Estimation of functionals of sparse covariance matrices

The Annals of Statistics
2015, Vol. 43, No. 6, 2706–2737
DOI: 10.1214/15-AOS1357

ESTIMATION OF FUNCTIONALS OF SPARSE COVARIANCE MATRICES

By Jianqing Fan[1,*], Philippe Rigollet[2,†] and Weichen Wang[3,*]

Princeton University[*] and Massachusetts Institute of Technology[†]

High-dimensional statistical tests often ignore correlations to gain simplicity and stability leading to null distributions that depend on functionals of correlation matrices such as their Frobenius norm and other ℓ_r norms. Motivated by the computation of critical values of such tests, we investigate the difficulty of estimation the functionals of sparse correlation matrices. Specifically, we show that simple plug-in procedures based on thresholded estimators of correlation matrices are sparsity-adaptive and minimax optimal over a large class of correlation matrices. Akin to previous results on functional estimation, the minimax rates exhibit an elbow phenomenon. Our results are further illustrated in simulated data as well as an empirical study of data arising in financial econometrics.

1. Introduction. Covariance matrices are at the core of many statistical procedures such as principal component analysis or linear discriminant analysis. Moreover, not only do they arise as natural quantities to capture interactions between variables but, as we illustrate below, they often characterize the asymptotic variance of commonly used estimators. Following the original papers of Bickel and Levina (2008a, 2008b), much work has focused on the inference of high-dimensional covariance matrices under sparsity [Cai and Liu (2011), Cai, Ren and Zhou (2013), Cai and Yuan (2012), Cai, Zhang and Zhou (2010), Cai and Zhou (2012), El Karoui (2008), Lam and Fan (2009), Ravikumar et al. (2011)] and other structural as-

Received February 2015; revised June 2015.
[1]Supported in part by NSF Grants DMS-12-06464 and DMS-14-06266 and NIH grant R01-GM072611-9.
[2]Supported in part by NSF Grants DMS-13-17308, CAREER-DMS-1053987 and by the Howard B. Wentz Jr. Junior Faculty award.
[3]Supported in part by NSF Grant DMS-12-06464.
AMS 2000 subject classifications. Primary 62H12; secondary 62H15, 62C20, 62H25.
Key words and phrases. Covariance matrix, functional estimation, high-dimensional testing, minimax, elbow effect.

sumptions related to sparse principal component analysis [Amini and Wainwright (2009), Berthet and Rigollet (2013a, 2013b), Birnbaum et al. (2013), Cai, Ma and Wu (2013, 2015), Johnstone and Lu (2009), Levina and Vershynin (2012), Rothman, Levina and Zhu (2009), Ma (2013), Onatski, Moreira and Hallin (2013), Paul and Johnstone (2012), Fan, Fan and Lv (2008), Fan, Liao and Mincheva (2011, 2013), Jung and Marron (2009), Vu and Lei (2012), Zou, Hastie and Tibshirani (2006)]. This area of research is very active and, as a result, this list of references is illustrative rather than comprehensive. This line of work can be split into two main themes: estimation and detection. The former is the main focus of the present paper. However, while most of the literature has focused on estimating the covariance matrix itself, under various performance measures, we depart from this line of work by focusing on functionals of the covariance matrix rather than the covariance matrix itself.

Estimation of functionals of unknown signals such as regression functions or densities is known to be different in nature from estimation of the signal itself. This problem has received most attention in nonparametric estimation, originally in the Gaussian white noise model [Ibragimov, Nemirovskiĭ and Khas'minskiĭ (1987), Nemirovskiĭ and Khas'minskiĭ (1987), Fan (1991), Efromovich and Low (1996)] [see also Nemirovski (2000) for a survey of results in the Gaussian white noise model] and later extended to density estimation [Hall and Marron (1987), Bickel and Ritov (1988)] and various other models such as regression [Donoho and Nussbaum (1990), Cai and Low (2005, 2006), Klemelä (2006)] and inverse problems [Butucea (2007), Butucea and Meziani (2011)]. Most of these papers study the estimation of quadratic functionals and, interestingly, exhibit an elbow in the rates of convergence: there exists a critical regularity parameter below which the rate of estimation is nonparametric and above which, it becomes parametric. As we will see below the phenomenon also arises when regularity is measured by sparsity.

Over the past decade, sparsity has become the prime measure of regularity, both for its flexibility and generality. In particular, smooth functions can be viewed as functions with a sparse expansion in an appropriate basis. At a high level, sparsity assumes that many of the unknown parameters are equal to zero or nearly so, so that the few nonzero parameters can be consistently estimated using a small number of observations relative to the apparent dimensionality of the problem. Moreover, sparsity acts not only as a regularity parameter that stabilizes statistical procedures but also as key feature for interpretability. Indeed, it is often the case that setting many parameters to zero simply corresponds to a simpler sub-model. The main idea is to let data select the correct sub-model. This is the case in particular for covariance matrix estimation where zeros in the matrix correspond to

uncorrelated variables. Yet, while the value of sparsity for covariance matrix estimation has been well established, to the best of our knowledge, this paper provides the first analysis for the estimation of functionals of sparse covariance matrix. Indeed, the actual performance of many estimators critically depends on such functionals. Therefore, accurate functional estimation leads to a better understanding the performance of many estimators and can ultimately serve as a guide to selecting the best estimator. Applications of our results are illustrated in Section 2.

Our work is not only motivated by real applications, but also by a natural extension of the theoretical analysis carried out in the sparse Gaussian sequence model [Cai and Low (2005)]. In that paper, Cai and Low assume that the unknown parameter θ belong to an ℓ_q-ball, where $q > 0$ can be arbitrarily close to 0. Such balls are known to emulate sparsity and actually correspond to a more accurate notion of sparsity for signal θ that is encountered in applications [see, e.g., Foucart and Rauhut (2013)]. They also show that a nonquadratic estimator can be fully efficient to estimate quadratic functionals. We extend some of these results to covariance matrix estimation. Such an extension is not trivial since, unlike the Gaussian sequence model, covariance matrix lies at high-dimensional manifolds and its estimation exhibits complicated dependencies in the structure of the noise.

We also compare our results for optimal rates of estimating matrix functionals with that of estimating matrix itself. Many methods have been proposed to estimate covariance matrix in different sense of sparsity using different techniques including thresholding [Bickel and Levina (2008a)], tapering [Bickel and Levina (2008b), Cai, Zhang and Zhou (2010), Cai and Zhou (2012)] and penalized likelihood [Lam and Fan (2009)] to name only a few. These methods often lead to minimax optimal rates in various classes and under several metrics [Cai, Zhang and Zhou (2010), Cai and Zhou (2012), Rigollet and Tsybakov (2012)]. However, the optimal rates of estimating matrix functionals have not yet been covered by much literature. Intuitively, it should have faster rates of convergence on estimating a matrix functional than itself since it is just a one-dimensional estimating problem and the estimating error cancel with each other when we sum those elements together. We will see this is indeed the case when we compare the minimax rates of estimating matrix functionals with those of estimating matrices.

The rest of the paper is organized as follows. We begin in Section 2 by two motivating examples of high-dimensional hypothesis testing problems: a two-sample testing problem of Gaussian means that arises in genomics and validating the efficiency of markets based on the Capital Asset Pricing Model (CAPM). Next, in Section 3, we introduce an estimator of the quadratic functional of interest that is based on the thresholding estimator introduced in Bickel and Levina (2008a). We also prove its optimality in a minimax sense over a large class of sparse covariance matrices. The study is further extended

to estimating other measures of sparsity of covariance matrix. Finally, we study the numerical performance of our estimator in Section 5 on simulated experiments as well as in the framework of the two applications described in Section 2. Due to space restrictions, the proofs for the upper bounds are relegated to the Appendix in the supplementary material [Fan, Rigollet and Wang (2015)].

