Haar-Fisz wavelet method for interpretable estimation of large, sparse, time-varying volatility matrices

Piotr Fryzlewicz
Department of Statistics
London School of Economics, UK
p.fryzlewicz@lse.ac.uk

December 13, 2010

Abstract

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, with the number of jumps possibly increasing with the sample size. We show that the proposed model accurately reflects the typical stylised facts of multivariate returns. We propose a new wavelet-based technique for estimating the volatility matrix, which combines four essential ingredients: a Haar wavelet decomposition, variance stabilisation 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 polarisation identity, which permits a more precise derivation of the thresholds. Using the example of a stock index portfolio, we discuss practical selection of the parameters of our procedure.

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), whose concept of portfolio efficiency used the standard deviation of a portfolio as a measure of its uncertainty; consequently, allocating a Markowitz-efficient portfolio in practice requires accurate estimation of the associated volatility matrix. 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. (2010).

Not being directly observable, 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 (see e.g. Lunde and Hansen, 2005, for a review), volatility is understood as the variance of the returns process conditional on its own past values; in Stochastic Volatility modelling (Taylor, 1986; see Andersen et al., 2009, Part II for a review) it is the variance conditional on a possibly external random process; in the non-stationary deterministic approach by Starica and Granger (2005), Fryzlewicz (2005) and Fryzlewicz et al. (2006), it is the unconditional local variance of the returns process.

The interest of this work lies in multivariate volatility modelling and the associated estimation of the volatility matrix. There have been numerous attempts at extending the (G)ARCH framework to multiple dimensions, although the level of complexity of the proposed multivariate GARCH models has understandably fallen short of that of univariate models. The VEC (Bollerslev et al., 1988) and the BEKK (Engle and Kroner, 1995) models are both natural and direct generalisations of the univariate GARCH model to the multivariate case. Due to the need to parameterise the evolution of the covariances between the assets, the numbers of parameters in the general VEC and BEKK models are large, which, combined with the complicated form of the likelihoods, makes estimation challenging; it is for this reason that the simplest BEKK model of order $(1, 1)$ is by far the most popular one in literature. Some efforts have been made to alleviate the parameterisation problem, e.g. Engle et al. (1990) assume that the conditional covariance matrix is generated by only a few factors, each of which has a simple GARCH(1,1) structure. Alexander (2001) decorrelates the multivariate time series of returns via principal components (PC) and applies univariate GARCH modelling to each PC. Motivated by the fact that PCs are only unconditionally uncorrelated, Fan et al. (2008a) construct Conditionally Uncorrelated Components and model each one as univariate GARCH. Using a straightforward but elegant decomposition of the conditional covariance matrix into conditional standard deviations and correlations, Bollerslev (1990) models conditional correlations as constant and conditional variances as univariate GARCH processes, whereas Tse and Tsui (2002) and Engle (2002) introduce GARCH-type dynamics into the conditional correlation structure. The reader is also referred to a survey on multivariate GARCH models by Bauwens et al. (2006). There have been numerous attempts at extending the Stochastic Volatility (SV) framework to multiple dimensions: we refer the reader to the survey in Andersen et al. (2009), Part II. Estimation in multivariate Stochastic Volatility models is typically not straightforward and tends to be performed via computationally intensive approaches. We note that the fact that SV models tend to be "more difficult to estimate" than ARCH-type models was also noted by Robert Engle in his Nobel Prize lecture.

The overwhelming majority of existing multivariate GARCH and SV models, including those listed above, are stationary. Although stationarity is an attractive assumption from the point of view of estimation and interpretability, some authors point out that the typical "stylised 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 and Starica (2004), Starica and Granger (2005) and Fryzlewicz et al. (2008) for arguments in the univariate case. Underlying these approaches is the observation that given the changing pace of the world economy, it is unlikely that log-return series should stay stationary over long time intervals. Janeway (2009) goes further and claims that traditional models' stationarity (an attribute which can potentially translate into lack of flexibility) might have been a contributing factor in the recent financial crisis, since they could have been too slow to react to early warning signs of the crisis and therefore might have impeded efficient risk management. In the multivariate GARCH

setting, non-stationarity tends to be introduced in either of two ways. One is to assume that some of the parameters of the GARCH model evolve slowly over time, and estimate them nonparametrically, as in the Semi-Parametric Conditional Correlation GARCH model of Hafner et al. (2005). The other is to assume that some of the GARCH parameters change according to a (possibly stationary) regime-switching mechanism, as in the Regime Switching Dynamic Correlation GARCH model by Pelletier (2006). A computationally intensive method is required to perform full estimation in the latter modelling approach.

Multivariate GARCH models, while being an intellectually appealing extension of the popular and successful univariate framework, suffer from a number of issues, which must not be overlooked by the analyst wishing to apply them in practice. Estimation is not always easy or reliable (see e.g. Brooks et al. (2003)), and reducing the number of parameters to aid estimation can result in models that are unlikely to be able to describe the dynamics of the data accurately. Further, theoretical properties of multivariate GARCH models are not yet fully understood, and, in some cases, difficult to establish. Thus, it is meaningful, perhaps more so than in the univariate case, to look for alternatives to multivariate GARCH modelling.

One such alternative modelling avenue opens up once one relaxes the assumption of stationarity, as motivated above. An interesting question which can be asked in the non-stationary setting is whether non-linearity (as present in the data-generating mechanism of GARCH models) is still needed to model returns data accurately, or whether it is sufficient to stick to linear models, the latter being conceptually simpler and better understood. Locally stationary linear models (Dahlhaus (1997), Nason et al. (2000)) seem to be a particularly interesting option here, as they combine linearity with a modelling approach whereby the time-varying parameters are modelled as "well-behaved" functions defined on a compact interval, which enables a meaningful asymptotic estimation theory. Some authors have applied the locally stationary linear framework to the modelling of univariate returns data; see for example Clemencon and Slim (2004), who apply the locally stationary covariance estimation methodology of Donoho et al. (2003) to returns data, Fryzlewicz (2005), who provides an exploratory analysis of log-returns in the framework of Nason et al. (2000), or Fryzlewicz et al. (2006), who model volatility as a slow-varying deterministic function, with an associated "Haar-Fisz" estimation technique. This latter technique will be summarised in detail later on. The locally stationary linear approach has also been explored in the multivariate setting. Rodriguez-Poo and Linton (2001) and Herzel et al. (2006) both assume the locally stationary linear model in which multivariate volatility is understood to be the local unconditional covariance matrix, and use kernel smoothing to estimate it.

Although using kernel smoothing to estimate time-varying multivariate volatility certainly has its appeal as it leads to nonnegative-definite estimators without any further fine-tuning, it also has its drawbacks. Kernel estimators are known to be non-adaptive in that they do not automatically adapt to the unknown smoothness of the target function. For example, imagine the simplest possible case where the underlying "true" volatility is constant, or can be reasonably modelled as such from the point of view of the analyst. A kernel estimator of such constant volatility will always be oscillatory (however slightly) around the true value. This is undesirable from the point of view of the analyst (e.g. a trading firm), who will have to bear unnecessary trading costs in making extra adjustments to their positions at every time unit, driven by the oscillations of the kernel estimator of volatility, in a situation where no adjustments are necessary as the true volatility is constant. Thus, it makes perfect sense

to seek more "adaptive" estimators of multivariate volatility, which would only change in value whenever necessary, and which would estimate constant volatility as constant, with high probability. One such estimator is proposed by Haerdle et al. (2003), who search for the longest interval of approximate constancy of volatility via iterative hypothesis testing.

Our work is motivated by the three arguments summarised above: the need to look for alternatives to multivariate GARCH modelling; the appeal of non-stationarity in financial returns modelling; and the need to depart from traditional kernel smoothing in nonparametric estimation of volatility. The aims of this work can be briefly summarised as follows:

- To propose a locally stationary linear model for the evolution of multivariate financial returns, where the time-varying volatility is modelled as a function, defined in rescaled time over a compact interval, possessing some degree of regularity. In this work, our regularity assumption is that the volatility is piecewise-constant, where the number of the breakpoints can increase with the sample size and they can approach each other in rescaled time.

- To show that the proposed locally stationary model can accurately reflect the typical stylised facts of multivariate returns, e.g. those concerning the behaviour of sample cross-covariances of returns across markets.

- To propose a new wavelet-based technique for estimating the correlation matrix and the covariance matrix (= volatility matrix) of multivariate returns. The method combines Haar wavelets and the variance-stabilising Fisz transformation, which is described in more detail below. Our rationale for choosing wavelets in this setting is at least fourfold: firstly, wavelets provide an adaptive "bridge" between stationary and non-stationary estimation in the sense that if the true volatility is constant, then so will be the wavelet-estimated volatility, with high-probability, which is a desirable feature of any volatility estimator as described above. Secondly, being piecewise constant, Haar wavelets also guarantee this property interval-wise, which ensures a degree of interpretability of our estimator. Thirdly, wavelet estimators are fast to compute, which is important if the dimension of the volatility matrix is large. Finally, unlike many other wavelet families, Haar wavelets do not suffer from the boundary issue, which for our purposes means that our estimators will be theoretically tractable also at the right-hand end of the data, i.e. at the current time $t = T$. We describe the Fisz transform (and why it is needed) in more detail below.

In modern financial statistics, it is impossible to avoid the setting of high-dimensionality, i.e. the situation where the number of assets considered is high, perhaps even higher than the effective number of observations for each asset (note that the effective number of observations may be less than the actual number of observations in a non-stationary setting). In this challenging setting, one needs to choose paradigms and estimation tools carefully to ensure good theoretical and practical performance of the volatility matrix estimator, e.g. its consistency in the desired matrix norm or its stable invertibility. Our paradigm of choice is sparsity, where, at each time point, only a certain (small, in comparison to the number of assets) number of true cross-market covariances are assumed to be non-zero. We note that Bickel and Levina (2008) and El Karoui (2008) proposed thresholding estimators of a sparse stationary covariance matrix, and Wang and 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 (daily) data, but in a non-stationary setting.

Classical function estimation via wavelet thresholding in the function+noise setting requires that the standard deviation of the noise should be constant over time; the value of the threshold is then, usually, a multiple of this quantity. In our setting however, the standard deviation of the sample local cross-covariance is a non-trivial function of the local cross-covariance itself. Thus, in order to apply the wavelet thresholding machinery, variance stabilisation is required. Our methodology of choice is the Haar-Fisz technique (Fryzlewicz and Nason (2004), Fryzlewicz (2008)), in which, roughly speaking, empirical wavelet coefficients are standardised by the local maximum likelihood estimates of their own standard deviations, which ensures variance stabilisation prior to the application of wavelet thresholding. Thus this procedure can be viewed as a type of "studentisation" performed in the wavelet domain. We note that this technique was applied to univariate volatility estimation in Fryzlewicz et al. (2006); however, critical and interesting differences arise in the multivariate setting. Another crucial difference between our approach and classical wavelet thresholding (as well as between multivariate and univariate Haar-Fisz methodology) is that we apply a certain essential bias correction and perform additional thresholding in the time domain. These two actions ensure, in our sparse setting, that true zero correlations are estimated exactly as zero with high probability.

The paper is organised as follows. Section 2 introduces and motivates our model. Section 3 introduces our estimators and studies their theoretial properties. Section 4 offers insights as to the practical choice of the parameters of our estimators in the context of a stock index portfolio example. The proofs are in the Appendix.

2 Motivation

2.1 The multivariate model

This section introduces and motivates our locally stationary model for multivariate financial returns. Let $\mathbf{X}_{t,T}$, $t = 1, \ldots, T$, be a p-dimensional process of log-returns on financial instruments, with components $X_{j,t,T}$ for $j = 1, \ldots, p$, where p can be large and even 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 $\boldsymbol{\varepsilon}_t = (\varepsilon_{1,t}, \ldots, \varepsilon_{p,t})^T$ are independent, and distributed such that

$$\begin{aligned} \mathbb{E}(\boldsymbol{\varepsilon}_t) &= \mathbf{0} \\ \mathrm{Var}(\boldsymbol{\varepsilon}_t) &= \Gamma(t/T), \end{aligned}$$

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. Extra distributional assumptions on $\varepsilon_{j,t}$ will be specified in Section 3.1. Let $\Sigma(t/T) = (c_{i,j}(t/T))_{i,j=1}^p$ denote the variance matrix of $\mathbf{X}_{t,T}$, 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 obvious decomposition

$$\Sigma(t/T) = D(t/T)\, \Gamma(t/T)\, D(t/T).$$

Further, we assume that all $\varepsilon_{j,t}$'s are identically distributed; thus, marginally, each $X_{j,t,T}$ follows exactly the same univariate model as that proposed in 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, which needs to be estimated from a single stretch of observations $\mathbf{X}_{t,T}$. Note that $\Sigma(u)$ is defined over the interval $(0, 1]$, which is common practice in nonparametric regression and is done in order to enable meaningful estimation theory. In Section 2.2, we demonstrate that piecewise stationarity, which is arguably the simplest type of departure from stationarity, is already flexible enough to enable realistic modelling of multivariate volatilities. Note that piecewise-constant modelling of multivariate volatilities was also considered by Haerdle et al. (2003). Our nonparametric approach allows us to avoid the restrictions imposed by the parametric structures of ARCH/GARCH models.

