilde{\delta}_0 \geq 2\bar{c} \left\{ \frac{\log\, p + \log\, c_0(p) + \log a_{p,T}^{-1}}{T} \right\}^{1/2} \qquad \text{(A9)}$$

guarantees that the terms $\mathrm{pr}(A^c)$ in (A8) sum to at most a term of order $a_{p,T}$ across the entire matrix. Using the lower bound for $\tilde{\delta}_0$ from (A9), we bound

$$\begin{aligned}
&\mathrm{pr}\left[|\tilde{s}_{0,1}^{(i,l)}| < \lambda_1 \left\{(c_{i,i} + \tilde{\delta}_0)(c_{l,l} + \tilde{\delta}_0)\right\}^{1/2}\right] \\
&\quad \leq \mathrm{pr}\left[|\tilde{s}_{0,1}^{(i,l)} - c_{i,l}| > |c_{i,l}| - \lambda_1 \left\{(c_{i,i} + \tilde{\delta}_0)(c_{l,l} + \tilde{\delta}_0)\right\}^{1/2}\right] \\
&\quad = \mathrm{pr}\left(\frac{T^{1/2}|\tilde{s}_{0,1}^{(i,l)} - c_{i,l}|}{(c_{i,i}c_{l,l} + c_{i,l}^2)^{1/2}} > \frac{T^{1/2}\left[|\rho_{i,l}| - \lambda_1 \left\{(1 + \tilde{\delta}_0/c_{i,i})(1 + \tilde{\delta}_0/c_{l,l})\right\}^{1/2}\right]}{(1 + \rho_{i,l}^2)^{1/2}} \right) \\
&\quad \leq \mathrm{pr}\left[\frac{T^{1/2}|\tilde{s}_{0,1}^{(i,l)} - c_{i,l}|}{(c_{i,i}c_{l,l} + c_{i,l}^2)^{1/2}} > \frac{T^{1/2}}{2^{1/2}} \left\{ \min_{\rho_{i,l} \neq 0} |\rho_{i,l}| - \lambda_1 \max_i (1 + \tilde{\delta}_0/c_{i,i}) \right\} \right],
\end{aligned}$$

which, analogously to the proof of Lemma 1, leads to

$$2\left\{ \frac{\log\, p + \log\, c_0(p) + \log a_{p,T}^{-1}}{T} \right\}^{1/2} + \lambda_1 \max_i (1 + \tilde{\delta}_0/c_{i,i}) \leq \min_{\rho_{i,l} \neq 0} |\rho_{i,l}|,$$

which agrees with the assumptions of Lemma 3. This completes the proof of statement (a).

The proofs of statements (b) and (c) are like those of the analogous statements in Lemma 1, so we omit them here. □

Proof of Theorem 1. The proof uses Lemma 3 in the same way as the proof of Lemma 2 uses Lemma 1. The construction of the proof of Theorem 1 is analogous to that of Lemma 2. We omit the details. □

REFERENCES

ANDERSEN, T. G., DAVIS, R. A., KREISS, J.-P. & MIKOSCH, T., eds. (2009). *Handbook of Financial Time Series*. Springer.

BICKEL, P. J. & LEVINA, E. (2008). Covariance regularization by thresholding. *Ann. Stat.* **36**, 2577–2604.

CAI, T. T. & JIANG, T. (2011). Limiting laws of coherence of random matrices with applications to testing covariance structure and construction of compressed sensing matrices. *Annals of Statistics* **39**, 1496–1525.

CAI, T. T. & JIANG, T. (2012). Phase transition in limiting distributions of coherence of high-dimensional random matrices. *Journal of Multivariate Analysis* **107**, 24–39.

DEY, D. K. & SRINIVASAN, C. (1985). Estimation of a covariance matrix under Stein's loss. *Ann. Statist.* **13**, 1581–1591.

EL KAROUI, N. (2008). Operator norm consistent estimation of large-dimensional sparse covariance matrices. *Ann. Statist.* **36**, 2717–2756.

ENGLE, R. F. (1982). Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation. *Econometrica* **50**, 987–1007.

FAN, J., ZHANG, J. & YU, K. (2012). Vast portfolio selection with gross-exposure constraints. *Journal of the American Statistical Association* **107**, 592–606.

FISZ, M. (1955). The limiting distribution of a function of two independent random variables and its statistical application. *Colloquium Mathematicum* **3**, 138–146.