Notation: Let d be a positive integer. The space of $d \times d$ positive semi-definite matrices is denoted by $\mathbf{S}_d^+$. For any two integers $c < d$, define $[c : d] = \{c, c+1, \ldots, d\}$ to be the sequence of contiguous integers between c and d, and we simply write $[d] = \{1, \ldots, d\}$. I_d denotes the identity matrix of $\mathbb{R}^d$. Moreover, for any subset $S \subset [d]$, denote by $\mathbf{1}_S \in \{0,1\}^d$ the column vector with jth coordinate equal to one iff $j \in S$. In particular, $\mathbf{1}_{[d]}$ denotes the d dimensional vector of all ones.

We denote by tr the trace operator on square matrices and by diag (resp., off) the linear operator that sets to 0 all the off diagonal (resp., diagonal) elements of a square matrix. The Frobenius norm of a real matrix M is denoted by $\|M\|_{\mathsf{F}}$ and is defined by $\|M\|_{\mathsf{F}} = \sqrt{\mathrm{tr}(M^\top M)}$. Note that $\|M\|_{\mathsf{F}}$ is a the Hilbert–Schimdt norm associated with the inner product $\langle A, B\rangle = \mathrm{tr}(A^\top B)$ defined on the space of real rectangular matrices of the same size. Moreover, $|A|$ denotes the determinant of a square matrix A. The variance of a random variable X is denote by $\mathrm{var}(X)$.

In the proofs, we often employ C to denote a generic positive constant that may change from line to line.

2. Two motivating examples. In this section, we describe our main motivation for estimating quadratic functionals of a high-dimensional covariance matrix in the light of two applications to high-dimensional testing problems. The first one is a high-dimensional two-sample hypothesis testing with applications in gene-set testing. The second example is about testing the validity of the *capital asset pricing model (CAPM)* from financial economics.

2.1. *Two-sample hypothesis testing in high-dimensions.* In various statistical applications, in particular in genomics, the dimensionality of the problems is so large that statistical procedures involving inverse covariance matrices are not viable due to its lack of stability both from a statistical and numerical point of view. This limitation can be well illustrated on a showcase example: two-sample hypothesis testing [Bai and Saranadasa (1996)] in high-dimensions.

Suppose that we observe two independent samples $X_1^{(1)}, \ldots, X_{n_1}^{(1)} \in \mathbb{R}^p$ that are i.i.d. $\mathcal{N}(\mu_1, \Sigma_1)$ and $X_1^{(2)}, \ldots, X_{n_2}^{(2)} \in \mathbb{R}^p$ that are i.i.d. $\mathcal{N}(\mu_2, \Sigma_2)$. Let $n = n_1 + n_2$. The goal is to test $H_0 : \mu_1 = \mu_2$ vs. $H_1 : \mu_1 \neq \mu_2$.

Assume first that $\Sigma_1 = \Sigma_2 = \Sigma$. In this case, Hotelling's test is commonly employed when p is small. Nevertheless, when p is large, Bai and Saranadasa

(1996) showed that the test based on Hotelling's T^2 has low power and suggest a new statistics M for the random matrix asymptotic regime where $n, p \to \infty, \frac{n}{p} \to \gamma > 0, \frac{n_1}{n_1+n_2} \to \kappa \in (0,1)$. The statistics, implementing the naive Bayes rule, is defined as

$$M = (\bar{X}^{(1)} - \bar{X}^{(2)})^\top (\bar{X}^{(1)} - \bar{X}^{(2)}) - \frac{n}{n_1 n_2} \operatorname{tr}(\hat{\Sigma}),$$

and is proved to be asymptotically normal under the null hypothesis with

$$\operatorname{var}(M) = 2\frac{n(n-1)}{(n_1 n_2)^2} \|\Sigma\|_{\mathsf{F}}^2 (1 + o(1)).$$

Clearly, the asymptotic variance of M depends on the unknown covariance matrix Σ through its quadratic functional, and in order to compute the critical value of the test, Bai and Saranadasa suggest to estimate $\|\Sigma\|_{\mathsf{F}}^2$ by the quantity

$$B^2 = \frac{n^2}{(n+2)(n-1)} \left[\|\hat{\Sigma}\|_{\mathsf{F}}^2 - \frac{1}{n} (\operatorname{tr}(\hat{\Sigma}))^2 \right].$$

They show that B^2 is a ratio-consistent estimator of $\|\Sigma\|_{\mathsf{F}}^2$ in the sense that $B^2 = (1 + o_P(1))\|\Sigma\|_{\mathsf{F}}^2$. Clearly, this solution does not leverage any sparsity assumption and may suffer from power deficiency if the matrix Σ is indeed sparse. Rather, if the covariance matrix Σ is believed to be sparse, one may prefer to use a thresholded estimator for Σ as in Bickel and Levina (2008a) rather than the empirical covariance matrix $\hat{\Sigma}$. In this case, we estimate $\|\Sigma\|_{\mathsf{F}}^2$ by $\widehat{\|\Sigma\|_{\mathsf{F}}^2} = \sum_{i,j=1}^p \hat{\sigma}_{ij}^2 \mathbb{1}\{|\hat{\sigma}_{ij}| > \tau\}$, where $\{\hat{\sigma}_{ij}, i, j \in [p]\}$ could be any consistent estimator of σ_{ij} and $\tau > 0$ is a threshold parameter.

More recently, Chen and Qin (2010) took into account the case $\Sigma_1 \neq \Sigma_2$ and proposed a test statistic based on an unbiased estimate of each of the three quantities in $\|\mu_1 - \mu_2\|^2 = \|\mu_1\|^2 + \|\mu_2\|^2 - 2\mu_1^\top \mu_2$. In this case, the quantities $\|\Sigma_i\|_{\mathsf{F}}^2$, $i = 1, 2$ and $\langle \Sigma_1, \Sigma_2 \rangle$ appear in the asymptotic variance. The detailed formulation and assumptions of this statistic, as well as discussions about other testing methods such as Srivastava and Du (2008), are provided in the supplementary material [Fan, Rigollet and Wang (2015)] for completeness. If Σ_1 and Σ_2 are indeed sparse, akin to the above reasoning, we can also estimate $\|\Sigma_i\|_{\mathsf{F}}^2$, $i = 1, 2$ and $\langle \Sigma_1, \Sigma_2 \rangle$ using thresholding to leverage sparsity assumption. It is not hard to derive a theory for estimating quadratic functionals involving two covariance matrices but the details of this procedure are beyond the scope of the present paper.

2.2. *Testing high-dimensional CAPM model.* The capital asset pricing model (CAPM) is a simple financial model that postulates how individual asset returns are related to the market risks. Specifically, the individual

excessive return $Y_t^{(i)}$ of asset $i \in [N]$ over the risk-free rate at time $t \in [T]$ can be expressed as an affine function of a vector of K risk factors $f_t \in \mathbb{R}^K$:

$$Y_t^{(i)} = \alpha_i + \beta_i^\top f_t + \varepsilon_t^{(i)}, \tag{2.1}$$

where we assume for any $t \in [T]$, $f_t \in \mathbb{R}^K$ are observed. The case $K = 1$ with f_t being the excessive return of the market portfolio corresponds to the CAPM [Sharpe (1964), Lintner (1965), Mossin (1966)]. It is nowadays more common to employ the Fama–French three-factor model [see Fama and French (1993) for a definition] for the US equity market, corresponding to $K = 3$.

For simplicity, let us rewrite the model (2.1) in the vectorial form

$$Y_t = \alpha + Bf_t + \varepsilon_t, \qquad t \in [T].$$

The multi-factor pricing model postulates $\alpha = 0$. Namely, all returns are fully compensated by their risks: no extra returns are possible and the market is efficient. This leads us to naturally consider the hypothesis testing problem $H_0 : \alpha = 0$ vs. $H_1 : \alpha \neq 0$.