2.2 Stylised facts of multivariate asset returns

Fryzlewicz et al. (2006) argued how the *univariate* model (1) was able to capture typical stylised facts of univariate financial returns. More specifically, by Proposition 1 of Fryzlewicz et al. (2006), if the number of jumps in $\sigma_j(t/T)$ is finite and if $\mathbb{E}(\varepsilon_{j,t}^8) < \infty$, then the following holds (the reader is referred to the above work for precise mathematical statements).

- The sample mean of $X_{j,t,T}$ converges to zero in mean-square.
- The sample kurtosis of $X_{j,t,T}$ converges in probability to a limit larger than or equal to $\mathbb{E}(\varepsilon_{j,t}^4)$, where the equality holds if and only if $\sigma_j(t/T)$ is constant; this implies that $X_{j,t,T}$ "spuriously" appears as heavy-tailed if it is non-stationary, when its kurtosis is estimated by the global sample kurtosis.
- For each fixed lag, the sample autocovariance of $X_{j,t,T}$ converges to zero in mean-square.
- For each fixed lag, the sample autocovariance of $X_{j,t,T}^2$ converges in mean-square to a non-negative quantity, which is zero if and only if $\sigma_j(t/T)$ is constant; this implies that $X_{j,t,T}^2$ "spuriously" appears as a long memory process if it is non-stationary, when the global sample autocovariance is used to estimate its autocovariance structure.

The purpose of this section is to demonstrate interesting and important "stylised facts" of multivariate returns, and argue how our non-stationary linear model can capture them. The cross-market analogy of squared returns are product returns, defined as $X_{i,t,T}X_{j,t,T}$ for the pair of assets (i, j). The global sample autocovariance of product returns at lag h for markets (i, j) is defined as

$$\gamma_{i,j}^T(h) = \frac{1}{T}\sum_{t=1}^{T-h} X_{i,t,T}X_{j,t,T}\, X_{i,t+h,T}X_{j,t+h,T} - \left(\frac{1}{T}\sum_{t=1}^{T} X_{i,t,T}X_{j,t,T}\right)^2.$$

Obviously, $\gamma_{i,i}^T(h)$ reduces to the global sample autocovariance of squared returns $X_{i,t,T}^2$.

In this example, we let $\mathbf{X}_{t,T} = (X_{1,t,T}, \ldots, X_{10,t,T})^T$ be the series of logged and differenced daily closing values, from 3 January 1995 to 10 March 2010, of 10 major stock indices: All Ordinaries, Bovespa, CAC 40, DAX, Dow Jones Industrial Average, FTSE 100, Hang

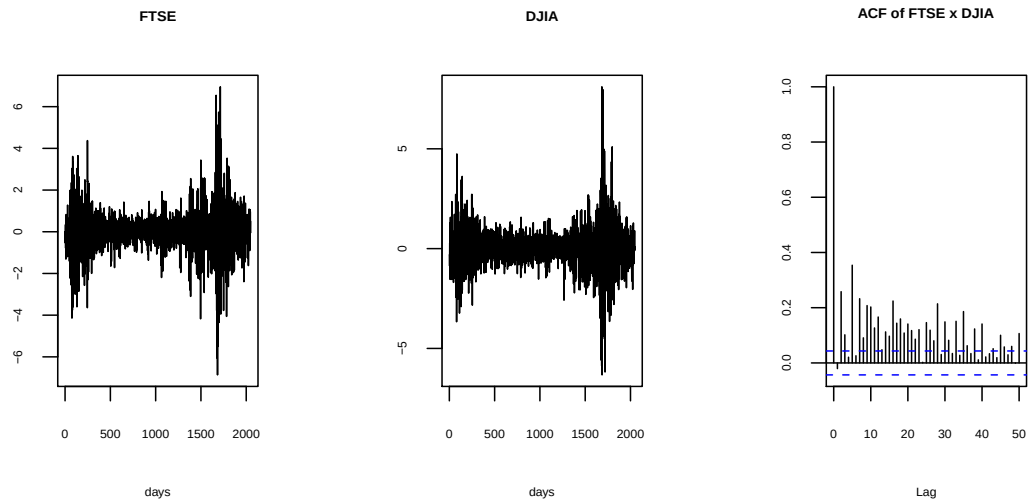

Figure 1: Daily returns on the FTSE index ($X_{6,t}$, left) and the DJIA index ($X_{5,t}$, middle) from 3 January 1995 to 10 March 2010, normalised so that sample variance is one. Right: sample acf of $X_{1,t}X_{2,t}$.

Seng, NASDAQ, Nikkei and S&P 500 (respectively). $X_{5,t,T}$, $X_{6,t,T}$ and $\gamma^T_{5,6}(\cdot)$ (on the autocorrelation, rather than autocovariance scale) are plotted in Figure 1. If $X_{5,t,T}X_{6,t,T}$ were generated by a stationary, short-memory time series model with appropriate moment conditions satisfied, we would expect $\gamma^T_{5,6}(h)$ to decay quickly to zero as h increased. However, empirically, we observe a high degree of persistence of $\gamma^T_{5,6}(h)$, which appears to be mostly large positive even for lags around $h = 50$ and possibly beyond. This is by no means an isolated case. Table 1 lists the quantity $\frac{1}{30\gamma^T_{i,j}(0)}\sum_{h=1}^{30}\gamma^T_{i,j}(h)$, computed empirically for all pairs (i, j) (multiplied by 100 and rounded). The degree of persistence of the global sample autocovariance of product returns is high for the overwhelming majority of asset pairs: indeed, our quantity of interest is less than 5 only for 6 (out of 55) asset pairs: (All Ordinaries, DJIA), (All Ordinaries, NASDAQ), (All Ordinaries, S&P 500), (Nikkei, DJIA), (Nikkei, NASDAQ), (Nikkei, S&P 500). To see how our model can reproduce this persistence, we consider the following proposition.

Proposition 2.1 *Let the number of jumps in $c_{i,j}(u)$ be finite, and let $\mathbb{E}(\varepsilon^8_{i,t}) < \infty$. Let h be fixed. We then have*

$$\gamma^T_{i,j}(h) \to \int_0^1 c^2_{i,j}(u)du - \left(\int_0^1 c_{i,j}(u)du\right)^2 \tag{2}$$

in mean-square, as $T \to \infty$.

The proof is similar to Proposition 1 in Fryzlewicz et al. (2006) and so we omit it. The difference of the integrals on the right-hand side of formula (2) is always positive, unless $c_{i,j}(u)$ is constant, when it is equal to zero. This provides an explanation for the empirically noted persistence of the global sample autocorrelation of product returns: if the underlying process in non-stationary (in the sense that the cross-covariance $c_{i,j}(u)$ varies over time),

Table 1: The quantity $\left[\frac{100}{30\gamma^T_{i,j}(0)}\sum_{h=1}^{30}\gamma^T_{i,j}(h)\right]$ for all pairs (i,j) of assets.

$i \setminus j$	1	2	3	4	5	6	7	8	9	10
1	22	5	12	10	1	12	17	-1	18	1
2		21	15	12	17	16	12	18	9	19
3			21	18	12	21	11	9	12	12
4				20	13	18	9	11	10	14
5					24	13	7	24	0	25
6						22	13	10	13	13
7							20	6	18	7
8								23	-1	24
9									21	1
10										26

Table 2: P-values of our test of constancy of cross-market correlations, multiplied by 100 and rounded.

$i \setminus j$	1	2	3	4	5	6	7	8	9	10
1	92	97	25	2	67	11	0	99	2	69
2		91	0	0	1	0	13	0	31	1
3			77	76	3	38	51	13	14	3
4				100	1	28	19	1	2	0
5					97	10	3	80	85	99
6						90	32	16	22	11
7							84	23	13	8
8								96	52	99
9									92	88
10										100

then the use of the global sample acf of the product returns (a measure of dependence which is unsuitable in this non-stationary situation due to its "non-local" nature) results in the spurious effect of autocorrelation persistence.

Going out step further, recalling that $c_{i,j}(u) = \sigma_i(u)\sigma_j(u)\rho_{i,j}(u)$, it is interesting to investigate whether the postulated time variation in $c_{i,j}(u)$ (exhibiting itself via the persistent global acf of the product returns) can be attributed to time variations in $\sigma_i(u)$ and $\sigma_j(u)$, in $\rho_{i,j}(u)$, or in both. Firstly, noting that $c_{i,i}(u) = \sigma_i^2(u)$, we can see from the diagonal elements of Table 1 that the squared returns $X^2_{i,t,T}$ indeed exhibit a high degree of persistence of their global sample acf, which, in our model, provides evidence for $\sigma_i^2(u)$ (and hence $\sigma_i(u)$) varying over time for each i (by Proposition 2.1).

To investigate the constancy (or otherwise) of $\rho_{i,j}(u)$, we firstly note that $\rho_{i,j}(t/T) = \mathbb{E}(\varepsilon_{i,t}\varepsilon_{j,t})$. As the $\varepsilon_{i,t}$'s are not directly observable, it is sensible to construct their empirical counterparts $\hat{\varepsilon}_{i,t}$ and then test for the constancy of the mean of the vector $\{\hat{\varepsilon}_{i,t}\hat{\varepsilon}_{j,t}\}_{t=1}^T$. We construct $\hat{\varepsilon}_{i,t}$ as

$$\hat{\varepsilon}_{i,t} = \frac{X_{i,t,T}}{\hat{\sigma}_i(t/T)},$$

where $\hat{\sigma}_i^2(t/T)$ is our estimator of $\sigma_i^2(t/T)$ described in Section 3.2 and denoted as $\tilde{c}_{i,i}(t/T)$

therein. We use a simple CUSUM-type test based on the following well-known fact (see e.g. Corollary 29.7 of Davidson (1994)): if $\{Z_t\}_{t=1}^T$ is a sequence of iid variables with $\mathbb{E}|Z_t|^r < \infty$ for some $r > 2$, then, for $u \in [0,1]$,

$$\begin{aligned} W_T(u) &= \mathrm{Var}^{-1/2}(Z_t) T^{-1/2} \sum_{t=1}^{\lceil Tu \rceil} Z_t - \mathbb{E}(Z_t) \quad \xrightarrow{d} \quad B_u \\ U_T(u) &= W_T(u) - \frac{\lceil Tu \rceil}{T} W_T(1) \quad \xrightarrow{d} \quad B_u^o, \end{aligned}$$

where B_u and B_u^o are, respectively, standard Brownian motion and Brownian bridge on $[0,1]$. With $Z_t := \hat{\varepsilon}_{i,t}\hat{\varepsilon}_{j,t}$ as input, we construct the empirical counterpart of $U_T(u)$ by replacing the expectation and variance in $W_T(u)$ with their sample counterparts. We compare the range of the resulting sample path with the theoretical range of Brownian bridge, whose cumulative distribution is well known (Kennedy (1976)) and given by

$$F_{B^o}(x) = 1 + 2\sum_{k=1}^{\infty}(1 - 4k^2x^2)\exp(-2k^2x^2).$$

P-values of this test (multiplied by 100 and rounded), for each pair of assets individually, are given in Table 2. While the multiple hypothesis setting obviously obscures the picture, the low p-values of some of the tests of constancy constitute strong evidence that some of the functions $\rho_{i,j}(u)$ are indeed time-varying, or in other words that some of the cross-asset correlations evolve over time. On the other hand, the high p-values of some of the tests provide evidence that it may not always be the case and that some cross-asset correlations can safely be modelled as constant. One advantage of our adaptive estimation method for $\rho_{i,j}(u)$, proposed in Section 3.1, is that it will automatically and naturally "test" for the constancy of $\rho_{i,j}(u)$ as part of the estimation process (while properly accounting for multiple hypotheses) and estimate $\rho_{i,j}(u)$ as constant with a high probability, if and only if the underlying true function $\rho_{i,j}(u)$ is constant. This is in contrast to, say, kernel-based methods, which always produce estimators which oscillate (however slightly) even if the target function is constant.

