The Annals of Statistics
2016, Vol. 44, No. 3, 928–953
DOI: 10.1214/15-AOS1393

NONPARAMETRIC EIGENVALUE-REGULARIZED PRECISION OR COVARIANCE MATRIX ESTIMATOR

BY CLIFFORD LAM

London School of Economics and Political Science

We introduce nonparametric regularization of the eigenvalues of a sample covariance matrix through splitting of the data (NERCOME), and prove that NERCOME enjoys asymptotic optimal nonlinear shrinkage of eigenvalues with respect to the Frobenius norm. One advantage of NERCOME is its computational speed when the dimension is not too large. We prove that NERCOME is positive definite almost surely, as long as the true covariance matrix is so, even when the dimension is larger than the sample size. With respect to the Stein's loss function, the inverse of our estimator is asymptotically the optimal precision matrix estimator. Asymptotic efficiency loss is defined through comparison with an ideal estimator, which assumed the knowledge of the true covariance matrix. We show that the asymptotic efficiency loss of NERCOME is almost surely 0 with a suitable split location of the data. We also show that all the aforementioned optimality holds for data with a factor structure. Our method avoids the need to first estimate any unknowns from a factor model, and directly gives the covariance or precision matrix estimator, which can be useful when factor analysis is not the ultimate goal. We compare the performance of our estimators with other methods through extensive simulations and real data analysis.

1. Introduction. Thanks to the rapid development of computing power and storage in recent years, large data sets are becoming more readily available. Analysis of large data sets for knowledge discovery thus becomes more important in various fields. One basic and important input in data analysis is the covariance matrix or its inverse, called the precision matrix. For a large data set, the high dimensionality of the data adds much difficulty for estimating these matrices. One of the main difficulties is having an ill-conditioned sample covariance matrix when the dimension p is large relative to the sample size n [Ledoit and Wolf (2004)].

Realizing the ill-conditioned nature of the sample covariance matrix, many efforts are devoted to imposing special structures in estimating the covariance or the precision matrix. It is hoped that the true underlying matrices indeed conform to these structures, so that convergence results can be proved. Examples include a banded structure [Bickel and Levina (2008b)], sparseness of the covariance matrix [Bickel and Levina (2008a), Rothman, Levina and Zhu (2009), Lam and Fan

Received January 2015; revised September 2015.
MSC2010 subject classifications. Primary 62H12; secondary 62G20, 15B52.
Key words and phrases. High dimensional data analysis, covariance matrix, Stieltjes transform, data splitting, nonlinear shrinkage, factor model.

(2009), Cai and Zhou (2012)], sparseness of the precision matrix related to a graphical model [Meinshausen and Bühlmann (2006), Friedman, Hastie and Tibshirani (2008)], sparseness related to the modified Cholesky decomposition of a covariance matrix [Pourahmadi (2007)], a spiked covariance matrix from a factor model [Fan, Fan and Lv (2008), Fan, Liao and Mincheva (2011)] or combinations of them [see Fan, Liao and Mincheva (2013), e.g.].

Without a particular structure assumed, Stein (1975) and lecture 4 of Stein (1986) proposed the use of the class of rotation-equivariant estimators that retains the eigenvectors of the sample covariance matrix, but shrinks its eigenvalues. Hence, the smaller eigenvalues are made larger, while the larger eigenvalues are made smaller. This proposal indeed makes perfect sense, since Bai and Yin (1993) provided solid theoretical justification that, when the dimension p grows with the sample size n, the extreme eigenvalues of the sample covariance matrix are more extreme than the population counterparts. Shrinkage of the eigenvalues of a sample covariance matrix then becomes an important branch for covariance matrix estimation. Ledoit and Wolf (2004) proposed a well-conditioned covariance matrix estimator based on a weighted average of the identity and the sample covariance matrix. In effect, it shrinks the eigenvalues toward their grand mean. Won et al. (2013) proposed a condition number regularized estimator, which has the middle portion of the sample eigenvalues unchanged, and the more extreme eigenvalues are winsorized at certain constants. All these methods can be considered as a branch of eigenvalues-stabilizing covariance matrix estimator.

Recently, using random matrix theory, Ledoit and Wolf (2012) proposed a class of rotation-equivariant covariance estimator with nonlinear shrinkage of the eigenvalues. They demonstrate great finite sample performance of their estimator in difficult settings. Ledoit and Wolf (2013b) extends their results to allow for $p > n$, and proposes the estimation of the spectrum of a covariance matrix. At the same time, an independent idea of regularization is proposed in Abadir, Distaso and Žikeš (2014) for a class of rotation-equivariant estimator, where the data is split into two, and the eigenvalues are regularized by utilizing the two independent sets of split data. However, their theoretical analysis is based on the assumption that the dimension of the covariance matrix is fixed, or growing slower than the sample size.

In this paper, we investigate the theoretical properties of the regularized eigenvalues in Abadir, Distaso and Žikeš (2014). As a first contribution, we show that these eigenvalues are in fact asymptotically the same as those nonlinearly shrunk ones in Ledoit and Wolf (2012), when the observations $\mathbf{y}_i$ can be written as $\mathbf{y}_i = \mathbf{\Sigma}_p^{1/2}\mathbf{z}_i$, with $\mathbf{\Sigma}_p$ being the true covariance matrix, and $\mathbf{z}_i$ a vector of independent and identically distributed entries [see Section 2 and Assumption (A1) therein for more details]. The method for estimating the nonlinearly shrunk eigenvalues in Ledoit and Wolf (2012) involves nonconvex optimizations, and is solved via a commercial package. On the other hand, the method in Abadir, Distaso and

Žikeš (2014) involves only eigen-decompositions of $p \times p$ matrices, and can be much quicker to do when p is not too large, while commercial packages are not required. Although recently the code for Ledoit and Wolf (2012) is updated to interior point algorithm which does not require a commercial package, we tested it to be much slower than the one using a commercial package. The speed of our method can still be particularly attractive for practitioners.

As a second contribution, we also show that, if the data is from a factor model, while the low dimensional factor can render the nonlinear shrinkage formula in Ledoit and Wolf (2012) incorrect, the regularized eigenvalues using the data splitting idea from Abadir, Distaso and Žikeš (2014) are still asymptotically optimal when we consider minimizing the Frobenius loss for estimating the covariance matrix, or the inverse Stein's loss for estimating the precision matrix (see Sections 2.3, 2.4 and 3 for more details). It means that, in estimating the covariance matrix for data from a factor model, we do not need to explicitly estimate the number of factors, the factor loading matrix and the unknown factor series, which can be difficult tasks themselves. This can be particularly important when factor analysis is not the final goal, and estimating the covariance or precision matrix is just an intermediate step. Section 5.1 demonstrates this with a simulation involving a real macroeconomic data set.

As a third contribution, we define a notion of efficiency for our estimator through comparing it with an "ideal" one, which is also rotation-equivariant itself and assumed the knowledge of the true covariance matrix. We prove that, almost surely, our estimator does not lose any efficiency compared to such an ideal estimator asymptotically. We also show that our estimator is almost surely positive definite. All these proofs are under the high dimensional setting $p/n \to c > 0$, where n is the sample size. This provides an important relaxation to the assumptions in Abadir, Distaso and Žikeš (2014), where p is either fixed or $p/n \to 0$. See Section 4 for more details.

Certainly, under the high dimensional setting $p/n \to c > 0$, the eigenvectors of a rotation-equivariant estimator that based on sample covariance eigenvectors are not converging to the population ones. It means that if the covariance matrix is estimated for the purpose of principal component analysis, for instance, the estimated principal directions are not converging to what we want. Hence our covariance estimator should be applied when the eigenvectors themselves are not of primary interests, but it can be envisaged that stabilization of eigenvalues of the sample covariance matrix can lead to better performance in a data analysis, in particular when the covariance matrix itself is not the ultimate aim or the structure of the population covariance matrix is not surely known. Sections 5.1, 5.2 and 5.3 present three data examples to show that our proposed covariance estimator helps in various tasks. NERCOME even ultimately outperforms methods which exploit the approximate structure of the population covariance matrix in Sections 5.1 and 5.3.

The rest of the paper is organized as follows. We first present the framework for the data together with the notation and the main assumptions to be used in Section 2. We then present the generalization to the Marčenko–Pastur law proposed

in Ledoit and Péché (2011) in Section 2.1, followed by the sample splitting regularization idea in Abadir, Distaso and Žikeš (2014) in Section 2.2. Together, these inspire us to propose our covariance matrix estimator in Section 2.3, where we also present a theorem to show that our estimator is asymptotically optimal in a certain sense. We also show that our precision matrix estimator is asymptotically optimal with respect to the inverse Stein's loss function in Section 2.4. Extension of our results to data from a factor model is introduced in Section 3. Efficiency loss of an estimator is defined in Section 4, with asymptotic efficiency of our estimator shown. Improvement to finite sample properties, and the choice of the split of data for sample splitting are discussed in Section 4. Simulation results comparing with the performance of other state-of-the-art methods is given in the supplementary material [Lam (2015)]. Computing time of different methods are given in Section 5. Three real data analysis are given in Sections 5.1, 5.2 and 5.3, respectively. Finally, the conclusion is given in Section 6. All of the proofs in our results are given in the supplementary material [Lam (2015)].

2. The framework and overview of relevant results. Let $\mathbf{y}_i$, $i = 1, \ldots, n$ be an independent and identically distributed sample that we can observe, where each $\mathbf{y}_i$ is of length p. In this paper, we assume that $p = p_n$, such that $p/n \to c > 0$ as $n \to \infty$. The subscript n will be dropped if no ambiguity arises.

We let $\mathbf{Y} = (\mathbf{y}_1, \ldots, \mathbf{y}_n)$, and define the sample covariance matrix as

$$\mathbf{S}_n = \frac{1}{n}\mathbf{Y}\mathbf{Y}^{\mathrm{T}}. \tag{2.1}$$

We assume $\mathbf{y}_i$ has mean $\mathbf{0}$ (the vector of zeros with suitable size), and the covariance matrix is denoted as $\mathbf{\Sigma}_p = E(\mathbf{y}_i\mathbf{y}_i^{\mathrm{T}})$ for each i. We assume the following for Theorem 1 in Section 2.3 to hold. Related to asymptotic efficiency of our estimator, Theorem 5 in Section 4 has more restrictive moment assumptions. See Section 4 for more details.

(A1) Each observation can be written as $\mathbf{y}_i = \mathbf{\Sigma}_p^{1/2}\mathbf{z}_i$ for $i = 1, \ldots, n$, where each $\mathbf{z}_i$ is a $p \times 1$ vector of independent and identically distributed random variables z_{ij}. Each z_{ij} has mean 0 and unit variance, and $E|z_{ij}|^{2k} = O(p^{k/2-1})$, $k = 2, 3, 4, 5, 6$.

(A2) The population covariance matrix $\mathbf{\Sigma}_p$ is non-random and of size $p \times p$. Furthermore, $\|\mathbf{\Sigma}_p\| = O(p^{1/2})$, where $\|\cdot\|$ is the L_2 norm of a matrix.

(A3) Let $\tau_{n,1} \geq \cdots \geq \tau_{n,p}$ be the p eigenvalues of $\mathbf{\Sigma}_p$, with corresponding eigenvectors $\mathbf{v}_{n,1}, \ldots, \mathbf{v}_{n,p}$. Define $H_n(\tau) = p^{-1}\sum_{i=1}^{p} \mathbf{1}_{\{\tau_{n,i} \leq \tau\}}$ the empirical distribution function (e.d.f.) of the population eigenvalues, where $\mathbf{1}_A$ is the indicator function of the set A. We assume $H_n(\tau)$ converges to some non-random limit H at every point of continuity of H.

(A4) The support of H defined above is the union of a finite number of compact intervals bounded away from zero and infinity. Also, there exists a compact interval in $(0, +\infty)$ that contains the support of H_n for each n.

These four assumptions are very similar to Assumptions A1 to A4 in Ledoit and Wolf (2012) and (H_1) to (H_4) in Ledoit and Péché (2011). Instead of setting $E|z_{ij}|^{12} \le B$ for some constant B as needed in Assumption A1 in Ledoit and Wolf (2012), we have relaxed this, and only need $E|z_{ij}|^{2k} = O(p^{k/2-1})$, $k = 2, 3, 4, 5, 6$ in Assumption (A1). It means that we require the fourth-order moments to be bounded asymptotically, but the higher order moments (up to 12th order) can diverge to infinity as $n, p \to \infty$.