Let $\hat{\alpha}$ and $\hat{B}$ be the least-squares estimate and $\hat{\varepsilon}_t = Y_t - \hat{\alpha} - \hat{B}f_t$ be a residual vector. Then an unbiased estimator of $\Sigma = \mathrm{var}(\varepsilon_t)$ is

$$\tilde{\Sigma} = \frac{1}{T - K - 1} \sum_{t=1}^{T} \hat{\varepsilon}_t \hat{\varepsilon}_t^\top.$$

Let $\hat{D} = \mathrm{diag}(\tilde{\Sigma})$ and $M_F = I_T - F(F^\top F)^{-1}F^\top$ where $F = (f_1, \ldots, f_T)^\top$. Define $W_d = (\mathbf{1}_{[T]}^\top M_F \mathbf{1}_{[T]})\hat{\alpha}^\top \hat{D}^{-1}\hat{\alpha}$ the Wald-type of test statistics with correlation ignored, whose normalized version is given by

$$J_\alpha = \frac{W_d - \mathbb{E}(W_d)}{\sqrt{\mathrm{var}(W_d)}}. \tag{2.2}$$

Under some conditions, it was shown by Pesaran and Yamagata (2012) that, under H_0, $J_\alpha \to \mathcal{N}(0, 1)$ as $N \to \infty$. Moreover, if $\varepsilon_t^{(i)}$'s are i.i.d. Gaussian, it holds that $\mathbb{E}(W_d) = \nu N/(\nu - 2)$ and

$$\mathrm{var}(W_d) = \frac{2N(\nu - 1)}{\nu - 4} \left(\frac{\nu}{\nu - 2} \right)^2 [1 + (N - 1)\bar{\rho}^2 + O(\nu^{-1/2})],$$

where $\nu = T - K - 1$ is the degrees of freedom and

$$\bar{\rho}^2 = \frac{2}{N(N - 1)} \sum_{i=2}^{N} \sum_{j=1}^{i-1} \rho_{ij}^2,$$

where $\rho = D^{-1/2}\Sigma D^{-1/2}$ with $D = \mathrm{diag}(\Sigma)$ is the correlation matrix of the stationary process $(\varepsilon_t)_{t \in [T]}$. The authors go on to propose an estimator of

the quadratic functional $\bar{\rho}^2$ by replacing the correlation coefficients ρ_{ij} in the above expression by $\hat{\rho}_{i,j}\mathbb{1}(|\hat{\rho}_{ij}| > \tau)$ where $(\hat{\rho}_{ij})_{i,j\in[N]} = \hat{D}^{-1/2}\tilde{\Sigma}\hat{D}^{-1/2}$ and $\tau > 0$ is a threshold parameter. However, they did not provide any analysis of this method, nor any guidance to chose τ.

3. Optimal estimation of quadratic functionals. In the previous section, we have described rather general questions involving the estimation of quadratic functions of covariance or correlation matrices. We begin by observing that consistent estimation of $\|\Sigma\|_{\mathsf{F}}^2$ is impossible unless $p = o(n)$. This precludes in particular the high-dimensional framework that motivates our study.

Our goal is to estimate the Frobenius norm $\|\Sigma\|_{\mathsf{F}}^2$ of a sparse $p \times p$ covariance matrix Σ using n i.i.d. observations $X_1, \dots, X_n \sim \mathcal{N}(0, \Sigma)$. Observe that $\|\Sigma\|_{\mathsf{F}}^2$ can be decomposed as $\|\Sigma\|_{\mathsf{F}}^2 = Q(\Sigma) + D(\Sigma)$ where $Q(\Sigma) = \sum_{i\neq j}\sigma_{ij}^2$ corresponds to the off-diagonal elements and $D(\Sigma) = \sum_j \sigma_{jj}^2$ corresponds to the diagonal elements. The following theorem, implies that even if $\Sigma = \operatorname{diag}(\Sigma)$ is diagonal, the quadratic functional $\|\Sigma\|_{\mathsf{F}}^2$ cannot be estimated consistently in absolute error if $p \geq n$. Note that the situation is quite different when it comes to *relative error*. Indeed, the estimator of Bai and Saranadasa (1996) is consistent in relative error with no sparsity assumption even in the high-dimensional regime. Study of the relative error in the presence of sparsity is an interesting question that deserves further developments.This makes sense intuitively as the diagonal of Σ consists of p unknown parameters while we have only n observations.

PROPOSITION 3.1. *Fix $n, p \geq 1$ and let*

$$\mathcal{D}_p = \{\Sigma \in \mathbf{S}_p^+ : \Sigma = \operatorname{diag}(\Sigma), \Sigma_{ii} \leq 1\}$$

be the class of diagonal covariance matrices with diagonal elements bounded by 1. Then there exists a universal constant $C > 0$ such that

$$\inf_{\hat{D}} \sup_{\Sigma\in\mathcal{D}_p} \mathbb{E}[\hat{D} - D(\Sigma)]^2 \geq C\frac{p}{n}.$$

In particular, it implies that

$$\inf_{\hat{F}} \sup_{\Sigma\in\mathcal{D}_p} \mathbb{E}[\hat{F} - \|\Sigma\|_{\mathsf{F}}^2]^2 \geq C\frac{p}{n},$$

where the infima are taken with over all real valued measurable functions of the observations.

PROOF. Our lower bounds rely on standard arguments from minimax theory. We refer to Chapter 2 of Tsybakov (2009) for more details. In the

sequel, let $\mathsf{KL}(P, \bar{P})$ denote the Kullback–Leibler divergence between two distributions P and $\bar{P}$, where $P \ll \bar{P}$. It is defined by

$$\mathsf{KL}(P, \bar{P}) = \int \log\left(\frac{\mathrm{d}P}{\mathrm{d}\bar{P}}\right) \mathrm{d}P.$$

We are going to employ a simple two-point lower bound. Fix $\varepsilon \in (0, 1/2)$ and let P_p^n (resp., $\bar{P}_p^n$) denote the distribution of a sample $X_1, \ldots, X_n$ where $X_1 \sim \mathcal{N}(0, I_p)$ [resp., $X_1 \sim \mathcal{N}(0, (1-\varepsilon)I_p)$]. Next, observe that $I_p, (1-\varepsilon)I_p \subset \mathcal{D}_p$ so that

$$\sup_{\Sigma \in \mathcal{D}_p} \mathbb{E}|\hat{D} - D(\Sigma)| \geq \max_{\Sigma \in \{I_p, (1-\varepsilon)I_p\}} \mathbb{E}|\hat{D} - D(\Sigma)|. \tag{3.1}$$

Moreover, $|D(I_p) - D((1-\varepsilon)I_p)| = p(2\varepsilon - \varepsilon^2) > p\varepsilon$. Then it follows from the Markov inequality that

$$\begin{aligned} \frac{1}{p\varepsilon} \max_{\Sigma \in \{I_p, (1-\varepsilon)I_p\}} \mathbb{E}|\hat{D} - D(\Sigma)| &\geq \max_{\Sigma \in \{I_p, (1-\varepsilon)I_p\}} \mathbb{P}[|\hat{D} - D(\Sigma)| > p\varepsilon] \\ &\geq \frac{1}{4} \exp[-\mathsf{KL}(P_p^n, \bar{P}_p^n)], \end{aligned} \tag{3.2}$$

where the last inequality follows from Theorem 2.2(iii) of Tsybakov (2009).

Completion of the proof requires an upper bound on $\mathsf{KL}(P_p^n, \bar{P}_p^n)$. To that end, note that it follows from the chain rule and simple algebra that

$$\mathsf{KL}(P_p^n, \bar{P}_p^n) = np\mathsf{KL}(P_1^1, \bar{P}_1^1) = \frac{np}{2}\left[\log(1-\varepsilon) + \frac{\varepsilon}{1-\varepsilon}\right] \leq \frac{np}{2}\frac{\varepsilon^2}{1-\varepsilon} \leq np\varepsilon^2.$$

Taking now $\varepsilon = 1/(2\sqrt{np}) \leq 1/2$ yields $\mathsf{KL}(P_p^n, Q_p^n) \leq 1/4$. Together with (3.1) and (3.2), it yields

$$\inf_{\hat{D}} \sup_{\Sigma \in \mathcal{D}_p} \mathbb{E}|\hat{D} - D(\Sigma)| \geq \frac{1}{8e^{1/4}}\sqrt{\frac{p}{n}}.$$

To complete the proof, we square the above inequality and employ Jensen's inequality. □

To overcome the above limitation, we consider the following class of sparse covariance matrices (indeed correlation matrices). For any $q \in [0, 2), R > 0$ let $\mathcal{F}_q(R)$ denote the set of $p \times p$ covariance matrices defined by

$$\mathcal{F}_q(R) = \left\{\Sigma \in \mathbf{S}_p^+ : \sum_{i \neq j} |\sigma_{ij}|^q \leq R, \mathrm{diag}(\Sigma) = I_p\right\}. \tag{3.3}$$

Note that for this class of functions, we assume that the variance along each coordinate is normalized to 1. This normalization is frequently obtained by

sample estimates, as shown in the previous section. This simplified assumption is motivated also by Proposition 3.1 above which implies that $\|\Sigma\|_{\mathsf{F}}^2$ for general covariance matrix cannot be estimated accurately in absolute error in the large p small n regime since sparsity assumptions on the diagonal elements are implausible. Note that the condition $\operatorname{diag}(\Sigma) = I_p$ implies that diagonal elements $D(\Sigma)$ of matrices in $\mathcal{F}_q(R)$ can be estimated without error so that we could possibly achieve consistency even if the case of large p small n.