As an illustration, the right-hand plot of Figure 2 shows the absolute value of the sample path of $U_T(u)$, constructed for the following pair of markets: DAX and S&P 500 (for which the above p-value is close to zero). The vertical line indicates its maximum, which is also the most likely, according to our CUSUM test, location of a breakpoint in the mean of $\hat{\varepsilon}_{4,t}\hat{\varepsilon}_{10,t}$, if there is any. The left-hand plot shows the corresponding sequence $\hat{\varepsilon}_{4,t}\hat{\varepsilon}_{10,t}$, with the same potential breakpoint. Interestingly, the breakpoint corresponds to a date in August 2008, i.e. just before the recent financial crisis entered its most severe phase. This appears to confirm the common wisdom that markets tend to be more highly correlated to each other during times of crises: note that the sample mean of $\hat{\varepsilon}_{4,t}\hat{\varepsilon}_{10,t}$ before the breakpoint is approximately 0.51, whereas after the breakpoint, it reaches 0.72. It is important for any statistical estimation procedure to react adequately to such changes, and we will later argue that our estimation methodology meets this requirement.

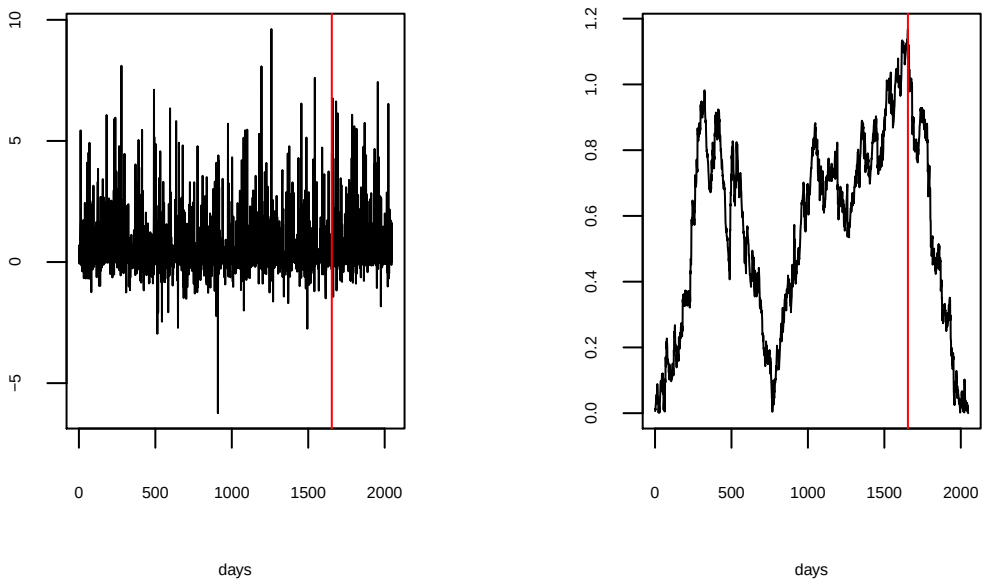

Figure 2: Left: $\hat{\varepsilon}_{4,t}\hat{\varepsilon}_{10,t}$ (corresponding to DAX (4) and S&P 500 (10)); right: the corresponding sample path of $|U_T(u)|$. Vertical line in both plots indicates the location of the maximum of $|U_T(u)|$.

3 Haar-Fisz estimation of the volatility matrix $\Sigma(u)$

3.1 Case $D(u) =$ identity

We start our exposition of the Haar-Fisz estimation methodology for the volatility matrix $\Sigma(u)$ by considering the idealised case where the marginal volatilities $\sigma_i(u)$ are known to be constant and equal to one, i.e. where the matrix $D(u)$ is the identity matrix. This is done firstly for clarity, so that the basic mechanics of the Haar-Fisz estimation procedure for the covariances $c_{i,l}(u)$, $i \neq l$, can be highlighted without the time variation in $\sigma_i(u)$ obscuring the picture. Secondly, from the point of view of interpretation, considerations of (piecewise) constancy of $c_{i,l}(u)$ are more informative when it is correlation rather than covariance: note that the fact that $c_{i,l}(u) = \sigma_i(u)\sigma_l(u)\rho_{i,l}(u)$ is constant, does not imply that any of the functions: $\sigma_i(u)$, $\sigma_l(u)$, $\rho_{i,l}(u)$ are constant, unless we assume the constancy of $\sigma_i(u)$, $\sigma_l(u)$, which is our assumption in this section. Estimation in the general case where $D(u)$ is not necessarily the identity matrix is detailed in Section 3.2.

Although this is an idealistic situation, it can also serve as a model in the case where the marginal volatilities have already been previously removed. This would have proceeded by estimating each marginal volatility using the analyst's preferred technique (e.g. our method, as described in Section 3.2, or a univariate GARCH fit) and computing the empirical residuals from this fit. The empirical residuals could then be modelled as having

marginal volatilities constant and equal to 1.

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 $\{X_{i,t,T}X_{l,t,T}\}_{t=1}^{T}$, recalling that $\sigma_i(u) \equiv \sigma_l(u) \equiv 1$. We note, however, that our theoretical results later in this section concern the quality of the estimation of the entire matrix $\Sigma(u)$ in the operator norm, rather than each component $c_{i,l}(u)$ separately.

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 $\mathbb{E}(\xi_{i,l,t,T}) = 0$. Thus the problem of estimating $c_{i,l}(u)$ can be viewed as the problem of "denoising" the sequence $X_{i,t,T}X_{l,t,T}$. As the target function $c_{i,l}(u)$ is piecewise constant, Haar wavelets are a natural estimation tool here, being themselves piecewise constant (the reader is referred to Vidakovic (1999) for a readable introduction to wavelets in statistics, although we summarise the basics of Haar wavelets below).

To apply Haar wavelet thresholding to estimate $c_{i,l}(u)$, we need to know or easily be able to estimate the standard deviation of the noise $\xi_{i,l,t,T}$. If the innovations $\varepsilon_{i,t}$ were Gaussian, then we would be able to use Isserlis' theorem to write

$$\{\mathrm{Var}(\xi_{i,l,t,T})\}^{1/2} = \{1 + c_{i,l}^2(t/T)\}^{1/2}. \tag{3}$$

On the right-hand side, $c_{i,l}(u)$ is obviously unknown (since this is what we are trying to estimate). One of the key points of our Haar-Fisz estimation methodology is the fact that we can *pre-estimate* it by taking a certain local average of the sequence $X_{i,t,T}X_{l,t,T}$ (more details below). Unfortunately, if $\varepsilon_{i,t}$ is not Gaussian, then similarly simple expressions for the variance of $\xi_{i,l,t,T}$ cannot be guaranteed to exist, which is why we assume Gaussianity of $\varepsilon_{i,t}$ in the remainder of this paper.

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 to two; techniques for adapting wavelet transforms to non-dyadic sample sizes are described e.g. in Wickerhauser (1994). The algorithm is valid for all pairs i, l except $i = l$ since we assume, in this section, that $\mathbb{E}(X_{i,t,T}^2)$ is known and equal to 1. Denote $J = \log_2 T$. The estimation algorithm proceeds as follows:

1. Compute the Haar decomposition of $\{X_{i,t,T}X_{l,t,T}\}_{t=1}^{T}$ using the following algorithm:
   (a) Let $s_{J,k}^{(i,l)} := X_{i,t,T}X_{l,t,T}$, $k = 1, 2, \ldots, 2^J$.
   (b) For each $j = J-1, J-2, \ldots, 0$, recursively form vectors $s_j^{(i,l)}$, $d_j^{(i,l)}$, $\tilde{s}_j^{(i,l)}$, and $f_j^{(i,l)}$ with elements:

$$\begin{aligned}
s_{j,k}^{(i,l)} &= 2^{-1/2}(s_{j+1,2k-1}^{(i,l)} + s_{j+1,2k}^{(i,l)}) \\
d_{j,k}^{(i,l)} &= 2^{-1/2}(s_{j+1,2k-1}^{(i,l)} - s_{j+1,2k}^{(i,l)}) \\
\tilde{s}_{j,k}^{(i,l)} &= 2^{(j-J)/2} s_{j,k}^{(i,l)} \\
f_{j,k}^{(i,l)} &= d_{j,k}^{(i,l)} \left\{ 1 + \left(\tilde{s}_{j,k}^{(i,l)} \right)^2 \right\}^{-1/2},
\end{aligned}$$

   where $k = 1, \ldots, 2^j$.

2. For each j and k, denote $\mu_{j,k}^{(i,l)} := \mathbb{E}(d_{j,k}^{(i,l)})$. For most levels j (in a sense to be made precise later), estimate $\mu_{j,k}^{(i,l)}$ by

$$\hat{\mu}_{j,k}^{(i,l)} = d_{j,k}^{(i,l)} \mathbb{I}(|f_{j,k}^{(i,l)}| > \lambda),$$

where $\mathbb{I}(\cdot)$ is the indicator function. In other words, we "kill" each $d_{j,k}^{(i,l)}$ if and only if the corresponding Haar-Fisz coefficient $f_{j,k}^{(i,l)}$ does not exceed in absolute value a certain threshold λ (to be specified later). Note that this is different from classical wavelet thresholding in that the thresholded quantity $d_{j,k}^{(i,l)}$ and the thresholding statistic $f_{j,k}^{(i,l)}$ are different.

3. 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 $c_{i,l}(t/T)$.

4. The initial estimate $\hat{c}_{i,l}(t/T)$ is a piecewise-constant function of t. 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. With some abuse of terminology, we refer to this step as "bias correction" and denote the bias-corrected estimate by $\tilde{c}_{i,l}(t/T)$.

5. Apply additional thresholding in the time domain, i.e. construct the final estimate by either of the two operations

$$\begin{aligned} \bar{c}_{i,l}^{(h)}(t/T) &= \tilde{c}_{i,l}(t/T)\mathbb{I}(|\tilde{c}_{i,l}(t/T)| > \lambda_1) \quad \text{(hard thresholding)}, \\ \bar{c}_{i,l}^{(s)}(t/T) &= \operatorname{sign}(\tilde{c}_{i,l}(t/T)) \max(|\tilde{c}_{i,l}(t/T)| - \lambda_1, 0) \quad \text{(soft thresholding)}, \end{aligned}$$

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

Some remarks are in order. Note first that the quantity $\tilde{s}_{j,k}^{(i,l)}$ is simply the local sample mean of the sequence $\{X_{i,t,T}X_{l,t,T}\}_{t=1}^{T}$, computed over the interval $t \in [(k-1)2^{J-j} + 1, \ldots, k\,2^{J-j}]$. Thus, if $c_{i,l}(u)$ is constant over the interval $u \in \left(\frac{k-1}{2^J}, \frac{k}{2^J}\right]$, then the quantity $\left\{1 + \left(\tilde{s}_{j,k}^{(i,l)}\right)^2\right\}^{1/2}$ is, by Isserlis' theorem, the maximum likelihood estimator of the standard deviation of $d_{j,k}^{(i,l)}$. Therefore, $f_{j,k}^{(i,l)}$ can be viewed as a variance-stabilised, or studentised, version of the Haar coefficient $d_{j,k}^{(i,l)}$. It is this variance-stabilisation step which permits us to use a threshold λ independent of scale j or location k, which facilitates its choice (recall that in classical wavelet thresholding, the value of the threshold is usually a multiple of the standard deviation of the thresholding statistic).

A similar trick (stabilising the Haar wavelet coefficient by the MLE estimate of its own standard deviation) was proposed in Fryzlewicz et al. (2006) for the estimation of univariate volatilities. However, the application to volatility matrix estimation is new. We term this methodology "Haar-Fisz" in recognition of Marek Fisz who studied, in the 1950s, the asymptotic behaviour of some random variables of the form $(X-Y)/h(X,Y)$, where h is a nonnegative-valued function (which $f_{j,k}^{(i,l)}$ is an instance of), albeit obviously outside of a wavelet context, see e.g. Fisz (1955).

As well as the above variance stabilisation, we also note that it is possible, and straightforward, to show the asymptotic normality of the term $f_{j,k}^{(i,l)}$, as we move away from the finest scales of the Haar decomposition in the sense that $J - j \to \infty$. However, in contrast to the univariate volatility case, it is not easy to identify the exact distribution of $f_{j,k}^{(i,l)}$, unless $i = l$. We return to this theme in Section 3.3 which describes the "polarised" version of our estimator.

Besides the fact that the Haar-Fisz methodology has not been previously applied in a setting of covariance estimation, other methodological novelties of this work, which set it apart from both classical wavelet thresholding and the previously proposed Haar-Fisz techniques, are in steps 4 and 5 of the above algorithm.