2.1. *The Marčenko–Pastur law and its generalization.* We introduce further notation and definitions in order to present the Marčenko–Pastur law and its generalizations. The Stieltjes transform of a nondecreasing function G is given by

$$m_G(z) = \int_{\mathbb{R}} \frac{1}{\lambda - z} G(\lambda), \qquad \text{for any } z \in \mathbb{C}^+. \tag{2.2}$$

Here, $\mathbb{C}^+$ represents the upper half of the complex plane, where the imaginary part of any complex numbers are strictly positive. If G is continuous at a and b, then the following inversion formula for the Stieltjes transform holds:

$$G(b) - G(a) = \lim_{\eta \to 0^+} \frac{1}{\pi} \int_a^b \operatorname{Im}\bigl[m_G(\lambda + i\eta)\bigr] d\lambda, \tag{2.3}$$

where $\operatorname{Im}(z)$ denotes the imaginary part of a complex number.

Suppose the eigen-decomposition of the sample covariance matrix in (2.1) is $\mathbf{S}_n = \mathbf{P}_n \mathbf{D}_n \mathbf{P}_n^{\mathrm{T}}$, where

$$\mathbf{P}_n = (\mathbf{p}_{n,1}, \ldots, \mathbf{p}_{n,p}), \qquad \mathbf{D}_n = \operatorname{diag}(\lambda_{n,1}, \ldots, \lambda_{n,p}), \tag{2.4}$$

so that $\lambda_{n,1} \ge \cdots \ge \lambda_{n,p}$ are the eigenvalues of $\mathbf{S}_n$ with corresponding eigenvectors $\mathbf{p}_{n,1}, \ldots, \mathbf{p}_{n,p}$. The notation $\operatorname{diag}(\cdot)$ denotes a diagonal matrix with the listed entries as the diagonal. We then define the e.d.f. of the sample eigenvalues in (2.4) as $F_p(\lambda) = p^{-1} \sum_{i=1}^p \mathbf{1}_{\{\lambda_{n,i} \le \lambda\}}$. Using (2.2), the Stieltjes transform of F_p will then be (we suppress the subscript n in $\lambda_{n,i}$ and $\mathbf{p}_{n,i}$ hereafter if no ambiguity arises)

$$m_{F_p}(z) = \frac{1}{p} \sum_{i=1}^p \frac{1}{\lambda_i - z} = p^{-1} \operatorname{tr}\bigl[(\mathbf{S}_n - z\mathbf{I}_p)^{-1}\bigr], \qquad z \in \mathbb{C}^+,$$

where $\operatorname{tr}(\cdot)$ is the trace of a square matrix, and $\mathbf{I}_p$ denotes the identity matrix of size p. It is proved in Marčenko and Pastur (1967) [see Marčenko and Pastur (1967) for the exact assumptions used] that F_p converges almost surely (abbreviated as a.s. hereafter) to some non-random limit F at all points of continuity of F. They also discovered the famous equation, later called the Marčenko–Pastur equation, which can be expressed as

$$m_F(z) = \int_{\mathbb{R}} \frac{1}{\tau[1 - c - czm_F(z)] - z} \, dH(\tau), \qquad \text{for any } z \in \mathbb{C}^+. \tag{2.5}$$

For the generalization of (2.5), consider $\Theta_p^g(z) = p^{-1}\operatorname{tr}[(\mathbf{S}_n - z\mathbf{I}_p)^{-1}g(\mathbf{\Sigma}_p)]$, where $g(\cdot)$ is in fact a scalar function on the eigenvalues of a matrix, such that if $\mathbf{\Sigma}_p = \mathbf{V}\operatorname{diag}(\tau_1, \ldots, \tau_p)\mathbf{V}^{\mathrm{T}}$, then

$$g(\mathbf{\Sigma}_p) = \mathbf{V}\operatorname{diag}\big(g(\tau_1), \ldots, g(\tau_p)\big)\mathbf{V}^{\mathrm{T}}.$$

Hence, $m_{F_p}(z) = \Theta_p^g(z)$ with $g \equiv 1$. In Theorem 2 of Ledoit and Péché (2011), under Assumptions (H_1) to (H_4) in their paper, it is proved that $\Theta_p^g(z)$ converges a.s. to $\Theta^g(z)$ for any $z \in \mathbb{C}^+$, where

$$\Theta^g(z) = \int_{\mathbb{R}} \frac{1}{\tau[1 - c - czm_F(z)] - z} g(\tau)\, dH(\tau), \qquad z \in \mathbb{C}^+. \tag{2.6}$$

By taking $g = Id$, the identity function, the inverse Stieltjes transform of $\Theta_p^g(z)$, denoted by $\Delta_p(x)$, is given by

$$\Delta_p(x) = \frac{1}{p}\sum_{i=1}^{p} \mathbf{p}_i^{\mathrm{T}}\mathbf{\Sigma}_p\mathbf{p}_i \mathbf{1}_{\{\lambda_i \le x\}}, \qquad x \in \mathbb{R},$$

on all points of continuity of $\Delta_p(x)$. In Theorem 4 of Ledoit and Péché (2011), using also Assumptions (H_1) to (H_4) in their paper, it is proved using (2.6) that $\Delta_p(x)$ converges a.s. to $\Delta(x) = \int_{-\infty}^{x} \delta(\lambda)\, dF(\lambda)$ for $x \in \mathbb{R}/\{0\}$, the inverse Stieltjes transform of $\Theta^{Id}(z)$. To present the function $\delta(\lambda)$, let $\breve{m}_F(\lambda) = \lim_{z \in \mathbb{C}^+ \to \lambda} m_F(z)$ for any $\lambda \in \mathbb{R}/\{0\}$, and $\underline{F}(\lambda) = (1 - c)\mathbf{1}_{\{\lambda \ge 0\}} + cF(\lambda)$ when $c > 1$, with $\breve{m}_{\underline{F}}(\lambda) = \lim_{z \in \mathbb{C}^+ \to \lambda} m_{\underline{F}}(z)$ for any $\lambda \in \mathbb{R}$. The quantities $\breve{m}_F(\lambda)$ and $\breve{m}_{\underline{F}}(\lambda)$ are shown to exist in Silverstein and Choi (1995). We then have for any $\lambda \in \mathbb{R}$,

$$\delta(\lambda) = \begin{cases} \dfrac{\lambda}{|1 - c - c\lambda\breve{m}_F(\lambda)|^2}, & \text{if } \lambda > 0; \\ \dfrac{1}{(c-1)\breve{m}_{\underline{F}}(0)}, & \text{if } \lambda = 0 \text{ and } c > 1; \\ 0, & \text{otherwise.} \end{cases} \tag{2.7}$$

This result means that the asymptotic quantity that corresponds to $\mathbf{p}_i^{\mathrm{T}}\mathbf{\Sigma}_p\mathbf{p}_i$ is $\delta(\lambda)$, provided that λ corresponds to λ_i, the ith largest eigenvalue of the sample covariance matrix $\mathbf{S}_n$. It means that we can calculate asymptotically the value of $\mathbf{p}_i^{\mathrm{T}}\mathbf{\Sigma}_p\mathbf{p}_i$ using the sample eigenvalue λ_i according to the nonlinear transformation $\delta(\lambda_i)$, which is an estimable quantity from the data. This is the basis of the nonlinear transformation of sample eigenvalues used in Ledoit and Wolf (2012). We shall come back to this result in Sections 2.3 and 2.4.

2.2. *Regularization by sample splitting.* In Abadir, Distaso and Žikeš (2014), the data $\mathbf{Y} = (\mathbf{y}_1, \ldots, \mathbf{y}_n)$ is split into two parts, say $\mathbf{Y} = (\mathbf{Y}_1, \mathbf{Y}_2)$, where $\mathbf{Y}_1$ has size $p \times m$ and $\mathbf{Y}_2$ has size $p \times (n - m)$. The sample covariance matrix for $\mathbf{Y}_i$

is calculated as $\widetilde{\mathbf{\Sigma}}_i = n_i^{-1}\mathbf{Y}_i\mathbf{Y}_i^{\mathrm{T}}$, $i = 1, 2$, where $n_1 = m$ and $n_2 = n - m$. They propose to estimate the covariance matrix by

$$\check{\mathbf{\Sigma}}_m = \mathbf{P}\,\mathrm{diag}(\mathbf{P}_1^{\mathrm{T}}\widetilde{\mathbf{\Sigma}}_2\mathbf{P}_1)\mathbf{P}^{\mathrm{T}}, \tag{2.8}$$

where $\mathbf{P} = \mathbf{P}_n$ as in (2.4), $\mathrm{diag}(A)$ denotes the diagonal matrix with diagonal entries as in A, and $\mathbf{P}_1$ is an orthogonal matrix such that $\widetilde{\mathbf{\Sigma}}_1 = \mathbf{P}_1\mathbf{D}_1\mathbf{P}_1^{\mathrm{T}}$. Their idea is to use the fact that $\mathbf{P}_1$ and $\widetilde{\mathbf{\Sigma}}_2$ are independent to regularize the eigenvalues. Writing $\mathbf{P}_1 = (\mathbf{p}_{11}, \mathbf{p}_{12}, \ldots, \mathbf{p}_{1p})$, the diagonal values in $\mathbf{P}_1^{\mathrm{T}}\widetilde{\mathbf{\Sigma}}_2\mathbf{P}_1$ are $\mathbf{p}_{1i}^{\mathrm{T}}\widetilde{\mathbf{\Sigma}}_2\mathbf{p}_{1i}$, $i = 1, \ldots, p$, and they become the eigenvalues of $\check{\mathbf{\Sigma}}_m$ in (2.8).

On the other hand, in light of (2.7) and the descriptions thereafter, $\mathbf{p}_{1i}^{\mathrm{T}}\mathbf{\Sigma}_p\mathbf{p}_{1i}$ will have the asymptotic nonlinear transformation

$$\delta_1(\lambda) = \begin{cases} \dfrac{\lambda}{|1 - c_1 - c_1\lambda\check{m}_{F_1}(\lambda)|^2}, & \text{if } \lambda > 0; \\ \dfrac{1}{(c_1 - 1)\check{m}_{\underline{F_1}}(0)}, & \text{if } \lambda = 0 \text{ and } c_1 > 1; \\ 0, & \text{otherwise,} \end{cases} \tag{2.9}$$

where $c_1 > 0$ is a constant such that $p/n_1 \to c_1$. The distribution function $F_1(\lambda)$ is the non-random limit of $F_{1p}(\lambda) = p^{-1}\sum_{i=1}^{p}\mathbf{1}_{\{\lambda_{1i}\le\lambda\}}$, with $\lambda_{11} \ge \lambda_{12} \ge \cdots \ge \lambda_{1p}$ being the eigenvalues of $\widetilde{\mathbf{\Sigma}}_1$. The quantities $\check{m}_{F_1}(\lambda)$ and $\check{m}_{\underline{F_1}}(0)$ are defined in parallel to those in (2.7). We show in Theorem 1 that the quantities $\mathbf{p}_{1i}^{\mathrm{T}}\widetilde{\mathbf{\Sigma}}_2\mathbf{p}_{1i}$ and $\mathbf{p}_{1i}^{\mathrm{T}}\mathbf{\Sigma}_p\mathbf{p}_{1i}$ are in fact asymptotically the same. Hence, they both correspond to $\delta_1(\lambda_{1i})$, with $\delta_1(\lambda)$ defined in (2.9). This forms the basis for the covariance matrix estimator in our paper, and will be explained in full detail in Section 2.3.

REMARK 1. If $n_1/n \to 1$, then p/n_1, p/n both go to the same limit $c_1 = c > 0$. Theorem 4.1 of Bai and Silverstein (2010) tells us then both F_p and F_{1p} converges to the same limit almost surely under Assumptions (A1) to (A4). That is, $F = F_1$ almost surely, and hence $\delta_1(\cdot) = \delta(\cdot)$. This implies that $\mathbf{p}_i^{\mathrm{T}}\mathbf{\Sigma}_p\mathbf{p}_i$ and $\mathbf{p}_{1i}^{\mathrm{T}}\mathbf{\Sigma}_p\mathbf{p}_{1i}$ are asymptotically almost surely the same.