Matrices in $\mathcal{F}_q(R)$ have many small coefficients for small values of q and R. In particular, when $q = 0$, there are no more than R entries of nonvanishing correlations. Following a major trend in the estimation of sparse covariance matrices [Bickel and Levina (2008a, 2008b), Cai and Liu (2011), Cai and Yuan (2012), Cai, Zhang and Zhou (2010), Cai and Zhou (2012), El Karoui (2008), Lam and Fan (2009)], we employ a thresholding estimator of the covariance matrix as a running horse to estimate the quadratic functionals. From the n i.i.d. observations $X_1, \dots, X_n \sim \mathcal{N}(0, \Sigma)$, we form the empirical covariance matrix $\hat{\Sigma}$ that is defined by

$$\hat{\Sigma} = \frac{1}{n} \sum_{k=1}^{n} X_k X_k^\top \tag{3.4}$$

with elements $\hat{\Sigma} = \{\hat{\sigma}_{ij}\}_{ij}$ and for any threshold $\tau > 0$, let $\tilde{\Sigma}_\tau = \{\tilde{\sigma}_{ij}\}_{ij}$ denote the thresholding estimator of Σ defined by $\tilde{\sigma}_{ij} = \hat{\sigma}_{ij} \mathbb{1}\{|\hat{\sigma}_{ij}| > \tau\}$ if $i \neq j$ and $\tilde{\sigma}_{ii} = \hat{\sigma}_{ii}$.

Next, we employ a simple plug-in estimator for $Q(\Sigma)$:

$$\widehat{Q(\Sigma)} = Q(\tilde{\Sigma}_\tau) = \sum_{i \neq j} \hat{\sigma}_{ij}^2 \mathbb{1}\{|\hat{\sigma}_{ij}| > \tau\}. \tag{3.5}$$

Note that no value of the diagonal elements is used to estimate $Q(\Sigma)$.

In the rest of this section, we establish that $\widehat{Q(\Sigma)}$ is minimax adaptive over the scale $\{\mathcal{F}_q(R), q \in [0, 2), R > 0\}$. Interestingly, we will see that the minimax rate presents an elbow as often in quadratic functional estimation.

THEOREM 3.1. *Assume that $\gamma \log(p) < n$ for some constant $\gamma > 8$ and fix $C_0 \geq 4$. Consider the threshold*

$$\tau = 2C_0 \sqrt{\frac{\gamma \log p}{n}},$$

and assume that $\tau \leq 1$. Then, for any $q \in [0, 2), R > 0$, the plug-in estimator $Q(\tilde{\Sigma}_\tau)$ satisfies

$$\mathbb{E}[(Q(\tilde{\Sigma}_\tau) - Q(\Sigma))^2] \leq C_1 \psi_{n,p}(q, R) + C_2 p^{4-\gamma/2},$$

where

$$\psi_{n,p}(q,R) = \frac{R^2}{n} \vee R^2 \left(\frac{\log p}{n}\right)^{2-q},$$

and C_1, C_2 are positive constants depending on γ, C_0, q.

The proof is postponed to the supplementary material.

Note that the rates $\psi_{n,p}(q,R)$ present an elbow at $q = 1 - \log\log p/\log n$ as usually the case in functional estimation. We now argue that the rates $\psi_{n,p}(q,R)$ are optimal in a minimax sense for a wide range of settings. In particular, the elbow effect arising from the maximum in the definition of ψ is not an artifact. In the following theorem, we emphasize the dependence on Σ by using the notation $\mathbb{E}_\Sigma$ for the expectation with respect to the distribution of the sample $X_1, \ldots, X_n$, where $X_i \sim \mathcal{N}(0, \Sigma)$.

THEOREM 3.2. *Fix $q \in [0,2), R > 0$ and assume $2\log p < n$ and $R^2 < (p-1)n^{-q}/2$. Then there exists a positive constant $C_3 > 0$ such that*

$$\inf_{\hat{Q}} \sup_{\Sigma \in \mathcal{F}_q(R)} \mathbb{E}_\Sigma[(\hat{Q} - Q(\Sigma))^2] \geq C_3 \phi_{n,p}(q,R),$$

where $\phi_{n,p}(q,R)$ is defined by

$$\phi_{n,p}(q,R) = \frac{R^2}{n} \vee \left\{ R^2 \left(\frac{\log((p-1)/(R^2 n^q) + 1)}{2n}\right)^{2-q} \wedge R^{4/q} \wedge 1 \right\} \tag{3.6}$$

and the infimum is taken over all measurable functions $\hat{Q}$ of the sample $X_1, \ldots, X_n$.

Before proceeding to the proof, a few remarks are in order.

1. The additional term of order $p^{4-\gamma/2}$ in Theorem 3.1 can be made negligible by taking γ large enough. To show this tradeoff explicitly, we decided keep this term.

2. When $1 \leq R^2 < p^\alpha n^{-q}$ for some constant $\alpha < 1$, a slightly stronger requirement than Theorem 3.2, the lower bound there can be written as

$$\phi_{n,p}(q,R) = \frac{R^2}{n} \vee \left\{ R^2 \left(\frac{\log p}{n}\right)^{2-q} \wedge 1 \right\}. \tag{3.7}$$

Observe that the above lower bound matches the upper bound presented in Theorem 3.1 when $R^{2/(2-q)} \log p \leq n$. Arguably, this is the most interesting range as it characterizes rates of convergence (to zero) rather than rates of divergence, that may be of different nature [see, e.g., Verzelen (2012)]. In other words, the rates given in (3.7) are minimax adaptive with respect to n, R, p and q. In our formulation, we allow $R = R_{n,p}$ to depend on other parameters of the problem. We choose here to keep the notation light.

3. The reason we choose correlation matrix class to present the elbow effect is just for simplicity. Actually, we can replace the constraint $\operatorname{diag}(\Sigma) = I_p$ in the definition of $\mathcal{F}_q(R)$ by boundedness of diagonal elements of Σ. Then for estimating off-diagonal elements $Q(\Sigma)$, following exactly the same derivation, the same elbow phenomenon has been noticed. Meanwhile, the optimal rate for estimating diagonal elements $D(\Sigma)$ is again of the order p/n. This optimal rate can be attained by the estimator

$$\widehat{D(\Sigma)} = \frac{1}{n(n-1)} \sum_{i=1}^{p} \sum_{k \neq j} X_{k,i}^2 X_{j,i}^2. \tag{3.8}$$

We omitted the proof here. Thus, if we do not have prior information about diagonal elements, we could still estimate optimally the quadratic functional of a covariance matrix by applying the thresholding method (3.5) for off-diagonal elements, together with (3.8) for diagonal elements.

4. The rate $\phi_{n,p}(q, R)$ presents the same elbow phenomenon at $q = 1$ observed in the estimation of functionals, starting independently with work of Bickel and Ritov (1988) and Fan (1991). Closer to the present setup is the work of Cai and Low (2005) who study the estimation of functionals of "sparse" sequences in the infinite Gaussian sequence model. There, a parameter controls the speed of decay of the unknown coefficients. Note that while smaller values q lead to sparser matrices Σ, no estimator can benefit further from sparsity below $q = 1$ [the estimator has a rate of convergence $O(R^2/n)$ for any $q < 1$], unlike in the case of estimation of Σ. Again, this is inherent to estimating functionals.