In step 4, each piecewise constant section of the estimate is replaced by the local sample mean of the data computed over the same sub-interval. This "bias correction" is essential in establishing interval-wise consistency results (in the operator matrix norm; more details in Theorems 3.1 – 3.4) of our estimator. The reason why such a bias correction is not normally performed in traditional Haar wavelet thresholding estimators is that the usual error norm of choice when analysing wavelet estimators is the mean integrated square error, in which case no bias correction is needed. However, we are more interested in the interval-wise convergence rates, a choice which we motivate in more detail below. Finally, note that the thresholding performed in step 2 of the algorithm is of the "hard" type. Applying the bias correction in step 4 means that we could equally have used "soft" thresholding in step 2 but that would have led to exactly the same values of the bias-corrected estimate $\tilde{c}_{i,l}(t/T)$ as both hard and soft thresholding estimates based on Haar wavelets lead to exactly the same set of breakpoints.

In step 5, extra thresholding is applied in the time domain. This is done to ensure that zero covariances are estimated as exactly zero with high probability. Heuristically speaking, because our true volatility matrix will be assumed to be sparse, estimating zero covariances as exactly zero will help to significantly reduce the overall estimation error, in the operator norm. Here, it is essential to consider both hard and soft thresholding: while hard thresholding produces consistent estimators under slighly less strict assumptions (see Theorems 3.1 – 3.4), soft thresholding tends to produce better practical performance. Roughly speaking, the reason for the latter is that $\bar{\Sigma}^{(s)}(t/T)$ is a continuous function of the threshold λ_1, which means that it is guaranteed to be positive-definite for a certain range of thresholds λ_1. For completeness, we mention that shrinkage- (as opposed to hard-thresholding-) type estimators for stationary covariance matrices were considered e.g. in Haff (1980), Dey and Srinivasan (1985) and Ledoit and Wolf (2003).

Finally, it is worth making a general note that in some nonparametric models, one route of 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.

We first analyse the behaviour of our estimator in the "null" case where the true volatility matrix $\Sigma(u)$ is constant with respect to u, i.e. $\Sigma(u) = \Sigma$. In order to do this, 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 $\lambda_{\max}(M) = \lambda_1(M) \geq \ldots \geq \lambda_p(M) = \lambda_{\min}(M)$. With $\|\mathbf{v}\|_2$ denoting the l_2 norm of a vector $\mathbf{v}$, the operator norm of M is defined as

$$\|M\| = \sup\{\|M\mathbf{v}\|_2 \,:\, \|\mathbf{v}\|_2 = 1\},$$

and for symmetric matrices (e.g. covariance matrices) is given by $\|M\| = \max_{1\le i\le p} |\lambda_i(M)|$. It is well known (see e.g. Golub and Van Loan (1989)) that for symmetric matrices, we have

$$\|M\| \le \max_l \sum_i |m_{i,l}|.$$

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

$$\mathcal{U}(c_0(p)) = \{\Sigma : c_{i,i} = 1, \max_i \sum_{l=1}^p \mathbb{I}(c_{i,l} \neq 0) \le c_0(p)\},$$

and a class of invertible sparse constant volatility matrices as

$$\mathcal{U}(c_0(p), \epsilon_0) = \{\Sigma : \Sigma \in \mathcal{U}(c_0(p)),\ \lambda_{\min}(\Sigma) \ge \epsilon_0 > 0\}.$$

The results follow.

Theorem 3.1 *(Hard thresholding.) Let our hard-thresholding estimator $\bar{\Sigma}^{(h)}(u)$ of $\Sigma(u)$ be constructed as in the algorithm of Section 3.1, where $\hat{\mu}_{j,k}^{(i,l)} = d_{j,k}^{(i,l)} \mathbb{I}(|f_{j,k}^{(i,l)}| > \lambda)$ 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. 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*

$$\lambda \ge \sqrt{2\left(2\log\, p + (1-\delta)\log\, T + \log\frac{1}{a_{p,T}}\right)},$$

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

for some $C > 0$, where $a_{p,T}$ is a certain sequence, tending 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 $\bar{c}_{i,l}^{(h)}(t/T) \equiv \frac{1}{T}\sum_{t=1}^T X_{i,t,T} X_{l,t,T}$ if $c_{i,l} \neq 0$.

(b) We have $\|\bar{\Sigma}^{(h)} - \Sigma\| \le 2c_0(p)\sqrt{\frac{\log\, p + \log\, c_0(p) + \log\frac{1}{a_{p,T}}}{T}}$.

(c) If, in addition, $\Sigma \in \mathcal{U}(c_0(p), \epsilon_0)$, then $\|(\bar{\Sigma}^{(h)})^{-1} - \Sigma^{-1}\| \le C_2 c_0(p)\sqrt{\frac{\log\, p + \log\, c_0(p) + \log\frac{1}{a_{p,T}}}{T}}$ for some positive C_2.

Theorem 3.2 *(Soft thresholding.) Assume the set-up of Theorem 3.1. The following holds with probability of at least $1 - C_1\, a_{p,T}$ for some positive C_1:*

(a) Our estimator $\bar{\Sigma}^{(s)}(u) = \bar{\Sigma}^{(s)}$ is constant and such that $\bar{c}_{i,l}^{(s)}(t/T) \equiv 0$ if $c_{i,l} = 0$ and $\bar{c}_{i,l}^{(s)}(t/T) \equiv \text{sign}(\bar{c}_{i,l}^{(h)}(t/T))(|\bar{c}_{i,l}^{(h)}(t/T)| - \lambda_1)$ if $c_{i,l} \neq 0$.

(b) We have $\|\bar{\Sigma}^{(s)} - \Sigma\| \leq c_0(p)\left(2\sqrt{\frac{\log p + \log c_0(p) + \log \frac{1}{a_{p,T}}}{T}} + \lambda_1\right)$.

(c) If, in addition, $\Sigma \in \mathcal{U}(c_0(p), \epsilon_0)$, *then*

$$\|(\bar{\Sigma}^{(s)})^{-1} - \Sigma^{-1}\| \leq C_2 c_0(p)\left(2\sqrt{\frac{\log p + \log c_0(p) + \log \frac{1}{a_{p,T}}}{T}} + \lambda_1\right)$$

for some positive C_2.

Finally, if $\lambda_1 = O\left(T^{-1/2}(\log p + \log \frac{1}{a_{p,T}})^{1/2}\right)$, *then the above convergence rates for* $\bar{\Sigma}^{(s)}$ *and* $(\bar{\Sigma}^{(s)})^{-1}$ *are the same as those for* $\bar{\Sigma}^{(h)}$ *and* $(\bar{\Sigma}^{(h)})^{-1}$.

We now quote the results in the case when $\Sigma(u)$ is not necessarily constant. In the following theorem, a dyadic interval is defined as any interval of the form $(\{k-1\}2^{-j}, k2^{-j}]$ where $j = 0, \ldots, J-1$ and $k = 1, \ldots, 2^j$.

Theorem 3.3 *(Hard thresholding.) Let our hard-thresholding estimator* $\bar{\Sigma}^{(h)}(u)$ *of* $\Sigma(u)$ *be constructed as in the algorithm of Section 3.1, where* $\hat{\mu}_{j,k}^{(i,l)} = d_{j,k}^{(i,l)}\mathbb{I}(|f_{j,k}^{(i,l)}| > \lambda)$ *for scales* $j = 0, \ldots J^*$ *with* $2^{J^*} = T^{1-\delta}$ *for some* $\delta \in (0,1)$, *and* $\hat{\mu}_{j,k}^{(i,l)} = 0$ *otherwise. 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, l, *the function* $c_{i,l}(u)$ *is constant for all* $u \in \mathcal{I}$.

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

$$\min_{i,l,m} \int_{\mathcal{J}_{i,l}^m} \left\{ c_{i,l}(u) - |\mathcal{J}_{i,l}^m| \int_{\mathcal{J}_{i,l}^m} c_{i,l}(z)dz \right\}^2 du \geq C_3 T^{-\beta} \tag{4}$$

for $\beta \in [0, 1-\delta)$. *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*

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

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

for some $C > 0$, *where* $a_{p,T}$ *is a certain sequence, tending 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 $\|\bar{\Sigma}^{(h)}(\mathcal{I}) - \Sigma(\mathcal{I})\| \leq 2c_0(p)\sqrt{\frac{\log p + \log c_0(p) + \log \frac{1}{a_{p,T}}}{T|\mathcal{I}|}}$*.*

(c) If, in addition, $\Sigma(\mathcal{I}) \in \mathcal{U}(c_0(p), \epsilon_0)$*, then*

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

for some positive C_2*.*

Theorem 3.4 *(Soft thresholding.) Assume the set-up of Theorem 3.3. The following holds with probability of at least* $1 - C_1\, a_{p,T}$ *for some positive* C_1*:*

(a) Our estimator $\bar{\Sigma}^{(s)}(u)$ *is constant for* $u \in \mathcal{I}$ *and such that* $\bar{c}_{i,l}^{(s)}(\mathcal{I}) = 0$ *if* $c_{i,l}(\mathcal{I}) = 0$ *and* $\bar{c}_{i,l}^{(s)}(\mathcal{I}) \equiv \text{sign}(\bar{c}_{i,l}^{(h)}(\mathcal{I}))(|\bar{c}_{i,l}^{(h)}(\mathcal{I})| - \lambda_1)$ *if* $c_{i,l}(\mathcal{I}) \neq 0$*.*

(b) We have $\|\bar{\Sigma}^{(s)}(\mathcal{I}) - \Sigma(\mathcal{I})\| \leq c_0(p)\left(2\sqrt{\frac{\log p + \log c_0(p) + \log \frac{1}{a_{p,T}}}{T|\mathcal{I}|}} + \lambda_1\right)$*.*

(c) If, in addition, $\Sigma(\mathcal{I}) \in \mathcal{U}(c_0(p), \epsilon_0)$*, then*

$$\|(\bar{\Sigma}^{(s)}(\mathcal{I}))^{-1} - \Sigma(\mathcal{I})^{-1}\| \leq C_2 c_0(p)\left(2\sqrt{\frac{\log p + \log c_0(p) + \log \frac{1}{a_{p,T}}}{T|\mathcal{I}|}} + \lambda_1\right)$$

for some positive C_2*.*

Finally, if $\lambda_1 = O\left(T^{-1/2}|\mathcal{I}|^{-1/2}(\log p + \log \frac{1}{a_{p,T}})^{1/2}\right)$*, then the above convergence rates for* $\bar{\Sigma}^{(s)}(u)$ *and* $(\bar{\Sigma}^{(s)}(u))^{-1}$ *are the same as those for* $\bar{\Sigma}^{(h)}(u)$ *and* $(\bar{\Sigma}^{(h)}(u))^{-1}$*.*

We first note that Theorems 3.1 and 3.2 are special cases of Theorems 3.3 and 3.4, respectively, with $\mathcal{I} = (0, 1]$. We chose to present them separately to emphasise the "modularity" of their proof structure, in the sense that one essential part of the proofs of the latter two theorems is a localised, interval-wise application of techniques from the proofs of the former two theorems. This elegant localisation can be achieved thanks to the multiscale structure of our Haar-wavelet-based estimator.

The most commonly used measure of quality of wavelet-based function estimators in literature is the mean integrated square error (MISE). The reason for this is partly technical: wavelet provide an orthonormal series expansion, so measuring MISE in the time domain is equivalent to measuring MISE in the wavelet domain, the latter often being theoretically tractable. However, in the above Theorems 3.3 and 3.4, we show instead localised, interval-wise convergence of our estimators (which is stronger than and implies pointwise

convergence). The reason for this is that the practitioner is likely to be interested in the local, pointwise quality of the volatility estimator at the "current" time $u = 1$ ($t = T$). Note that MISE convergence would not be able to guarantee pointwise convergence at any particular point u.

We now comment on the various quantities appearing in the theorems. The parameter δ is required to be less than 1 for various important technical reasons, including guaranteeing uniform strong asymptotic normality of quantities such as $d_{j,k}^{(i,l)}$ and $s_{j,k}^{(i,l)}$, which is a core part of the proof. The lower its value, the less strict the assumptions (i) and (ii) of Theorems 3.3 and 3.4 (i.e. the larger the class of volatilities $\Sigma(u)$ for which our method is applicable), but the (potentially) worse the error bounds in statements (b) and (c); thus δ is a parameter describing balance between applicability and performance. Note that 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 achieve consistency of the estimators of the volatility matrix and its inverse; however, the only assumption about $\mathcal{I}$ is that $|\mathcal{I}| \geq T^{\delta-1}$ with δ being possibly arbitrarily close to zero.

