Econometric Theory, **37**, 2021, 1100–1134.
doi:10.1017/S0266466620000535

ESTIMATION OF TIME-VARYING COVARIANCE MATRICES FOR LARGE DATASETS*

YIANNIS DENDRAMIS
Athens University of Economics and Business

LIUDAS GIRAITIS
Queen Mary University of London

GEORGE KAPETANIOS
King's College London

Time variation is a fundamental problem in statistical and econometric analysis of macroeconomic and financial data. Recently, there has been considerable focus on developing econometric modelling that enables stochastic structural change in model parameters and on model estimation by Bayesian or nonparametric kernel methods. In the context of the estimation of covariance matrices of large dimensional panels, such data requires taking into account time variation, possible dependence and heavy-tailed distributions. In this paper, we introduce a nonparametric version of regularization techniques for sparse large covariance matrices, developed by Bickel and Levina (2008) and others. We focus on the robustness of such a procedure to time variation, dependence and heavy-tailedness of distributions. The paper includes a set of results on Bernstein type inequalities for dependent unbounded variables which are expected to be applicable in econometric analysis beyond estimation of large covariance matrices. We discuss the utility of the robust thresholding method, comparing it with other estimators in simulations and an empirical application on the design of minimum variance portfolios.

1. INTRODUCTION

This paper considers estimation of large covariance matrices under structural change, possible dependence and heavy-tailed distributions. The problem of structural change in the econometric literature has been mainly addressed within either univariate or relatively small multivariate models and settings. Existing research is primarily focused on factor models where datasets are summarised by a finite set of unobserved time series, usually referred to as factors. Examples of this literature include Stock and Watson (2002a, 2002b). There has been little work

*We thank the Editor, the Co-Editor and the Referees for constructive comments. Yiannis Dendramis acknowledges support from the European Union's Horizon 2020 research and innovation program under the Marie Sklodowska-Curie grant agreement No 708501. Corresponding address: School of Economics and Finance, Queen Mary University of London, London E1 4NS, UK: e-mail: L.Giraitis@qmul.ac.uk.

© The Author(s), 2021. Published by Cambridge University Press.

https://doi.org/10.1017/S0266466620000535 Published online by Cambridge University Press

done on structural change in large datasets. This paper addresses the issue without recourse to factor modeling.

Estimation of large covariance matrices is particularly demanding since the number of estimated objects rises as a square of the dimension of the dataset leading to a large amount of aggregated estimation error. Several regularization techniques for improved estimation of large covariance matrices have been proposed which are summarized briefly in the body of the paper. These include Ledoit and Wolf (2004), Bickel and Levina (2008), Cai and Liu (2011) and Abadir, Distaso, and Zikes (2014). See also the excellent review article by Fan, Liao, and Liu (2016) and references therein. That review article focuses on structure-based estimators of covariance and precision matrices—that is estimators that assume sparsity or a factor model. For structure-free estimators, see Pourahmadi (2013).

In this paper, we develop regularized thresholding estimation in the presence of structural change of a general unspecified form. The characterization of the change is an important aspect of the problem. In contrast to the majority of the work we allow for smooth deterministic or stochastic change of the covariance matrix rather than structural breaks. Some new characterizations of smooth change introduced in this paper expand upon the notion of smoothness used in nonparametric inference.

There is limited literature on regularized estimation under deterministic structural change, see Chen, Xu, and Wu (2013), Zhou, Lafferty, and Wasserman (2010) and Kolar and Xing (2011); strong mixing, see Fan, Liao, and Mincheva (2013), and heavy-tailed data (independent, modelled by elliptical distribution), see, for example, Wegkamp and Zhao (2016), Han and Liu (2017), and Fan, Wang, and Zhong (2016). All this work assumes that the volatility process is a deterministic function or a constant which is considered extremely restrictive for economic and financial data, both from a theoretical and an empirical point of view. In a different stochastic setup than the one considered in the paper, Bickel, Wang, and Zhou (2013) have provided a theoretical analysis for Bickel–Levina thresholding in the context of realized covariance, estimated from noisy data.

We provide a unified framework for estimation of the paths of large covariance matrices, that change over time, in a potentially stochastic way, with temporal dependence, characterized by mixing, and with potentially heavy tails. The key characteristic of such regularized estimation is robustness to all those data features, which are highly likely to be present in financial and economic data.

Deterministic change of covariance corresponds to the typical characterization of heteroscedasticity. Stochastic (persistent) change of volatility, also considered in our paper, is a standard vehicle for the modeling of structural change in the economy. It differs from ARCH type volatility models that have dominated the financial econometrics literature since their introduction in Engle (1982). It aligns with the empirically established fact that stochastic change in volatility may be more persistent and smooth than that allowed by stationary conditional volatility models, see Kapetanios (2010).

https://doi.org/10.1017/S0266466620000535 Published online by Cambridge University Press

In order to achieve such a characterization of persistent volatility, we follow the modelling framework of Giraitis, Kapetanios, and Yates (2014, 2018) and specify that the volatility process is, potentially stochastic, bounded and has small increments. That framework, in turn, extended the work of, for example, Dahlhaus (1997) on locally stationary processes, where change was assumed to be smooth and deterministic, in nature.

This paper is organised as follows. In Section 2, we present the main results. We show that the thresholding estimation procedure developed by Bickel and Levina (2008) and others, in general, is robust to dependence, random scaling (volatility) and heteroscedasticity and the type of distribution of the data. In Section 3, we discuss cross-validation methods for the selection of tuning parameters and use simulations to compare the performance of regularized estimation methods. In that Section, we further provide an application of regularized estimation methods for designing minimum variance portfolios. Section 4 contains new results on Bernstein type inequalities for dependent data, that are of independent interest. They are particularly important for the rigorous analysis of penalised regression methods, like Lasso, in the presence of time series dependence. The online supplement contains proofs and technical results.

2. THRESHOLDING ESTIMATION

Given a sample $(\boldsymbol{y}_1, \ldots, \boldsymbol{y}_T)$ of a p-variate process $\boldsymbol{y}_t = (y_{1,t}, ..., y_{p,t})'$, estimation of the population covariance $p \times p$ matrix $\boldsymbol{\Sigma}$ by the sample covariance is a well defined procedure when p is fixed. For large covariance matrices, when p increases with T the poor performance of the sample covariance matrix estimate $\widehat{\boldsymbol{\Sigma}}$ can be improved by various regularization procedures which include the thresholding methods developed by Bickel and Levina (2008), Cai and Liu (2011), Fan et al. (2013) and others.

The aim of this paper is to investigate the impact of dependence, heteroscedasticity and distribution of the data on consistency rates for regularised estimation of the covariance matrix using the Bickel and Levina (2008) thresholding procedure. Under heteroscedasticity, the covariance matrix $\boldsymbol{\Sigma}_t = \text{var}(\boldsymbol{y}_t) = E[(\boldsymbol{y}_t - E\boldsymbol{y}_t)(\boldsymbol{y}_t' - E\boldsymbol{y}_t')]$ varies over time, and can be estimated by the local sample covariance matrix $\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{t}}$, see (10). For data with a random scaling (volatility) component, the limit of $\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{t}}$ is less obvious. As shown in the following analysis, such a limit comprises a time-varying deterministic matrix and a random scale. The novelty of our theoretical findings is showing that the Bickel–Levina thresholding procedure is robust to dependence, light or heavy tailedness of the data and it can be adjusted for heteroscedasticity.

To reflect the robustness properties of regularized estimation of $\boldsymbol{\Sigma}_{\boldsymbol{t}}$, we consider three settings for $\boldsymbol{y}_t$. Before proceeding further, we provide assumptions, notation and definitions.

https://doi.org/10.1017/S0266466620000535 Published online by Cambridge University Press

Assumption M. (i) The centered stochastic process $\boldsymbol{y}_t - E\boldsymbol{y}_t$ is an α-mixing (but not necessarily stationary) process with mixing coefficients α_k[1] such that for some $0 < \phi < 1$ and $c > 0$,

$$\alpha_k \leq c\phi^k, \ k \geq 1. \tag{1}$$

(ii) The elements of $\mathrm{E}\boldsymbol{y}_t = (\mu_{1,t}, ..., \mu_{p,t})'$ and $\mathrm{var}(\boldsymbol{y}_t) = \boldsymbol{\Sigma_t} = (\sigma_{jk,t})_{j,k=1,...,p}$ are such that

$$|\mu_{j,t}| \leq C, \quad |\mu_{j,t} - \mu_{j,s}| \leq C\frac{|t-s|}{\max(t,s)}, \quad |\sigma_{jk,t}| \leq C, \quad |\sigma_{jk,t} - \sigma_{jk,s}| \leq C\frac{|t-s|}{\max(t,s)} \tag{2}$$

for $1 \leq t, s \leq T$ where $C < \infty$ does not depend on j, k and t, s and T.

The components of $\boldsymbol{y}_t = (y_{1,t}, ..., y_{p,t})'$ can have a wide variety of tail behavior. We write $(\boldsymbol{y}_t) \in \mathcal{E}(s)$, $s > 0$ to denote a thin-tailed distribution for $y_{j,t}$:

$$\max_{j,t} E\exp(a|y_{j,t}|^s) < \infty \quad \text{for some } a > 0. \tag{3}$$

The notation $(\boldsymbol{y}_t) \in \mathcal{H}(\theta)$, $\theta > 2$ denotes a heavy-tailed distribution property:

$$\max_{j,t} E|y_{j,t}|^\theta < \infty. \tag{4}$$

The definitions (3) and (4) imply that there exists $c_0 > 0$, $c_1 > 0$ such that for all $\zeta > 0$, $t \geq 1$,

$$P(|y_{i,t}| \geq \zeta) \leq \begin{cases} c_0\exp(-c_1\zeta^s) & \text{if } (y_{j,t}) \in \mathcal{E}(s), s > 0, \\ c_0\zeta^{-\theta} & \text{if} (y_{j,t}) \in \mathcal{H}(\theta), \theta > 2. \end{cases}$$

Thresholding under stationarity. Suppose that $\boldsymbol{y}_t$ is a stationary α-mixing process. Then, $E\boldsymbol{y}_t = \boldsymbol{\mu}$ and $\boldsymbol{\Sigma} = \mathrm{var}(\boldsymbol{y}_t) = [\sigma_{ij}]$ do not depend on t. Denote by $\widehat{\boldsymbol{\Sigma}} = [\widehat{\sigma}_{ij}]$ the sample covariance estimate of $\boldsymbol{\Sigma}$,

$$\widehat{\boldsymbol{\Sigma}} = T^{-1}\sum_{j=1}^{T} \boldsymbol{y}_j\boldsymbol{y}_j' - \bar{\boldsymbol{y}}\bar{\boldsymbol{y}}', \qquad \bar{\boldsymbol{y}} = T^{-1}\sum_{j=1}^{T} \boldsymbol{y}_j. \tag{5}$$

Hard and adaptive thresholding, introduced by Bickel and Levina (2008) and Cai and Liu (2011), are two standard procedures to regularize $\widehat{\boldsymbol{\Sigma}}$, when p increases with T. Hard thresholding is based on the idea of setting the elements of $\widehat{\boldsymbol{\Sigma}}$, whose absolute values are smaller than some threshold λ, to zero. It yields the estimate

$$T_\lambda\big(\widehat{\boldsymbol{\Sigma}}\big) = \big(\widehat{\sigma}_{ij}I(|\widehat{\sigma}_{ij}| > \lambda)\big).$$

[1] Let $\mathcal{F}_{-\infty}^{j}$, $\mathcal{F}_{j}^{\infty}$ denote σ-fields generated by $\{\boldsymbol{y}_t, t \leq j\}$ and $\{\boldsymbol{y}_t, t \geq j\}$ respectively. Define the α-mixing coefficient as $\alpha_k = \sup_j \sup_{A\in\mathcal{F}_{-\infty}^{j},\, B\in\mathcal{F}_{j+k}^{\infty}} |P(A)P(B) - P(A\cap B)|$.

https://doi.org/10.1017/S0266466620000535 Published online by Cambridge University Press

Procedures based on other thresholding operators can be defined, but they have similar properties to hard thresholding, asymptotically, although they may differ in finite samples.

In their seminal work Bickel and Levina (2008) showed that if $\boldsymbol{y}_t$ is an i.i.d. Gaussian process, then the regularized estimator $T_\lambda(\widehat{\boldsymbol{\Sigma}})$ of $\boldsymbol{\Sigma}$ under a sparsity assumption is consistent,

$$\left\| T_\lambda\left(\widehat{\boldsymbol{\Sigma}}\right) - \boldsymbol{\Sigma} \right\| = O_P(n_p\lambda) \qquad \text{with} \quad \lambda = \kappa\sqrt{\frac{\log p}{T}}, \tag{6}$$

$$\left\| T_\lambda\left(\widehat{\boldsymbol{\Sigma}}\right)^{-1} - \boldsymbol{\Sigma}^{-1} \right\| = O_P(n_p\lambda) \quad \text{if } n_p\lambda = o(1) \text{ and } ||\boldsymbol{\Sigma}|| \geq c > 0 \tag{7}$$

where $\|.\|$ denotes the spectral norm and κ is a tuning parameter. Thresholding estimation as a rule assumes that $\boldsymbol{\Sigma}$ is approximately sparse, that is the sparsity parameter, n_p, which is the maximum number of nonzero elements in a row of $\boldsymbol{\Sigma}$,

$$\max_{i=1,\ldots,p} \sum_{j=1}^{p} I(\sigma_{ij} \neq 0) = n_p, \tag{8}$$

does not grow too fast with p.