2.3. *The covariance matrix estimator.* In this paper, we use the sample splitting idea in Section 2.2 and split the data into $\mathbf{Y} = (\mathbf{Y}_1, \mathbf{Y}_2)$, so that $\mathbf{Y}_1$ and $\mathbf{Y}_2$ are independent of each other by our assumption of independence. We calculate $\widetilde{\mathbf{\Sigma}}_i = n_i^{-1}\mathbf{Y}_i\mathbf{Y}_i^{\mathrm{T}}$ with $n_1 = m$ and $n_2 = n - m$, and carry out the eigendecomposition $\widetilde{\mathbf{\Sigma}}_1 = \mathbf{P}_1\mathbf{D}_1\mathbf{P}_1^{\mathrm{T}}$ as in Section 2.2. Using $\|A\|_F = \mathrm{tr}^{1/2}(AA^{\mathrm{T}})$ to denote the Frobenius norm of a matrix A, we then consider the following optimization problem:

$$\min_{\mathbf{D}} \|\mathbf{P}_1\mathbf{D}\mathbf{P}_1^{\mathrm{T}} - \mathbf{\Sigma}_p\|_F \qquad \text{where } \mathbf{D} \text{ is a diagonal matrix.} \tag{2.10}$$

Essentially, we are considering the class of rotation-equivariant estimator $\widehat{\mathbf{\Sigma}}_m = \mathbf{P}_1\mathbf{D}\mathbf{P}_1^{\mathrm{T}}$, where m is the location we split the data matrix $\mathbf{Y}$ into two. Basic calculus shows that the optimum is achieved at $\mathbf{D} = \operatorname{diag}(d_1, \ldots, d_p)$, where

$$d_i = \mathbf{p}_{1i}^{\mathrm{T}}\mathbf{\Sigma}_p\mathbf{p}_{1i}, \qquad i = 1, \ldots, p. \tag{2.11}$$

This is however unknown to us since $\mathbf{\Sigma}_p$ is unknown. We can see now that why we are interested in the asymptotic properties of $\mathbf{p}_{1i}^{\mathrm{T}}\mathbf{\Sigma}_p\mathbf{p}_{1i}$, which corresponds asymptotically to $\delta_1(\lambda_{1i})$ as defined in (2.9). The important thing about $\delta_1(\lambda_{1i})$ is that this function depends on the sample eigenvalues λ_{1i} of $\widetilde{\mathbf{\Sigma}}_1$, which are immediately available, and other quantities inside the definition of $\delta_1(\lambda_{1i})$ are also estimable from data.

Instead of estimating $\breve{m}_{F_1}(\lambda_{1i})$ and $\breve{m}_{\underline{F_1}}(0)$ as contained in the expression in $\delta_1(\lambda_{1i})$ in (2.9), which is basically what the paper Ledoit and Wolf (2012) is about [they estimate $\breve{m}_F(\lambda_i)$ and $\breve{m}_{\underline{F}}(0)$ in $\delta(\lambda_i)$ in (2.7) in their paper, that is, without splitting the data at all], we consider the asymptotic properties of $\mathbf{p}_{1i}^{\mathrm{T}}\widetilde{\mathbf{\Sigma}}_2\mathbf{p}_{1i}$. To this end, define the function

$$\Psi_m^{(1)}(z) = \frac{1}{p}\operatorname{tr}\big[(\widetilde{\mathbf{\Sigma}}_1 - z\mathbf{I}_p)^{-1}\widetilde{\mathbf{\Sigma}}_2\big], \qquad z \in \mathbb{C}^+. \tag{2.12}$$

We can show that the inverse Stieltjes transform of this function is, on all points of continuity of $\Phi_m^{(1)}$,

$$\Phi_m^{(1)}(x) = \frac{1}{p}\sum_{i=1}^{p}\mathbf{p}_{1i}^{\mathrm{T}}\widetilde{\mathbf{\Sigma}}_2\mathbf{p}_{1i}\mathbf{1}_{\{\lambda_{1i}\le x\}}, \qquad x \in \mathbb{R}. \tag{2.13}$$

Similar to Section 2.1, we can carry out asymptotic analysis on $\Phi_m^{(1)}(x)$ in order to study the asymptotic behavior of $\mathbf{p}_{1i}^{\mathrm{T}}\widetilde{\mathbf{\Sigma}}_2\mathbf{p}_{1i}$ for each i. The results are shown in the following theorem.

THEOREM 1. *Let Assumptions* (A1) *to* (A4) *be satisfied. Suppose* $p/n_1 \to c_1 > 0$ *and* $c_1 \neq 1$. *Assume also* $\sum_{n\ge1} n_2^{-3} < \infty$. *We have the following*:

(i) *The function* $\Psi_m^{(1)}(z)$ *defined in* (2.12) *converges a.s. to a non-random limit* $\Psi^{(1)}(z)$ *for any* $z \in \mathbb{C}^+$, *defined by*

$$\Psi^{(1)}(z) = \frac{1}{c_1(1 - c_1 - c_1 z m_{F_1}(z))} - \frac{1}{c_1},$$

where F_1 *and* c_1 *are defined in equation* (2.9) *and the descriptions therein.*

(ii) *The inverse Stieltjes transform of* $\Psi^{(1)}(z)$ *is* $\Phi^{(1)}(x) = \int_{-\infty}^{x}\delta_1(\lambda)\,dF_1(\lambda)$ *on all points of continuity of* $\Phi^{(1)}(x)$, *where* $\delta_1(\lambda)$ *is given in* (2.9).

(iii) *The function* $\Phi_m^{(1)}(x)$ *defined in* (2.13) *converges a.s. to* $\Phi^{(1)}(x)$ *on all points of continuity of* $\Phi^{(1)}(x)$.

Moreover, if $c_1 = 1$, *we still have* $\mathbf{\Phi}_m^{(1)}(x) - p^{-1}\sum_{i=1}^{p}\mathbf{p}_{1i}^{\mathrm{T}}\mathbf{\Sigma}_p\mathbf{p}_{1i}\mathbf{1}_{\{\lambda_{1i}\le x\}}$ *converges a.s. to* 0 *as* $n_1, p \to \infty$, *so that* $p/n_1 \to 1$ *and* $\sum_{n\ge 1} n_2^{-3} < \infty$.

Part (iii) of this theorem shows that the asymptotic nonlinear transformation for $\mathbf{p}_{1i}^{\mathrm{T}}\widetilde{\mathbf{\Sigma}}_2\mathbf{p}_{1i}$ is given by (2.9), which is the same asymptotic nonlinear transformation for $d_i = \mathbf{p}_{1i}^{\mathrm{T}}\mathbf{\Sigma}_p\mathbf{p}_{1i}$ in (2.11). This means that $\mathbf{p}_{1i}^{\mathrm{T}}\widetilde{\mathbf{\Sigma}}_2\mathbf{p}_{1i}$ and $\mathbf{p}_{1i}^{\mathrm{T}}\mathbf{\Sigma}_p\mathbf{p}_{1i}$ are asymptotically the same. Note that this conclusion is true even when $c_1 = 1$ by the very last part of the theorem, which is a case excluded in Ledoit and Wolf (2012) and Ledoit and Wolf (2013a). With this, we propose our covariance matrix estimator as

$$\widehat{\mathbf{\Sigma}}_m = \mathbf{P}_1 \operatorname{diag}(\mathbf{P}_1^{\mathrm{T}}\widetilde{\mathbf{\Sigma}}_2\mathbf{P}_1)\mathbf{P}_1^{\mathrm{T}}. \tag{2.14}$$

This is almost the same as $\check{\mathbf{\Sigma}}_m$ in (2.8), except that $\mathbf{P}$ there is replaced by $\mathbf{P}_1$. This makes sense with respect to minimizing the Frobenius loss in (2.10), since Theorem 1 shows that $\mathbf{p}_{1i}^{\mathrm{T}}\widetilde{\mathbf{\Sigma}}_2\mathbf{p}_{1i}$ is asymptotically the same as the minimizer d_i in (2.11).

It appears that the estimator $\widehat{\mathbf{\Sigma}}_m$ is not using as much information as $\check{\mathbf{\Sigma}}_m$, since the eigen-matrix $\mathbf{P}_1$ uses only information on $\mathbf{Y}_1$ but not the full set of data $\mathbf{Y}$. However, coupled with averaging to be introduced in Section 4.1, simulation experiments in Lam (2015) show that our estimator can have comparable or even better performance than $\check{\mathbf{\Sigma}}_m$ or its averaging counterparts introduced in Abadir, Distaso and Žikeš (2014). Figure 4 in Lam (2015) shows explicitly why averaging using a slightly smaller data set $\mathbf{Y}_1$ is better than not averaging while using the full set of data $\mathbf{Y}$. Please see the descriptions therein for more details.

While constructing $\widehat{\mathbf{\Sigma}}_m$ involves only splitting the data into two portions and carrying out eigen-analysis for one of them, the estimator proposed in Ledoit and Wolf (2012) requires the estimation of $\check{m}_F(\lambda_i)$ for each $i = 1, \ldots, p$, which can be computationally expensive. By inspecting the form of the nonlinear transformation in (2.7), when $c > 1$, the term $\check{m}_{\underline{F}}(0)$ has to be estimated as well, which requires special attention, and is not dealt with in Ledoit and Wolf (2012), although Ledoit and Wolf (2013b) has addressed this and extended their package to deal with $c > 1$. The estimator $\widehat{\mathbf{\Sigma}}_m$, on the other hand, is calculated the same way no matter $c_1 < 1$ or $c_1 \ge 1$. In Section 4, we propose an improvement to $\widehat{\mathbf{\Sigma}}_m$ through averaging, and compare the performance and speed of calculating this improved version of $\widehat{\mathbf{\Sigma}}_m$ with the one in Ledoit and Wolf (2013b). Comparisons will also be carried out with the grand average estimator proposed in equation (18) of Abadir, Distaso and Žikeš (2014).

Intuitively, the choice of m is important for the performance of the estimator, and it seems that we should use as much data to estimate $\widetilde{\mathbf{\Sigma}}_1$ as possible since $\mathbf{P}_1$ plays an important role. However, in Theorem 5, the sample size n_2 for constructing $\widetilde{\mathbf{\Sigma}}_2$ has to go to infinity with n, albeit at a slower rate, in order for our estimator to be asymptotically efficient. We also demonstrate empirically in the simulations in Lam (2015) that n_2 has to be reasonably large for the estimator to perform well in practice.

2.4. *The precision matrix estimator.* We use the inverse of $\widehat{\mathbf{\Sigma}}_m$ in (2.14) as our precision matrix estimator, that is,

$$\widehat{\mathbf{\Omega}}_m = \widehat{\mathbf{\Sigma}}_m^{-1}. \tag{2.15}$$

With respect to the inverse Stein's loss function SL for estimating $\mathbf{\Omega}_p = \mathbf{\Sigma}_p^{-1}$, where

$$\mathrm{SL}(\mathbf{\Omega}_p, \widehat{\mathbf{\Omega}}) = \mathrm{tr}(\mathbf{\Omega}_p^{-1}\widehat{\mathbf{\Omega}}) - \log|\mathbf{\Omega}_p^{-1}\widehat{\mathbf{\Omega}}| - p, \tag{2.16}$$

an asymptotically optimal estimator is indeed $\widehat{\mathbf{\Omega}}_m = \widehat{\mathbf{\Sigma}}_m^{-1}$ given in (2.15), as shown in Proposition 2 below. This inverse Stein's loss function is also introduced in Ledoit and Wolf (2013a). In Theorem 4.1 of their paper, they proved that the nonlinear transformation depicted in equation (2.7) of our paper is in fact optimal with respect to asymptotically minimizing this loss function.

The Stein's loss function first appeared in James and Stein (1961) for measuring the error of estimating $\mathbf{\Sigma}_p$ by $\widehat{\mathbf{\Sigma}}$. The inverse Stein's loss function is also a scaled version of the Kullback–Leibler divergence of the normal distribution $N(\mathbf{0}, \mathbf{\Sigma}_p)$ relative to $N(\mathbf{0}, \widehat{\mathbf{\Sigma}})$. We provide an alternative formulation from the results of Theorem 4.1 of Ledoit and Wolf (2013a), showing that the precision matrix estimator $\widehat{\mathbf{\Omega}}_m$ in equation (2.15) does indeed minimize the loss in (2.16) asymptotically.

PROPOSITION 2. *Consider the class of estimators* $\widehat{\mathbf{\Omega}} = \mathbf{P}_1\mathbf{D}_1^{-1}\mathbf{P}_1^{\mathrm{T}}$, *where* $\mathbf{D}_1^{-1} = \mathrm{diag}(d_1^{-1}, \ldots, d_p^{-1})$. *Then the optimization problem*

$$\min_{d_i} \mathrm{SL}(\mathbf{\Omega}_p, \widehat{\mathbf{\Omega}})$$

has a solution given by $d_i = \mathbf{p}_{1i}^{\mathrm{T}}\mathbf{\Sigma}_p\mathbf{p}_{1i}$, $i = 1, \ldots, p$.