We have already argued before that it is the application of the variance-stabilising "Fisz transform" (i.e. the division of $d_{j,k}^{(i,l)}$ by an estimate of its own standard deviation to obtain $f_{j,k}^{(i,l)}$) that allows the threshold λ to be independent of $c_{i,l}(u)$. Heuristically speaking, the reason why λ_1 is also independent of $c_{i,l}(u)$ is that it is calibrated under the hypothesis that, locally, $c_{i,l}(u) = 0$. Theorems 3.3 and 3.4 suggest a permitted range of the parameter λ, and a valid practical choice could be made e.g. by setting λ to be the lower bound of its permitted range. As for the selection of λ_1, the parameter $|\mathcal{I}|$ is obviously unknown, but again recalling that $|\mathcal{I}| \geq T^{\delta-1}$, we could set $\lambda_1 = \sqrt{2T^{-\delta}(2\log p + \log \frac{1}{a_{p,T}})}$ under certain additional assumptions on $\min_{c_{i,l}(\mathcal{I})\neq 0} |c_{i,l}(\mathcal{I})|$. In Section 4, we describe ways of selecting λ, λ_1 and δ in practice. Although there is no theoretical advantage in choosing λ_1 to be large from the point of view of the error bounds of our soft thresholding estimator, we will demonstrate in that section that it might be advantageous to do so from the point of view of their practical performance.

The variance-type condition (4) specifies how large, or how isolated, the nearest breakpoint needs to be before our estimator reacts to it. Parameter $a_{p,T}$ determines the probability with which the results of the theorem hold: the higher the desired probability, the worse the error bounds. As in the stationary set-up of Bickel and Levina (2008), the magnitude of the error bounds specifies how fast the sparsity parameter $c_0(p)$ is permitted to grow before consistency breaks down.

Finally, we note that the convergence rates in Theorems 3.3 and 3.4 are adaptive to the length of the interval $\mathcal{I}$. Since the results hold for any interval of constancy satisfying the assumptions, they also hold for the largest such interval. Thus the estimators exhibit interval-wise adaptation to the features of the target matrix. In particular, if the breakpoints are fixed in rescaled time, then the convergence rates are near-parametric, subject of course to $c_0(p)$ being small. However, our results demonstrate that consistent interval-wise estimation is still possible in set-ups where breakpoints approach each other in rescaled time.

3.2 The general case

In this section, we describe our estimation algorithm for the volatility matrix $\Sigma(u)$ in the general case where we have no knowledge of the marginal volatilities $\sigma_i^2(u) = c_{i,i}(u)$ and where they also need to be estimated. As before, we decompose

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

where, by Isserlis' theorem,

$$\{\mathrm{Var}(\xi_{i,l,t,T})\}^{1/2} = \{c_{i,i}(t/T)c_{l,l}(t/T) + c_{i,l}^2(t/T)\}^{1/2}. \tag{5}$$

The difference in comparison to the case $c_{i,i}(u) = \sigma_i^2(u) \equiv 1$ is that $c_{i,i}(u)$ and $c_{l,l}(u)$ appear non-trivially in the standard deviation formula (5). The form of the standard deviation leads to the following algorithm.

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 $\tilde{s}_{j,k}^{(\eta,\upsilon)}$.

2. Obtain the variance-stabilised coefficients

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

3. Estimate $\mu_{j,k}^{(i,l)}$ by $\hat{\mu}_{j,k}^{(i,l)} = d_{j,k}^{(i,l)} \mathbb{I}(|f_{j,k}^{(i,l)}| > \lambda)$ for scales $j = 0, \ldots J^*$ with $2^{J^*} = T^{1-\delta}$ for some $\delta \in (0,1)$, and $\hat{\mu}_{j,k}^{(i,l)} = 0$ otherwise.

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 $c_{i,l}(t/T)$.

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 $\tilde{c}_{i,l}(t/T)$.

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

$$\begin{aligned}
\bar{c}_{i,l}^{(h)}(t/T) &= \tilde{c}_{i,l}(t/T)\mathbb{I}(|\tilde{c}_{i,l}(t/T)| > \lambda_1 \tilde{c}_{i,i}^{1/2}(t/T)\tilde{c}_{l,l}^{1/2}(t/T)) \quad \text{(hard thresholding)}, \\
\bar{c}_{i,l}^{(s)}(t/T) &= \mathrm{sign}(\tilde{c}_{i,l}(t/T)) \max(|\tilde{c}_{i,l}(t/T)| - \lambda_1 \tilde{c}_{i,i}^{1/2}(t/T)\tilde{c}_{l,l}^{1/2}(t/T), 0) \quad \text{(soft thresholding)},
\end{aligned}$$

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

Interval-wise consistency results for $\bar{\Sigma}^{(h)}(t/T)$ and $\bar{\Sigma}^{(s)}(t/T)$, similar to those in Theorems 3.1 – 3.4, are also possible to obtain in this case, using a similar set of tools and techniques; we skip them for brevity. Apart from the different form of $f_{j,k}^{(i,l)}$ to reflect the standard deviation formula (5), another difference between the above algorithm and the algorithm of Section 3.1 is the form of the time-domain threshold in step 6. Here, the threshold takes

the form $\lambda_1 \tilde{c}_{i,i}^{1/2}(t/T)\tilde{c}_{l,l}^{1/2}(t/T)$, rather than simply λ_1. This is again done to correct for the different scale of variability of the estimator $\tilde{c}_{i,l}(t/T)$, arising from the fact that the marginal volatilities $c_{i,i}(t/T)$ and $c_{l,l}(t/T)$ are not necessarily equal to one.

In the special case when $i = l$, the above algorithm enables the estimation of each marginal volatility $\sigma_i^2(t/T) = c_{i,i}(t/T)$, or in other words the local variance of each $X_{i,t,T}$. Note that our initial estimator of this quantity before the bias correction, denoted by $\hat{c}_{i,i}(t/T)$ in the above algorithm, is a special case of the estimator proposed in Fryzlewicz et al. (2006) (the latter work did not deal at all with the multivariate case).

3.3 Alternative approach via polarisation identity

In this section, we propose an alternative to the initial estimator $\hat{c}_{i,l}(t/T)$, based on the polarisation identity

$$X_{i,t,T}X_{l,t,T} = \frac{1}{4}\left\{(X_{i,t,T} + X_{l,t,T})^2 - (X_{i,t,T} - X_{l,t,T})^2\right\}.$$

To start with, 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 polarisation identity, the following *in*equality will hold

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

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

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

Our motivation for this proposal is as follows. Firstly note that both $X_{i,t,T} + X_{l,t,T}$ and $X_{i,t,T} - X_{l,t,T}$ follow the multiplicative models

$$\begin{aligned}X_{i,t,T} + X_{l,t,T} &= \sigma_{i,l}^+(t/T)\varepsilon_{i,l,t}^+ \\ X_{i,t,T} - X_{l,t,T} &= \sigma_{i,l}^-(t/T)\varepsilon_{i,l,t}^-,\end{aligned}$$

where the functions $\sigma_{i,l}^{\pm}(t/T)$ are piecewise constant, $\varepsilon_{i,l,t}^{\pm}$ are i.i.d. Gaussian, and $\sigma_{i,l}^{\pm\,2}(t/T) = \sigma_i^2(t/T)+\sigma_l^2(t/T)\pm 2c_{i,l}(t/T)$. Thus, to estimate $\sigma_{i,l}^{\pm\,2}(t/T)$ (and therefore compute $\hat{c}_{i,l}^P(t/T)$), we can use the algorithm of Section 3.2 with $\{(X_{i,t,T} \pm X_{l,t,T})^2\}_{t=1}^T$ as input. The advantage of proceeding in this way is that in this case, 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 $f_{j,k}^{(i,l,\pm)}$ here to differentiate them from $f_{j,k}^{(i,l)}$) under the null hypothesis of the local constancy of $\sigma_{i,l}^{\pm\,2}(t/T)$ over the corresponding sub-interval, which leads to a more accurate, non-asymptotic selection of the threshold λ. To see this, first note that $f_{j,k}^{(i,l,\pm)}$ simplify to

$$f_{j,k}^{(i,l,\pm)} = \frac{d_{j,k}^{(i,l,\pm)}}{2^{\frac{1+j-J}{2}} s_{j,k}^{(i,l,\pm)}} = 2^{\frac{J-j-1}{2}} \frac{\sum_{t=(k-1)2^{J-j}+1}^{(k-1/2)2^{J-j}} \sigma_{i,l}^{\pm\,2}(t/T)\varepsilon_{i,l,t}^{\pm\,2} - \sum_{t=(k-1/2)2^{J-j}+1}^{k\,2^{J-j}} \sigma_{i,l}^{\pm\,2}(t/T)\varepsilon_{i,l,t}^{\pm\,2}}{\sum_{t=(k-1)2^{J-j}+1}^{(k-1/2)2^{J-j}} \sigma_{i,l}^{\pm\,2}(t/T)\varepsilon_{i,l,t}^{\pm\,2} + \sum_{t=(k-1/2)2^{J-j}+1}^{k\,2^{J-j}} \sigma_{i,l}^{\pm\,2}(t/T)\varepsilon_{i,l,t}^{\pm\,2}},$$

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

$$2^{\frac{j+1-J}{2}} f_{j,k}^{(i,l,\pm)} = \frac{\sum_{t=(k-1/2)2^{J-j}+1}^{(k-1/2)2^{J-j}} \varepsilon_{i,l,t}^{\pm 2} - \sum_{t=(k-1/2)2^{J-j}+1}^{k\,2^{J-j}} \varepsilon_{i,l,t}^{\pm 2}}{\sum_{t=(k-1/2)2^{J-j}+1}^{(k-1/2)2^{J-j}} \varepsilon_{i,l,t}^{\pm 2} + \sum_{t=(k-1/2)2^{J-j}+1}^{k\,2^{J-j}} \varepsilon_{i,l,t}^{\pm 2}}.$$

However, note that by Lemma 1 of Fryzlewicz et al. (2006), $2^{\frac{j+1-J}{2}} f_{j,k}^{(i,l,\pm)}$ in the above 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 is in contrast to the results of Theorems 3.1 – 3.4 where the choice of λ is based on strong asymptotic normality arguments. We emphasise that the distribution of the Haar-Fisz coefficients is only readily available in the case of the polarised estimator $\hat{c}_{i,l}^P(t/T)$; indeed, it is not clear how to obtain the exact distribution of $f_{j,k}^{(i,l)}$ (i.e. the Haar-Fisz coefficients in the computation of the non-polarised estimator $\hat{c}_{i,l}(t/T)$) in the case $i \neq l$.

As an example of how the knowledge of the distribution of $f_{j,k}^{(i,l,\pm)}$ 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 again the case where the true volatility is constant, i.e. $\Sigma(u) = \Sigma$. To ensure that our initial polarised estimator $\hat{\Sigma}^P(u) = \{\hat{c}_{i,l}^P(t/T)\}_{i,l=1}^p$ is also constant with probability no less than $1 - a_{p,T}$, it is sufficient to require that

$$\mathbb{P}\left(\bigcup_{i,l} \bigcup_{j,k} \bigcup_{s\in\{+,-\}} |f_{j,k}^{(i,l,s)}| \geq \tilde{\lambda}_j \right) \leq a_{p,T}.$$

Setting $\mathbb{P}(|f_{j,k}^{(i,l,s)}| \geq \tilde{\lambda}_j)$ to be independent of j and using the Bonferroni inequality, the above is implied by

$$2p^2 T^{1-\delta} \mathbb{P}(|2^{\frac{j+1-J}{2}} f_{j,k}^{(i,l,s)}| \geq 2^{\frac{j+1-J}{2}} \tilde{\lambda}_j) = a_{p,T},$$

which can easily be solved numerically for each j separately using the quantiles of the relevant beta distribution. The final estimators, using the thus-constructed initial estimator $\hat{\Sigma}^P(u)$, are denoted by $\bar{\Sigma}^{P,(h)}(u)$ and $\bar{\Sigma}^{P,(s)}(u)$.

4 Practical parameter choice using the example of a stock index portfolio

In this section, we discuss the practical choice of the parameters δ, $a_{p,T}$, λ (for the non-polarised estimators $\bar{\Sigma}^{(h)}(u)$ and $\bar{\Sigma}^{(s)}(u)$), $\tilde{\lambda}_j$ (for the polarised estimators $\bar{\Sigma}^{P,(h)}(u)$ and $\bar{\Sigma}^{P,(s)}(u)$) and λ_1, using the example of the volatility matrix of the 10 major stock indices introduced earlier in the paper.