FRYZLEWICZ, P. (2005). Modelling and forecasting financial log-returns as locally stationary wavelet processes. *Journal of Applied Statistics* **32**, 503–528.

FRYZLEWICZ, P. (2008). Data-driven wavelet–Fisz methodology for nonparametric function estimation. *Electronic Journal of Statistics* **2**, 863–896.

FRYZLEWICZ, P. & NASON, G. P. (2004). A Haar–Fisz algorithm for Poisson intensity estimation. *Journal of Computational and Graphical Statistics* **13**, 621–638.

FRYZLEWICZ, P., SAPATINAS, T. & SUBBA RAO, S. (2006). A Haar–Fisz technique for locally stationary volatility estimation. *Biometrika* **93**, 687–704.

FRYZLEWICZ, P., SAPATINAS, T. & SUBBA RAO, S. (2008). Normalised least-squares estimation in time-varying ARCH models. *Annals of Statistics* **36**, 742–786.

GOLUB, G. H. & VAN LOAN, C. F. (1989). *Matrix Computations, 2nd ed.* Baltimore, MD: Johns Hopkins Univ. Press.

HAFF, L. R. (1980). Empirical Bayes estimation of the multivariate normal covariance matrix. *Ann. Statist.* **8**, 586–597.

HÄRDLE, W., HERWARTZ, H. & SPOKOINY, V. (2003). Time inhomogeneous multiple volatility modeling. *Journal of Financial Econometrics* **1**, 55–95.

HERZEL, S., STARICA, C. & TUTUNCUC, R. (2006). A non-stationary paradigm for the dynamics of multivariate financial returns. In *Dependence in Probability and Statistics*, Lecture Notes in Statistics, vol. 187, Part III. pp. 391–429.

JANEWAY, W. (2009). Six impossible things before breakfast: lessons from the crisis. *Significance* **6**, 28–31.

JIANG, T. (2004). The asymptotic distribution of the largest entries of sample correlation matrices. *Annals of Applied Probability* **14**, 865–880.

LEDOIT, O. & WOLF, M. (2003). A well-conditioned estimator for large-dimensional covariance matrices. *J. Multivariate Anal.* **88**, 365–411.

LI, D., QI, Y. & ROSALSKY, A. (2012). On Jiang's asymptotic distribution of the largest entry of a sample correlation matrix. *Journal of Multivariate Analysis* **111**, 256–270.

LUNDE, A. & HANSEN, P. R. (2005). A forecast comparison of volatility models: does anything beat a GARCH(1,1)? *J. Appl. Econometrics* **20**, 873–889.

MARKOWITZ, H. (1952). Portfolio selection. *Journal of Finance* **7**, 77–91.

MARKOWITZ, H. (1959). *Portfolio Selection: Efficient Discrimination of Investments*. New York: Wiley.

MIKOSCH, T. & STARICA, C. (2004). Non-stationarities in financial time series, the long range dependence and the IGARCH effects. *Review of Economics and Statistics* **86**, 378–390.

MOTTA, G., HAFNER, C. M. & VON SACHS, R. (2011). Locally stationary factor models: identification and non-parametric estimation. *Econometric Theory* **27**, 1279–1319.

RODRIGUEZ-POO, J. M. & LINTON, O. (2001). Nonparametric factor analysis of residual time series. *Test* **10**, 161–182.

ROTHMAN, A. J., LEVINA, E. & ZHU, J. (2009). Generalized thresholding of large covariance matrices. *Journal of the American Statistical Association* **104**, 177–186.

RUDZKIS, R., SAULIS, L. & STATULEVICIUS, V. (1978). A general lemma on probabilities of large deviations. *Lithuanian Math. J.* **18**, 226–238.

STARICA, C. & GRANGER, C. (2005). Non-stationarities in stock returns. *Review of Economics and Statistics* **87**, 503–522.

TAYLOR, S. J. (1986). *Modelling Financial Time Series*. Chichester: Wiley.

TIBSHIRANI, R., SAUNDERS, M., ROSSET, S., ZHU, J. & KNIGHT, K. (2005). Sparsity and smoothness via the fused lasso. *J. Roy. Stat. Soc. B* **67**, 91–108.

VIDAKOVIC, B. (1999). *Statistical Modeling by Wavelets*. New York: Wiley.

WANG, Y. & ZOU, J. (2010). Vast volatility matrix estimation for high-frequency financial data. *Ann. Stat.* **38**, 943–978.