As in Bickel and Levina (2008) and Fan et al. (2013), we consider, as a leading case, the one where $\boldsymbol{\Sigma}$ is sparse, although the theory can accommodate large $n_p \leq p$. We assume that $T, p \to \infty$.

THEOREM 1. *Let $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_T$ be a sample from a stationary process $(\boldsymbol{y}_t)$ which satisfies Assumption M. Let $c > 0$, $\varepsilon > 0$ and*

$$T \geq cp^{\varepsilon}, \quad T, p \to \infty. \tag{9}$$

Then, for sufficiently large $\kappa > 0$, the regularized estimate $T_\lambda\left(\widehat{\boldsymbol{\Sigma}}\right)$ of $\boldsymbol{\Sigma}$ satisfies (6) and (7) in the following cases.

(i) *If $(\boldsymbol{y}_t) \in \mathcal{E}(s)$, $s > 0$.*
(ii) *If $(\boldsymbol{y}_t) \in \mathcal{H}(\theta)$, $\theta > 4$ and $\varepsilon > 8/(\theta - 4)$ in (9).*

This theorem shows that the Bickel–Levina thresholding procedure is valid for a stationary α-mixing process $(\boldsymbol{y}_t)$ which may have a non-Gaussian, heavy-tailed distribution. For $\boldsymbol{y}_t$ with a heavy-tailed distribution, it requires at least four finite moments and p should be relatively small compared to T.

Thresholding under heteroscedesticity. In economic applications, heteroscedasticity may arise due to smooth changes of the mean and unconditional variance of observations. Next we consider the case when $\boldsymbol{y}_t$ is a heteroscedastic process, that is $\text{var}(\boldsymbol{y}_t) = \boldsymbol{\Sigma}_t = [\sigma_{ij,t}]$ and $E[\boldsymbol{y}_t] = \boldsymbol{\mu}_t$ vary with t. To account for variation in $\boldsymbol{\Sigma}_t$, we use a kernel type sample covariance estimator $\widehat{\boldsymbol{\Sigma}}_t = [\widehat{\sigma}_{ij,t}]$ of $\boldsymbol{\Sigma}_t$ at time

https://doi.org/10.1017/S0266466620000535 Published online by Cambridge University Press

$t = 1, ..., T$ given by

$$\widehat{\mathbf{\Sigma}}_t \equiv \widehat{\mathbf{\Sigma}}_{T,t} = K_t^{-1} \sum_{k=1}^{T} b_{H,|t-k|} \boldsymbol{y}_k \boldsymbol{y}_k' - \bar{\boldsymbol{y}}_t \bar{\boldsymbol{y}}_t', \qquad \bar{\boldsymbol{y}}_t = K_t^{-1} \sum_{k=1}^{T} b_{H,|t-k|} \boldsymbol{y}_k \tag{10}$$

where $K_t = \sum_{k=1}^{T} b_{H,|t-k|}$ and $b_{H,|t-k|}$ are kernel weights defined in (12). The regularized sample covariance estimate of $\mathbf{\Sigma}_t$ is defined as

$$T_\lambda\big(\widehat{\mathbf{\Sigma}}_t\big) = \big(\widehat{\sigma}_{ij,t} I(|\widehat{\sigma}_{ij,t}| > \lambda)\big). \tag{11}$$

Below we modify the thresholding procedure to account for heteroscedasticity.

We assume that $\mathbf{\Sigma}_t$ is approximately sparse, that is the maximum number of nonzero elements in each row of $\mathbf{\Sigma}_t$ for all t, n_p, is either finite, or does not grow too fast with p. The rate of change of $\mathbf{\Sigma}_t$ and $\boldsymbol{\mu}_t$ will be controlled by (2) of Assumption M.

We will consider kernel estimates (10) with weights

$$b_{H,|t-k|} = K(|t-k|/H), \tag{12}$$

where $H \to \infty$, $H = o(T)$. $K(x)$, $x \in (0, a)$ is a non-negative continuous function with finite or infinite support, such that, for some $C > 0$ and $\nu > 3$,

$$K(x) \le C(1+x^\nu)^{-1}, \quad |(d/dx)K(x)| \le C(1+x^\nu)^{-1}, \quad x \in (0, a). \tag{13}$$

In particular, the functions $K(x) = I(0 \le x \le 1)$, $K(x) = (1+x^\nu)^{-1}$ with $\nu > 2$ and $K(x) = \exp(-cx^\alpha)$ with $\alpha > 0$ satisfy (13).

The following theorem establishes the consistency properties of the regularised estimate $T_\lambda\big(\widehat{\mathbf{\Sigma}}_t\big)$ of $\mathbf{\Sigma}_t$. Under heteroscedasticity the threshold parameter λ in (15) is larger than under stationarity in Theorem 1. It depends on the bandwidth H ("window size") and accounts for the bias-variance tradeoff of change in mean $\boldsymbol{\mu}_t = E[\boldsymbol{y}_t]$ and covariance $\mathbf{\Sigma}_t = \mathrm{var}(\boldsymbol{y}_t)$. We denote by n_p the sparsity parameter of $\mathbf{\Sigma}_t$ defined as in (8).

THEOREM 2. *Suppose that $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_T$ is a sample from a p-dimensional heteroscedastic process $(\boldsymbol{y}_t)$ which satisfies Assumption M. Let $c_0, c, \delta, \varepsilon > 0$.*

Then, the regularised estimate $T_\lambda\big(\widehat{\mathbf{\Sigma}}_t\big)$ of $\mathbf{\Sigma}_t$, $\delta T \le t \le T$, with bandwidth H such that

$$c_0 p^\varepsilon \le H = o(T/\sqrt{\log T}) \quad \text{as } T, p \to \infty \tag{14}$$

for sufficiently large κ has the properties

$$\big\| T_\lambda\big(\widehat{\mathbf{\Sigma}}_t\big) - \mathbf{\Sigma}_t \big\| = O_P(n_p \lambda) \quad \text{with } \lambda = \kappa \sqrt{\log p}\, \max\Big(\frac{1}{\sqrt{H}}, \frac{H}{T}\Big), \tag{15}$$

$$\big\| T_\lambda\big(\widehat{\mathbf{\Sigma}}_t\big)^{-1} - \mathbf{\Sigma}_t^{-1} \big\| = O_P(n_p \lambda) \quad \text{if } n_p \lambda = o(1) \text{ and } ||\mathbf{\Sigma}_t|| \ge c > 0 \tag{16}$$

in the following two cases.

(i) *If $(\boldsymbol{y}_t) \in \mathcal{E}(s)$, $s > 0$.*
(ii) *If $(\boldsymbol{y}_t) \in \mathcal{H}(\theta)$, $\theta > 4$ and $\varepsilon > 8/(\theta - 4)$ in (14).*

https://doi.org/10.1017/S0266466620000535 Published online by Cambridge University Press

The bandwidth $H_{opt} = T^{2/3}$ *yields the lowest threshold* $\lambda_{opt} = \kappa\sqrt{\frac{\log p}{H_{opt}}} = \kappa\frac{\sqrt{\log p}}{T^{1/3}}$ *in (15).*

Condition (14) on H is similar to that on T in equation (9), Theorem 1. The choice of λ reflects the bias–variance tradeoff of nonparametric inference. Under heteroscedasticity, dependence and thin-tailed or heavy-tailed distribution of $(\boldsymbol{y}_t)$, the optimal bandwidth, $H_{opt} = T^{2/3}$ yields the threshold $\lambda_{opt} = \kappa\sqrt{\frac{\log p}{H_{opt}}}$ which is the Bickel–Levina threshold (6) with T replaced by H_{opt}. The tuning parameter κ can be selected by cross-validation.

We complete this subsection with typical examples of μ_t.

EXAMPLE 3. *The following two sequences of real numbers* $\mu_1, \ldots, \mu_T$ *satisfy assumption (2).*

a) $\mu_t = t^{-1}\sum_{j=1}^{t} a_j$, $t \geq 1$ *where* a_j *is a bounded sequence of real numbers.*
b) $\mu_t = \mu_{Tt} = g(t/T)$, $1 \leq t \leq T$ *where* $g(x)$, $x \in [0,1]$ *is a bounded function with a bounded derivative. Then,*

$$|\mu_t - \mu_s| \leq C\frac{|t-s|}{T} \leq C\frac{|t-s|}{\max(t,s)} \text{ for } 1 \leq t,s \leq T.$$

Thresholding under random scaling. Allowing for heteroscedasticity that involves stochastic scaling is crucial in modelling macroeconomic and financial data.

Such data can be put in the form

$$\boldsymbol{y}_t = \boldsymbol{H}_t\boldsymbol{x}_t, \tag{17}$$

where $\boldsymbol{x}_t$ is a heteroscedastic α-mixing process with time-varying mean $\boldsymbol{\mu}_t^{(x)} = E\boldsymbol{x}_t$ and covariance matrix $\boldsymbol{\Sigma}_t^{(x)} = \text{var}(\boldsymbol{x}_t)$, and $\boldsymbol{H}_t = (h_{ij,t})$ is a random persistent $p \times p$ matrix-valued scaling process. We show below, that for such $\boldsymbol{y}_t$ the limit of $T_\lambda(\widehat{\boldsymbol{\Sigma}}_t)$ is a time varying matrix

$$\boldsymbol{\Sigma}_t = \boldsymbol{H}_t\boldsymbol{\Sigma}_t^{(x)}\boldsymbol{H}_t'$$

which itself is a random process. Although no restriction on the dependence between $\boldsymbol{H}_t$ and $\boldsymbol{x}_t$ is imposed, the elements $(h_{ij,t})$ of $\boldsymbol{H}_t$ are assumed to be persistent (smooth) and thin-tailed processes. The latter is formalized as follows.

Assumption H. The elements $(h_{ij,t})$ of $\boldsymbol{H}_t$ have the following property. For all $j,k = 1,\ldots,p$, $1 \leq t,s \leq T$, $T \geq 2$,

$$|h_{jk,t} - h_{jk,s}| \leq \left(\frac{|t-s|}{\max(t,s)}\right)^{1/2} h_{jk,ts} \tag{18}$$

where $(h_{jk,t}) \in \mathcal{E}(\alpha)$, $(h_{jk,ts}) \in \mathcal{E}(\alpha)$ for some $\alpha > 0$.

https://doi.org/10.1017/S0266466620000535 Published online by Cambridge University Press

Under Assumption H, all processes $h_{jk,t}$ and $h_{jk,ts}$ satisfy the thin tail property (3) with the same parameters a.

In Theorem 4 we assume that the sparsity parameter n_H of $\boldsymbol{H}_t$ is bounded for all t and p, while the sparsity parameter n_p of $\boldsymbol{\Sigma}_t^{(x)}$ may increase with p.

THEOREM 4. *Suppose that $\boldsymbol{y}_1,\ldots,\boldsymbol{y}_T$ is a sample from a p-dimensional process, $\boldsymbol{y}_t$ as in (17). Assume that $\boldsymbol{x}_t$ satisfies Assumption M and $\boldsymbol{H}_t$ satisfies Assumption H with parameter $\alpha > 0$. Let $\nu = (\alpha+4)/(2\alpha)$ and $c_0, c, \delta, \varepsilon > 0$.*

Then, the regularised estimate $T_\lambda(\widehat{\boldsymbol{\Sigma}}_t)$ of $\boldsymbol{\Sigma}_t = \boldsymbol{H}_t\boldsymbol{\Sigma}_t^{(x)}\boldsymbol{H}_t'$, $\delta T \le t \le T$, with the bandwidth H such that

$$c_0 p^{\varepsilon} \le H = o(T/(\log T)^{\nu}) \quad \text{as } T,p \to \infty \tag{19}$$

for sufficiently large κ has the property that

$$\left\| T_\lambda(\widehat{\boldsymbol{\Sigma}}_t) - \boldsymbol{\Sigma}_t \right\| = O_P(n_p\lambda) \quad \text{with } \lambda = \kappa(\log p)^{\nu}\max\left(\frac{1}{\sqrt{H}}, \left(\frac{H}{T}\right)^{1/2}\right), \tag{20}$$

$$\left\| T_\lambda(\widehat{\boldsymbol{\Sigma}}_t)^{-1} - \boldsymbol{\Sigma}_t^{-1} \right\| = O_P(n_p\lambda) \quad \text{if } n_p\lambda = o(1) \text{ and } ||\boldsymbol{\Sigma}_t|| \ge c > 0 \tag{21}$$

in the following two cases.

(i) *If $(\boldsymbol{x}_t) \in \mathcal{E}(s)$, $s > 0$.*
(ii) *If $(\boldsymbol{x}_t) \in \mathcal{H}(\theta)$, $\theta > 4$ and $\varepsilon > 8/(\theta-4)$ in (19).*

The bandwidth $H_{opt} = T^{1/2}$ yields the lowest threshold $\lambda_{opt} = \kappa\sqrt{\frac{(\log p)^{2\nu}}{H_{opt}}} = \kappa\frac{(\log p)^{\nu}}{T^{1/4}}$ in (20).

In the presence of random scaling, the threshold, λ, in Theorem 4 is larger and the consistency rate slower than in Theorem 2 under heteroscedasticity. The optimal threshold λ_{opt} depends on the unknown parameter $\nu = (\alpha+4)/(2\alpha)$ where α is the tail parameter of $\boldsymbol{H}_t$. If $\alpha = 2$, then $\nu = 3/2$. If $\boldsymbol{H}_t$ is bounded, then $\alpha = \infty$ and $\nu = 1/2$. In this case, $\lambda_{opt} = \kappa\sqrt{(\log p)/H_{opt}}$. Parameter κ can be selected by cross-validation.