Since Theorem 1(iii) shows that the asymptotic nonlinear transformation for $\mathbf{p}_{1i}^{\mathrm{T}}\widetilde{\mathbf{\Sigma}}_2\mathbf{p}_{1i}$ is the same as that for $d_i = \mathbf{p}_{1i}^{\mathrm{T}}\mathbf{\Sigma}_p\mathbf{p}_{1i}$, the above proposition immediately implies that

$$\widehat{\mathbf{\Omega}}_m = \mathbf{P}_1\big[\mathrm{diag}(\mathbf{P}_1^{\mathrm{T}}\widetilde{\mathbf{\Sigma}}_2\mathbf{P}_1)\big]^{-1}\mathbf{P}_1^{\mathrm{T}} = \widehat{\mathbf{\Sigma}}_m^{-1}$$

is an asymptotically optimal estimator for $\mathbf{\Omega}_p$.

PROOF OF PROPOSITION 2. With $\widehat{\mathbf{\Omega}} = \mathbf{P}_1\mathbf{D}_1^{-1}\mathbf{P}_1^{\mathrm{T}}$ where $\mathbf{P}_1 = (\mathbf{p}_{11}, \ldots, \mathbf{p}_{1p})$, we have

$$\begin{aligned}
\mathrm{SL}(\mathbf{\Omega}_p, \widehat{\mathbf{\Omega}}) &= \mathrm{tr}\big(\mathbf{\Sigma}_p\mathbf{P}_1\mathbf{D}_1^{-1}\mathbf{P}_1^{\mathrm{T}}\big) - \log|\mathbf{\Sigma}_p\mathbf{P}_1\mathbf{D}_1^{-1}\mathbf{P}_1^{\mathrm{T}}| - p \\
&= \sum_{i=1}^{p} \frac{\mathbf{p}_{1i}^{\mathrm{T}}\mathbf{\Sigma}_p\mathbf{p}_{1i}}{d_i} - \log|\mathbf{D}_1^{-1}| - \log|\mathbf{P}_1\mathbf{P}_1^{\mathrm{T}}| - \log|\mathbf{\Sigma}_p| - p \\
&= \sum_{i=1}^{p}\left(\frac{\mathbf{p}_{1i}^{\mathrm{T}}\mathbf{\Sigma}_p\mathbf{p}_{1i}}{d_i} + \log(d_i)\right) - \big(p + \log|\mathbf{\Sigma}_p|\big),
\end{aligned}$$

which is clearly minimized at $d_i = \mathbf{p}_{1i}^{\mathrm{T}} \boldsymbol{\Sigma}_p \mathbf{p}_{1i}$ for $i = 1, \ldots, p$. $\square$

In Section 4, we propose an improvement on the estimator $\widehat{\boldsymbol{\Sigma}}_m$, and we use its inverse as an improvement to the precision matrix estimator $\widehat{\boldsymbol{\Omega}}_m$.

3. Extension to data from a factor model. Results from previous sections rely heavily on Assumption (A1) in Section 2, that $\mathbf{y}_i = \boldsymbol{\Sigma}_p^{1/2} \mathbf{z}_i$ with $\mathbf{z}_i$ having p independent and identically distributed components, and that p goes to infinity together with n. This is also an assumption on which the results from Ledoit and Wolf (2012) rely. However, even for a random sample of p-dimensional vectors, this may not always be true. For instance, consider the factor model defined by

$$\mathbf{y}_i = \mathbf{A}\mathbf{x}_i + \boldsymbol{\varepsilon}_i, \qquad i = 1, \ldots, n, \tag{3.1}$$

with $\mathbf{A}$ being a $p \times r$ factor loading matrix, $\mathbf{x}_i$ an $r \times 1$ vector of factors and $\boldsymbol{\varepsilon}_i$ a $p \times 1$ vector of noise series. We assume that r is much smaller than p, so that a large number of p components in $\mathbf{y}_i$ have dynamics driven by the small number of r factors. In Lam, Yao and Bathia (2011), they consider $\{\mathbf{x}_i\}$ being a stationary time series and $\{\boldsymbol{\varepsilon}_i\}$ being a white noise. In this paper, we assume that $\{\mathbf{x}_i\}$ has independent and identically distributed vectors, and the same goes for $\{\boldsymbol{\varepsilon}_i\}$. Hence, the $\mathbf{y}_i$'s are independent and identically distributed. A factor model is often used in modeling a panel of stock returns in finance. See Fan, Fan and Lv (2008), for example. Also, see the data analysis of the return series of S&P500 constituents in Lam and Yao (2012). There, two factors are found, and 97.7% of the variation of the return series can be explained by a linear combination of the two factors.

The covariance matrix for $\mathbf{y}_i$ in model (3.1) can be written as

$$\boldsymbol{\Sigma}_p = \mathbf{A}\boldsymbol{\Sigma}_x \mathbf{A}^{\mathrm{T}} + \boldsymbol{\Sigma}_\varepsilon, \tag{3.2}$$

where $\boldsymbol{\Sigma}_x = \operatorname{var}(\mathbf{x}_i)$ and $\boldsymbol{\Sigma}_\varepsilon = \operatorname{var}(\boldsymbol{\varepsilon}_i)$. We can easily see that with the low dimensionality of $\mathbf{x}_i$, if some of the factors in $\mathbf{x}_i$ are strong factors [i.e., the corresponding columns in $\mathbf{A}$ have the majority of coefficients being non-zero; see Lam, Yao and Bathia (2011) and Lam and Yao (2012) for the formal definitions of strong and weak factors], we cannot really write $\mathbf{y}_i = \boldsymbol{\Sigma}_p^{1/2} \mathbf{y}_i^*$ with $\mathbf{y}_i^*$ having p independent components, since some factors in $\mathbf{x}_i$ are shared in most of the components of $\mathbf{y}_i$. Hence, Assumption (A1) in Section 2 cannot be satisfied, and the method in Ledoit and Wolf (2012) cannot be asymptotically optimal. We show in Theorem 3 below that our estimator $\widehat{\boldsymbol{\Sigma}}_m$ in (2.14), on the other hand, is still asymptotically optimal in the same sense as in Sections 2.3 and 2.4.

Before presenting the main result of this section, we present the following assumptions for the factor model in (3.1). They are parallel to the Assumptions (A1) to (A2) in Section 2.

(F1) The series $\{\boldsymbol{\varepsilon}_i\}$ has $\boldsymbol{\varepsilon}_i = \boldsymbol{\Sigma}_\varepsilon^{1/2}\boldsymbol{\xi}_i$, where $\boldsymbol{\xi}_i$ is a $p \times 1$ vector of independent and identically distributed random variables ξ_{ij}. Each ξ_{ij} has mean 0 and unit variance, and $E|\xi_{ij}|^{2k} = O(p^{k/2-1})$ for $k = 2, 3, 4, 5, 6$. The factor series $\{\mathbf{x}_t\}$ has a constant dimension r, and $\mathbf{x}_t = \boldsymbol{\Sigma}_x^{1/2}\mathbf{x}_t^*$ where $\mathbf{x}_t^*$ is a $r \times 1$ vector of independent and identically distributed random variables x_{ti}^*. Also, $E|x_{tj}^*|^{2k} < \infty$ for each t, j for $k = 2, 3, 4, 5, 6$.

(F2) The covariance matrix $\boldsymbol{\Sigma}_x = \operatorname{var}(\mathbf{x}_i)$ is such that $\|\boldsymbol{\Sigma}_x\| = O(1)$. The covariance matrix $\boldsymbol{\Sigma}_\varepsilon = \operatorname{var}(\boldsymbol{\varepsilon}_i)$ also has $\|\boldsymbol{\Sigma}_\varepsilon\| = O(1)$. Both covariance matrices are non-random. The factor loading matrix $\mathbf{A}$ is such that $\|\mathbf{A}\|_F^2 = \operatorname{tr}(\mathbf{A}\mathbf{A}^{\mathrm{T}}) = O(p)$.

The assumption $\|\mathbf{A}\|_F^2 = O(p)$ entails both strong and weak factors as defined in Lam, Yao and Bathia (2011).

THEOREM 3. *Let Assumptions* (F1) *and* (F2) *be satisfied. Assume that we split the data like that in Section 2.3, with p and n_2 both go to infinity together and $p/n_1 \to c_1 > 0$. Assume also $\sum_{n\geq 1} n_2^{-3} < \infty$. Then with $\boldsymbol{\Sigma}_p$ the covariance matrix of $\mathbf{y}_i$ as defined in* (3.2), *we have for almost all $x \in \mathbb{R}$,*

$$\frac{1}{p}\sum_{i=1}^{p} \mathbf{p}_{1i}^{\mathrm{T}}\widetilde{\boldsymbol{\Sigma}}_2\mathbf{p}_{1i}\mathbf{1}_{\{\lambda_{1i}\leq x\}} - \frac{1}{p}\sum_{i=1}^{p} \mathbf{p}_{1i}^{\mathrm{T}}\boldsymbol{\Sigma}_p\mathbf{p}_{1i}\mathbf{1}_{\{\lambda_{1i}\leq x\}} \xrightarrow{\text{a.s.}} 0.$$

Like Theorem 1, this theorem says that $\mathbf{p}_{1i}^{\mathrm{T}}\widetilde{\boldsymbol{\Sigma}}_2\mathbf{p}_{1i}$ is asymptotically equal to $\mathbf{p}_{1i}^{\mathrm{T}}\boldsymbol{\Sigma}_p\mathbf{p}_{1i}$ for each $i = 1, \ldots, p$. Since $d_i = \mathbf{p}_{1i}^{\mathrm{T}}\boldsymbol{\Sigma}_p\mathbf{p}_{1i}$ [see (2.11)] is the optimal solution for the optimization problem (2.10), it means that the covariance matrix estimator $\widehat{\boldsymbol{\Sigma}}_m$ in (2.14), under the factor model setting in this section, is still asymptotical optimal for estimating $\boldsymbol{\Sigma}_p$ in (3.2) with respect to the Frobenius loss, when considering the class of rotation-equivariant estimators $\widehat{\boldsymbol{\Sigma}}(\mathbf{D}) = \mathbf{P}_1\mathbf{D}\mathbf{P}_1^{\mathrm{T}}$. At the same time, by Proposition 2, the inverse estimator $\widehat{\boldsymbol{\Omega}}_m = \widehat{\boldsymbol{\Sigma}}_m^{-1}$ is also an asymptotically optimal estimator of the precision matrix $\boldsymbol{\Omega}_p = \boldsymbol{\Sigma}_p^{-1}$ with respect to the inverse Stein's loss function (2.16).

The optimality result in Theorem 3 means that we do not need to know the exact asymptotic nonlinear transformation to which $\mathbf{p}_{1i}^{\mathrm{T}}\boldsymbol{\Sigma}_p\mathbf{p}_{1i}$ converges in order to find an optimal covariance or precision matrix estimator. On the other hand, the form of the nonlinear transformation is crucial for the method in Ledoit and Wolf (2012) to work. In this sense, our estimators are more robust to changes in the structure of the data. We demonstrate the performance of the estimator when the data follows a factor model in the supplementary material [Lam (2015)].