5. The condition $R^2 < (p-1)n^{-q}/2$ corresponds to the high-dimensional regime and allows us to keep clean terms in the logarithm. Similar assumptions are made in related literature [see, e.g., Cai and Zhou (2012)].

6. The optimal rates obtained here cannot be implied by existing ones for estimating sparse covariance matrices. In particular, the latter do not admit an elbow phenomenon. Specifically, Rigollet and Tsybakov (2012) showed the optimal rate for estimating Σ for $\Sigma \in \mathcal{F}_q(R)$ under the Frobenius norm is $\sqrt{R}(\log p/n)^{1/2-q/4}$ for $0 \leq q < 2$. Using this, it is not hard to derive with high probability,

$$|Q(\hat{\Sigma}) - Q(\Sigma)| \leq C_1 R \left(\frac{\log p}{n}\right)^{1/2-q/4} + C_2 R \left(\frac{\log p}{n}\right)^{1-q/2},$$

since $\|Q(\Sigma)\|_F = O(\sqrt{R})$ if nonvanishing correlations are bounded away from zero. On one hand, when $q < 2$ the first term always dominates so that we do not observe the elbow effect. In addition, the rate so obtained is not optimal.

We now turn to the proof of Theorem 3.2

PROOF OF THEOREM 3.2. To prove minimax lower bounds, we employ a standard technique that consists of reducing the estimation problem to a testing problem. We split this proof into two parts and begin by proving

$$\inf_{\hat{Q}} \sup_{\Sigma \in \mathcal{F}_q(R)} \mathbb{E}_\Sigma [\hat{Q} - Q(\Sigma)]^2 \geq C \frac{R^2}{n},$$

for some positive constant $C > 0$. To that end, for any $A \in \mathbf{S}_p^+$, let $\mathbb{P}_A$ denote the distribution of $X \sim \mathcal{N}(0, A)$. It is not hard to show if $|A| > 0$ and $|B| > 0$, $A, B \in \mathbf{S}_p^+$, then the Kullback–Leibler divergence between $\mathbb{P}_A$ and $\mathbb{P}_B$ is given by

$$\mathsf{KL}(\mathbb{P}_A, \mathbb{P}_B) = \frac{1}{2}\left[\log\left(\frac{|B|}{|A|}\right) + \operatorname{tr}(B^{-1}A) - p\right]. \tag{3.9}$$

Next, take A and B to be of the form

$$A^{(k)} = \begin{pmatrix} \mathbf{1}\mathbf{1}^\top & a\mathbf{1}\mathbf{1}^\top & 0 \\ a\mathbf{1}\mathbf{1}^\top & \mathbf{1}\mathbf{1}^\top & 0 \\ 0 & 0 & I_{p-k} \end{pmatrix}, \qquad B^{(k)} = \begin{pmatrix} \mathbf{1}\mathbf{1}^\top & b\mathbf{1}\mathbf{1}^\top & 0 \\ b\mathbf{1}\mathbf{1}^\top & \mathbf{1}\mathbf{1}^\top & 0 \\ 0 & 0 & I_{p-k} \end{pmatrix},$$

where $a, b \in (0, 1/2)$, 0 is a generic symbol to indicate that the missing space is filled with zeros, and $\mathbf{1}$ denotes a vector of ones of length $k/2$. Note that if we have random variables $(X, Y, Z_1, \ldots, Z_{p-2})$ chosen from distribution $\mathcal{N}(0, A^{(2)})$ meaning that Z_k's are independent with X, Y but the correlation between X and Y is a, then random vector $(X, \ldots, X, Y, \ldots, Y, Z_1, \ldots, Z_{p-k})$ with $k/2$ X's and Y's in it follows $\mathcal{N}(0, A^{(k)})$. It is obvious that these two matrices are degenerate and comes from perfectly correlated random variables. Since perfectly correlated random variables do not add new information, for such matrices, an application of (3.9) yields

$$\mathsf{KL}(\mathbb{P}_{A^{(k)}}, \mathbb{P}_{B^{(k)}}) = \mathsf{KL}(\mathbb{P}_{A^{(2)}}, \mathbb{P}_{B^{(2)}}) = \frac{1 - ab}{1 - b^2} - \frac{1}{2}\log\left(\frac{1 - a^2}{1 - b^2}\right) - 1.$$

Next, using the convexity inequality $\log(1 + x) \geq x - x^2/2$ for all $x > 0$, we get that

$$\mathsf{KL}(\mathbb{P}_{A^{(k)}}, \mathbb{P}_{B^{(k)}}) \leq \frac{(a-b)^2}{2(1-b^2)}\left[1 + \frac{(a+b)^2}{2(1-b^2)}\right] \leq 2(a-b)^2,$$

using the fact that $a, b \in (0, 1/2)$. Take now if $R > 4$

$$a = \frac{1}{4}, \qquad b = a + \frac{1}{4\sqrt{n}}, \qquad k = \sqrt{R}$$

so that we indeed have $a, b \in (0, 1/2)$ and also $A^{(k)}, B^{(k)} \in \mathcal{F}_q(R)$ obviously. If $R < 4$, take $k = 2, a = \sqrt{R}/8, b = a + \sqrt{R/64n}$ instead. Moreover, this choice

leads to $n\mathsf{KL}(\mathbb{P}_A, \mathbb{P}_B) \le 1/5$. Using standard techniques to reduce estimation problems to testing problems [see, e.g., Theorem 2.5 of Tsybakov (2009)], we find that

$$\inf_{\hat{Q}} \max_{\Sigma \in \{A,B\}} \mathbb{E}_\Sigma[(\hat{Q} - Q(\Sigma))^2] \ge C(Q(A) - Q(B))^2.$$

For the above choice of A and B, we have

$$(Q(A^{(k)}) - Q(B^{(k)}))^2 = \frac{k^4}{4}(a^2 - b^2)^2 \ge C\frac{R^2}{n}.$$

Since $A^{(k)}, B^{(k)} \in \mathcal{F}_q(R)$, the above two displays imply that

$$\inf_{\hat{Q}} \max_{\Sigma \in \mathcal{F}_q(R)} \mathbb{E}_\Sigma[(\hat{Q} - Q(\Sigma))^2] \ge C\frac{R^2}{n},$$

which completes the proof of the first part of the lower bound.

For the second part of the lower bound, we reduce our problem to a testing problem of the same flavor as Arias-Castro, Bubeck and Lugosi (2015), Berthet and Rigollet (2013b). Note, however, that our construction is different because the covariance matrices considered in these papers do not yield large enough lower bounds. We use the following construction.

Fix an integer $k \in [p-1]$ and let $\mathcal{S} = \{S \subset [p-1] : |S| = k\}$ denote the set of subsets of $[p-1]$ that have cardinality k. Fix $a \in (0,1)$ to be chosen later and for any $S \in \mathcal{S}$, recall that $\mathbf{1}_S$ is the column vector in $\{0,1\}^{p-1}$ with support given by S. For each $S \in \mathcal{S}$, we define the following $p \times p$ covariance matrix:

$$\Sigma_S = \begin{pmatrix} 1 & a\mathbf{1}_S^\top \\ a\mathbf{1}_S & I_{p-1} \end{pmatrix}. \tag{3.10}$$

Let $\mathbb{P}_0$ denote the distribution of $X \sim \mathcal{N}_p(0, I_p)$ and $\mathbb{P}_S$ denote the distribution of $X \sim \mathcal{N}_p(0, \Sigma_S)$. Let $\mathbb{P}_0^n$ (resp., $\mathbb{P}_S^n$) denote the distribution of $\mathbf{X} = (X_1, \dots, X_n)$ of a collection n i.i.d. random variables drawn from $\mathbb{P}_0$ (resp., $\mathbb{P}_S$). Moreover, let $\bar{\mathbb{P}}^n$ denote the distribution of $\mathbf{X}$ where the X_i's are drawn as follows: first draw S uniformly at random from $\mathcal{S}$ and then, conditionally on S, draw $X_1, \dots, X_n$ independently from $\mathbb{P}_S$. Note that $\bar{\mathbb{P}}^n$ is the mixture of n independent samples rather the distribution of n independent random vectors drawn from a mixture distribution. Consider the following testing problem:

$$H_0 : \quad \mathbf{X} \sim \mathbb{P}_0^n \quad \text{vs.} \quad H_1 : \quad \mathbf{X} \sim \bar{\mathbb{P}}^n.$$