Next, we provide two examples of processes satisfying (18).

EXAMPLE 5. *Let $\xi_j \sim IIDN(0,1)$. The following sequences ζ_t, $t \ge 1$ satisfy (18) with $\alpha = 2$.*

a) $\zeta_t = t^{-1/2}\sum_{j=1}^{t}\xi_j$, $1 \le t \le T$,
b) $\zeta_t = T^{-1/2}\sum_{j=1}^{t}\xi_j$.

To verify a), let $t > s$. Write

$$|\zeta_t - \zeta_s| = \left| t^{-1/2}\sum_{j=s+1}^{t}\xi_j + (t^{-1/2} - s^{-1/2})\sum_{j=1}^{s}\xi_j \right|$$

https://doi.org/10.1017/S0266466620000535 Published online by Cambridge University Press

$$\leq \frac{(t-s)^{1/2}}{t^{1/2}}\left|\frac{1}{(t-s)^{1/2}}\sum_{j=s+1}^{t}\xi_j\right| + \frac{|t^{1/2}-s^{1/2}|}{t^{1/2}}\left|\frac{1}{s^{1/2}}\sum_{j=1}^{s}\xi_j\right|$$

$$\leq \frac{(t-s)^{1/2}}{t^{1/2}}\zeta_{ts}, \quad \zeta_{ts} = \frac{1}{(t-s)^{1/2}}\left|\sum_{j=s+1}^{t}\xi_j\right| + \frac{1}{s^{1/2}}\left|\sum_{j=1}^{s}\xi_j\right|$$

noting that $t^{1/2}-s^{1/2} \leq (t-s)^{1/2}$. *Since* $\zeta_t \sim N(0,1)$, *then* $(t-s)^{-1/2}\sum_{j=s+1}^{t}\xi_j \sim N(0,1)$ *which implies* $(\zeta_t) \sim \mathcal{E}(2)$, $(\zeta_{ts}) \sim \mathcal{E}(2)$ *and verifies (18) with* $\alpha = 2$.

b) Observe that $|\zeta_t| \leq t^{-1/2}|\sum_{j=1}^{t}\xi_j|$ *and* $|\zeta_t - \zeta_s| \leq (t-s)^{1/2}T^{-1/2}|(t-s)^{1/2}\sum_{j=s+1}^{t}\xi_j|$ *which implies (18) with* $\alpha = 2$.

Throughout our analysis, we have assumed a significant degree of sparsity for the covariance matrix. The issue of sparsity can be addressed by extending the method developed by Fan et al. (2013), based on a factor structure, to the time varying case. To do so note that by Stock and Watson (2002a), factors can be consistently estimated, using principal component analysis, in the presence of structural change. Then, one can use time varying regression analysis, based on Giraitis, Kapetanios, and Yates (2018) to obtain time varying factor loadings and associated residuals. The (sparse) covariance matrix of these residuals can then be analysed using our proposed method. Due to the considerable technical arguments needed to implement this approach rigorously we leave this to future research.

3. IMPLEMENTATION OF REGULARIZED ESTIMATION

In this section, we compare the finite sample performance of various approaches for the estimation of large dimensional covariance matrices. We also examine their usefulness for one step ahead out-of-sample forecasting of such matrices.

Besides thresholding, another popular method of regularizing the sample covariance estimator is based on shrinkage. Ledoit and Wolf (2003, 2004) have promoted this approach in a series of papers. We consider the full sample and time-varying version of shrinkage estimators, defined as

$$\widehat{\mathbf{\Sigma}}_{LW} = \rho\,\mu_T \boldsymbol{I} + (1-\rho)\widehat{\mathbf{\Sigma}}, \qquad \widehat{\mathbf{\Sigma}}_{LW,t} = \rho\,\mu_{t,T}\boldsymbol{I} + (1-\rho)\widehat{\mathbf{\Sigma}}_t,$$

where $\mu_T = p^{-1}\mathrm{tr}(\widehat{\mathbf{\Sigma}})$, $\mu_{t,T} = p^{-1}\mathrm{tr}(\widehat{\mathbf{\Sigma}}_t)$. Here, the matrices $\mu_t \boldsymbol{I}$ and $\mu_{t,T}\boldsymbol{I}$ approximate the shrinkage target $\mathrm{diag}(\widehat{\mathbf{\Sigma}})$ and $\mathrm{diag}(\widehat{\mathbf{\Sigma}}_t)$ and the shrinkage intensity parameter $\rho \in [0,1]$ can be obtained through cross-validation. Ledoit and Wolf (2004) suggested to evaluate $\widehat{\mathbf{\Sigma}}_{LW}$ with the theoretically optimal

$$\rho_{opt} = \frac{b_T^2}{d_T^2}, \quad d_T^2 = p^{-1}\mathrm{tr}(\widehat{\mathbf{\Sigma}}^2) - \mu_T^2, \quad b_T^2 = \min(\bar{b}_T^2, d_T^2),$$

$$\bar{b}_T^2 = \frac{1}{pT^2}\sum_{t=1}^{T}\left(\sum_{i=1}^{p} y_{it}^2\right)^2 - \frac{1}{pT}\mathrm{tr}(\widehat{\mathbf{\Sigma}}^2).$$

https://doi.org/10.1017/S0266466620000535 Published online by Cambridge University Press

It is of relevance and interest to see how linear shrinkage and its time-varying version compares to thresholding. Therefore, we implement the linear shrinkage estimator for both simulation and empirical exercise. For i.i.d. data, the linear shrinkage estimator has by now been superseded by the nonlinear shrinkage estimator of Ledoit and Wolf (2015). It is important to note, however, that the latter, unlike its linear counterpart, depends crucially on random matrix theoretical arguments. Therefore, it requires assuming that data are i.i.d. and p/T has a finite limit, both of which we consider as too restrictive for our purposes. While the analysis of the linear shrinkage estimator has also been carried out under similar assumptions, it is likely that a number of desirable properties of this estimator carry over to less strict settings than the i.i.d. one. In our empirical application, we also consider this type of shrinkage estimator.

In the literature, other popular methods for large covariance estimation have been proposed, which do not account explicitly for dependent data. For instance, Cai and Liu (2011) in a recent paper provide adaptive threshold estimators for the large covariance matrix $\mathbf{\Sigma}$. In their approach the threshold λ_{ij} depends on the i,jth entry of the matrix $\mathbf{\Sigma} = \text{var}(\boldsymbol{y}_t)$ as follows:

$$\lambda_{ij} = \delta\sqrt{\frac{\widehat{\theta}_{ij}\log p}{T}},\ \delta > 0, \tag{22}$$

$$\widehat{\theta_{ij}} = T^{-1}\sum_{t=1}^{T}\left(\left(y_{it}-\bar{y}_i\right)\left(y_{jt}-\bar{y}_j\right)-\widehat{\sigma}_{ij}\right)^2,\ \bar{y}_i = T^{-1}\sum_{j=1}^{T} y_{it}.$$

Now, the lower bound for an off diagonal element of $\mathbf{\Sigma}$ is not global, but depends on the variability of the individual point estimate of the i,j th element of the sample covariance matrix $\widehat{\mathbf{\Sigma}}$.

The adaptive thresholding approach of Cai and Liu is further extended in another stimulating paper by Fan et al. (2013). The authors introduce the Principal Orthogonal complement Thresholding (POET) estimator, to account for non sparsity, using an approximate factor model. This is done by combining the factor based covariance matrix estimator and the thresholding approach developed in Cai and Liu (2011). To this end, consider the approximate factor model for $\boldsymbol{y}_t$,

$$\boldsymbol{y}_t = \mathbf{b}\mathbf{f}_t + \mathbf{u}_t, \tag{23}$$

where $\mathbf{b}$ is the $p \times K$ matrix of factor loadings, $\mathbf{f}_t$ is the $K \times 1$ vector of unobserved factors, with $K << p$, and $\mathbf{u}_t$ is the idiosyncratic error which is uncorrelated with the factor $\mathbf{f}_t$. The factors $\mathbf{f}_t$ summarize information of the large vector $\boldsymbol{y}_t$. Model (23) implies that

$$\mathbf{\Sigma} = \boldsymbol{b}\mathbf{\Sigma}_f\mathbf{b}' + \mathbf{\Sigma}_u$$

where $\mathbf{\Sigma}_f$, $\mathbf{\Sigma}_u$ are defined as $\mathbf{\Sigma}_f = \text{var}(\mathbf{f}_t)$, and $\mathbf{\Sigma}_u = \text{var}(\mathbf{u}_t)$. When (23) is an approximate factor model, $\mathbf{\Sigma}_u$ is a non diagonal but sparse matrix. A natural

https://doi.org/10.1017/S0266466620000535 Published online by Cambridge University Press

estimator is defined as

$$\widehat{\mathbf{\Sigma}}_{poet} = \widehat{\mathbf{b}}\widehat{\mathbf{\Sigma}}_f\widehat{\mathbf{b}}' + T_{\lambda_{ij}}(\widehat{\mathbf{\Sigma}}_u)$$

with $\widehat{\mathbf{b}}, \widehat{\mathbf{f}}_t$ estimated by PCA, $\widehat{\mathbf{\Sigma}}_f = T^{-1}\widehat{\mathbf{f}}'\widehat{\mathbf{f}}$, $\widehat{\mathbf{\Sigma}}_u = T^{-1}\widehat{\mathbf{u}}'\widehat{\mathbf{u}}$, $\widehat{\mathbf{u}}_t = \boldsymbol{y}_t - \widehat{\mathbf{f}}_t\widehat{\mathbf{b}}$, and $T_{\lambda_{ij}}(\widehat{\mathbf{\Sigma}}_u)$ is the regularized estimate of $\widehat{\mathbf{\Sigma}}_u$, which is adaptively thresholded (see (22)). In practice, the number of factors K is chosen by information criteria, K_{opt}, as the one used in (34).

3.1. Cross Validation Methods

Estimation of large covariance matrices requires the use of a number of tuning parameters. All the estimators presented above can be generically denoted as $m_\gamma(\widehat{\mathbf{\Sigma}})$ and $m_\gamma(\widehat{\mathbf{\Sigma}}_t)$, where m_γ is some function and γ is a vector of tuning parameters, for example, ρ for the Ledoit and Wolf estimator, κ for the Bickel and Levina, (ρ, H) for time varying Ledoit and Wolf, (κ, H) or only κ for the time varying Bickel and Levina, and δ for the Cai and Liu and POET estimators. A popular approach to obtaining values for tuning parameters is via cross-validation. Cross-validation schemes are especially well-suited to the time-varying framework. We propose for this framework two objective functions that the optimal choice of parameter(s) γ minimize over a sensible parameter space:

$$Q^{(1)}_{T,\gamma} := \Big\| \frac{1}{T-T_a} \sum_{t=T_a+1}^{T} \big(m_\gamma(\widehat{\mathbf{\Sigma}}_{t|t-1}) - \boldsymbol{y}^0_t\boldsymbol{y}^{0\prime}_t\big)\big(m_\gamma(\widehat{\mathbf{\Sigma}}_{t|t-1}) - \boldsymbol{y}^0_t\boldsymbol{y}^{0\prime}_t\big)' \Big\|_F, \tag{24}$$

$$Q^{(2)}_{T,\gamma} := \Big\| \frac{1}{T-T_a} \sum_{t=T_a+1}^{T} \big(m_\gamma(\widehat{\mathbf{\Sigma}}_{t|t-1})^{-1/2}\boldsymbol{y}^0_t\boldsymbol{y}^{0\prime}_t m_\gamma(\widehat{\mathbf{\Sigma}}_{t|t-1})^{-1/2} - I\big) \tag{25}$$

$$\times \big(m_\gamma(\widehat{\mathbf{\Sigma}}_{t|t-1})^{-1/2}\boldsymbol{y}^0_t\boldsymbol{y}^{0\prime}_t m_\gamma(\widehat{\mathbf{\Sigma}}_{t|t-1})^{-1/2} - I\big)' \Big\|_F$$

where $\boldsymbol{y}^0_t = \boldsymbol{y}_t - \bar{\boldsymbol{y}}_t$ with $\bar{\boldsymbol{y}}_t$ as in (10). They are based on the estimate of $\mathrm{var}(\boldsymbol{y}_t|\boldsymbol{y}_{t-1})$, $\widehat{\mathbf{\Sigma}}_{t|t-1}$, defined as

$$\widehat{\mathbf{\Sigma}}_{t|t-1} = \Big(\sum_{j=1}^{t-1} b_{H,|j-t|}\Big)^{-1} \sum_{j=1}^{t-1} b_{H,|j-t|}\boldsymbol{y}_j\boldsymbol{y}'_j - \bar{\boldsymbol{y}}_{t|t-1}\bar{\boldsymbol{y}}'_{t|t-1},$$

$$\bar{\boldsymbol{y}}_{t|t-1} = \Big(\sum_{j=1}^{t-1} b_{H,|j-t|}\Big)^{-1} \sum_{k=1}^{t-1} b_{H,|t-k|}\boldsymbol{y}_k. \tag{26}$$

The optimal choice of γ is then given by $\widehat{\gamma} = \arg\min_\gamma Q^{(j)}_{T,\gamma}$ for $j = 1$ or 2. In (25), the regularized covariance $m_\gamma(\widehat{\mathbf{\Sigma}}_t)$ matrix is constrained to be positive definite. If $E\boldsymbol{y}_t = 0$ (or $E\boldsymbol{x}_t = 0$ in (17)), then $\boldsymbol{y}^0_t$ can be replaced by $\boldsymbol{y}_t$. For full sample methods, $m_\gamma(\mathbf{\Sigma})$, that do not account for time variation (i.e., no H parameter), to select γ, we use the same objective functions with $\widehat{\mathbf{\Sigma}}_{t|t-1}$ computed setting $b_{H,|j-t|} = 1$ in (26).

https://doi.org/10.1017/S0266466620000535 Published online by Cambridge University Press

The choice of the objective function depends on the empirical application at hand, while in different situations, the objectives (24) and (25) can give slightly different estimation results. In practice, one can use the (25) for choices of γ that yield a positive definite covariance matrix $m_\gamma(\widehat{\mathbf{\Sigma}}_t)$, while dropping all other alternatives, that do not guarantee this property. To ensure this property, the eigenvalues of $m_\gamma(\widehat{\mathbf{\Sigma}}_t)$ should always be positive. Giraitis, Kapetanios, and Price (2013) have shown that cross-validation methods can return tuning parameters that optimize the forecasting mean squared error in time-varying settings.