4. Asymptotic efficiency loss and practical implementation. In this section, we introduce the following ideal estimator:

$$\widehat{\boldsymbol{\Sigma}}_{\text{Ideal}} = \mathbf{P}\operatorname{diag}(\mathbf{P}^{\mathrm{T}}\boldsymbol{\Sigma}_p\mathbf{P})\mathbf{P}^{\mathrm{T}}. \tag{4.1}$$

This is also called the finite-sample optimal estimator in Ledoit and Wolf (2012). Compared to $\widehat{\mathbf{\Sigma}}_m$ in (2.14), this ideal estimator used the full set of data for calculating the eigenmatrix $\mathbf{P}$, and it assumed the knowledge of $\mathbf{\Sigma}_p$ itself instead of using $\widetilde{\mathbf{\Sigma}}_2$ to estimate it like our estimator does. With this ideal estimator, we define the efficiency loss of an estimator $\widehat{\mathbf{\Sigma}}$ as

$$EL(\mathbf{\Sigma}_p, \widehat{\mathbf{\Sigma}}) = 1 - \frac{L(\mathbf{\Sigma}_p, \widehat{\mathbf{\Sigma}}_{\text{Ideal}})}{L(\mathbf{\Sigma}_p, \widehat{\mathbf{\Sigma}})}, \tag{4.2}$$

where $L(\mathbf{\Sigma}_p, \widehat{\mathbf{\Sigma}})$ is a loss function for estimating $\mathbf{\Sigma}_p$ by $\widehat{\mathbf{\Sigma}}$. The two loss functions we focus on in this paper are the Frobenius loss

$$L(\mathbf{\Sigma}_p, \widehat{\mathbf{\Sigma}}) = \|\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}_p\|_F^2, \tag{4.3}$$

and the inverse Stein's loss introduced in (2.16), which, in terms of $\widehat{\mathbf{\Sigma}}$ and $\mathbf{\Sigma}_p$, is

$$L(\mathbf{\Sigma}_p, \widehat{\mathbf{\Sigma}}) = \operatorname{tr}\big(\mathbf{\Sigma}_p \widehat{\mathbf{\Sigma}}^{-1}\big) - \log\det\big(\mathbf{\Sigma}_p \widehat{\mathbf{\Sigma}}^{-1}\big) - p. \tag{4.4}$$

If $EL(\mathbf{\Sigma}_p, \widehat{\mathbf{\Sigma}}) \leq 0$, it means that the estimator $\widehat{\mathbf{\Sigma}}$ is doing at least as good as the ideal estimator $\widehat{\mathbf{\Sigma}}_{\text{Ideal}}$ in terms of the loss function L, and vice versa. To present the asymptotic efficiency results with respect to these two loss functions, we need to assume the following set of Assumptions (A1)′ and (A2)′:

(A1)′ Each observation can be written as $\mathbf{y}_i = \mathbf{\Sigma}_p^{1/2}\mathbf{z}_i$ for $i = 1, \ldots, n$, where each $\mathbf{z}_i$ is a $p \times 1$ vector of independent and identically distributed random variables z_{ij}. Each z_{ij} has mean 0 and unit variance, and $E|z_{ij}|^k \leq B < \infty$ for some constant B and $2 < k \leq 20$.

(A2)′ The population covariance matrix is non-random and of size $p \times p$. Furthermore, $\|\mathbf{\Sigma}_p\| = O(1)$.

Or, if the data follows a factor model $\mathbf{y}_i = \mathbf{A}\mathbf{x}_i + \boldsymbol{\varepsilon}_i$, we need to assume the following set of Assumptions (F1)′ and (F2)′:

(F1)′ The series $\{\boldsymbol{\varepsilon}_i\}$ has $\boldsymbol{\varepsilon}_i = \mathbf{\Sigma}_\varepsilon^{1/2}\boldsymbol{\xi}_i$, where $\boldsymbol{\xi}_i$ is a $p \times 1$ vector of independent and identically distributed random variables ξ_{ij}. Each ξ_{ij} has mean 0 and unit variance, and $E|\xi_{ij}|^k \leq B < \infty$ for some constant B and $k \leq 20$. The factor series $\{\mathbf{x}_t\}$ has a constant dimension r, and $\mathbf{x}_t = \mathbf{\Sigma}_x^{1/2}\mathbf{x}_t^*$ where $\mathbf{x}_t^*$ is a $r \times 1$ vector of independent and identically distributed random variables x_{ti}^*. Also, $E|x_{tj}^*|^k \leq B < \infty$ for some constant B and $2 < k \leq 20$.

(F2)′ Same as (F2), meaning that $\|\mathbf{\Sigma}_x\|, \|\mathbf{\Sigma}_\varepsilon\| = O(1)$ and $\|\mathbf{A}\|_F^2 = O(p)$.

Assumptions (A1)′ and (A2)′ are parallel to (A1) and (A2), respectively. The more restrictive moments assumptions are needed for the proof of Lemma 1, which is important for proving Corollary 4 and the asymptotic efficiency results in Theorem 5. Assumption (F1)′ is parallel to (F1), and is for data with a factor structure.

LEMMA 1. *Let Assumption* (A1)′ *be satisfied. If the split location* m *is such that* $\sum_{n\geq 1} p(n-m)^{-5} < \infty$, *we have*

$$\max_{1\leq i\leq p}\left|\frac{\mathbf{p}_{1i}^{\mathrm{T}}\widetilde{\boldsymbol{\Sigma}}_2\mathbf{p}_{1i} - \mathbf{p}_{1i}^{\mathrm{T}}\boldsymbol{\Sigma}_p\mathbf{p}_{1i}}{\mathbf{p}_{1i}^{\mathrm{T}}\boldsymbol{\Sigma}_p\mathbf{p}_{1i}}\right| \xrightarrow{\text{a.s.}} 0.$$

The same holds true if the data is from a factor model, with Assumption (F1)′ *satisfied together with* $\sum_{n\geq 1} p(n-m)^{-5} < \infty$.

The proof of this lemma is in the supplementary material [Lam (2015)]. With the result in Lemma 1, it is easy to see the following.

COROLLARY 4. *Let the assumptions in Lemma 1 hold. Then as* $n, p \to \infty$ *almost surely,* $\widehat{\boldsymbol{\Sigma}}_m$ *is positive definite as long as* $\boldsymbol{\Sigma}_p$ *is in the following.*

PROOF. Note that $\widehat{\boldsymbol{\Sigma}}_m$ is always positive semi-definite by construction, since all the eigenvalues $\mathbf{p}_{1i}^{\mathrm{T}}\widetilde{\boldsymbol{\Sigma}}_2\mathbf{p}_{1i}$, $i = 1, \ldots, p$, are non-negative. The convergence result in Lemma 1 ensures that all the eigenvalues of $\widehat{\boldsymbol{\Sigma}}_m$ are almost surely larger than $(1-\varepsilon)\mathbf{p}_{1i}^{\mathrm{T}}\boldsymbol{\Sigma}_p\mathbf{p}_{1i} \geq (1-\varepsilon)\lambda_{\min}(\boldsymbol{\Sigma}_p) > 0$ for large enough n and a fixed $0 < \varepsilon < 1$, if $\boldsymbol{\Sigma}_p$ is positive definite. □

This result is not formally proved in Ledoit and Wolf (2012, 2013a, 2013b), and in fact the conclusion can be wrong for their nonlinear shrinkage estimator when the data follows a factor model. Our simulation results in Lam (2015) do show that the nonlinear shrinkage estimator is singular in some simulation runs for data following a factor model, when our estimator is still positive definite.

Corollary 4 is also different from Proposition 1 in Abadir, Distaso and Žikeš (2014). For their proof to be valid, they in fact need p to be smaller than $n-m$ (in our notation), otherwise the estimator is only at most positive semi-definite. Our Corollary 4, on the other hand, allows for p to be even larger than n.

With Lemma 1, we can also prove that $\widehat{\boldsymbol{\Sigma}}_m$ is asymptotically efficient relative to $\widehat{\boldsymbol{\Sigma}}_{\text{Ideal}}$ with respect to both the Frobenius and the inverse Stein's losses, as long as the split location m satisfies some conditions.

THEOREM 5. *Let Assumptions* (A1)′, (A2)′, (A3) *and* (A4) *be satisfied. Assume the split location* m *for* $\widehat{\boldsymbol{\Sigma}}_m$ *is such that* $m/n \to 1$ *and* $n-m \to \infty$ *as* $n \to \infty$, *with* $\sum_{n\geq 1} p(n-m)^{-5} < \infty$. *We then have* $EL(\boldsymbol{\Sigma}_p, \widehat{\boldsymbol{\Sigma}}_m) \xrightarrow{\text{a.s.}} 0$ *with respect to both the Frobenius and the inverse Stein's loss functions, as long as* $\boldsymbol{\Sigma}_p \neq \sigma^2\mathbf{I}_p$.

The proof is in the supplementary material [Lam (2015)]. Note that we exclude the case $\boldsymbol{\Sigma}_p = \sigma^2\mathbf{I}_p$, which has zero loss for the ideal estimator. In Lam (2015),

we include this case to compare the performance of various methods including our estimator.

The results from the above theorem show that $\widehat{\mathbf{\Sigma}}_m$ is asymptotically the same as the ideal estimator $\widehat{\mathbf{\Sigma}}_{\text{Ideal}}$ with respect to the Frobenius or inverse Stein's losses. Since $p/n \to c > 0$, a choice of m that satisfies all the conditions is $m = n - an^{1/2}$ for some constant $a > 0$. In practice, this form of split location works well, and we provide a way in Section 4.2 to identify an m for good performance.

In Theorem 5, data from a factor model is excluded. We do not pursue the proof here, although we still conjecture that $EL(\mathbf{\Sigma}_p, \widehat{\mathbf{\Sigma}}_m) \xrightarrow{\text{a.s.}} 0$ for the inverse Stein's loss. See the empirical results in Lam (2015) for details.

4.1. *Improvement with averaging.* We can improve the performance of $\widehat{\mathbf{\Sigma}}_m$ in (2.14) by noting that each vector $\mathbf{y}_i$ in $\mathbf{Y}$ is independent of each other and identically distributed. We can permute the data, form another data matrix $\mathbf{Y}^{(j)}$, and split the data into two independent parts $\mathbf{Y}^{(j)} = (\mathbf{Y}_1^{(j)}, \mathbf{Y}_2^{(j)})$ as in Section 2.3. Then we can form another estimator

$$\widehat{\mathbf{\Sigma}}_m^{(j)} = \mathbf{P}_{1j} \operatorname{diag}(\mathbf{P}_{1j}^{\mathrm{T}} \widetilde{\mathbf{\Sigma}}_2^{(j)} \mathbf{P}_{1j}) \mathbf{P}_{1j}^{\mathrm{T}}, \qquad \text{where } \widetilde{\mathbf{\Sigma}}_i^{(j)} = n_i^{-1} \mathbf{Y}_i^{(j)} \mathbf{Y}_i^{(j)\mathrm{T}}, \tag{4.5}$$

with $m = n_1$, $n = n_1 + n_2$, for $j = 1, \ldots, M$. Each j represents a permutation of the data so that no two $\mathbf{Y}_1^{(j)}$'s contain exactly the same data, thus $M \le \binom{n}{m}$. The matrix $\mathbf{P}_{1j}$ contains the orthonormal set of eigenvectors such that $\widetilde{\mathbf{\Sigma}}_1^{(j)} = \mathbf{P}_{1j} \mathbf{D}_{1j} \mathbf{P}_{1j}^{\mathrm{T}}$.

In Abadir, Distaso and Žikeš (2014), they improve the performance of their estimator by averaging the regularized eigenvalues over different j and different split location m. However, we know from Theorem 1 and Proposition 2 that the regularized eigenvalues $\mathbf{p}_{1i}^{\mathrm{T}} \widetilde{\mathbf{\Sigma}}_2 \mathbf{p}_{1i}$ are asymptotically optimal only when coupled with $\mathbf{P}_1$ to form the estimator $\widehat{\mathbf{\Sigma}}_m$. Hence, for each j, the regularized eigenvalues in $\operatorname{diag}(\mathbf{P}_{1j}^{\mathrm{T}} \widetilde{\mathbf{\Sigma}}_2^{(j)} \mathbf{P}_{1j})$ are asymptotically optimal only when coupled with $\mathbf{P}_{1j}$ to calculate $\widehat{\mathbf{\Sigma}}_m^{(j)}$, as in (4.5). This forbid us from averaging the eigenvalues in $\operatorname{diag}(\mathbf{P}_{1j}^{\mathrm{T}} \widetilde{\mathbf{\Sigma}}_2^{(j)} \mathbf{P}_{1j})$ over the index j, since each set is only asymptotically optimal when coupled with $\mathbf{P}_{1j}$, and it can be suboptimal to couple the averaged set of eigenvalues with other orthogonal matrix $\mathbf{P}$.

In light of the above argument, we average the sum of $\widehat{\mathbf{\Sigma}}_m^{(j)}$ to improve the performance by setting

$$\widehat{\mathbf{\Sigma}}_{m,M} = \frac{1}{M} \sum_{j=1}^{M} \widehat{\mathbf{\Sigma}}_m^{(j)}, \tag{4.6}$$

where $\widehat{\mathbf{\Sigma}}_m^{(j)}$ is as given in (4.5). This is different from equation (15) in Abadir, Distaso and Žikeš (2014) as we have argued in the paragraph before. However, they have proved in Proposition 3 in their paper that the expected element-wise loss in L_1 and L_2 norm for their estimator is asymptotically optimized if the split

location m is such that $m, n-m \to \infty$ with $m/n \to \gamma \in (0,1)$. This choice of split is certainly excluded from our results in Theorem 5, where we need $m/n \to 1$. The reason for this major difference is that $p \to \infty$ as $n \to \infty$ such that $p/n \to c > 0$ in our paper, whereas p is treated as fixed in Abadir, Distaso and Žikeš (2014) (except for their Proposition 5, where p can diverge to infinity, but still need to be at a rate slower than n). In our simulations in Lam (2015), we demonstrate that the split m with $m/n \to \gamma \in (0,1)$ can be suboptimal if p is indeed growing with n.