Using Theorem 2.2, part (iii) of Tsybakov (2009), we get that for any test $\psi = \psi(\mathbf{X})$, we have

$$\mathbb{P}_0^n(\psi = 0) \vee \max_{S \in \mathcal{S}} \mathbb{P}_S^n(\psi = 1) \ge \mathbb{P}_0^n(\psi = 0) \vee \bar{\mathbb{P}}^n(\psi = 1) \ge \frac{1}{4}\exp(-\chi^2(\bar{\mathbb{P}}^n, \mathbb{P}_0)),$$

where we recall that the χ^2-divergence between two probability distributions P and Q is defined by

$$\chi^2(P,Q) = \begin{cases} \displaystyle\int \left(\frac{\mathrm{d}P}{\mathrm{d}Q} - 1\right)^2 \mathrm{d}Q, & \text{if } P \ll Q, \\ \infty, & \text{otherwise.} \end{cases}$$

Lemma A.1 implies that for suitable choices of the parameters a and k, we have $\chi^2(\bar{\mathbb{P}}^n, \mathbb{P}_0) \le 2$ so that the test errors are bounded below by a constant $C = e^{-2}/4$. Since $Q(\Sigma_S) = 2ka^2$ for any $S \in \mathcal{S}$, it follows from a standard reduction from hypothesis testing to estimation [see, e.g., Theorem 2.5 of Tsybakov (2009)] that the above result implies the following lower bound:

$$\inf_{\hat{Q}} \max_{\Sigma \in \mathcal{H}} \mathbb{E}_\Sigma[(\hat{Q} - Q(\Sigma))^2] \ge Ck^2a^4, \tag{3.11}$$

for some positive constant C, where the infimum is taken over all estimators $\hat{Q}$ of $Q(\Sigma)$ based on n observations and $\mathcal{H}$ is the class of covariance matrices defined by

$$\mathcal{H} = \{I_p\} \cup \{\Sigma_S : S \in \mathcal{S}\}.$$

To complete the proof, observe that the values of a and k prescribed in Lemma A.1 imply that $\mathcal{H} \subset \mathcal{F}_q(R)$ and give the desired lower bound. Note first that, for any choice of a and k, the following holds trivially: $I_p \in \mathcal{F}_q(R)$ and $\operatorname{diag}(\Sigma_S) = I_p$ for any $S \in \mathcal{S}$. Write $\Sigma_S = (\sigma_{ij})$ and observe that

$$\sum_{i \ne j} |\sigma_{ij}|^q = 2ka^q.$$

Next, we treat each case of Lemma A.1 separately.

Case 1. Note first that $2ka^q = R/2 < R$ so that $\Sigma_S \in \mathcal{F}_q(R)$. Moreover, $k^2a^4 = CR^{4/q}$.

Case 2. Note first that $2ka^q \le R/2 < R$ so that $\Sigma_S \in \mathcal{F}_q(R)$. Since $k \ge 2$ and $k^2 \le R^2n^q$, we have

$$k \ge \frac{R}{4}\left(\frac{\log((p-1)/k^2 + 1)}{2n}\right)^{-q/2}.$$

Therefore,

$$k^2a^4 \ge \frac{R^2}{16}\left(\frac{\log((p-1)/(R^2n^q) + 1)}{2n}\right)^{2-q} \wedge \frac{1}{4}.$$

Combining the two cases, we get

$$k^2a^4 \ge C\left[R^2\left(\frac{\log((p-1)/(R^2n^q) + 1)}{2n}\right)^{2-q} \wedge R^{4/q} \wedge 1\right].$$

Together with (3.11), this completes the proof of the second part of the lower bound. $\square$

4. Extension to nonquadratic functionals. Closely related to quadratic functional is the ℓ_r functional of covariance matrices, which is defined by

$$\ell_r(\Sigma) = \max_{i\le p} \sum_{j\le p} |\sigma_{ij}|^r. \tag{4.1}$$

It is often used to measure the sparsity of a covariance matrix and plays an important role in estimating sparse covariance matrix. This along the theoretical interest on the difficulty of estimating such a functional give rise to this study. Note that $\ell_1(\Sigma)$ functional is indeed the ℓ_1-norm of the covariance matrix Σ, whereas when $r = 2$, ℓ_r functional is the maximal row-wise quadratic functional. Thus, the nonquadratic ℓ_r functional is just a natural extension of such a maximal quadratic functional, whose optimal estimation problem will be the main focus of this section.

4.1. *Optimal estimation of ℓ_r functionals.* We consider a class of matrix with row-wise sparsity structure as follows:

$$\mathcal{G}_q(R) = \left\{\Sigma \in \mathbf{S}_p^+ : \max_{i\le p} \sum_{j\le p} |\sigma_{ij}|^q \le R, \operatorname{diag}(\Sigma) = I_p\right\}, \tag{4.2}$$

for $q \in [0, r)$ and $R > 0$ which can depend on n and p. A similar class of covariance matrices has been considered by Bickel and Levina (2008a) and Cai and Zhou (2012).

THEOREM 4.1. *Fix $q \in [0, r), R > 0$ and assume that $2\log p < n$ and $R^2 < (p-1)n^{-q}/2$. Then there exists a positive constant $C_4 > 0$ such that,*

$$\inf_{\hat{L}} \sup_{\Sigma\in\mathcal{G}_q(R)} \mathbb{E}_\Sigma[(\hat{L} - \ell_r(\Sigma))^2] \ge C_4\tilde{\phi}_{n,p}(q, R),$$

where $\tilde{\phi}_{n,p}(q, R)$ is defined by

$$\begin{aligned}\tilde{\phi}_{n,p}(q, R) &= R^2\frac{\log p}{n} \\ &\quad \vee\left\{R^2\left(\frac{\log((p-1)/(R^2n^q)+1)}{2n}\right)^{r-q} \wedge R^{2r/q} \wedge 1\right\}\end{aligned} \tag{4.3}$$

and the infimum is taken over all measurable functions $\hat{L}$ of the sample $X_1, \ldots, X_n$.

The proof is similar to that of Theorem 3.2 and is relegated to the Appendix.

As in (3.7), when $1 < R^2 < p^\alpha n^{-q}$ for some $\alpha < 1$, the lower bound in Theorem 4.1 can be written as

$$\tilde{\phi}_{n,p}(q, R) = R^2\frac{\log p}{n} \vee \left\{R^2\left(\frac{\log p}{n}\right)^{r-q} \wedge 1\right\}. \tag{4.4}$$

To establish the upper bound, we consider again a thresholding estimator. Naturally, we estimate ℓ_r functional of each single row, denoted by $\ell_r^{(i)}(\Sigma) = \sum_j |\sigma_{ij}|^r$, using the thresholding technique. Following the same notation as the previous section, the estimator is defined by

$$\widehat{\ell_r(\Sigma)} = \ell_r(\tilde{\Sigma}_\tau) = \max_i \ell_r^{(i)}(\tilde{\Sigma}_\tau) = \max_i \sum_{j \le p} |\hat{\sigma}_{ij}|^r \mathbb{1}\{|\hat{\sigma}_{ij}| > \tau\}, \tag{4.5}$$

for a threshold $\tau > 0$. We will see in the next theorem that this estimator achieves the adaptive minimax optimal rate.