All tuning parameters are calibrated over a sensible parameter space. For the time varying estimator $\widehat{\mathbf{\Sigma}}_t$ we search for tuning parameter $H = T^h$, with h on the interval $[0.45, 1]$. For the thresholding methods, the parameter space depends on the data at hand. For instance, κ depends on the individual entries of the $\widehat{\mathbf{\Sigma}}_t$ and the specific thresholding approach that is considered. To this end, we focus on values of $\kappa \in [\kappa_{\min}, \kappa_{\max}]$, where $0 < \kappa_{\min}$, and $\kappa_{\max}$ is the minimum value that implies $m_\gamma(\widehat{\mathbf{\Sigma}}_t) \equiv \text{diag}(\widehat{\mathbf{\Sigma}}_t)$ for $H \in [T^{0.45}, T]$. In the adaptive thresholding approach, as suggested by Cai and Liu, we search for δ on $[0, 12]$. In this case, one theoretically reasonable value is $\delta_{opt} = 2$. This result relies on the fact the $\widehat{\theta}_{ij}$ (see (22)) provides an estimate of the variability of each individual entry of $\widehat{\mathbf{\Sigma}}_t$. The same parameter space is used for the POET estimator. For the Ledoit and Wolf methods, we search for ρ in the interval $[0, 1]$. We look for optimal values using a grid of 20 points for h, 100 points for κ, 50 points for δ, and 20 points for ρ in all the above parameter spaces. In the simulation experiments we use $Q^{(1)}_{T,\gamma}$ although results do not alter significantly when we use the $Q^{(2)}_{T,\gamma}$.

3.2. Monte Carlo Experiments

Following the theoretical exposition, we carry out a Monte Carlo study to explore the properties of estimation and forecasting of large deterministic and stochastic time-varying covariance matrices. We compare the performance of the estimates $\widehat{\mathbf{\Sigma}}$, $\widehat{\mathbf{\Sigma}}_t$, Ledoit-Wolf shrinkage estimates $\widehat{\mathbf{\Sigma}}_{LW}$, $\widehat{\mathbf{\Sigma}}_{LW,t}$ and Bickel–Levina thresholding estimates $T_\lambda(\widehat{\mathbf{\Sigma}})$, $T_\lambda(\widehat{\mathbf{\Sigma}}_t)$, $T_{\lambda_{opt}}(\widehat{\mathbf{\Sigma}}_t)$, Cai-Liu thresholding estimates $T_{\lambda_{ij}}(\widehat{\mathbf{\Sigma}})$ and POET estimators $\widehat{\mathbf{\Sigma}}_{poet}$.

Monte Carlo design. The generation of $\mathbf{\Sigma}_t$ presents particular challenges. We wish to have a reasonably realistic generation mechanism for $\mathbf{\Sigma}_t$ that corresponds to economic or financial data. Previous work on large dimensional covariance matrices of stationary processes offers little guidance since its designs for $\mathbf{\Sigma}$ with no time variation are rather simplistic for our purposes.

We choose to have a design for $\mathbf{\Sigma}_t$ based on the one factor setting used in Bailey, Kapetanios, and Pesaran (2016). It allows for varying degrees of sparsity $n_p \leq p$. We consider the data generating model

$$\boldsymbol{y}_t = (\mathbf{\Sigma}_t)^{1/2}\boldsymbol{e}_t, \quad t = 1, \ldots, T \tag{27}$$

https://doi.org/10.1017/S0266466620000535 Published online by Cambridge University Press

with two settings for $\mathbf{\Sigma}_t$: deterministic and random. We consider two different settings for $\boldsymbol{\varepsilon}_t$: in the first one $\boldsymbol{\varepsilon}_t$ is an i.i.d. p-dimensional standard normal variable while in the second $\boldsymbol{\varepsilon}_t$ is an i.i.d. p-dimensional random variable where each element follows a Student-$t(12)$ distribution with 12 degrees of freedom. As we consider relatively large values of p compared to T, we feel that our choice of the number of degrees of freedom is reasonable. Denote

$$\boldsymbol{V}_t = \text{diag}(\boldsymbol{e}_t) + \boldsymbol{b}_t \boldsymbol{b}_t' = \{v_{ij,t}\}$$

where $\mathbf{b}_t=(b_{1t}, b_{2t}, \ldots, b_{n_p t}, 0, \ldots, 0)'$, $\mathbf{e}_t=(e_{1t}, e_{2t}, \ldots, e_{pt})'$ are $p \times 1$ vectors. Then, we set $\mathbf{\Sigma}_t = \{v_{ij,t}/(v_{ii,1} v_{jj,1})^{1/2}\}$. We assume $e_{it} = h_{it} d_i$ where b_{it}, e_{it} and d_i will be specified below. This simulation design can be seen as a one factor model with time-varying factor loadings. Stock and Watson (2002a) argue that macroeconomic and/or financial time series do possess small instabilities, which can be amplified when these span a significant long time period, and the proposed simulation scheme is consistent with this idea.

We set $T = 400$ and consider three cases: $p = 10$ with sparsity parameter $3, 5, 10$, $p = 50$ with $n_p = 5, 20, 50$, and $p = 100$ with $n_p = 10, 40, 100$, respectively.

Deterministic $\mathbf{\Sigma}_t$. We consider two settings for deterministic time-varying covariance matrices $\mathbf{\Sigma}_t$ generated respectively by

$$b_{it} = 4 + 10(t/T), \quad h_{it} = 10 + 25(t/T), \tag{28}$$

$$b_{it} = 4 + 2\sin(2\pi t/T)(1 + 2t/T), \quad h_{it} = 10 + 2\sin(2\pi t/T)(1 + 2t/T). \tag{29}$$

In these settings, we set $d_i = 2$, $i = 1, ..., p$. The first one is a linear trend, while the second is a sine function of time. All settings provide considerable, smooth change for $\mathbf{\Sigma}_t$ over time.

Random $\mathbf{\Sigma}_t$. We consider one setting for a stochastic time-varying $\mathbf{\Sigma}_t$ which is generated by the following rescaled unit root processes:

$$b_{it} = \big(2.4|u_{it}/\sqrt{t}| + 0.04\big)\big(1 + 2t/T\big), \quad h_{it} = \big(9|u_{it}/\sqrt{t}| + 16\big)(1 + 2t/T), \tag{30}$$

where $(u_{i\cdot})$ are independent random walk processes: $u_{it} = u_{i,t-1} + \xi_{it}$ and ξ_{it} are i.i.d. $N(0, 1)$. In both settings, we assume that d_i are i.i.d. χ^2_2 random variables, and in (27) we assume $(\mathbf{\Sigma}_t)$ to be independent of $(\boldsymbol{\varepsilon}_t)$.

Finally, and for reference, we also consider a case where the covariance matrix $\mathbf{\Sigma}_t$ is actually constant and equal to the identity matrix. Notice that all the above considered simulation designs are in accordance with our theoretical formulations. For example the processes defined in (30) involve scaled random walk processes which satisfy Assumption H. Similar arguments can be used to verify that all the simulated processes we consider satisfy our theoretical assumptions. As expected, in such cases, forecast methods that allow for time-varying $\mathbf{\Sigma}_t$ outperform forecast methods based on non-time varying estimates of $\mathbf{\Sigma}$.

https://doi.org/10.1017/S0266466620000535 Published online by Cambridge University Press

3.3. Monte Carlo Results

In this section, we discuss performance of the one-step-ahead out-of-sample forecasts of $\mathbf{\Sigma}_T$ for various estimation methods, $m_\gamma(\mathbf{\Sigma}_t)$, of a time-varying covariance matrix $\mathbf{\Sigma}_t$ using several Monte Carlo experiments.

For the forecasting experiment, given the sample $\boldsymbol{y}_t$, $t = 1, ..., T$, we use the estimate of $\mathbf{\Sigma}_T$ as the forecast for $\mathbf{\Sigma}_{T+1}$. Formally, we define the one-step-ahead out-of-sample forecast of $\mathbf{\Sigma}_{T+1}$ based on estimation method $m_\gamma(\widehat{\mathbf{\Sigma}}_t)$ as $\widehat{\mathbf{\Sigma}}_{T+1|T} = m_\gamma(\widehat{\mathbf{\Sigma}}_{T,T})$. In our Monte Carlo experiments, we evaluate the forecast error using the Frobenius norm

$$\left\| \widehat{\mathbf{\Sigma}}_{T|T-1} - \mathbf{\Sigma}_T \right\|_F \tag{31}$$

and denote by frmse(m_γ) its Monte Carlo average over 500 replications. To compare the quality of forecasts based on different estimation methods, as a benchmark, we choose the standard sample covariance estimate $m_{bench} = \widehat{\mathbf{\Sigma}}$ which does not account for time variation. We report the relative frmse, frmse(m_γ)/frmse(m_{bench}). The smaller the latter is, the better is the performance of the method. Since in-sample forecasting of $\mathbf{\Sigma}_t$ by $\widehat{\mathbf{\Sigma}}_{t|t-1}$ reduces to estimation of $\mathbf{\Sigma}_{t-1}$ by $m_\gamma(\widehat{\mathbf{\Sigma}}_{t-1,t-1})$ using data $\boldsymbol{y}_t$, $t = 1, \ldots, t-1$, parameter γ can be selected using cross-validation.

For each replication the tuning parameter γ in (31) is estimated by cross-validation method using objective function (24) over the last 24 observations ($T_a = T - 24$ in (24)). We denote by subscript "cv" tuning parameters γ_{cv} obtained by cross-validation, set $H_{opt} = T^{2/3}$ and $H_{opt} = T^{1/2}$ in estimation of a deterministic and stochastic $\mathbf{\Sigma}_t$, respectively, whereas K_{opt} is as in Section 3, and ρ_{opt} as above.

Tables 1–7 present simulation results on the out-of-sample forecast error of large covariance matrices for various models of Σ_t and its estimation methods. They report the average relative frmse(m_γ)/frmse(m_{bench}), over 500 replications. Smaller numbers indicate superior performance. The best performing method is bolded and the second best is underlined.

Tables 1–6 provide comparisons of the quality of out-of-sample forecasting of a time-varying covariance matrix Σ_t based on the sample covariances $\widehat{\mathbf{\Sigma}}$, $\widehat{\mathbf{\Sigma}}_t$, and their regularized versions by Ledoit–Wolf shrinkage and Bickel–Levina thresholding methods. For the latter, the threshold λ is adapted to time variation via bandwidth H as described in (15) and (20), while tuning parameters are selected by cross-validation as described in Section 3.1. For comparison reasons we also include the Cai–Liu adaptive thresholding estimator and the POET estimator of $\mathbf{\Sigma}_t$. We use cross-validation to choose the tuning parameter, δ, in Cai–Liu adaptive thresholding estimator, as well as the theoretically optimal value ($\delta = 2$). In POET estimator we use information criteria to choose the optimal number of factors K denoted by K_{opt}.

https://doi.org/10.1017/S0266466620000535 Published online by Cambridge University Press

https://doi.org/10.1017/S0266466620000535 Published online by Cambridge University Press

TABLE 1. Relative frmse of one step ahead forecasts for deterministic time varying Σ_t. $T = 400$, Σ_t generated by (29): sine function with a drift, ε_t $iid N(0, 1)$

"dimension p"		10	10	10	50	50	50	100	100	100
"sparsity n_p"		3	5	10	5	20	50	10	40	100
Method	Tuning prmt	Relative frmse								
Forecasts based on non time-varying estimates of Σ										
Bickel–Levina	κ_{cv}	1	1	1.01	0.97	1	1	0.97	1	1
Cai–Liu	$\delta = 2$	0.99	1	1	0.94	0.99	1	0.93	0.99	1
	δ_{cv}	1	1	1	0.97	1	1	0.97	1	1
POET	κ_{cv}, K_{opt}	1	0.95	0.96	0.94	0.98	0.99	0.93	0.99	1
Ledoit–Wolf	ρ_{opt}	1.03	1.02	1.01	1.09	1.02	1.01	1.11	1.02	1.01
	ρ_{cv}	1.01	1	1.01	1	1	1.01	1	1	1
Forecasts based on time-varying estimates of Σ_t										
$\widehat{\Sigma}_t$	H_{cv}	0.83	0.71	**0.62**	1.36	0.78	**0.61**	1.41	0.79	**0.59**
Tv–Ledoit–Wolf	H_{cv}, ρ_{cv}	0.86	0.72	0.62	1.33	0.78	0.62	1.41	0.79	0.59
Tv–Bickel–Levina	κ_{cv}, H_{opt}	**0.71**	**0.69**	0.82	**0.81**	**0.77**	0.61	**0.81**	**0.88**	0.89
	κ_{cv}, H_{opt}	0.83	0.85	0.85	0.89	0.86	0.85	0.91	0.87	0.95

https://doi.org/10.1017/S0266466620000535 Published online by Cambridge University Press