We provide an efficiency result for the averaged estimator $\widehat{\mathbf{\Sigma}}_{m,M}$ below.

THEOREM 6. *Let the assumptions in Theorem 5 be satisfied. Assume further that M is finite, and $\tau_{n,p}$, the smallest eigenvalue of $\mathbf{\Sigma}_p$, is bounded uniformly away from 0. We then have $EL(\mathbf{\Sigma}_p, \widehat{\mathbf{\Sigma}}_{m,M}) \le 0$ almost surely with respect to the Frobenius loss or the inverse Stein's loss function as $n, p \to \infty$, as long as $\mathbf{\Sigma}_p \neq \sigma^2 \mathbf{I}_p$.*

This theorem shows that the estimator $\widehat{\mathbf{\Sigma}}_{m,M}$ also enjoys asymptotic efficiency with respect to the ideal estimator. The proof is in the supplementary material [Lam (2015)].

A larger M usually gives better performance, but becomes computationally expensive when it is too large. Luckily, our simulation results in Lam (2015) demonstrate that $M = 50$ attains a very good performance already in a variety of settings, when a random permutation of the data is used.

Depending on the data application, one can construct cross-validation criterion for finding a good M which has good performance but not computationally too expensive. For instance, if $\widehat{\mathbf{\Sigma}}_{m,M}$ or $\widehat{\mathbf{\Sigma}}_{m,M}^{-1}$ are needed so that forecasts for the data $\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_n$ can be made, we can split the data into a training and a test set, and construct $\widehat{\mathbf{\Sigma}}_{m,M}$ using the training data set given an M (with M, the way to choose m is described in Section 4.2). Forecasts error can be obtained by using the test set. We can then repeat the above using another larger M, to see if forecasts error is reduced significantly. If so, we may prefer a larger M, but would choose a smaller M otherwise to balance out forecasting accuracy and computational cost. This is particularly important if the data analysis involves a moving window of data and multiple forecasts have to be done continually. See also the call center data application in Section 5.3.

4.2. *Choice of split location.* Rather than averaging over different split locations m like the grand average estimator (15) of Abadir, Distaso and Žikeš (2014), Theorem 5 suggests that $m = n - an^{1/2}$ can be a choice to achieve asymptotic efficiency when $p/n \to c > 0$, but not so when m/n goes to a constant smaller than 1. Indeed, we find that our estimator does not perform well when m is small, and when m is too close to n, the performance suffers also, which is also demonstrated in Lam (2015). We propose to minimize the following criterion for a good

choice of m:

$$g(m) = \left\| \frac{1}{M} \sum_{s=1}^{M} (\widehat{\mathbf{\Sigma}}_m^{(s)} - \widetilde{\mathbf{\Sigma}}_2^{(s)}) \right\|_F^2, \tag{4.7}$$

where $\widehat{\mathbf{\Sigma}}_m^{(s)}$ and $\widetilde{\mathbf{\Sigma}}_2^{(s)}$ are defined in (4.5). This criterion is inspired by the one used in Bickel and Levina (2008b) for finding a good banding number, where the true covariance matrix is replaced by a sample one. We demonstrate the performance of the criterion $g(m)$ in Lam (2015) under various settings. Although Theorem 5 suggests m to be such that $m/n \to 1$, finite sample performance may be the best for smaller values of m. Hence, we suggest to search the following split locations in practice in order to minimize $g(m)$:

$$m = [2n^{1/2}, 0.2n, 0.4n, 0.6n, 0.8n, n - 2.5n^{1/2}, n - 1.5n^{1/2}]. \tag{4.8}$$

The smaller splits are actually better when $\mathbf{\Sigma}_p = \sigma^2 \mathbf{I}_p$. This case is excluded in Theorem 5, and we study this by simulations in Lam (2015).

5. Empirical results. In Lam (2015), we have created five profiles of simulations, each with a different population covariance matrix and data generating mechanism. We present the computing time for different methods here. Please refer to Lam (2015) for all other details and simulation results.

Sections 5.1, 5.2 and 5.3 illustrate and compare our method with other state-of-the-art methods using real data, or simulation with real data. Hereafter, we abbreviate our method as NERCOME for estimating $\mathbf{\Sigma}_p$ or $\mathbf{\Omega}_p$, as in Nonparametric Eigenvalue–Regularized Covariance Matrix Estimator. The method proposed in Abadir, Distaso and Žikeš (2014) is abbreviated as CRC (Condition number Regularized Covariance estimator), while the nonlinear shrinkage method in Ledoit and Wolf (2012) is abbreviated as NONLIN. We call the grand average estimator (15) in Abadir, Distaso and Žikeš (2014) the CRC grand average. The method in Fan, Liao and Mincheva (2013) is abbreviated as POET. The graphical LASSO in Friedman, Hastie and Tibshirani (2008) is abbreviated as GLASSO, and finally, the adaptive SCAD thresholding, which is a special case of POET without any factors, is abbreviated as SCAD.

We look at the computing time for NERCOME, NONLIN, SCAD and GLASSO for profile (I). The computing times for all the methods are similar to other profiles, and are not shown here. While all methods are Matlab coded, only NONLIN requires a third-party SLP optimizer, since NONLIN involves solving non-convex optimization problems, which is done using a commercial package called SNOPT in Matlab [see Ledoit and Wolf (2012) for more details]. Recently, the code for NONLIN is updated to use interior-point methods which do not require a commercial package. However, our extensive testing show that the commercial package is in fact much faster than the updated interior-point version, and hence we are still using the commercial package for all the NONLIN simulations.

TABLE 1

Mean time (in seconds) for computing a covariance matrix estimator for NERCOME [*including the time for finding the best split using* (4.7)], NONLIN, *SCAD and GLASSO for profile* (*I*). *Standard deviation is in bracket. Refer to Lam* (2015) *for all simulation details*

		$p = 50$	$p = 100$	$p = 200$	$p = 500$
$n = 200$	NERCOME	$0.43_{(0.0)}$	$1.8_{(0.2)}$	$8.4_{(0.4)}$	$78.3_{(2.6)}$
	NONLIN	$21.4_{(1.3)}$	$29.3_{(6.0)}$	$39.0_{(5.9)}$	$38.9_{(7.4)}$
	SCAD	$0.1_{(0.0)}$	$0.7_{(0.0)}$	$4.8_{(0.0)}$	–
	GLASSO	$6.5_{(0.9)}$	$25.1_{(4.2)}$	$93.2_{(3.3)}$	–
$n = 400$	NERCOME	$0.50_{(0.0)}$	$2.3_{(0.1)}$	$9.9_{(0.5)}$	$70.5_{(9.5)}$
	NONLIN	$25.3_{(12.0)}$	$33.4_{(10.1)}$	$38.4_{(3.5)}$	$55.1_{(10.1)}$
	SCAD	$0.2_{(0.0)}$	$1.2_{(0.0)}$	$8.6_{(0.1)}$	–
	GLASSO	$6.3_{(0.2)}$	$31.9_{(4.4)}$	$104.6_{(4.6)}$	–
$n = 800$	NERCOME	$0.64_{(0.0)}$	$2.7_{(0.1)}$	$11.7_{(0.6)}$	$76.4_{(25.4)}$
	NONLIN	$24.8_{(9.8)}$	$34.0_{(12.1)}$	$40.0_{(8.4)}$	$83.6_{(135.9)}$
	SCAD	$0.3_{(0.0)}$	$2.2_{(0.1)}$	$16.3_{(0.4)}$	–
	GLASSO	$6.3_{(0.3)}$	$31.7_{(0.8)}$	$150.8_{(454.8)}$	–

From Table 1, it is clear that SCAD thresholding is the fastest, albeit we have set favorable values of C for the thresholding in advance. NERCOME is the second fastest for $p \le 200$. When $p \ge 500$, NONLIN is faster than NERCOME for smaller values of n, but the computational cost increases quickly with the increase of n. NERCOME, on the other hand, remains similar in computational costs for a wide range of values of n. This is because the major computational cost for NERCOME comes from the M eigen-decompositions of a $p \times p$ sample covariance matrix, with each eigen-decomposition being computationally expensive when p is large. Increasing n only marginally increases the computational cost of an eigen-decomposition.

5.1. *Bias reduction with generalized least squares* (*GLS*). In this simulation, we aim to demonstrate that for data with a potential factor structure, if covariance or precision matrix estimation is just an intermediate step instead of the final goal, then our method can do well in the end even compared to methods that exploit the factor structure through factor analysis.

Consider a linear model

$$\mathbf{y}_i = \mathbf{X}_i \boldsymbol{\beta} + \boldsymbol{\varepsilon}_i, \qquad i = 1, \ldots, n,$$

where the $\mathbf{X}_i$'s are known covariates and the $\boldsymbol{\varepsilon}_i$'s have $E(\boldsymbol{\varepsilon}_i \boldsymbol{\varepsilon}_i^{\mathrm{T}}) = \boldsymbol{\Sigma}_p$. Then the generalized least squares estimator

$$\widehat{\boldsymbol{\beta}}_{\mathrm{GLS}} = \left(\sum_{i=1}^{n} \mathbf{X}_i^{\mathrm{T}} \boldsymbol{\Sigma}_p^{-1} \mathbf{X}_i \right)^{-1} \sum_{i=1}^{n} \mathbf{X}_i^{\mathrm{T}} \boldsymbol{\Sigma}_p^{-1} \mathbf{y}_i$$

is more efficient than the least squares one in general. However, we need the precision matrix $\mathbf{\Sigma}_p^{-1}$ as an input. We demonstrate the effectiveness of bias reduction of $\widehat{\boldsymbol{\beta}}_{\mathrm{GLS}}$ using different methods in estimating $\mathbf{\Sigma}_p^{-1}$.

To this end, we run 500 simulations. We set $\boldsymbol{\beta} = (-0.5, 0.5, 0.3, -0.6)^{\mathrm{T}}$. In each simulation run, and we generate $\mathbf{X}_i$ with independent $N(0,1)$ entries. For the noise, we use the standardized macroeconomic data $\mathbf{w}_i$ analyzed in Stock and Watson (2005). The data consists of $p = 132$ monthly U.S. macroeconomic time series running from January 1959 to December 2003 ($n = 526$), and is categorized into 14 categories of different size. In Stock and Watson (2005), they argue that there are 7 factors in the data. We set $\boldsymbol{\varepsilon}_i = 2\mathbf{w}_i + \mathbf{z}_i$, where $\mathbf{z}_i$ consists of independent $N(0, 0.2^2)$ entries and they are generated in each simulation run. Hence, we have $\mathbf{y}_t = \mathbf{X}_t\boldsymbol{\beta} + 2\mathbf{w}_i + \mathbf{z}_i$. We choose to use $\boldsymbol{\varepsilon}_i = 2\mathbf{w}_i + \mathbf{z}_i$ in order to add challenges to NERCOME, as it does not only contain potentially many factors, but also exhibits certain degree of serial correlation, which is violating our assumptions of independence. For NERCOME, we choose m to minimize (4.7) on a grid of 7 split locations as in (4.8). Figure 1 shows the mean sum of absolute bias $\|\widehat{\boldsymbol{\beta}}_{\mathrm{GLS}} - \boldsymbol{\beta}\|_1$, as well as the mean root-average-square prediction error (RASE) $n^{-1}\sum_{t=1}^{n} p^{-1/2}\|\mathbf{y}_t - \mathbf{X}_t\widehat{\boldsymbol{\beta}}_{\mathrm{GLS}}\|$, for different methods. In Fan, Liao and Mincheva (2013), POET, a covariance estimator through exploiting the factor structure, is described as being sensitive to underestimating the number of factors. This is indeed the case, as both the upper and lower left panels in the figure show a lot of fluctu-