THEOREM 4.2. *Assume that $\gamma \log(p) < n$ for some constant $\gamma > 8$ and fix $C_0 \ge 4$. Consider the threshold*

$$\tau = 2C_0 \sqrt{\frac{\gamma \log p}{n}}$$

and assume that $\tau \le 1$. Then, for any $q \in [0, r), R > 0$, the plug-in estimator $\ell_r(\tilde{\Sigma}_\tau)$ satisfies

$$\mathbb{E}[(\ell_r(\tilde{\Sigma}_\tau) - \ell_r(\Sigma))^2] \le C_5 \tilde{\psi}_{n,p}(q, R) + C_6 p^{4-\gamma/2},$$

where

$$\tilde{\psi}_{n,p}(q, R) = \begin{cases} \dfrac{R^2 \log p}{n}, & \text{if } q < \max\{r-1, 0\}, \\ R^2 \left(\dfrac{\gamma \log p}{n} \right)^{r-q}, & \text{if } q \ge \max\{r-1, 0\} \end{cases}$$

and C_5 and C_6 are positive constants.

The proof of this theorem is a generalization of the proof of Theorem 3.1 but some aspects that have independent value are presented here. In the proof of Theorem 3.1, we used the decomposition

$$\hat{\sigma}_{ij}^2 - \sigma_{ij}^2 = 2\sigma_{ij}(\hat{\sigma}_{ij} - \sigma_{ij}) + (\hat{\sigma}_{ij} - \sigma_{ij})^2,$$

which is actually the Taylor expansion of $\hat{\sigma}_{i,j}^2$ at $\sigma_{i,j}$. Carefully scrutinizing the proof, we find that the first term has the parametric rate $O(R^2/n)$ whereas the second term contributes to the rate $O(R^2(\log p/n)^{2-q})$. This phenomenon can be generalized to the ℓ_r-functional. In the latter case, we will apply the Taylor expansion of $|\hat{\sigma}_{ij}|^r$ at $|\sigma_{ij}|$, and the first-order term will contribute to the parametric rate of $O(R^2 \log p/n)$ while the second-order term has the rate $O(R^2(\log p/n)^{r-q})$. The elbow effect stems from the dominance of estimation errors of the first- and second-order terms of Taylor's expansion. We relegate the complete proof to the supplementary material.

A few remarks should be mentioned:

1. The combination of the two theorems imply that the estimator $\ell_r(\tilde{\Sigma}_\tau)$ is minimax adaptive over the space $\{\mathcal{G}_q(R), q \in [0, r), R > 0\}$ under very mild conditions. The adaptive minimax optimal rate of convergence is given by (4.4). The term $p^{4-\gamma/2}$ can be made arbitrarily small by choosing large enough γ.

2. The ℓ_r functional involves the maxima of the row sums. Compared it with estimating the quadratic functional, we need to pay the price of an extra $\log p$ term in the parametric rate.

3. The rate $\tilde{\phi}_{n,p}(q, R)$ presents the elbow phenomenon at $q = r - 1$ if $r > 1$. So quadratic row-wise functional $\ell_2(\tilde{\Sigma}_\tau)$ bears the same elbow behavior as the quadratic functional $Q(\tilde{\Sigma}_\tau)$.

4.2. *Optimal detection of correlations.* In this subsection, we illustrate the intrinsic link between functional estimation and hypothesis testing. To that end, consider the following hypothesis testing problem:

$$H_0: \quad X \sim \mathcal{N}(0, I_p),$$
$$H_1: \quad X \sim \mathcal{N}(0, I_p + \kappa \cdot \operatorname{off}(\Sigma)), \qquad \Sigma \in \bigcup_{q \in [0,r)} \{\mathcal{G}_q(R) : \ell_r(\operatorname{off}(\Sigma)) = 1\}.$$

This problem is intimately linked to sparse principal component analysis [Berthet and Rigollet (2013a, 2013b)]. A natural question associated with this problem is to find the minimal signal strength κ such that these hypotheses can be tested with high accuracy.

The previous subsection provides the optimal estimate for $\ell_r(\operatorname{off}(\Sigma))$. However, we need a result with high probability rather than in expectation. Using Lemma 4.2 in the supplementary material [Fan, Rigollet and Wang (2015)] and arguments similar to those employed to prove Theorem 4.2, it is not hard to show that

$$|\ell_r(\tilde{\Sigma}_\tau) - \ell_r(\Sigma)| \le CR\left(\frac{\gamma \log p}{n}\right)^{(r-q)/2} = CR\left(\frac{2\log p + \log(4/\delta)}{n}\right)^{(r-q)/2},$$

with probability larger than $1 - 4p^{-(\gamma-2)} =: 1 - \delta$. Therefore, letting

$$s_0 = 1 + CR\left(\frac{2\log p + \log(4/\delta)}{n}\right)^{(r-q)/2},$$
$$s_1 = 1 + \kappa^r - CR\left(\frac{2\log p + \log(4/\delta)}{n}\right)^{(r-q)/2},$$

we get $\mathbb{P}_{H_0}(\ell_r(\tilde{\Sigma}_\tau) \le s_0) \ge 1 - \delta$ and $\mathbb{P}_{H_1}(\ell_r(\tilde{\Sigma}_\tau) \ge s_1) \ge 1 - \delta$. Here, $\mathbb{P}_{H_0}$ denotes the probability under the null hypothesis and $\mathbb{P}_{H_1}$ denotes the largest probability over the composite alternative. To build a hypothesis test, note

that if $s_1 > s_0$, then for any $s \in [s_0, s_1]$, the test $\psi = \mathbb{1}\{\ell_r(\tilde{\Sigma}_\tau) \geq s\}$ satisfies $\mathbb{P}_{H_0}(\psi = 1) \vee \mathbb{P}_{H_1}(\psi = 0) \leq \delta$. We say that the test ψ discriminates between H_0 and H_1 with accuracy δ.

THEOREM 4.3. *Assume that n, p, R, q, r and δ are such that $\bar{\kappa} < 1$ where*

$$\bar{\kappa} := 2CR^{1/r}\left(\frac{2\log p + \log(4/\delta)}{n}\right)^{(r-q)/(2r)}.$$

Then, for any $\kappa > \bar{\kappa}$ and for any $s \in [s_0, s_1]$, the test $\psi = \mathbb{1}\{\ell_r(\tilde{\Sigma}_\tau) \geq s\}$ discriminates between H_0 and H_1 with accuracy δ.

The minimax risk for the correlation detection is given in the next theorem, which will be proved in the Appendix.

THEOREM 4.4. *For fixed $\nu > 0$, define $\underline{\kappa} > 0$ by*

$$\underline{\kappa} := R^{1/r}\left(\frac{\log(\nu p/(R^2 n^q))}{2n}\right)^{(r-q)/(2r)}.$$

Then for any $\kappa < \underline{\kappa}$,

$$\inf_{\psi}\{\mathbb{P}_{H_0}(\psi = 1) \vee \mathbb{P}_{H_1}(\psi = 0)\} \geq C_\nu,$$

where the infimum is taken over all possible tests and $C_\nu > 0$ is a continuous function of ν that tends to $1/2$ as $\nu \to 0$.

If we assume the high-dimensional regime $R^2 < p^\alpha n^{-q}$ for some $\alpha < 1$ as discussed before, then the lower bound matches the upper bound. So the theorem concludes that no test has asymptotic power for correlation detection unless κ is of higher order than $R^{1/r}(\log p/n)^{(r-q)/(2r)}$ and the detection method based on optimal $\ell_r(\Sigma)$ estimation is also optimal for testing existence of correlation.

5. Numerical experiments. Simulations are conducted in this section to evaluate the numerical performance of our plug-in estimator for quadratic functionals. Then the proposed method is applied to two high-dimensional testing problems: simulated two-sample data and real financial equity market data.

5.1. *Quadratic functional estimation.* We first study the behavior of estimators $\widehat{Q(\Sigma)} + \widehat{D(\Sigma)}$ for the total quadratic functional and $\widehat{Q(\Sigma)} = Q(\tilde{\Sigma}_\tau)$ for its off-diagonal part. To that end, four sparse covariance matrix structures were used in the simulations:

(M1) auto-correlation AR(1) covariance matrix $\sigma_{ij} = 0.25^{|i-j|}$;
(M2) banded correlation matrix with $\sigma_{ij} = 0.3$ if $|i-j| = 1$ and 0 otherwise;
(M3) sparse matrix with a block, size $p/20$ by $p/20$, of correlation 0.3;
(M4) identity matrix (it attains the maximal level of sparsity).