TABLE 2. Relative frmse of one step ahead forecasts for deterministic time varying Σ_t. $T = 400$, Σ_t generated by (28): trend function, ε_t $iid N(0,1)$

"dimension p"		10	10	10	50	50	50	100	100	100
"sparsity n_p"		3	5	10	5	20	50	10	40	100
Method	Tuning prmt	Relative frmse								
Forecasts based on non time-varying estimates of Σ										
Bickel–Levina	κ_{cv}	1	1	1	0.97	1	1	0.96	1	1
Cai–Liu	$\delta = 2$	0.99	1	1	0.94	0.99	1	0.93	0.99	1
	δ_{cv}	1	1	1	0.97	1	1	0.96	1	1
POET	κ_{cv}, K_{opt}	1	0.95	0.97	0.94	0.98	0.99	0.93	0.99	1
Ledoit–Wolf	ρ_{opt}	1.03	1.02	1.01	1.09	1.02	1.01	1.11	1.02	1.01
	ρ_{cv}	1.01	1.01	1.01	1	1	1.01	1	1	1.01
Forecasts based on time-varying estimates of Σ_t										
$\widehat{\Sigma}_t$	H_{cv}	0.62	0.57	0.5	1.07	0.6	0.5	1.11	0.6	0.5
Tv–Ledoit–Wolf	ρ_{cv}, H_{cv}	0.68	0.61	0.52	1.09	0.61	0.51	1.12	0.61	0.52
Tv–Bickel–Levina	κ_{cv}, H_{cv}	0.55	0.55	0.59	0.59	0.58	0.6	0.52	0.55	0.59
	κ_{cv}, H_{opt}	**0.53**	**0.47**	**0.49**	**0.57**	**0.49**	**0.5**	**0.51**	**0.43**	**0.44**

https://doi.org/10.1017/S0266466620000535 Published online by Cambridge University Press

TABLE 3. Relative frmse of one step ahead forecasts for stochastic time varying Σ_t. $T = 400$, Σ_t stochastic generated by (30), ε_t $iid N(0,1)$

"dimension p"		10	10	10	50	50	50	100	100	100
"sparsity n_p"		3	5	10	5	20	50	10	40	100
Method	Tuning prmt	Relative frmse								
Forecasts based on non time-varying estimates of Σ										
Bickel–Levina	κ_{cv}	1	1	1.01	0.96	1	1	0.95	1	1
Cai–Liu	$\delta = 2$	0.99	1	1.01	0.93	1.01	1.03	0.91	1.02	1.03
	δ_{cv}	1	1	1.01	0.96	1	1	0.95	1	1
POET	κ_{cv}, K_{opt}	1.01	0.99	0.98	0.97	0.99	0.99	0.93	0.99	1
Ledoit–Wolf	ρ_{opt}	1.02	1.01	1.01	1.01	1.02	1.01	1.02	1.02	1.01
	ρ_{cv}	1.01	1	1	0.99	0.99	1	1	1	1
Forecasts based on time-varying estimates of Σ_t										
$\widehat{\Sigma}_t$	H_{cv}	0.66	0.5	**0.58**	1.13	0.61	**0.63**	1.23	0.74	**0.54**
Tv–Ledoit–Wolf	ρ_{cv}, H_{cv}	0.66	0.52	0.6	0.98	0.61	0.64	1.12	0.75	0.54
Tv–Bickel–Levina	κ_{cv}, H_{cv}	**0.61**	0.5	0.63	**0.70**	**0.60**	0.74	**0.77**	**0.7**	0.8
	κ_{cv}, H_{opt}	0.65	**0.48**	0.66	0.75	0.63	0.71	0.79	0.78	0.81

https://doi.org/10.1017/S0266466620000535 Published online by Cambridge University Press

TABLE 4. Relative frmse of one step ahead forecasts for deterministic time varying Σ_t. $T = 400$, Σ_t generated by (29): sine function with a drift, ε_t *iid* $t(12)$

"dimension p"		10	10	10	50	50	50	100	100	100
"sparsity n_p"		3	5	10	5	20	50	10	40	100
Method	Tuning prmt	Relative frmse								
Forecasts based on non time-varying estimates of Σ										
Bickel–Levina	κ_{cv}	0.99	1	1.03	0.93	1	1	0.91	0.99	1.02
Cai–Liu	$\delta = 2$	0.98	0.99	1	0.88	0.99	1	0.87	0.99	1
	δ_{cv}	0.99	1	1	0.94	1	1	0.92	0.99	1
POET	κ_{cv}, K_{opt}	1	0.93	0.95	0.89	0.97	0.99	0.88	0.98	0.99
Ledoit–Wolf	ρ_{opt}	1.04	1.03	1.01	1.12	1.03	1.01	1.13	1.03	1.01
	ρ_{cv}	1	1	1.03	1	1.01	1	1	1	1.02
Forecasts based on time-varying estimates of Σ_t										
$\widehat{\Sigma}_t$	H_{cv}	0.8	0.69	**0.52**	1.27	0.71	0.54	1.07	0.67	0.5
Tv–Ledoit–Wolf	ρ_{cv}, H_{cv}	0.79	0.7	0.58	1.26	0.72	0.56	1.07	0.69	0.54
Tv–Bickel–Levina	κ_{cv}, H_{cv}	0.71	0.65	0.61	0.76	0.61	0.55	0.59	0.55	0.53
	κ_{cv}, H_{opt}	**0.62**	**0.56**	0.57	**0.74**	**0.58**	**0.53**	**0.57**	**0.49**	**0.48**

https://doi.org/10.1017/S0266466620000535 Published online by Cambridge University Press

TABLE 5. Relative frmse of one step ahead forecasts for deterministic time varying Σ_t. $T = 400$, Σ_t generated by (28): trend function, ε_t *iid t*(12)

"dimension p"		10	10	10	50	50	50	100	100	100
"sparsity n_p"		3	5	10	5	20	50	10	40	100
Method	Tuning prmt	Relative frmse								
Forecasts based on non time-varying estimates of Σ										
Bickel–Levina	κ_{cv}	0.99	1	1	0.93	1	1	0.93	0.99	1
Cai–Liu	$\delta = 2$	0.98	0.99	1	**0.89**	0.99	1	**0.88**	0.99	1
	δ_{cv}	0.99	1	1	0.93	1	1	0.93	1	1
POET	κ_{cv}, K_{opt}	1	0.93	0.95	0.89	0.97	0.99	0.88	0.98	0.99
Ledoit–Wolf	ρ_{opt}	1.04	1.02	1.01	1.12	1.03	1.01	1.14	1.03	1.01
	ρ_{cv}	1	1	1.01	1	1	1.01	1	1	1
Forecasts based on time-varying estimates of Σ_t										
$\widehat{\Sigma}_t$	H_{cv}	1	0.88	**0.78**	1.18	0.94	**0.8**	1.15	0.96	**0.74**
Tv–Ledoit–Wolf	ρ_{cv}, H_{cv}	1.01	0.88	0.79	1.17	0.95	0.82	1.14	0.97	0.75
Tv–Bickel–Levina	κ_{cv}, H_{cv}	0.95	0.83	0.78	0.92	0.87	0.8	0.95	0.86	0.74
	κ_{cv}, H_{opt}	**0.86**	**0.78**	0.82	0.95	**0.85**	0.8	0.89	**0.84**	0.78

https://doi.org/10.1017/S0266466620000535 Published online by Cambridge University Press

TABLE 6. Relative frmse of one step ahead forecasts for stochastic time varying Σ_t. $T = 400$, Σ_t stochastic generated by (30), ε_t $iid\,t(12)$

"dimension p"		10	10	10	50	50	50	100	100	100
"sparsity n_p"		3	5	10	5	20	50	10	40	100
Method		Relative frmse								
Forecasts based on non time-varying estimates of Σ										
Bickel–Levina	κ_{cv}	0.99	1.02	1.01	0.93	1.02	1	0.97	1	1
Cai–Liu	$\delta = 2$	0.97	1	1.02	0.88	1.02	1.02	0.93	1.03	1.05
	δ_{cv}	0.98	1.01	1.01	0.93	1.01	1	0.96	1	1
POET	κ_{cv}, K_{opt}	1.03	0.99	0.97	0.95	0.97	1	0.96	0.98	1
Ledoit–Wolf	ρ_{opt}	1.02	1.01	1.02	0.98	1.03	1.01	1.01	1.03	1.01
	ρ_{cv}	1	1.03	1	1	1.01	0.99	0.99	1	1
Forecasts based on time-varying estimates of Σ_t										
$\widehat{\Sigma}_t$	H_{cv}	1.01	**0.73**	0.67	1.03	0.9	0.57	0.85	0.84	0.71
Tv–Ledoit–Wolf	ρ_{cv}, H_{cv}	1.01	0.76	**0.66**	1.04	0.9	**0.56**	0.84	0.84	**0.69**
Tv–Bickel–Levina	κ_{cv}, H_{cv}	**0.93**	0.73	0.73	**0.77**	**0.88**	0.59	**0.65**	**0.81**	0.72
	κ_{cv}, H_{opt}	1.19	0.79	0.82	0.9	1.04	0.58	0.7	0.88	0.73

Monte Carlo results, for deterministic Σ_t, reported in Tables 1 and 2, 4 and 5 and for stochastic Σ_t, reported in Tables 3 and 6 clearly show what to expect from the considered methods. In general, they indicate a significant impact from deterministic or stochastic change of Σ_t on the quality of forecasting of Σ_t by various methods.

A number of interesting conclusions can be drawn from the tables. It is unlikely that the time invariant sample covariance estimate $\widehat{\mathbf{\Sigma}}$ will produce good forecasts of $\mathbf{\Sigma}_t$ for all t. We use it as a benchmark. Clearly, estimation and forecasting need to be adapted to both time variation and sparsity of $\mathbf{\Sigma}_t$. The relative forecasting error, reported in the tables shows that regularization of the sample covariance estimate $\widehat{\mathbf{\Sigma}}$ by Ledoit–Wolf shrinkage, Bickel–Levina, Cai–Liu and POET thresholding methods is not effective. However, the time varying sample covariance estimate $\widehat{\mathbf{\Sigma}}_t$ seems to be able to adapt to time variation effectively: in some cases it significantly improves the quality of forecasting of $\mathbf{\Sigma}_t$. Regularization of $\widehat{\mathbf{\Sigma}}_t$ by Ledoit–Wolf shrinkage does not materially improve forecasting quality, while regularization of $\widehat{\mathbf{\Sigma}}_t$ by Bickel–Levina thresholding method significantly improves it in all experiments. As expected, improvement tends to be *slightly* stronger when $\mathbf{\Sigma}_t$ is sparse, however, the effect of time variation in $\mathbf{\Sigma}_t$ on forecasting quality is more severe than that of sparsity.

The thresholding procedure $T_{\lambda_{opt}}(\widehat{\mathbf{\Sigma}}_t)$ with λ_{opt} based on the theoretically optimal H_{opt} and cross-validated κ and thresholding procedure $T_{\lambda}(\widehat{\mathbf{\Sigma}}_t)$ where both κ and H are chosen using cross-validation, produce similar improvements of the forecasting quality. Hence, the choice of theoretically optimal H_{opt} given in Theorems 2 and 4, is a decent alternative to cross-validated H and the $T_{\lambda_{opt}}(\widehat{\mathbf{\Sigma}}_t)$ thresholding method can be recommended for empirical work.

Finally, in Table 7, we consider a base case scenario in which the true covariance matrix $\mathbf{\Sigma} = I$ is the identity matrix and does not vary with t. Now, the sample variance estimate $\widehat{\mathbf{\Sigma}}$ performs better, especially for small p. The methods that do not account for time variation, outperform the time varying methods, as expected, while the Ledoit–Wolf shrinkage estimator is the best performing method in all experiments. This is not surprising given that the Ledoit–Wolf method actually shrinks the estimate towards the identity matrix. As expected, our proposed time-varying forecasting methods suffer somewhat when p is small, while when p increases, they can deliver benefits over sample variance estimate $\widehat{\mathbf{\Sigma}}$. This indicates that the sparsity feature can be captured over the methods in this case too.

3.4. Empirical Application

In this section, we present an empirical demonstration of the potential gains from using the proposed covariance matrix estimators to design minimum variance portfolios. The illustration is motivated by the availability of large datasets on stock returns, and the increased demand for portfolios with lower risk exposure.

The literature on portfolio allocation is well grounded on the mean variance efficient portfolio frontier proposed by Markowitz (1952). The minimum variance

https://doi.org/10.1017/S0266466620000535 Published online by Cambridge University Press

TABLE 7. Relative frmse of one step ahead forecasts for fixed deterministic Σ. $T = 400$, $\Sigma = I$, ε_t $iid\,t(12)$

"dimension p"		10	50	100
"sparsity n_p"		1	1	1
Method	Tuning prmt	Relative frmse		
Forecasts based on non time-varying estimates of Σ				
Bickel–Levina	κ_{cv}	0.9	0.7	0.61
Cai–Liu	$\delta = 2$	0.78	0.47	0.35
	δ_{cv}	0.88	0.68	0.58
POET	κ_{cv}, K_{opt}	1.16	0.86	0.71
Ledoit–Wolf	ρ_{opt}	**0.7**	**0.42**	**0.31**
	ρ_{cv}	0.98	1	1
Forecasts based on time-varying estimates of Σ_t				
$\widehat{\Sigma}_t$	H_{cv}	1.29	1.15	1.08
Tv–Ledoit–Wolf	ρ_{cv}, H_{cv}	1.02	1.09	1.05
Tv–Bickel–Levina	κ_{cv}, H_{cv}	1.05	0.72	0.58
	κ_{cv}, H_{opt}	1.18	0.87	0.68

portfolio offers a suitable device for examining the possible superiority of portfolio weights based on the proposed large time-varying covariance estimators, as it neutralizes the impact of the estimated expected return parameters, focusing solely on the covariance matrix estimation (see e.g., Jagannathan and Ma (2003)).