Parameter δ determines how many finest scales in the Haar wavelet decomposition of the data are ignored in the estimation procedure, and thus determines the length l_c of the shortest permitted intervals of constancy of the volatility matrix estimator. From the point of view of a large (and hence: slow) trader, it is probably impractical to set l_c to be less than 8 days. We use δ leading to $l_c = 8$ in the example below for yet another reason: with this choice of of l_c, the smallest "effective sample size" of our volatility matrix estimators is 8,

but the size of the matrix is $p = 10 > 8$. This means, in particular, that we certainly cannot use the non-time-thresholded estimator $\tilde{\Sigma}(t/T) = \{\tilde{c}_{i,l}(t/T)\}_{i,l=1}^{p}$ as our final estimator of the volatility matrix, as it will not be invertible when the effective sample size is 8. This emphasises the importance of the extra time-domain thresholding, which leads to invertible estimators. In other words, in this setting, $p = 10$ already counts as "large" as it can become larger than the effective sample size, and special techniques (in our case: time-domain thresholding) are required to estimate the volatility matrix effectively in this set-up.

The parameter $a_{p,T}$ determines the rate of the probability with which our results hold converging to one. In the classical univariate wavelet thresholding in the function + Gaussian i.i.d. noise set-up with the universal threshold, an analogue of $a_{p,T}$ is of order $O(\log^{-1/2} T)$, and we also set $a_{p,T} = \log^{-1/2} T$ for simplicity.

In our numerical study, we were reassured to find that the lower end of the permitted theoretical range for λ, that is the choice

$$\lambda = \sqrt{2\left(2\log\, p + (1-\delta)\log\, T + \log\frac{1}{a_{p,T}}\right)}$$

led to estimators which were visually convincing, and, in particular, tended to produce, in estimating the marginal volatilities $c_{i,i}(u)$, empirical residuals with no significant serial dependence in the squares. As for the polarised estimators, our thresholds $\tilde{\lambda}_j$ were computed exactly as specified in Section 3.3 and were also satisfactory in terms of their fit to the data, while leading to more "wiggly" estimates than the non-polarised ones due to the fact that $\tilde{\lambda}_j$'s were computed based on the exact beta quantiles and were therefore less conservative than the threshold λ, the latter stemming from approximate normality arguments.

As an example, Figure 3 shows our estimators of the covariance between $X_{1,t,T}$ (All Ordinaries) and $X_{3,t,T}$ (CAC 40), on 512 trading days (approximately 2 years) ending 10 March 2010, with an arbitrarily chosen value of $\lambda_1 = 0.2$ (more on the choice of λ_1 below). It is interesting to observe that the time-thresholding leads to the value of the estimators being zero towards the end of the time interval for the non-polarised estimators, but not for the polarised ones. The polarised estimators are more wiggly than the non-polarised ones, and their shortest intervals of constancy are of length $l_c = 8$. Of course, the practitioner will be the most interested in the value of the estimator at the right-hand end of the data, that is at the current time $t = T$, where our estimators admit one of two values: (a) the local mean of the last portion of the data computed over an interval whose length is chosen adaptively from the data (shrunk towards zero if soft thresholding is used), or (b) zero; whether (a) or (b) applies depends on the result of the time-domain thresholding.

Selection of λ_1 is a less straightforward issue and must be done with care. The theoretical result of Theorem 3.3 would be able to provide a recipe for the choice of λ_1 if we knew the value of $|\mathcal{I}|$, which is obviously unknown. However, it could in principle be pre-estimated from the intermediate estimator $\tilde{\Sigma}(u)$. Another option, setting $\lambda_1 = \sqrt{2T^{-\delta}(2\log\, p + \log\frac{1}{a_{p,T}})}$, was already discussed in Section 3.1, but would lead to non-adaptive consistency rates in Theorem 3.4. In this numerical study, we demonstrate an alternative "practical" way of selecting λ_1, which will be of interest to portfolio construction practitioners. The algorithm proceeds as follows:

1. Given a realisation of the process $\mathbf{X}_{t,T}$, $t = 1, \ldots, T$, select the length n of a moving

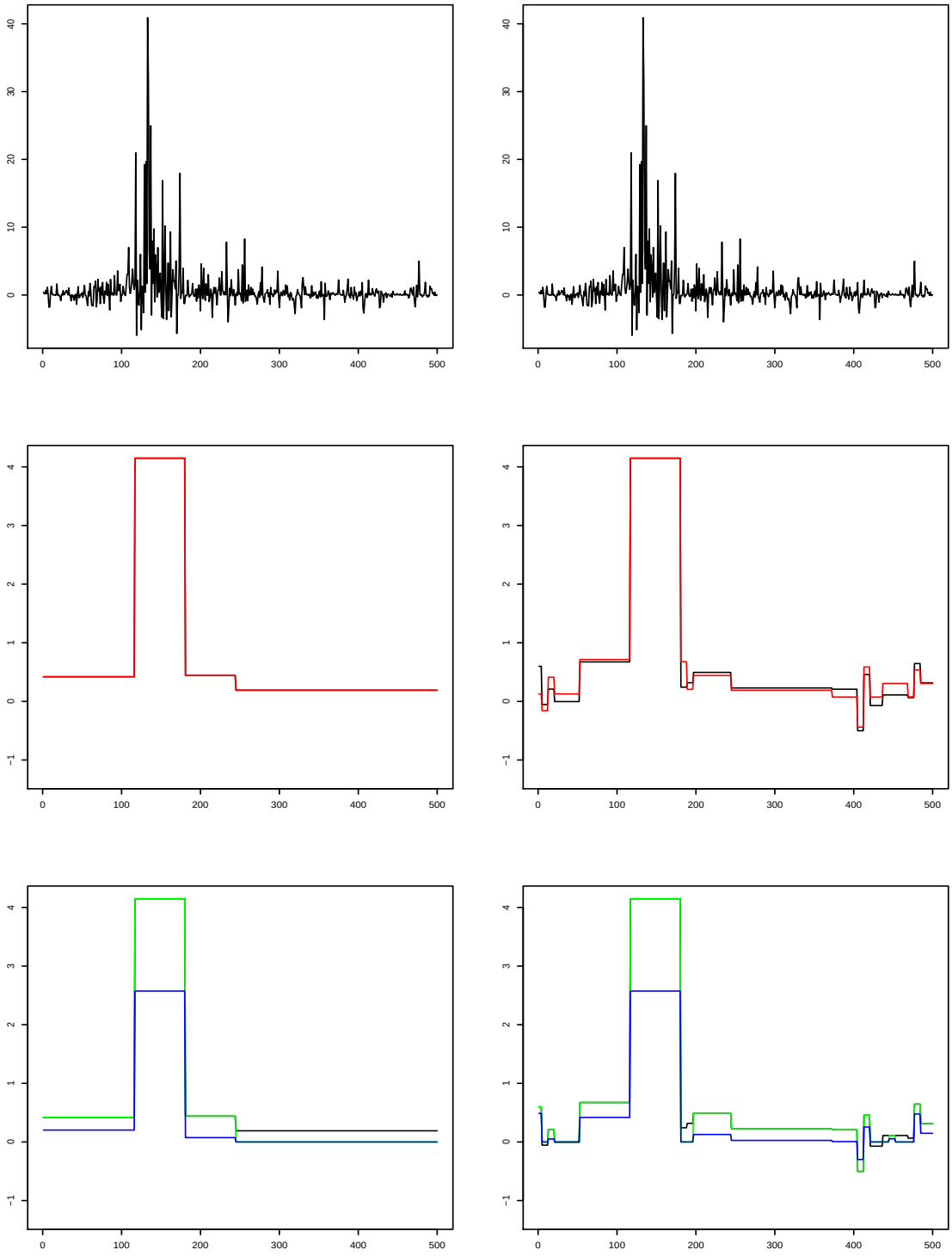

Figure 3: Top row: $\{X_{1,t,T}X_{3,t,T}\}_t$; middle left: $\tilde{c}_{1,3}(t/T)$ (black), $\hat{c}_{1,3}(t/T)$ (red); middle right: $\tilde{c}^P_{1,3}(t/T)$ (black), $\hat{c}^P_{1,3}(t/T)$ (red); bottom left: $\tilde{c}_{1,3}(t/T)$ (black), $\bar{c}^{(h)}_{1,3}(t/T)$ (green), $\bar{c}^{(s)}_{1,3}(t/T)$ (blue); bottom right: $\tilde{c}^P_{1,3}(t/T)$ (black), $\bar{c}^{P,(h)}_{1,3}(t/T)$ (green), $\bar{c}^{P,(s)}_{1,3}(t/T)$ (blue).

window. In our example below, $T = 2048$ and $n = 1024$.

2. For each sub-sequence $\{\mathbf{X}_{t,T}\}_{t=k}^{k+n-1}$, with $k = 1, \ldots, T - n + 1$, compute its bias-corrected volatility matrix estimate $\tilde{\Sigma}_k(t/n)$ (for clarity, we skip the superscript ^{P} for the polarised version) and record its value at the "last" time point $t = n$, denoting it by $\tilde{\Sigma}_k$.

3. Normalise $\tilde{\Sigma}_k$ so that it becomes a correlation matrix, denoted by $\tilde{\tilde{\Sigma}}_k$.

4. Threshold the off-diagonal elements of $\tilde{\tilde{\Sigma}}_k$ by λ_1, denoting the result by $\tilde{\tilde{\Sigma}}_k^{(s)}(\lambda_1)$ ($\tilde{\tilde{\Sigma}}_k^{(h)}(\lambda_1)$) if soft (hard) thresholding is used. We shall use the soft thresholding example from now on.

5. Denote by w_k^i the proportion of funds, allocated to each of the assets $i = 1, \ldots, p$ at time $k = n, \ldots, T$. After removing the effect of the marginal volatilities $c_{i,i}(u)$ and setting the expected return of each asset to zero, Markowitz portfolio theory states that the optimal weights w_k^i should be proportional to the inverse of the correlation matrix at time k, acting on a column vector of ones. We estimate this by $\hat{w}_k^i = (\tilde{\tilde{\Sigma}}_k^{(s)}(\lambda_1))^{-1}\mathbf{1}$, where $\mathbf{1}$ is a column vector of ones.

6. For each k, inspect the estimated weights $\hat{w}_k^i$ and investigate whether they satisfy the desired practical constraints. For example, it may be desirable to the analyst to avoid large relative positions on any particular market, in which case the analyst may wish to monitor the quantity $W_k = \max_i |\hat{w}_k^i / \sum_{j=1}^p \hat{w}_k^j|$. Note that due to the continuity of the soft thresholding operation, W_k converges to $1/p$ as $\lambda_1 \to 1$, i.e. as the correlation matrix becomes closer and closer to the diagonal, the allocation of assets becomed closer and closer to uniform. It may be of interest to e.g. select the lowest value of λ_1 for which W_k does not exceed a fixed multiple of $1/p$. This is a type of "exposure constraint", a different form of which was also discussed e.g. in Fan et al. (2008b).

In the following example, k runs over the 1025 trading days up to 10 March 2010. The top row in Figure 4 shows the sequence W_k for the non-polarised estimator (left, $\lambda_1 = 0.5$) and for the polarised one (right, $\lambda_1 = 0.7$). In either case, we chose the lowest value of λ_1 (in multiples of 0.1) such that W_k stays under 0.2 for all k. Note that since $p = 10$, if assets were allocated uniformly, each weight would equal 0.1. Thus, in this case, we make sure that the maxiumum funds allocation for each asset does not exceed twice the amount for the uniformly allocated portfolio.

The middle row shows the minimum eigenvalue of $\tilde{\tilde{\Sigma}}_k^{(s)}(\lambda_1)$ for the non-polarised estimator (left) and the polarised one (right). Note that this quantity would equal one if all cross-asset correlations were zero.

Finally, the bottom row shows the proportion of zeros in each matrix $\tilde{\tilde{\Sigma}}_k^{(s)}(\lambda_1)$. Roughly speaking, the period between $k = 600$ and $k = 800$ corresponds to the most severe phase of the recent financial crisis. It is interesting to note the decline in the proportion of zeros over that period (which is more evident for the non-polarised estimator than for the polarised one). This again serves as evidence for the common wisdom that markets tend to become more correlated to each other in times of crises.