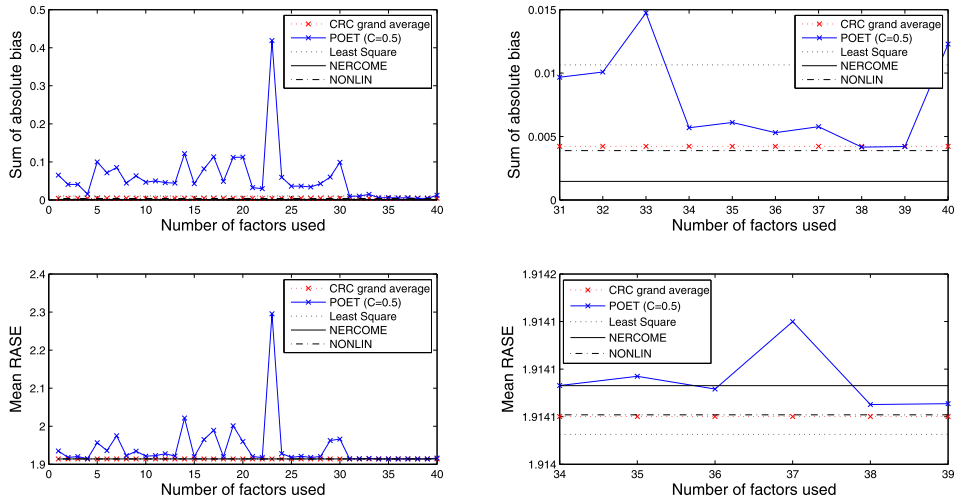

FIG. 1. *Upper row*: *Mean sum of absolute bias for estimating* $\boldsymbol{\beta}$ *against the number of factors* K *for POET. Lower row*: *The mean root-average-square prediction error against* K. *Least squares*, *NERCOME*, *CRC grand average and NONLIN stay constant throughout*. *Left*: *Number of factors used in POET from* 1 *to* 40. *Right*: *Number of factors used in POET from* 31 *to* 40 (*upper right*) *or from* 34 *to* 39 (*lower right*).

ations as the number of factors K used in POET varies, until around $K \geq 31$. This indicates that the number of factors in $\mathbf{w}_i$ is likely to be around 31, rather than 7 as suggested in Stock and Watson (2005). If factor analysis has to be done first before bias reduction can be performed, then it is very likely that the number of factors is grossly underestimated, thus giving a bad covariance matrix estimator that can affect the effectiveness of bias reduction. Even for the POET method which does not actually need an accurate input of the number of factors, it is still very unusual for one to input the number of factors to be this large at over 30.

The upper right panel of Figure 1 shows the sum of absolute bias when $K \geq 31$. Clearly, all the methods can improve upon the least squares one, which incurs the largest bias. CRC grand average, NONLIN and POET at $K = 38, 39$ have similar performance, while NERCOME has the best performance throughout. Nevertheless, apart from the least squares, all of the other methods perform well in absolute terms. It is clear that NERCOME, with theoretical support from Theorem 3, can produce good bias reduction results even without estimating the number of factors, the factor loading matrix and the factor series themselves. These can be difficult tasks, particularly in this scenario where there are a lot of potential factors.

The lower right panel of the figure shows the mean root-average-square prediction error when $K = 34, \ldots, 39$. This time, the least squares method incurs the smallest error, followed by CRC grand average and NONLIN. NERCOME and POET perform slightly worse than other methods. Again, all the methods perform well in absolute terms, especially the difference in errors are actually very small.

5.2. *Application*: *Risk minimization of a portfolio*. We consider risk minimization for a portfolio of $p = 100$ stocks, which is also considered in Section 7.2 of Fan, Liao and Mincheva (2013). The data consists of 2640 annualized daily excess returns $\{r_t\}$ for the period January 1, 2001, to December 31, 2010 (22 trading days each month). Five portfolios are created at the beginning of each month using five different methods in estimating the covariance matrix of returns. A typical setting here is $n = 264$, $p = 100$, that is, one year of past returns to estimate a covariance matrix of 100 stocks. Each portfolio has weights given by

$$\widehat{\mathbf{w}} = \frac{\widehat{\mathbf{\Sigma}}^{-1}\mathbf{1}_p}{\mathbf{1}_p^{\mathrm{T}}\widehat{\mathbf{\Sigma}}^{-1}\mathbf{1}_p},$$

where $\mathbf{1}_p$ is the vector of p ones, and $\widehat{\mathbf{\Sigma}}^{-1}$ is an estimator of the $p \times p$ precision matrix for the stock returns, using strict factor model (covariance of error forced to be diagonal, abbreviated as SFM), POET (constant C determined by cross-validation), CRC grand average, NERCOME [m automatically chosen by minimizing (4.7)] and NONLIN respectively. This weight formula solves the risk minimization problem

$$\min_{\mathbf{w}:\mathbf{w}^{\mathrm{T}}\mathbf{1}_p=1} \mathbf{w}^{\mathrm{T}}\widehat{\mathbf{\Sigma}}\mathbf{w}.$$

TABLE 2
Performance of different methods. SFM represents the strict factor model, with diagonal covariance matrix. CRC represents the CRC grand average

	SFM	POET	NERCOME	CRC	NONLIN
Total excess return	153.9	109.5	128.0	127.9	124.8
Out-of-sample variance	0.312	0.267	0.264	0.264	0.264
Mean Sharpe Ratio	0.224	0.197	0.212	0.211	0.205

At the end of each month, for each portfolio, we compute the total excess return, the out-of-sample variance and the mean Sharpe ratio, given respectively by [see also Demiguel and Nogales (2009)]:

$$\widehat{\mu} = \sum_{i=12}^{119} \sum_{t=22i+1}^{22i+22} \mathbf{w}^{\mathrm{T}} \mathbf{r}_t, \qquad \widehat{\sigma}^2 = \frac{1}{2376} \sum_{i=12}^{119} \sum_{t=22i+1}^{22i+22} (\mathbf{w}^{\mathrm{T}} \mathbf{r}_t - \widehat{\mu}_i)^2,$$

$$\widehat{\mathrm{sr}} = \frac{1}{108} \sum_{i=12}^{119} \frac{\widehat{\mu}_i}{\widehat{\sigma}_i^2}.$$

Table 2 shows the results. Clearly, the strict factor model has the highest return, followed by NERCOME, which is similar to slightly lower CRC grand average and NONLIN. POET has the lowest return of all. The out-of-sample variance, which is a measure of risk, is the smallest for NERCOME, CRC grand average and NONLIN, while the strict factor model has the highest risk. In essence, NERCOME, CRC grand average and NONLIN have risk minimization done well while maintaining a certain level of return.

5.3. *Application*: *Forecasting the number of calls for a call center.* We analyze the call center data considered in Huang et al. (2006) and Bickel and Levina (2008b). Phone calls to a call center are recorded from 7am to midnight everyday in 2002, except for weekends, holiday and when equipments are malfunctioning, leaving $n = 239$ days in total. In each day, a 17-hour recording period is divided into 10-minute intervals, resulting in 102 intervals. Let N_{ij} be the number of calls at the jth interval on the ith day, $i = 1, \ldots, 239$, $j = 1, \ldots, 102$, and let $y_{ij} = (N_{ij} + 1/4)^{1/2}$, which is a transformation used for bringing the variables closer to normal.

One particular interest is the prediction of the number of calls in a particular period of time from past data. One can in fact divide the data into intervals shorter than 10 minutes, and perform prediction for a certain number of intervals ahead from a moving window of past data. This usually involves substantial computational resources to do. Given its computational speed, NERCOME makes a competitive viable choice when the moving window of past data has sample size comparable to its dimension which is not very large.

Here, we simplify our task by considering prediction of a period of 30 days of arrival counts in the second half of the day ($j = 52, \ldots, 102$) using data from the first half of the day ($j = 1, \ldots, 51$). Let $\mathbf{y}_i = (\mathbf{y}_i^{(1)}, \mathbf{y}_i^{(2)})$, where $\mathbf{y}_i^{(1)} = (y_{i,1}, \ldots, y_{i,51})$ and $\mathbf{y}_i^{(2)} = (y_{i,52}, \ldots, y_{i,102})$. Partitioned accordingly, let the mean of $\mathbf{y}_i$ be $\boldsymbol{\mu} = (\boldsymbol{\mu}_1^{\mathrm{T}}, \boldsymbol{\mu}_2^{\mathrm{T}})$ and the covariance matrix of $\mathbf{y}_i$ be $\boldsymbol{\Sigma} = (\boldsymbol{\Sigma}_{ij})_{1\le i,j\le 2}$. We then use the best linear predictor of $\mathbf{y}_i^{(2)}$ from $\mathbf{y}_i^{(1)}$ for prediction:

$$\widehat{\mathbf{y}}_i^{(2)} = \boldsymbol{\mu}_2 + \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}(\mathbf{y}_i^{(1)} - \boldsymbol{\mu}_1). \tag{5.1}$$

Prediction result depends on two things. First, it depends on the estimator we use for $\boldsymbol{\Sigma}$. We cannot assume a particular structure of $\boldsymbol{\Sigma}$ for the data, other than the fact that for each i, $(y_{i,1}, \ldots, y_{i,102})$ is a time series, and if serial dependence is not too strong, then the Cholesky factor in the modified Cholesky decomposition should be approximately banded; see Bickel and Levina (2008b) for more details. Banding from Bickel and Levina (2008b) is then a natural choice. We also compare with NERCOME, NONLIN, CRC grand average and the sample covariance matrix to see if stabilization of the eigenvalues from NERCOME, CRC grand average and NONLIN can outperform banding.

Second, result depends on the window of past data we use for estimating $\boldsymbol{\Sigma}$, and the 30 day period we choose for forecasting. We compare results from using past data at a multiple of 30 days (not necessarily starting from day 1), and forecasting the 30 day period immediately following the past data (since $n = 239$, the last forecast will be using 210 days of past data and only forecast the next 29 days).

We compare the average absolute forecast error, defined by

$$E_{i,k} = \frac{1}{1530}\sum_{r=30i+1}^{30(i+1)}\sum_{j=52}^{102}|\widehat{y}_{rj,k}^{(2)} - y_{rj}^{(2)}|, \qquad E_{7,k} = \frac{1}{1479}\sum_{r=211}^{239}\sum_{j=52}^{102}|\widehat{y}_{rj,k}^{(2)} - y_{rj}^{(2)}|,$$

where $\widehat{y}_{rj,k}^{(2)}$ is a component of $\widehat{\mathbf{y}}_r^{(2)}$ defined in (5.1). The index k represents the start day of the past data used to calculate estimators of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ in (5.1), which can be $1, 31, \ldots, 211$. For NERCOME, CRC grand average and banding, since they have random fluctuations (random permutation and split of data for CRC grand average, in addition finding a good split for NERCOME, and finding a good banding number for banding), we estimate $\boldsymbol{\Sigma}$ 150 times and report the average forecast errors. For the choice of M for NERCOME and CRC grand average, we followed the procedure described at the end of Section 4.1 and split the past data into training and test sets, and calculate absolute average forecast errors according to (5.1). We found that $M = 50$ is working as good as $M = 75$ or 100 for all cases, and hence the data analysis is carried out with $M = 50$.