Noting that $\mathbf{\Sigma}_t$ is the $p \times p$ covariance matrix of p-dimensional vector of returns, collected in $\boldsymbol{y}_t$, the minimum variance portfolio is designed to minimize investors' exposure to risk, regardless of preferences. The vector of optimal weights given by

$$w_t(\mathbf{\Sigma}_t) = \frac{\mathbf{\Sigma}_t^{-1} 1_p}{1_p' \mathbf{\Sigma}_t^{-1} 1_p},$$

where 1_p is a p-dimensional vector of ones. In practice, one has to choose an estimator for $\mathbf{\Sigma}_t$ and routinely compute the minimum variance weights. In our analysis, we compute portfolio weights, employing our large covariance matrix estimators, that is $\widehat{w}_t = w_t(\widehat{\mathbf{\Sigma}}_t)$.

Since our proposals involve sparse covariance estimators, we choose our sample to accommodate this assumption as much as possible. To this end, we obtained data from the Center for Research in Security Prices database and focus on daily returns from 10 different industry sectors. The sectors are: Consumer NonDurables, Utilities, Healthcare-Medical Equipment-Drugs, Telephone–Television Transmission, Business Equipment, Oil-Gas-Coal Extraction, Manufacturing, Consumer

https://doi.org/10.1017/S0266466620000535 Published online by Cambridge University Press

Durables, Consumer NonDurables, Other. The smallest portfolio examined is a 10 stocks portfolio, comprised of 1 stock from each of the 10 sectors. The 20 stocks portfolio is comprised of 2 stocks from each of the 10 sectors and so forth until the 100 stocks portfolio which is the largest portfolio examined. With this selection of stocks, our aim is to accommodate a block diagonal structure for the large covariance matrix of returns. The sample period starts at June 10, 2005 and ends at August 23, 2019 ($T = 3967$ daily observations).

To examine the out of sample risk performance, we will proceed as follows: every 5 days portfolio weights are selected as a function of the large covariance matrix estimate $m_\gamma(\widehat{\mathbf{\Sigma}}_t)$. These are kept constant for the subsequent five trading days while the corresponding portfolio returns are computed as $\boldsymbol{y}_s^{port} = \sum_{i=1}^{p} \widehat{w}_{i,s} y_{i,s}$ for $s = t+1, .., t+5$. Then the large covariance matrix is reestimated and the portfolio weights are updated (see Table 8 Panels A, B, C), or it is not reestimated and the porfolio weights remain unchanged (see Table 8, Panel D). The above steps are repeated until the end of the sample, that is for the last 775 trading days of our sample (period from July 28, 2016 until August 23, 2019). At the end, we use the 775 out of sample portfolio returns $\boldsymbol{y}_s^{port}, s = 1, .., 775$, to compute the associated variance

$$PortfolioVariance = \widehat{var(y^{port})}.$$

The method that provides the minimum out of sample portfolio variance is considered as the best performing one.

To accommodate our minimum risk objective we modify the cross-validation criteria developed in Section 3.1 as follows: we choose tuning parameter(s) γ, used to compute $\widehat{w}_t$, that minimize the objective function

$$Q_{t,\gamma}^{emp} := \frac{1}{80} \sum_{s=t-79}^{t} \left(\widehat{w}_s \mathbf{y}_s - \frac{1}{80} \sum_{s=t-79}^{t} \widehat{w}_s \mathbf{y}_s \right)^2$$

where $\boldsymbol{y}_t$ is the p-vector of portfolio returns at time t. The tuning parameter ranges and positive definiteness constraints discussed in Section 3.1 also apply here.

To account for nonsparse covariance matrices, we further extend our proposals following the approach developed in Fan et al. (2013). To this end, we first extract a number of factors from the portfolio returns and then regularize the remaining part by our proposed estimator. More formally, we consider the approximate factor model, for returns $\boldsymbol{y}_t$,

$$\boldsymbol{y}_t = \mathbf{b}\mathbf{f}_t + \mathbf{u}_t$$

where $\mathbf{b}$ is the $p \times K$ matrix of factor loadings, $\mathbf{f}_t$ is the $K \times 1$ vector of unobserved factors, with $K << p$, and $\mathbf{u}_t$ is the idiosyncratic error. We can then conclude that

$$\mathbf{\Sigma}_t = \mathbf{b}' \mathbf{\Sigma}_f \mathbf{b} + \mathbf{\Sigma}_{u,t} \tag{32}$$

https://doi.org/10.1017/S0266466620000535 Published online by Cambridge University Press

https://doi.org/10.1017/S0266466620000535 Published online by Cambridge University Press

TABLE 8. Performance of out of sample minimum variance portfolio forecast

"dimension p"		10	20	30	40	50	70	100
"dimension T"		3967	3967	3967	3967	3967	3967	3967
Method	Tuning prmt	Relative Variance						
Panel A: Forecasts based on non time-varying estimates of Σ (with reestimation)								
1/p		1.166	1.636	1.486	1.572	2.175	1.993	1.76
Rolling 6 months		0.96	1.075	0.927	0.931	0.82	1.014	1.953
Rolling 12 months		0.893	0.952	0.812	0.823	0.654	0.658	0.73
Bickel–Levina	κ_{cv}	0.934	0.978	0.975	0.961	1.047	1.014	1.019
Cai–Liu	$\kappa = 2$	1	0.999	0.993	1.01	1.041	1.196	1.238
	κ_{cv}	0.932	0.984	0.971	0.96	0.986	0.982	1.03
POET	κ_{cv}, K_{opt}	1.011	1.012	1.006	1.016	1.014	1.019	1.011
Ledoit–Wolf	ρ_{opt}	0.993	0.978	0.982	0.992	1.012	1.009	1.007
	ρ_{cv}	0.929	0.925	0.843	0.893	1.023	0.989	0.969
Ledoit–Wolf (nonlinear)		1	0.995	0.993	0.991	0.999	1.002	0.998
Panel B: Forecasts based on time-varying estimates of Σ_t								
TV	H_{cv}	0.919	0.959	0.814	**0.746**	0.625	0.637	0.771
TV–LW	H_{cv}, ρ_{cv}	0.879	0.959	0.822	0.799	0.659	0.619	0.599
Tv–Bickel–Levina^S	κ_{cv}, H_{cv}	0.877	0.981	0.877	0.831	0.616	0.605	**0.597**
	κ_{cv}, H_{opt}	0.936	1.073	0.994	0.929	0.829	0.958	0.951
Tv–Bickel–Levina–Factor^S	κ_{cv}, H_{cv}	1.026	1.144	1.057	1.041	0.924	0.844	0.824
	κ_{cv}, H_{opt}	1.397	1.301	1.232	1.182	0.992	0.979	0.947

(Continued)

https://doi.org/10.1017/S0266466620000535 Published online by Cambridge University Press

TABLE 8. (Continued)

"dimension p"		10	20	30	40	50	70	100
"dimension T"		3967	3967	3967	3967	3967	3967	3967
Tv–Bickel–Levina^D	κ_{cv}, H_{cv}	**0.87**	0.987	0.87	0.832	**0.61**	**0.604**	0.597
	κ_{cv}, H_{opt}	0.884	**0.899**	**0.755**	0.789	0.648	0.641	0.634
Tv–Bickel–Levina–Factor^D	κ_{cv}, H_{cv}	1.031	1.139	1.046	1.056	0.916	0.846	0.835
	κ_{cv}, H_{opt}	1.279	1.36	1.108	1.144	0.95	0.926	0.905
Panel C: Forecasts based on conditional covariance matrix estimates of $\Sigma_{t\|t-1}$								
RARCH		0.92	0.909	0.801	0.8	0.726	0.994	1
Panel D: Forecasts based on non time-varying estimates of Σ (no reestimation)								
Rolling 6 months		1.113	1.028	0.929	1.142	1.223	1.718	2.494
Rolling 12 months		0.966	0.938	0.964	1.376	1.1	1.197	1.215
Bickel–Levina	κ_{cv}	1.027	1.038	1.036	1.146	1.075	1.096	1.091
Cai–Liu	$\kappa = 2$	1.027	1.037	1.039	1.068	1.109	1.562	1.576
	κ_{cv}	1.128	1.077	0.995	1.097	1.125	1.152	1.092
POET	κ_{cv}, K_{opt}	1.027	1.05	1.065	1.055	1.131	1.163	1.124
Ledoit–Wolf	ρ_{opt}	1.018	1.008	1.029	1.047	1.09	1.104	1.098
	ρ_{cv}	0.948	0.942	0.833	0.894	1.192	1.109	1.038
Ledoit–Wolf (nonlinear)		1.028	1.032	1.042	1.049	1.078	1.101	1.09

Notes: Out of sample portfolio variance relative to the performance of the portfolio derived from the full sample covariance estimate. The method that provides the minimum out of sample risk exposure is marked in bold (blue). The second best is underlined (red). The examined period starts at 10-June-2005 and ends at August 23, 2019 ($T = 3967$ daily observations), and the out of sample evaluation period covers the last 775 observations of our sample. Data correspond to daily returns from 10 different industry sectors. The superscript "S" in "Tv–Bickel–Levina" and "Tv–Bickel–Levina–Factor" methods indicate thresholding a stochastic Σ_t using λ given in (20) while the superscript "D" corresponds to thresholding a deterministic Σ_t using λ as in (15).

where $\mathbf{\Sigma}_f$, $\mathbf{\Sigma}_{u,t}$ are defined as $\mathbf{\Sigma}_f = \text{var}(\mathbf{f}_t)$, and $\mathbf{\Sigma}_{u,t} = \text{var}(\mathbf{u}_t)$. Assuming that $\mathbf{\Sigma}_{u,t}$ is sparse and possesses significant time variation, a natural estimate of (32) is

$$\widehat{\mathbf{\Sigma}}_t = \widehat{\mathbf{b}}'\widehat{\mathbf{\Sigma}}_f\widehat{\mathbf{b}} + T_\lambda(\widehat{\mathbf{\Sigma}}_{u,t}) \tag{33}$$

with $\widehat{\mathbf{b}}$, $\widehat{\mathbf{f}}_t$ estimated by PCA, $\widehat{\mathbf{\Sigma}}_f = T^{-1}\widehat{\mathbf{f}}'\widehat{\mathbf{f}}$, and $T_\lambda(\widehat{\mathbf{\Sigma}}_{u,t})$ is the time varying regularized large covariance estimator of the residuals $\widehat{\mathbf{u}}_t = \boldsymbol{y}_t - \widehat{\mathbf{b}}\widehat{\mathbf{f}}_t$. In practice, the number of factors K can be chosen by information criteria (see e.g., Bai and Ng (2002)) or set equal to a fixed number. In our empirical exercise we estimate K according to

$$\widehat{K} = \arg\min_{0\le K\le M}\left\{\ln\left(\frac{1}{p}\sum_{i=1}^{p}\widehat{\sigma}_i^2\right) + K\left(\frac{p+T}{pT}\right)\ln\left(\frac{pT}{p+T}\right)\right\} \tag{34}$$

where $\widehat{\sigma}_i^2 = \frac{1}{T}\sum_{t=1}^{T}\widehat{\mathbf{u}}_{it}^2$, and $\widehat{\mathbf{u}}_{it}$ is the ith element of $\mathbf{u}_t$.

In addition to all the above mentioned models, we consider two more methods for estimating the large covariance matrices. The first model is due to Ledoit and Wolf (2015) who extend the previously discussed linear shrinkage estimator to the non linear case. To this end, let u be an eigenvector of $\mathbf{\Sigma} = \text{var}(\boldsymbol{y}_t)$ and $\widehat{\mathbf{\Sigma}}$ be an associated estimate of $\mathbf{\Sigma}$. The corresponding sample eigenvalue is equal to $u'\widehat{\mathbf{\Sigma}}u$. Nonlinear shrinkage replaces this quantity with a consistent estimator of $u'\mathbf{\Sigma}u$. Recovering the population eigenvalues from the sample eigenvalues requires inverting the Marchenko–Pastur equation (see Theorem 1 of Marchenko and Pastur (1968)). Ledoit and Wolf (2015) introduced an effective numerical method for inverting the Marchenko–Pastur equation.

The second model stems for the large conditional covariance modelling literature. Noureldin, Shephard, and Sheppard (2014) propose a model that allows estimation of flexible GARCH type dynamics in moderately large dimensions. They refer to it as the multivariate rotated ARCH model. The main idea of this approach is to undertake a transformation (in particular, a rotation) of the raw returns, and then use a BEKK-type parametrization of the time-varying covariance matrix. Inference is computationally attractive and based on the quasi-maximum likelihood (QML).