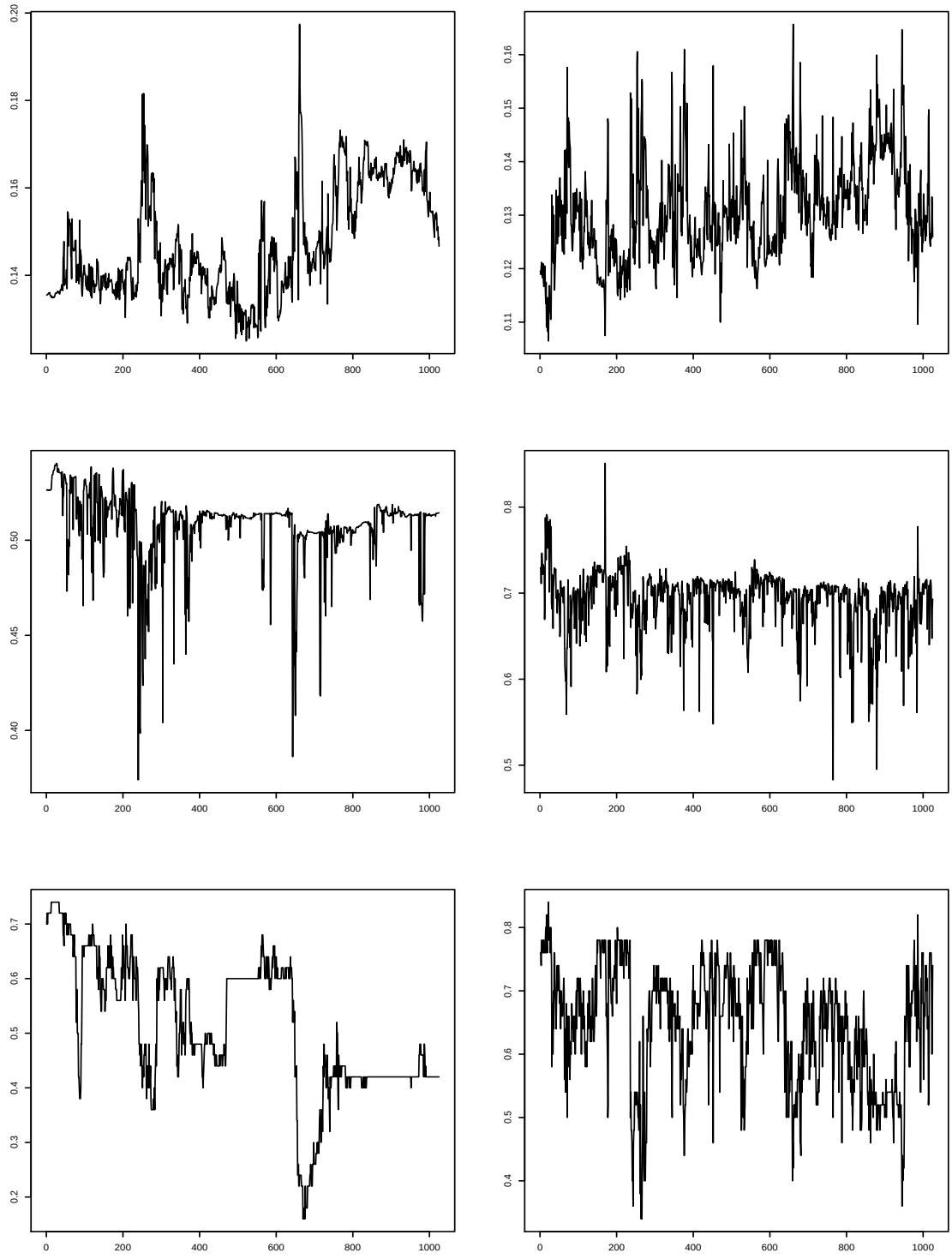

Figure 4: Top row: W_k for the non-polarised estimator (left) and polarised (right). Middle row: minimum eigenvalues of $\tilde{\tilde{\Sigma}}_k^{(s)}(\lambda_1)$ for the non-polarised estimator (left) and polarised (right). Bottom row: proportions of zeros in matrix $\tilde{\tilde{\Sigma}}_k^{(s)}(\lambda_1)$ for the non-polarised estimator (left) and polarised (right).

A Proofs

Proof of Theorem 3.1. 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

$$\mathbb{P}\left(\bigcup_{j=0}^{J^*}\bigcup_{k=1}^{2^j}\{|f_{j,k}^{(i,l)}|>\lambda\}\right) \leq \sum_{j=0}^{J^*} 2^j \mathbb{P}\left(|f_{j,k}^{(i,l)}|>\lambda\right) \leq \max_j \mathbb{P}\left(|f_{j,k}^{(i,l)}|>\lambda\right) CT^{1-\delta}, \quad (6)$$

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 $\lambda \leq C\sqrt{\log T}$. Of course the same bound will be also valid for higher values of λ. Assessing first the probability term, we have

$$\begin{aligned}\mathbb{P}\left(|f_{j,k}^{(i,l)}|>\lambda\right) &= \mathbb{P}\left(|d_{j,k}^{(i,l)}|>\lambda\sqrt{1+(\tilde{s}_{j,k}^{(i,l)})^2} \;\middle|\; |\tilde{s}_{j,k}^{(i,l)}-c_{i,l}|<\tilde{\delta}_j\right)\mathbb{P}(|\tilde{s}_{j,k}^{(i,l)}-c_{i,l}|<\tilde{\delta}_j)\\ &+ \mathbb{P}\left(|d_{j,k}^{(i,l)}|>\lambda\sqrt{1+(\tilde{s}_{j,k}^{(i,l)})^2} \;\middle|\; |\tilde{s}_{j,k}^{(i,l)}-c_{i,l}|\geq\tilde{\delta}_j\right)\mathbb{P}(|\tilde{s}_{j,k}^{(i,l)}-c_{i,l}|\geq\tilde{\delta}_j).\end{aligned}$$

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

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

Using this, we bound the above by

$$\mathbb{P}\left(|d_{j,k}^{(i,l)}|>\lambda\sqrt{1+c_{i,l}^2-2|c_{i,l}|\tilde{\delta}_j}\right)+\mathbb{P}(|\tilde{s}_{j,k}^{(i,l)}-c_{i,l}|\geq\tilde{\delta}_j) =: I+II.$$

Starting with I, we have

$$I=\mathbb{P}\left(\frac{|d_{j,k}^{(i,l)}|}{\sqrt{1+c_{i,l}^2}}>\lambda\sqrt{\frac{1+c_{i,l}^2-2|c_{i,l}|\tilde{\delta}_j}{1+c_{i,l}^2}}\right). \quad (7)$$

Since $\varepsilon_{i,t}$ are Gaussian, there exist $K>0$, $\gamma\geq 0$ such that

$$\mathbb{E}\left(\left|\frac{\varepsilon_{i,t}\varepsilon_{l,t}-c_{i,l}}{\sqrt{1+c_{i,l}^2}}\right|^n\right) \leq K^{n-2}(n!)^{1+\gamma} \quad \forall\, n\geq 3.$$

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 $\Delta = 2^{\frac{J-j}{2}}/(2\max\{K,1\})$. Since λ is logarithmic in T, $\tilde{\delta}_j\to 0$ uniformly over j (details below) and $2^{\frac{J-j}{2}}\geq T^{\frac{\delta}{2}}$, we have that

$$\lambda\sqrt{\frac{1+c_{i,l}^2-2|c_{i,l}|\tilde{\delta}_j}{1+c_{i,l}^2}} = o\left(\left\{\frac{2^{\frac{J-j}{2}}}{2\max\{K,1\}}\right\}^a\right), \quad \text{as} \quad T\to\infty,$$

uniformly over j, for all $a>0$. By Theorem 1 from Rudzkis et al. (1978), we uniformly bound (7) from above by the Gaussian tail probability

$$C\exp\left\{-\frac{\lambda^2}{2}\left(\frac{1+c_{i,l}^2-2|c_{i,l}|\tilde{\delta}_j}{1+c_{i,l}^2}\right)\right\}.$$

Turning now to II, we have

$$II = \mathbb{P}\left(\frac{2^{(J-j)/2}|\tilde{s}_{j,k}^{(i,l)} - c_{i,l}|}{\sqrt{1+c_{i,l}^2}} \geq \frac{2^{(J-j)/2}\tilde{\delta}_j}{\sqrt{1+c_{i,l}^2}}\right) \tag{8}$$

The random variable on the LHS of the argument of the probability function in (8) is almost the same as that in (7), 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 exactly the same conditions as I. In fact, we can choose $\tilde{\delta}_j$ to be such that the thresholds in (7) and (8) are exactly the same, 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 $\tilde{\delta}_j = 2^{j-J}\lambda\left(\sqrt{\lambda^2 c_{i,l}^2 + 2^{J-j}(1+c_{i,l}^2)} - \lambda|c_{i,l}|\right)$. Since λ is logarithmic in T, $\tilde{\delta}_j$ is of order $O(T^{-\alpha})$ for $\alpha > 0$ uniformly over j. Thus, we bound $I + II$ from above by

$$\begin{aligned} I + II &= 2\,I \leq 2C\exp\left\{-\frac{\lambda^2}{2}\left(\frac{1+c_{i,l}^2 - 2|c_{i,l}|\tilde{\delta}_j}{1+c_{i,l}^2}\right)\right\} \\ &= 2C\exp\left\{-\frac{\lambda^2}{2}\right\}\exp\left\{\frac{\lambda^2|c_{i,l}|\tilde{\delta}_j}{1+c_{i,l}^2}\right\} \leq C_1\exp\left\{-\frac{\lambda^2}{2}\right\}. \end{aligned}$$

Substituting this in (6), 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

$$\lambda \geq \sqrt{2\left(2\log p + (1-\delta)\log T + \log\frac{1}{a_{p,T}}\right)},$$

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 the Theorem.

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 $\mathbb{P}(|\tilde{s}_{0,1}^{(i,l)}| > \lambda_1) \leq C_4\exp(-\lambda_1^2 T/2)$. Thus, using the Bonferroni inequality again, we have

$$\mathbb{P}(\exists\, i,l\ \ |\tilde{s}_{0,1}^{(i,l)}| > \lambda_1) \leq C_5 p^2\exp(-\lambda_1^2 T/2).$$

Bounding this by $C_5 a_{p,T}$, we obtain

$$\lambda_1 \geq \sqrt{2T^{-1}\left(2\log\, p + \log\frac{1}{a_{p,T}}\right)}. \tag{9}$$

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

$$\begin{aligned}\mathbb{P}(|\tilde{s}_{0,1}^{(i,l)}| < \lambda_1) &= \mathbb{P}(|\tilde{s}_{0,1}^{(i,l)} - c_{i,l}| > |c_{i,l}| - \lambda_1) \\ &\leq \mathbb{P}\left(\frac{\sqrt{T}|\tilde{s}_{0,1}^{(i,l)} - c_{i,l}|}{\sqrt{1+c_{i,l}^2}} > \frac{\sqrt{T}(|c_{i,l}| - \lambda_1)}{\sqrt{1+c_{i,l}^2}}\right).\end{aligned}$$

Assuming that the threshold on the RHS is so low that the normal approximation still works (it is sufficient to consider this as this is the worst-case scenario), we bound the above by

$$C_6 \exp\left\{-\frac{T(|c_{i,l}| - \lambda_1)^2}{2(1+c_{i,l}^2)}\right\} \leq C_6 \exp\left\{-\frac{T(|c_{i,l}| - \lambda_1)^2}{4}\right\}. \tag{10}$$

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

$$\sum_i \sum_l \mathbb{I}(c_{i,l} \neq 0) \leq \sum_i \max_i \sum_l \mathbb{I}(c_{i,l} \neq 0) \leq p c_0(p).$$

Thus, by the Bonferroni inequality, we have

$$\mathbb{P}(\exists\, i, l \;\; |\tilde{s}_{0,1}^{(i,l)}| < \lambda_1) \leq C_6 p c_0(p) \exp\left\{-\frac{T \min_{c_{i,l}\neq 0}(|c_{i,l}| - \lambda_1)^2}{4}\right\}.$$

Bounding the above by $C_6 a_{p,T}$, we get

$$2\sqrt{\frac{\log\, p + \log\, c_0(p) + \log \frac{1}{a_{p,T}}}{T}} + \lambda_1 \leq \min_{c_{i,l}\neq 0} |c_{i,l}|,$$

which is satisfied as the LHS has a lower order of magnitude than the RHS by the assumptions of the Theorem. This completes the proof of statement (a) of the Theorem.