Table 3 shows the results. In general, sample covariance performed the worst as expected, followed by banding. Exception is the last 29-day forecast period (Start Day + Sample Size = 211), where sample covariance performed the same as NERCOME, NONLIN and CRC grand average. Banding performs better for

TABLE 3

Mean absolute forecast error (standard deviation in bracket) at different past data and forecasting period. Sample size is the length of past data used including the start day. Blocks with start day plus sample size being equal, that is, blocks on a lower left to upper right diagonal, are forecasting the same 30-day (or 29-day) period

Start day	$\widehat{\Sigma}$	Sample size						
		30	**60**	**90**	**120**	**150**	**180**	**210**
1	Sample	–	–	–	$0.85_{(0.20)}$	$0.79_{(0.19)}$	$0.89_{(0.18)}$	$1.53_{(0.48)}$
	Banding	$0.86_{(0.23)}$	$0.86_{(0.22)}$	$0.78_{(0.18)}$	$0.74_{(0.16)}$	$0.71_{(0.17)}$	$0.81_{(0.15)}$	$1.46_{(0.46)}$
	NERCOME	$0.72_{(0.13)}$	$0.73_{(0.14)}$	$0.65_{(0.15)}$	$0.72_{(0.16)}$	$0.71_{(0.19)}$	$0.77_{(0.14)}$	$1.53_{(0.51)}$
	CRC	$0.71_{(0.13)}$	$0.73_{(0.13)}$	$0.65_{(0.14)}$	$0.72_{(0.16)}$	$0.71_{(0.19)}$	$0.77_{(0.14)}$	$1.54_{(0.51)}$
	NONLIN	$0.72_{(0.14)}$	$0.73_{(0.13)}$	$0.66_{(0.15)}$	$0.72_{(0.17)}$	$0.71_{(0.19)}$	$0.78_{(0.14)}$	$1.53_{(0.51)}$
31	Sample	–	–	–	$0.81_{(0.18)}$	$0.91_{(0.19)}$	$1.60_{(0.51)}$	–
	Banding	$0.79_{(0.20)}$	$0.79_{(0.18)}$	$0.81_{(0.19)}$	$0.73_{(0.17)}$	$0.83_{(0.16)}$	$1.56_{(0.51)}$	–
	NERCOME	$0.66_{(0.13)}$	$0.64_{(0.14)}$	$0.73_{(0.15)}$	$0.70_{(0.19)}$	$0.79_{(0.15)}$	$1.60_{(0.55)}$	–
	CRC	$0.65_{(0.12)}$	$0.64_{(0.14)}$	$0.74_{(0.15)}$	$0.69_{(0.18)}$	$0.79_{(0.16)}$	$1.60_{(0.55)}$	–
	NONLIN	$0.66_{(0.12)}$	$0.65_{(0.15)}$	$0.73_{(0.15)}$	$0.69_{(0.18)}$	$0.81_{(0.16)}$	$1.60_{(0.55)}$	–
61	Sample	–	–	–	$0.97_{(0.22)}$	$1.59_{(0.51)}$	–	–
	Banding	$0.79_{(0.18)}$	$0.87_{(0.20)}$	$0.78_{(0.21)}$	$0.87_{(0.19)}$	$1.54_{(0.51)}$	–	–
	NERCOME	$0.64_{(0.12)}$	$0.77_{(0.16)}$	$0.72_{(0.20)}$	$0.82_{(0.17)}$	$1.60_{(0.53)}$	–	–
	CRC	$0.64_{(0.13)}$	$0.77_{(0.16)}$	$0.72_{(0.20)}$	$0.83_{(0.17)}$	$1.60_{(0.53)}$	–	–
	NONLIN	$0.65_{(0.13)}$	$0.78_{(0.16)}$	$0.72_{(0.21)}$	$0.86_{(0.19)}$	$1.61_{(0.54)}$	–	–
91	Sample	–	–	–	$1.60_{(0.47)}$	–	–	–
	Banding	$0.85_{(0.23)}$	$0.85_{(0.30)}$	$0.87_{(0.18)}$	$1.52_{(0.47)}$	–	–	–
	NERCOME	$0.77_{(0.21)}$	$0.73_{(0.21)}$	$0.82_{(0.17)}$	$1.59_{(0.51)}$	–	–	–
	CRC	$0.77_{(0.21)}$	$0.73_{(0.22)}$	$0.82_{(0.17)}$	$1.59_{(0.52)}$	–	–	–
	NONLIN	$0.77_{(0.21)}$	$0.75_{(0.23)}$	$0.85_{(0.18)}$	$1.60_{(0.52)}$	–	–	–
121	Sample	–	–	–	–	–	–	–
	Banding	$0.82_{(0.25)}$	$0.88_{(0.20)}$	$1.55_{(0.48)}$	–	–	–	–
	NERCOME	$0.69_{(0.17)}$	$0.76_{(0.13)}$	$1.59_{(0.50)}$	–	–	–	–
	CRC	$0.69_{(0.17)}$	$0.75_{(0.13)}$	$1.59_{(0.50)}$	–	–	–	–
	NONLIN	$0.70_{(0.19)}$	$0.78_{(0.14)}$	$1.60_{(0.51)}$	–	–	–	–
151	Sample	–	–	–	–	–	–	–
	Banding	$0.82_{(0.15)}$	$1.84_{(0.58)}$	–	–	–	–	–
	NERCOME	$0.73_{(0.12)}$	$1.50_{(0.47)}$	–	–	–	–	–
	CRC	$0.72_{(0.13)}$	$1.50_{(0.47)}$	–	–	–	–	–
	NONLIN	$0.73_{(0.13)}$	$1.50_{(0.48)}$	–	–	–	–	–
181	Sample	–	–	–	–	–	–	–
	Banding	$1.91_{(0.74)}$	–	–	–	–	–	–
	NERCOME	$1.45_{(0.50)}$	–	–	–	–	–	–
	CRC	$1.46_{(0.50)}$	–	–	–	–	–	–
	NONLIN	$1.48_{(0.52)}$	–	–	–	–	–	–

this period when sample size is larger than or equal to 90. Yet banding is not as good as others when sample size is 30 or 60. In fact, the best average performance for this forecast period is achieved by NERCOME when sample size is just 30, that is, using just the 30 days of data prior to the 29-day forecast period. Moreover, from the table, it is clear that for other forecasting periods, using just the prior 30 days of data for forecasting is in general better than using more past data for NERCOME, CRC grand average and NONLIN, which significantly outperform banding. This tells us that the underlying independence assumption for the $\mathbf{y}_i$'s could be only true locally rather than across the whole data set.

Note also that on all of the forecast periods and sample sizes, NERCOME, CRC grand average and NONLIN are essentially the same, with either NERCOME and CRC grand average outperforming NONLIN very slightly on average.

6. Conclusion. We investigate the problem of high dimensional covariance matrix estimation through regularizing the eigenvalues. Considering the class of rotation-equivariant covariance matrix, we split the data into two parts, and obtain a regularization of eigenvalues which is proved to be asymptotically optimal in the sense of minimizing the Frobenius loss. Incidentally, the inverse of this estimator is also asymptotically optimal in estimating the precision matrix with respect to the inverse Stein's loss. We also proved that our estimator is almost surely positive definite, and is asymptotically efficient relative to an ideal estimator which assumes the knowledge of the true covariance matrix itself.

The method proposed is applicable to data from a factor model also, which considerably expands its scope of applicability. When estimating the covariance or the precision matrix for data from a factor model, one usually needs to estimate the number of unknown factors, the factors themselves and the factor loading matrix. Our method does not need these, and gives an estimator of the covariance and the precision matrix directly, while still enjoys the aforementioned asymptotic optimality. This is demonstrated to be particularly convenient and can result in better performance especially when factor analysis itself is not the ultimate aim. Simulation results demonstrate comparable or even better performance than other state-of-the-art methods in various scenarios.

SUPPLEMENTARY MATERIAL

Simulations and proofs of theorems in the paper (DOI: 10.1214/15-AOS1393SUPP; .pdf). We present five profiles of simulations and compare the performance of NERCOME to other state-of-the-art methods. We also present the proofs of Theorem 1, Theorem 3, Lemma 1, Theorem 5 and Theorem 6 in the paper.

REFERENCES

ABADIR, K. M., DISTASO, W. and ŽIKEŠ, F. (2014). Design-free estimation of variance matrices. *J. Econometrics* **181** 165–180. MR3209862

BAI, Z. and SILVERSTEIN, J. W. (2010). *Spectral Analysis of Large Dimensional Random Matrices*, 2nd ed. Springer, New York. MR2567175

BAI, Z. D. and YIN, Y. Q. (1993). Limit of the smallest eigenvalue of a large-dimensional sample covariance matrix. *Ann. Probab.* **21** 1275–1294. MR1235416

BICKEL, P. J. and LEVINA, E. (2008a). Covariance regularization by thresholding. *Ann. Statist.* **36** 2577–2604. MR2485008

BICKEL, P. J. and LEVINA, E. (2008b). Regularized estimation of large covariance matrices. *Ann. Statist.* **36** 199–227. MR2387969

CAI, T. T. and ZHOU, H. H. (2012). Optimal rates of convergence for sparse covariance matrix estimation. *Ann. Statist.* **40** 2389–2420. MR3097607

DEMIGUEL, V. and NOGALES, F. J. (2009). A generalized approach to portfolio optimization: Improving performance by constraining portfolio norms. *Management Science* **55** 798–812.

FAN, J., FAN, Y. and LV, J. (2008). High dimensional covariance matrix estimation using a factor model. *J. Econometrics* **147** 186–197. MR2472991

FAN, J., LIAO, Y. and MINCHEVA, M. (2011). High-dimensional covariance matrix estimation in approximate factor models. *Ann. Statist.* **39** 3320–3356. MR3012410

FAN, J., LIAO, Y. and MINCHEVA, M. (2013). Large covariance estimation by thresholding principal orthogonal complements. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **75** 603–680. MR3091653

FRIEDMAN, J., HASTIE, T. and TIBSHIRANI, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics* **9** 432–441.

HUANG, J. Z., LIU, N., POURAHMADI, M. and LIU, L. (2006). Covariance matrix selection and estimation via penalised normal likelihood. *Biometrika* **93** 85–98. MR2277742

JAMES, W. and STEIN, C. (1961). Estimation with quadratic loss. In *Proc. 4th Berkeley Sympos. Math. Statist. and Prob. Contributions to the Theory of Statistics* **1** 361–379. Univ. California Press, Berkeley, CA. MR0133191

LAM, C. (2015). Supplement to "Nonparametric eigenvalue-regularized precision or covariance matrix estimator." DOI:10.1214/15-AOS1393SUPP.

LAM, C. and FAN, J. (2009). Sparsistency and rates of convergence in large covariance matrix estimation. *Ann. Statist.* **37** 4254–4278. MR2572459

LAM, C. and YAO, Q. (2012). Factor modeling for high-dimensional time series: Inference for the number of factors. *Ann. Statist.* **40** 694–726. MR2933663

LAM, C., YAO, Q. and BATHIA, N. (2011). Estimation of latent factors for high-dimensional time series. *Biometrika* **98** 901–918. MR2860332

LEDOIT, O. and PÉCHÉ, S. (2011). Eigenvectors of some large sample covariance matrix ensembles. *Probab. Theory Related Fields* **151** 233–264. MR2834718

LEDOIT, O. and WOLF, M. (2004). A well-conditioned estimator for large-dimensional covariance matrices. *J. Multivariate Anal.* **88** 365–411. MR2026339

LEDOIT, O. and WOLF, M. (2012). Nonlinear shrinkage estimation of large-dimensional covariance matrices. *Ann. Statist.* **40** 1024–1060. MR2985942

LEDOIT, O. and WOLF, M. (2013a). Optimal estimation of a large-dimensional covariance matrix under Stein's loss. ECON—Working Papers 122, Dept. Economics, Univ. Zürich.

LEDOIT, O. and WOLF, M. (2013b). Spectrum estimation: A unified framework for covariance matrix estimation and PCA in large dimensions. ECON—Working Papers 105, Dept. Economics, Univ. Zürich.

MARČENKO, V. and PASTUR, L. (1967). Distribution of eigenvalues for some sets of random matrices. *Math. USSR-Sb* **1** 457–483.

MEINSHAUSEN, N. and BÜHLMANN, P. (2006). High-dimensional graphs and variable selection with the lasso. *Ann. Statist.* **34** 1436–1462. MR2278363

POURAHMADI, M. (2007). Cholesky decompositions and estimation of a covariance matrix: Orthogonality of variance-correlation parameters. *Biometrika* **94** 1006–1013. MR2376812

ROTHMAN, A. J., LEVINA, E. and ZHU, J. (2009). Generalized thresholding of large covariance matrices. *J. Amer. Statist. Assoc.* **104** 177–186. MR2504372

SILVERSTEIN, J. W. and CHOI, S.-I. (1995). Analysis of the limiting spectral distribution of large-dimensional random matrices. *J. Multivariate Anal.* **54** 295–309. MR1345541

STEIN, C. (1975). Estimation of a covariance matrix. In *Rietz Lecture*, 39*th Annual Meeting IMS*. Atlanta, GA.

STEIN, C. (1986). Lectures on the theory of estimation of many parameters. *J. Sov. Math.* **34** 1373–1403.

STOCK, J. and WATSON, M. (2005). Implications of dynamic factor models for var analysis. NBER working papers. No. 11467.

WON, J.-H., LIM, J., KIM, S.-J. and RAJARATNAM, B. (2013). Condition-number-regularized covariance estimation. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **75** 427–450. MR3065474

DEPARTMENT OF STATISTICS
LONDON SCHOOL OF ECONOMICS
AND POLITICAL SCIENCE
HOUGHTON STREET
LONDON, WC2A 2AE
UNITED KINGDOM
E-MAIL: C.Lam2@lse.ac.uk
URL: http://stats.lse.ac.uk/lam