For the p-dimensional returns $\boldsymbol{y}_t = \mathbf{\Sigma}^{1/2}\varepsilon_t$, the unconditional covariance $\mathbf{\Sigma}$ is decomposed as $\mathbf{\Sigma} = P\Lambda P'$ where P is the matrix of eigenvectors and Λ is the matrix of non-negative eigenvalues. Then, since $\varepsilon_t = P(\Lambda)^{-1/2}P'y_t$, with $\text{var}(\varepsilon_t) = I_p$, the conditional variance of ε_t can be modelled as a BEKK type parametrization (see Engle and Kroner (1995))

$$\text{var}(\varepsilon_t|\varepsilon_{t-1}) = G_{t|t-1} = (I_p - AA' - BB') + A\varepsilon_{t-1}\varepsilon_{t-1}'A' + BG_{t-1|t-1}B', \quad G_0 = I_d. \tag{35}$$

In our empirical application we apply the scalar specification of (35) that assumes $A = a^{1/2}I_p$ and $B = b^{1/2}I_p$.

Finally, we also include in our analysis a number of portfolios which are commonly considered in the portfolio selection literature. The first is the equally

https://doi.org/10.1017/S0266466620000535 Published online by Cambridge University Press

weighted portfolio $\widehat{w}_t = \mathbf{1}/p$, where $\mathbf{1}$ is a vector of ones of dimension p. The second is the rolling sample, covariance matrix estimate based portfolio. We consider rolling sample estimates of size equal to 6 and 12 months.

In Table 8, we report out of sample portfolio variance results, as a ratio to the benchmark portfolio which is the full sample estimate, reestimated every 5 days. From Table 8, we can draw some important conclusions about the performance of our approaches. First, our study suggests that there are significant advantages from using the proposed time-varying covariance estimators to derive minimum variance portfolios. In all large portfolios considered, the proposed time-varying methodologies (see Panel B) outperform the full sample fixed ones (see Panel A and D). This is true when we estimate the fixed large covariance only once, using data up to July 28, 2016 (see Panel D) or when we reestimate them every 5 days in our out of sample evaluation period (see Panel A).

The unregularized time-varying estimator $\widehat{\mathbf{\Sigma}}_t$, seems to perform well, while its performance is affected by the size of the portfolio p as one would expect. For the proposed regularized estimator, the results remain satisfactory for both H selected by cross-validation and fixed H_{opt}, indicating the effectiveness of the derived theoretical value, H_{opt}. In this case, the deterministic estimator seems to outperform the stochastic one, providing larger risk improvements.

When we adjust our approach for non sparse covariance matrices, through the factor decomposition mechanism given in (33), the designed portfolios perform better than the fixed estimators (Panel A and D), especially when p is large. Having said that, it is important to emphasize that nonsparse versions of our proposals deteriorate, compared to their sparse analogues, indicating that the fixed component of estimator (33) captured by the factor part dominates the sparse time varying component, while the latter proves more important for minimum risk portfolios. On the other hand, time varying regularization through shrinkage (TV–LW), proves important as this method is among the top performing ones. Finally, comparing our methods with the large conditional covariance model RARCH (see Panel C), it becomes apparent that this outperforms the fixed estimators (Panel A and D), for small and medium p, while for $p = 100$ this becomes equivalent to the full sample estimate.

4. EXPONENTIAL INEQUALITIES

This section contains new results on Bernstein type inequalities for (weighted) sums of random variables (r.v.'s) (ξ_j) that are dependent, unbounded and have thin- or heavy-tailed distributions. We suppose that (ξ_j) satisfies the following α-mixing assumption.

Assumption A. (ξ_j) is such that $\xi_j - E\xi_j$ is an α-mixing (but not necessarily stationary) sequence with the mixing coefficients α_k such that for some $c_* > 0$ and $0 < \phi < 1$,

$$\alpha_k \leq c_* \phi^k, \quad k \geq 1. \tag{36}$$

https://doi.org/10.1017/S0266466620000535 Published online by Cambridge University Press

In addition, we assume that variables, (ξ_j), have thin- or heavy-tailed distributions.

The notation $(\xi_j) \in \mathcal{E}(s)$, $s > 0$ denotes thin tails and means that for some $a > 0$,

$$\max_j E \exp(a|\xi_j|^s) < \infty. \tag{37}$$

The notation $(\xi_j) \in \mathcal{H}(\theta)$, $\theta > 2$ corresponds to heavy tails and means that

$$\max_j E|\xi_j|^\theta < \infty. \tag{38}$$

Exponential inequalities for sums of r.v.'s, ξ_j with thin- and heavy-tailed distributions will be stated respectively using functions

$$f_t(\gamma_1, \gamma_2, c, \zeta) = c_0 \left\{ \exp\left(-c_1 \zeta^{\gamma_1}\right) + \exp\left(-c_2 \left(\frac{\zeta \sqrt{t}}{\log^2 t} \right)^{\gamma_2} \right) \right\}, \quad \zeta > 0,\ t \geq 2, \tag{39}$$

$$g_t(\gamma_1, \theta, c, \zeta) = c_0 \left\{ \exp(-c_1 \zeta^{\gamma}) + \zeta^{-\theta} t^{-(\theta/2-1)} \right\},$$

where $\gamma_1 > 0$, $\gamma_2 > 0$, $\theta > 2$ and non-negative constants $c = (c_0, c_1, c_2)$ do not depend on ζ, t.

Throughout the paper, we denote $a \vee b = \max(a, b)$ and $a \wedge b = \min(a, b)$, while C stands for generic constants.

4.1. Exponential Inequalities for Unbounded Variables

First, we establish Bernstein type inequalities for sums

$$S_T = T^{-1/2} \sum_{k=1}^{T} (\xi_k - E\xi_k)$$

of α-mixing and bounded variables ξ_k.

Lemma 1, (40), below significantly improves the bound for $P(S_T \geq \zeta)$ obtained in Theorem 3.5 of White and Wooldridge (1991). Its proof uses exponential inequalities for α-mixing bounded random variables obtained in Theorem 2 of Merlevede *et al.* (2009).

LEMMA 1. *Let the sequence (ξ_j) of r.v.'s satisfy Assumption A. Then, for all $\zeta > 0$, $T \geq 2$,*

$$P(|S_T| \geq \zeta) \leq \begin{cases} f_T(2, \gamma, c, \zeta) & \text{if } (\xi_j) \in \mathcal{E}(s), s > 0, \quad (40) \\ g_T(2, \theta', c, \zeta) & \text{if } (\xi_j) \in \mathcal{H}(\theta), \theta > 2 \quad (41) \end{cases}$$

with $\gamma = s/(s+1)$ and for any $2 < \theta' < \theta$ where c does not depend on ζ, T.

https://doi.org/10.1017/S0266466620000535 Published online by Cambridge University Press

4.2. Exponential Inequalities for Weighted Variables

Next, we obtain Bernstein type inequalities for sums

$$S_{T,t} := H^{-1/2} \sum_{k=1}^{T} b_{H,|t-k|}(\xi_k - E\xi_k) \tag{42}$$

of weighted α-mixing variables ξ_k with thin- or heavy-tailed distributions. The weights $b_{H,k}$ are defined in (12) and (13). Under the assumption (13), they satisfy

$$b_{H,k} \le C\big(1+(k/H)^{\nu}\big)^{-1}, \qquad |b_{H,k} - b_{H,k+1}| \le CH^{-1}\big(1+(k/H)^{\nu}\big)^{-1} \tag{43}$$

with $\nu > 3$. So, in (42), r.v.'s. ξ_k are strongly downweighted when k is distant from t.

In the next lemma we obtain exponential inequalities for $P(|S_{T,t}| \ge \zeta)$.

LEMMA 2. *Let (ξ_j) satisfy Assumption A and (43) holds.*
Then for all $\zeta > 0$, $1 \le t < T$,

$$P(|S_T| \ge \zeta) \le \begin{cases} f_H(2,\gamma,c,\zeta) & \text{if } (\xi_j) \in \mathcal{E}(s), s > 0, \quad (44) \\ g_H(2,\theta',c,\zeta) & \text{if } (\xi_j) \in \mathcal{H}(\theta), \theta > 2 \quad (45) \end{cases}$$

with $\gamma = s/(s+1)$ and for any $\theta' \in (2,\theta)$ where c does not depend on ζ, t, H, T.

To obtain exponential inequalities for sums

$$\widetilde{S}_{T,t} := H^{-1/2} \sum_{k=1}^{T} b_{H,|t-k|}(\xi_k - E\xi_t),$$

when the variables ξ_k are centered by $E\xi_t$, write

$$\widetilde{S}_{T,t} = S_{T,t} + r_{T,t}, \quad r_{T,t} := H^{-1/2} \sum_{k=1}^{T} b_{H,|t-k|}(E\xi_k - E\xi_t).$$

The next lemma provides bounds for $P(|\widetilde{S}_{T,t}| \ge \zeta)$ and $|r_{T,t}|$.

LEMMA 3. (a) *For any $|\zeta| > 2|r_{T,t}|$, $P\big(|\widetilde{S}_{T,t}| \ge \zeta\big) \le P\big(|S_{T,t}| \ge \zeta/2\big)$.*
(b) *If $|E\xi_k - E\xi_t| \le C|t-k|(t \vee k)^{-1}$ for $t,k = 1,2,\dots,$ then*

$$|r_{T,t}| \le C_1 H^{3/2}(H \vee t)^{-1}. \tag{46}$$

If $|E\xi_k - E\xi_t| \le C|t-k|T^{-1}$ for $t,k = 1,2,...,T,$ then

$$|r_{T,t}| \le C_1 H^{3/2} T^{-1}. \tag{47}$$

Constants C, C_1 do not depend on k, t, H, T.

https://doi.org/10.1017/S0266466620000535 Published online by Cambridge University Press

Exponential inequalities for sums $S_{T,t}$ allow to establish a bound for $\max_{t=1,..,T}|S_{T,t}|$ which is useful in applications. A bound for $\max_{t=1,..,T}|\widetilde{S}_{T,t}|$ can be obtained using

$$\max_t |\widetilde{S}_{T,t}| \le \max_t |S_{T,t}| + \max_t |r_{T,t}|.$$

COROLLARY 6. *Let (ξ_j) satisfy Assumption A and (43) hold. Assume that $(\xi_j) \in \mathcal{E}(s)$, $s > 0$ or $(\xi_j) \in \mathcal{H}(\theta)$, $\theta > 2$.*

(a) *Then, for any sequence $1 \le t = t_T \le T$, as $T \to \infty$,*

$$S_{T,t} = O_P(1). \tag{48}$$

(b) *If in addition,*

$$cT^{\delta} \le H \le T \quad \text{for some } c > 0, \delta > 0, \tag{49}$$

then for any $\varepsilon > 0$,

$$\max_{1\le t\le T}|S_{T,t}| = \begin{cases} O_P\big(\log^{1/2} T\big) & \text{if } (\xi_j) \in \mathcal{E}(s), s > 0, \quad (50) \\ O_P\big(\log^{1/2} T + (TH)^{1/\theta} H^{\varepsilon - 1/2}\big) & \text{if } (\xi_j) \in \mathcal{H}(\theta), \theta > 2. \quad (51) \end{cases}$$

Finally, we establish uniform bounds for sums of weighted variables $|\xi_k|$. Denote

$$v_{T,t} = H^{-1} \sum_{k=1}^{T} b_{H,|t-k|} \Big(\frac{|t-k|}{H}\Big)^{\nu} |\xi_k|, \quad \text{where } \nu \in [0,1].$$

COROLLARY 7. *Let (ξ_j) satisfy Assumption A and (43), (49) hold. Then,*

$$\max_{s=1,\dots,T} v_{T,s} = \begin{cases} O_P(1) & \text{if } (\xi_j) \in \mathcal{E}(s), s > 0, \quad (52) \\ O_P(1) & \text{if } (\xi_j) \in \mathcal{H}(\theta), \theta > 2 \text{ and } \delta > 1/(\theta - 1) \text{ in (49)}. \quad (53) \end{cases}$$

4.3. Exponential Inequalities with Random Scaling

In this section, we discuss exponential inequalities for sums

$$S_{T,t}^{(h)} := H^{-1/2} \sum_{k=1}^{T} b_{H,|t-k|} h_k (\xi_k - E\xi_k),$$

$$\widetilde{S}_{T,t}^{(h)} := H^{-1/2} \sum_{k=1}^{T} b_{H,|t-k|} (h_k \xi_k - h_t E\xi_t),$$

for products $h_k\xi_k$ of an α- mixing process (ξ_k) and a random scaling factor ("volatility") h_k. Differently from ARCH models, where h_k is a stationary process,

https://doi.org/10.1017/S0266466620000535 Published online by Cambridge University Press

here h_k is a persistent (nonstationary) process satisfying Assumption B below. It introduces smoothness and thin tail distribution restrictions on h_t, see Example 5. We impose no restrictions on the dependence between (h_k) and (ξ_k). We assume that (ξ_k) are α-mixing variables with thin- or heavy-tailed distributions.

Assumption B. (h_t) satisfies one of the following two assumptions:

$$|h_t - h_k| \le \begin{cases} \left(\dfrac{|t-k|}{t \vee k}\right)^{1/2} \xi_{tk}, \quad t,k = 1,2,\dots & \textbf{(54)} \\ \left(\dfrac{|t-k|}{T}\right)^{1/2} \xi_{tk}. \quad 1 \le t,k \le T & \textbf{(55)} \end{cases}$$

where (h_k) and (ξ_{tk}) are such that for some $\alpha > 0$,

$$(h_k) \in \mathcal{E}(\alpha), \quad (\xi_{tk}) \in \mathcal{E}(\alpha). \tag{56}$$

Set

$$d_{Ht} = \begin{cases} (t \vee H)^{1/2} H^{-1} & \text{if (54) holds,} \\ T^{1/2} H^{-1} & \text{if (55) holds.} \end{cases} \tag{57}$$

Denote $\zeta' = \zeta d_{Ht}$ for $\zeta > 0$. Notice that $\zeta \wedge \zeta' = (1 \wedge d_{Ht})\zeta$.