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

$$\begin{aligned}\mathbb{P}(\max_{i,l:c_{i,l}\neq 0} |\bar{c}_{i,l}^{(h)} - c_{i,l}| > \lambda_3) &\leq p c_0(p) \max_{i,l} \mathbb{P}(|\tilde{s}_{0,1}^{(i,l)} - c_{i,l}| > \lambda_3) \\ &= p c_0(p) \max_{i,l} \mathbb{P}\left(\frac{\sqrt{T}|\tilde{s}_{0,1}^{(i,l)} - c_{i,l}|}{\sqrt{1+c_{i,l}^2}} > \frac{\sqrt{T}\lambda_3}{\sqrt{1+c_{i,l}^2}}\right) \\ &\leq C_7 p c_0(p) \max_{i,l} \exp\left(-\frac{T\lambda_3^2}{2(1+c_{i,l}^2)}\right) \\ &\leq C_7 p c_0(p) \exp\left(-\frac{T\lambda_3^2}{4}\right).\end{aligned}$$

Equating this to $C_7 a_{p,T}$, we get

$$\lambda_3 = 2\sqrt{\frac{\log\, p + \log\, c_0(p) + \log \frac{1}{a_{p,T}}}{T}},$$

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

$$\begin{aligned} \|\bar{\Sigma}^{(h)} - \Sigma\| &\leq \max_l \sum_i |\bar{c}_{i,l}^{(h)} - c_{i,l}| = \max_l \sum_i |\bar{c}_{i,l}^{(h)} - c_{i,l}| \mathbb{I}(c_{i,l} \neq 0) \leq \lambda_3 c_0(p) \\ &= 2c_0(p) \sqrt{\frac{\log\, p + \log\, c_0(p) + \log \frac{1}{a_{p,T}}}{T}}, \end{aligned}$$

which completes the proof of statement (b) of the Theorem.

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 and Levina (2008). □

Proof of Theorem 3.2. Note that by definition, $\bar{c}_{i,l}^{(s)}(t/T) = \text{sign}(\bar{c}_{i,l}^{(h)}(t/T)) \max(|\bar{c}_{i,l}^{(h)}(t/T)| - \lambda_1, 0)$. However, by the proof of Theorem 3.1, if $c_{i,l} \neq 0$, then $\mathbb{P}(\exists_{i,l}\ |\bar{c}_{i,l}^{(h)}| < \lambda_1) \leq C_6 a_{p,T}$. This proves statement (a). For the proof of statement (b), assume without loss of generality that $\bar{c}_{i,l}^{(h)} > 0$, we then have $\bar{c}_{i,l}^{(s)} = \bar{c}_{i,l}^{(h)} - \lambda_1$. As in the proof of Theorem 3.1, we compute

$$\mathbb{P}(\max_{i,l:c_{i,l}\neq 0} |\bar{c}_{i,l}^{(s)} - c_{i,l}| > \lambda_4) \leq \mathbb{P}(\max_{i,l:c_{i,l}\neq 0} |\bar{c}_{i,l}^{(h)} - c_{i,l}| > \lambda_4 - \lambda_1),$$

which leads to

$$\lambda_4 = 2\sqrt{\frac{\log\, p + \log\, c_0(p) + \log \frac{1}{a_{p,T}}}{T}} + \lambda_1,$$

and thus to

$$\|\bar{\Sigma}^{(s)} - \Sigma\| \leq c_0(p) \left(2\sqrt{\frac{\log\, p + \log\, c_0(p) + \log \frac{1}{a_{p,T}}}{T}} + \lambda_1 \right),$$

which completes the proof of statement (b). The proof of statement (c) is identical to that in the proof of Theorem 3.1. □

Proof of Theorem 3.3. If there is a breakpoint 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 the following:

- Only up to one coefficient at each scale j is non-zero.
- By (4) 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 $f^{(i,l)}_{j_1(i,l),k_1(i,l)}$ survives thresholding. If it does, then with probability one, there will be a breakpoint in $\tilde{c}_{i,l}(u)$ at $u = u_0$ where u_0 is the right endpoint of $\mathcal{J}^1_{i,l}$; thus, there will be a breakpoint in $\tilde{c}_{i,l}(u)$ located between the interval $\mathcal{I}$ and its nearest breakpoint to the left. But, using exactly the same technique as in the proof of Theorem 3.1, we can show

$$\mathbb{P}(\exists\, i, l \;\; |f^{(i,l)}_{j_1(i,l),k_1(i,l)}| < \lambda) \leq C_7 a_{p,T}.$$

Moreover, since all coefficients $f^{(i,l)}_{j,k}$ 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 Theorem 3.1), we have that

$$\forall_{i,l} \quad \tilde{c}_{i,l}(\mathcal{I}) = \frac{1}{|T\mathcal{K}_{i,l}|} \sum_{t \in T\mathcal{K}_{i,l}} X_{i,t,T} X_{l,t,T},$$

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 Theorem 3.1, 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 a 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 Theorem 3.1, we can show that with probability at least $1 - C_9 a_{p,T}$, we have that $\bar{c}^{(h)}_{i,l}(\mathcal{I}) = 0$ for all those i, l for which $c_{i,l}(\mathcal{I}) = 0$ if

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

Similarly, with probability at least $1 - C_{10} a_{p,T}$, we have that $\bar{c}^{(h)}_{i,l}(\mathcal{I}) = \tilde{c}_{i,l}(\mathcal{I})$ for all those i, l for which $c_{i,l}(\mathcal{I}) \neq 0$ if

$$\lambda_1 \leq \min_{c_{i,l}(\mathcal{I}) \neq 0} |c_{i,l}(\mathcal{I})| - 2\sqrt{\frac{\log\, p + \log\, c_0(p) + \log \frac{1}{a_{p,T}}}{T|\mathcal{I}|}}.$$

This completes the proof of statement (a). The proofs of statements (b) and (c) proceed analogously to those of the corresponding statements in Theorem 3.1 by recalling that $|\mathcal{K}_{i,l}| \geq |\mathcal{I}|$ and replacing T with $T|\mathcal{I}|$ where appropriate. □

Proof of Theorem 3.4. The proof proceeds along similar lines to the proof of Theorem 3.2 and hence we omit the details. □

References

C. Alexander. Orthogonal GARCH. In *Mastering Risk, Vol. 2: Applications*, pages 21–38. Financial Times-Prentice Hall, London, 2001.

T.G. Andersen, R.A. Davis, J.-P. Kreiss, and T. Mikosch, editors. *Handbook of Financial Time Series*. Springer, 2009.

L. Bauwens, S. Laurent, and J. Rombouts. Multivariate GARCH models: a survey. *Journal of Applied Econometrics*, 21:79–109, 2006.

P. Bickel and E. Levina. Covariance regularization by thresholding. *Ann. Stat.*, 36:2577–2604, 2008.

T. Bollerslev. Modelling the coherence in short-run nominal exchange rates: amultivariate generalized ARCH model. *Review of Economics and Statistics*, 72:498–505, 1990.

T. Bollerslev, R. Engle, and J. Wooldridge. A capital asset pricing model with time-varying covariances. *J. Polit. Econ.*, 96:116–131, 1988.

C. Brooks, S. Burke, and G. Persand. Multivariate GARCH models: software choice and estimation issues. *Journal of Applied Econometrics*, 18:725–734, 2003.

S. Clemencon and S. Slim. Statistical analysis of financial time series under theassumption of local stationarity. *Quantitative Finance*, 4:208–220, 2004.

R. Dahlhaus. Fitting time series models to nonstationary processes. *Ann. of Stat.*, 25:1–37, 1997.

J. Davidson. *Stochastic Limit Theory*. Oxford University Press, 1994.

D.K. Dey and C. Srinivasan. Estimation of a covariance matrix under Stein's loss. *Ann. Statist.*, 13:1581–1591, 1985.

D. Donoho, S. Mallat, R. von Sachs, and Y. Samuelides. Locally stationary covariance and signal estimation withmacrotiles. *IEEE Trans. Signal Processing*, 53:614–627, 2003.

N. El Karoui. Operator norm consistent estimation of large-dimensional sparse covariance matrices. *Ann. Statist.*, 36:2717–2756, 2008.

R. Engle. Dynamic conditional correlation: a simple class of multivariate generalizedautoregressive conditional heteroskedasticity models. *Journal of Business and Economic Statistics*, 20:339–350, 2002.

R. Engle and K. Kroner. Multivariate simultaneous generalized ARCH. *Econometric Theory*, 11:122–150, 1995.

R. Engle, V.K. Ng, and M. Rothschild. Asset pricing with a factor ARCH covariance structure: empirical estimates for treasury bills. *Journal of Econometrics*, 45:213–238, 1990.

R. F. Engle. Autoregressive conditional heteroscedasticity with estimates ofthe variance of united kingdom inflation. *Econometrica*, 50:987–1007, 1982.

J. Fan, M. Wang, and Q. Yao. Modelling multivariate volatilities via conditionally uncorrelated components. *J. Roy. Statist. Soc. B*, 70:679–702, 2008a.

J. Fan, J. Zhang, and K. Yu. Asset allocation and risk assessment with gross exposure constraints for vast portfolios. *Preprint*, 2008b.

M. Fisz. The limiting distribution of a function of two independent random variablesand its statistical application. *Colloquium Mathematicum*, 3:138–146, 1955.

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

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

P. Fryzlewicz and G. P. Nason. A Haar-Fisz algorithm for Poisson intensity estimation. *Journal of Computational and Graphical Statistics*, 13:621–638, 2004.

P. Fryzlewicz, T. Sapatinas, and S. Subba Rao. A Haar-Fisz technique for locally stationary volatility estimation. *Biometrika*, 93:687–704, 2006.

P. Fryzlewicz, T. Sapatinas, and S. Subba Rao. Normalised least-squares estimation in time-varying ARCH models. *Annals of Statistics*, 36:742–786, 2008.

G. Golub and C. Van Loan. *Matrix Computations, 2nd ed.* Johns Hopkins Univ. Press, Baltimore, MD, 1989.

W. Haerdle, H. Herwartz, and V. Spokoiny. Time inhomogeneous multiple volatility modeling. *Journal of Financial Econometrics*, 1:55–95, 2003.

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

C. Hafner, D. van Dijk, and P. Franses. Semi-parametric modelling of correlation dynamics. In *Advances in Econometrics, vol. 20/A*, pages 59–103. Elsevier, Amsterdam, 2005.

S. Herzel, C. Starica, and R. Tutuncuc. 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, pages 391–429. 2006.

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

D. Kennedy. The distribution of the maximum Brownian excursion. *J. Appl. Prob.*, 13: 371–376, 1976.

O. Ledoit and M. Wolf. A well-conditioned estimator for large-dimensional covariance matrices. *J. Multivariate Anal.*, 88:365–411, 2003.

A. Lunde and P.R. Hansen. A forecast comparison of volatility models: does anything beata GARCH(1,1)? *J. Appl. Econometrics*, 20:873–889, 2005.

H. Markowitz. Portfolio selection. *Journal of Finance*, 7:77–91, 1952.

H. Markowitz. *Portfolio Selection: Efficient Discrimination of Investments*. John Wiley & Sons, New York, 1959.

T. Mikosch and C. Starica. Non-stationarities in financial time series, the long rangedependence and the IGARCH effects. *Review of Economics and Statistics*, 86:378–390, 2004.

G. Motta, C. Hafner, and R. von Sachs. Locally stationary factor models: identification and nonparametric estimation. *Econometric Theory,* to appear, 2010.

G. P. Nason, R. von Sachs, and G. Kroisandt. Wavelet processes and adaptive estimation of the evolutionary wavelet spectrum. *Journal of the Royal Statistical Society. Series B*, 62:271–292, 2000.

D. Pelletier. Regime switching for dynamic correlations. *Journal of Econometrics*, 131: 445–473, 2006.

J. Rodriguez-Poo and O. Linton. Nonparametric factor analysis of residual time series. *Test*, 10:161–182, 2001.

R. Rudzkis, L. Saulis, and V. Statulevicius. A general lemma on probabilities of large deviations. *Lithuanian Math. J.*, 18:226–238, 1978.

C. Starica and C. Granger. Non-stationarities in stock returns. *Review of Economics and Statistics*, 87:503–522, 2005.

S. J. Taylor. *Modelling Financial Time Series*. Wiley, Chichester, 1986.

R. Tibshirani, M. Saunders, S. Rosset, and J. Zhu. Sparsity and smoothness via the fused lasso. *J. Roy. Stat. Soc. B*, 67:91–108, 2005.

Y.K. Tse and K.C. Tsui. A multivariate generalized autoregressive conditional heteroscedasticity model with time-varying correlations. *Journal of Business and Economic Statistics*, 20:351–362, 2002.

B. Vidakovic. *Statistical Modeling by Wavelets*. Wiley, New York, 1999.

Y. Wang and J. Zou. Vast volatility matrix estimation for high-frequency financial data. *Ann. Stat.*, 38:943–978, 2010.

M.V. Wickerhauser. *Adapted Wavelet Analysis from Theory to Software*. AK Peters, 1994.