First, we establish exponential inequalities for sums where variables $h_k \xi_k$ are centered by $h_k E\xi_k$.

LEMMA 4. *Let (ξ_j) satisfy Assumption A, (h_j) satisfy Assumption B with $\alpha > 0$ and (43) holds.*

Then, there exists $c > 0$ such that for all $\zeta > 0$, $1 \le t, H \le T$, $T > 1$,

$$P(|S^{(h)}_{T,t}| \ge \zeta) \le \begin{cases} f_H\big(\gamma_1, \gamma_2, c, \zeta \wedge \zeta'\big) & \text{if } (\xi_j) \in \mathcal{E}(s), s > 0, \quad \textbf{(58)} \\ g_H\big(\gamma_1, \theta', c, \zeta \wedge \zeta'\big) & \text{if } (\xi_j) \in \mathcal{H}(\theta), \theta > 2, \quad \textbf{(59)} \end{cases}$$

with $\gamma_1 = 2\alpha/(2+\alpha)$, $\gamma_2 = \alpha s/(\alpha + s + 1)$ and for any $\theta' \in (2, \theta)$.

The same exponential inequalities hold for $\widetilde{S}^{(h)}_{T,t}$,
when $h_k \xi_k$ are centered by $h_t E\xi_t$. For that we need an additional assumption on $E\xi_k$.

Assumption B′. (h_k) satisfies Assumption B, and there exists $C < \infty$ such that for all t, k, T,

https://doi.org/10.1017/S0266466620000535 Published online by Cambridge University Press

$$|E\xi_t - E\xi_k| \le C \begin{cases} \left(\dfrac{|t-k|}{t \vee k}\right)^{1/2}, & t,k = 1,2,\dots \quad \text{if } (h_k) \text{ satisfies (54)}, \quad \textbf{(60)} \\ \left(\dfrac{|t-k|}{T}\right)^{1/2}, & 1 \le t,k \le T \quad \text{if } (h_k) \text{ satisfies (55)}. \quad \textbf{(61)} \end{cases}$$

LEMMA 5. *Let the assumptions of Lemma 4 hold and Assumption B′ be satisfied. Then*

$$P(|\widetilde{S}_{T,t}^{(h)}| \ge \zeta) \le \begin{cases} f_H\big(\gamma_1, \gamma_2, c, \zeta \wedge \zeta'\big) & \text{if } (\xi_j) \in \mathcal{E}(s), s > 0, \\ g_H\big(\gamma_1, \theta', c, \zeta \wedge \zeta'\big) & \text{if } (\xi_j) \in \mathcal{H}(\theta), \theta > 2, \end{cases}$$

where γ_1, γ_2 and θ' are the same as in Lemma 4.

Exponential inequalities allow to bound $S_{T,t}^{(h)}$ and $\max_{1\le s\le T} |\widetilde{S}_{T,s}^{(h)}|$.

COROLLARY 8. *Let Assumptions A and B with $\alpha > 0$ hold and (43) be satisfied. Assume that $(\xi_j) \in \mathcal{E}(s)$, $s > 0$ or $(\xi_j) \in \mathcal{H}(\theta)$, $\theta > 2$.*

(a) *Then, for any sequence $1 \le t = t_T \le T$, as $T \to \infty$,*

$$S_{T,t}^{(h)} = O_P(1 + d_{Ht}^{-1}). \quad \textbf{(62)}$$

In addition, if $E\xi_k$ satisfies Assumption B′ then $\widetilde{S}_{T,t}^{(h)} = O_P(1 + d_{Ht}^{-1})$.

(b) *Let h_k satisfy (55), $E\xi_k$ satisfy (61), and $cT^{\delta} \le H \le T$ where $c, \delta > 0$ do not depend on H, T. Then,*

$$\max_{1\le s\le T} |\widetilde{S}_{T,s}^{(h)}| = \begin{cases} O_P\big((1 + HT^{-1/2})(\log T)^{1/\gamma_1}\big) \\ \quad \text{if } (\xi_j) \in \mathcal{E}(s), s > 0, \quad \textbf{(63)} \\ O_P\left((1 + HT^{-1/2})\left\{(\log T)^{1/\gamma_1} + \dfrac{(TH)^{1/\theta}}{H^{-\varepsilon+1/2}}\right\}\right) \\ \quad \text{if } (\xi_j) \in \mathcal{H}(\theta), \theta > 2 \quad \textbf{(64)} \end{cases}$$

with $\gamma_1 = 2\alpha/(2+\alpha)$ and for any $\varepsilon > 0$.

The following auxiliary corollary provides bounds for sums of weighted variables $|h_k\xi_k|$. Denote

$$\Delta_{T,t} = H^{-1} \sum_{k=1}^{T} b_{H,|t-k|} \big|\beta_k - \beta_t\big| \, |h_k\xi_k|, \quad 1 \le t \le T.$$

To bound $\Delta_{T,t}$,we need an additional assumption on (β_k, h_k).

https://doi.org/10.1017/S0266466620000535 Published online by Cambridge University Press

Assumption C. (β_k, h_k) satisfy one of the following two assumptions:

$$|(\beta_k - \beta_t)h_k| \leq \begin{cases} C\dfrac{|t-k|}{T}, & 1 \leq t,k \leq T \quad (65) \\ \left(\dfrac{|t-k|}{T}\right)^{1/2} \nu_{tk} & 1 \leq t,k \leq T \quad (66) \end{cases}$$

where $C < \infty$ does not depend on t, k, T and $(\nu_{tk}) \in \mathcal{E}(\alpha)$ for some $\alpha > 0$.

COROLLARY 9. *Let Assumptions A, C and (43) be satisfied. Assume that* $(\xi_j) \in \mathcal{E}(s)$, $s > 0$ *or* $(\xi_j) \in \mathcal{H}(\theta)$, $\theta > 2$.

(a) *Then, for any* $1 \leq t = t_T \leq T$, *as* $T \to \infty$,

$$\Delta_{T,t} = \begin{cases} O_P\left(HT^{-1}\right) & \text{if (65) holds,} \quad (67) \\ O_P\left((H/T)^{1/2}\right) & \text{if (66) holds.} \quad (68) \end{cases}$$

(b) *Assume that* $cT^{\delta} \leq H \leq T$ *for some* $c, \delta > 0$, *and* $\delta > 1/(\theta - 1)$ *if* $(\xi_j) \in \mathcal{H}(\theta)$. *Then,*

$$\max_{1 \leq s \leq T} |\Delta_{T,s}| = \begin{cases} O_P(HT^{-1}) & \text{if (65) holds,} \quad (69) \\ O_P\left((H/T)^{1/2}(\log T)^{1/\alpha}\right) & \text{if (66) holds.} \quad (70) \end{cases}$$

5. CONCLUSION

The estimation of covariance matrices for large datasets has received considerable attention in recent years. Various regularization techniques for improved estimation of such matrices were developed, mainly for independent, identically distributed variables with exponentially declining probability tails. Dependence, heavy-tailed distributions and structural change are prevalent in large economic and financial data sets, and they may affect regularized estimation. The paper shows that the standard Bickel–Levina type thresholding procedure remains consistent for α-mixing stationary variables following various probability distributions. It takes a further step away from stationarity and allows for heteroscedasticity and stochastic change (volatility) of a very general form. It shows that the thresholding procedure can be aligned with kernel type estimates of time-varying covariances, and that the optimal threshold $\lambda_{opt} = \kappa\sqrt{(\log p)/H_{opt}}$ used in such a case is an intuitive adjustment for heteroscedasticity compared to the threshold, $\lambda_{opt} = \kappa\sqrt{(\log p)/T}$, that is appropriate for stationary data. The paper shows that the thresholding procedure, adjusted for heteroscedasticity, is robust to dependence and the type of distribution of the data, and its tuning parameters can be selected by cross-validation. Its finite sample performance, illustrated by a detailed Monte Carlo study and an empirical application on designing minimum variance

https://doi.org/10.1017/S0266466620000535 Published online by Cambridge University Press

portfolios, provides a clear rationale for the proposed theoretical methods. Finally, the paper derives Bernstein type exponential inequalities for weighted sums of dependent random variables with thin or heavy-tailed distributions, that are of independent theoretical interest.

SUPPLEMENTARY MATERIAL

To view the supplementary material for this article, please visit: https://doi.org/10.1017/S0266466620000535

REFERENCES

Abadir, K. M., W. Distaso, & F. Zikes (2014) Design-free estimation of variance matrices. *Journal of Econometrics* 181, 165–180.

Bai, J. & S. Ng (2002) Determining the number of factors in approximate factor models. *Econometrica* 70(1), 191–221.

Bailey, N., G. Kapetanios, & M. H. Pesaran (2016) Exponent of cross-sectional dependence: estimation and inference. *Journal of Applied Econometrics* 31, 929—960.

Bickel, M., Y. Wang, & H. Zhou (2013) Optimal sparse volatility matrix estimation for high-dimensional Itô processes with measurement errors. *The Annals of Statistics* 41, 1816–1864.

Bickel, P. & E. Levina (2008) Covariance regularization by thresholding. *The Annals of Statistics* 36, 2577–2604.

Cai, T. & W. Liu (2011) Adaptive thresholding for sparse covariance matrix estimation. *Journal of the American Statistical Association* 106, 672–684.

Chen, X., M. Xu, & W. B. Wu (2013) Covariance and precision matrix estimation for high-dimensional time series. *The Annals of Statistics* 41, 2994–3021.

Dahlhaus, R. (1997) Fitting time series models to nonstationary processes. *Annals of Statistics* 25, 1–37.

Engle, R. (1982) Autoregressive conditional heteroscedasticity with estimates of the variance of UK inflation. *Econometrica* 50, 987–1008.

Engle, R. F. & K. F. Kroner (1995) Multivariate simultaneous generalized Arch. *Econometric Theory* 11(1), 122–150.

Fan, J., Y. Liao, & H. Liu (2016) An overview of the estimation of large covariance and precision matrices. *Econometrics Journal* 19, C1–C32.

Fan, J., Y. Liao, & M. Mincheva (2013) Large covariance estimation by thresholding principal orthogonal complements. *Journal of the Royal Statistical Society: Series B* 75, 603–680.

Fan, J., W. Wang, & Y. Zhong (2016) Robust covariance estimation for approximate factor models. arXiv:1602.00719v1, pp. 1–31.

Giraitis, L., G. Kapetanios, & S. Price (2013) Adaptive forecasting in the presence of recent and ongoing structural change. *Journal of Econometrics* 177(2), 153–170.

Giraitis, L., G. Kapetanios, & T. Yates (2014) Inference on stochastic time-varying coefficient models. *Journal of Econometrics* 179, 46–65.

Giraitis, L., G. Kapetanios, & T. Yates (2018) Inference on multivariate heteroscedastic time varying random coefficient models. *Journal of Time Series Analysis* 39, 129–149.

Han, F., & H. Liu (2017) Statistical analysis of latent generalized correlation matrix estimation in transelliptical distribution. *Bernoulli* 23, 23–57.

Jagannathan, R. & T. Ma (2003) Risk reduction in large portfolios: why imposing the wrong constraints helps. *Journal of Finance* 58, 1651–1684.

Kapetanios, G. (2010) Testing for strict stationarity in financial variables. *Journal of Banking and Finance* 33, 2346–2362.

https://doi.org/10.1017/S0266466620000535 Published online by Cambridge University Press

Kolar, M. & E. Xing (2011) On time varying undirected graphs. *Proceedings of the 14th International Conference on Artificial Intelligence and Statistics (AISTATS) 2011 (JMLR)* 15, 407–415.

Ledoit, O. & M. Wolf (2003) Improved estimation of the covariance matrix of stock returns with an application to portfolio selection. *Journal of Empirical Finance* 10, 603–621.

Ledoit, O. & M. Wolf (2004) A well-conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis* 88, 365–411.

Ledoit, O. & M. Wolf (2015) Spectrum estimation: A unified framework for covariance matrix estimation and PCA in large dimensions. *Journal of Multivariate Analysis* 139, 360–384.

Marchenko, V. A. & L. A. Pastur (1968) Distribution of eigenvalues for some sets of random matrices. *Mathematics of the USSR, Sbornik* 1, 457–483.

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

Noureldin, D., N. Shephard, & K. Sheppard (2014) Multivariate rotated ARCH models. *Journal of Econometrics* 179(1), 16–30.

Pourahmadi, M. (2013) *High-Dimensional Covariance Estimation*. Hoboken: John Wiley.

Stock, J. H. & M. W. Watson (2002a) Forecasting using principal components from a large number of predictors. *Journal of the American Statistical Association* 97, 1167–1179.

Stock, J. H. & M. W. Watson (2002b) Macroeconomic forecasting using diffusion indexes. *Journal of Business & Economic Statistics* 20, 147–162.

Wegkamp, M. & Y. Zhao (2016) Adaptive estimation of the copula correlation matrix for semiparametric elliptical copulas. *Bernoulli* 22, 1184–1226.

White, H. & J. M. Wooldridge (1991) Some results on sieve estimation with dependent observations. In W. Barnett, J. Powell, & G. Tauchen (eds.), *Semiparametric and Nonparametric Methods in Econometrics and Statistics*, pp. 459–493. Cambridge University Press.

Zhou, S., J. Lafferty, & L. Wasserman (2010) Time varying undirected graphs. *Machine Learning* 80, 295–319.

https://doi.org/10.1017/S0266466620000535 Published online by Cambridge University Press