We chose $p = 500$ and let n vary from 30 to 100. For estimating the total quadratic functional, our proposed thresholding estimator, BS [Bai and Saranadasa (1996)] estimator and CQ [Chen and Qin (2010)] estimator were applied to each setting for repetition of 500 times. Their mean absolute estimation errors were reported in log scale (base 2) in Figure 1 with their

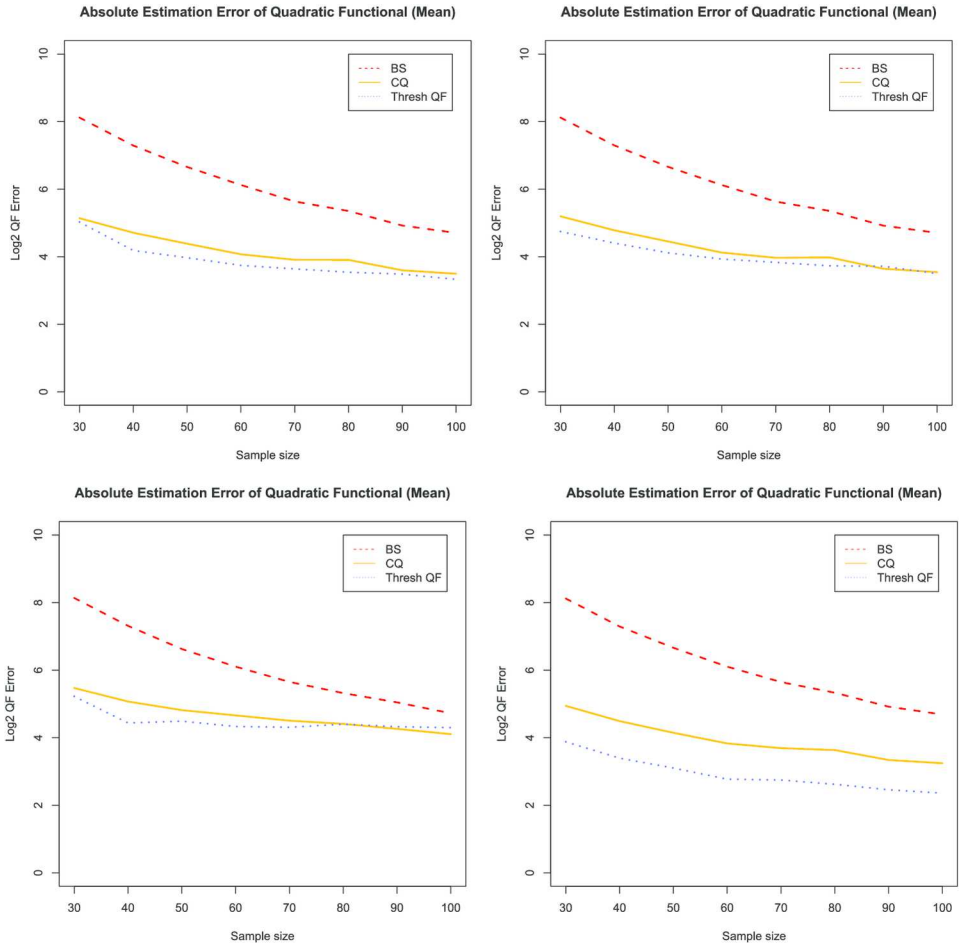

FIG. 1. *Performance of estimating* $\|\Sigma\|_F^2$ *using thresholded estimator* $\hat{Q} + \hat{D}$ *(dotted), CQ (solid) and BS (dashed). The mean of absolute errors over 500 repetitions in log scale (base 2) versus the sample size were reported for matrix* M1 *(top left),* M2 *(top right),* M3 *(bottom left),* M4 *(bottom right).*

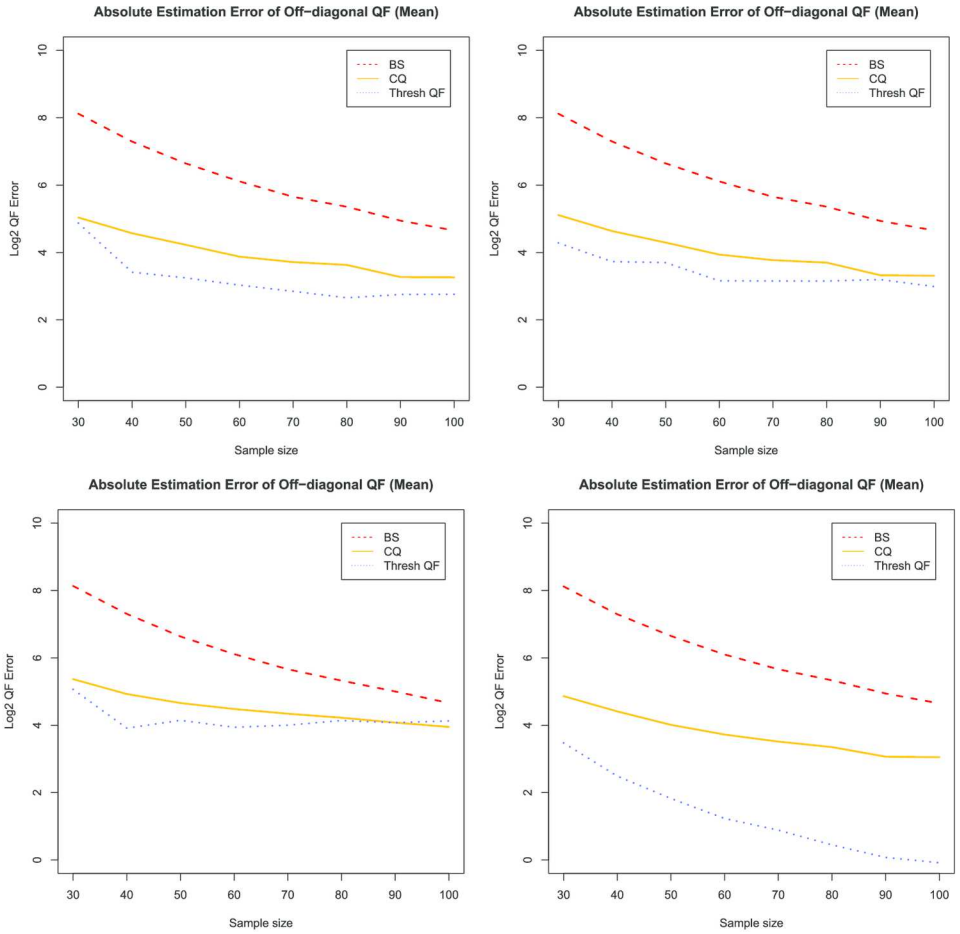

FIG. 2. *Performance of estimating $Q(\Sigma)$ using thresholded estimator $\hat{Q}$ (dotted), CQ-$\hat{D}$ (solid) and BS-$\hat{D}$ (dashed). The mean of absolute errors over 500 repetitions in log scale (base 2) versus the sample size were reported for matrix* M1 *(top left),* M2 *(top right),* M3 *(bottom left),* M4 *(bottom right).*

standard deviations omitted here. BS and CQ cannot be directly used for off-diagonal quadratic functional estimation, so we deducted $\widehat{D(\Sigma)}$ from both of them to serve as an estimator for only the off-diagonal part. The mean absolute estimation errors, compared with our proposed estimator $Q(\tilde{\Sigma}_\tau)$, are depicted in log scale (base 2) in Figure 2.

The four plots correspond to the aforementioned four covariance structures. We did not report the estimation error of directly using the naive plug-in which is an obvious disaster. In all the four cases, the BS (dashed line) method does not perform well in the "large p small n" regime. The method CQ (solid line) exhibits a relatively small estimation error in gen-

eral, but it can still be improved using the thresholding method. As theory shows, the method CQ is ratio-consistent [Chen and Qin (2010)], so our method (dotted line) is better only to a second order, which was captured by the small gap between dotted and solid curves. When estimating only off-diagonal quadratic functionals (Figure 2), the advantage of the thresholding method is even sharper since the error caused by nonsparse diagonal elements has been eliminated. The improved performance comes from the prior knowledge of sparsity, thus our method works best for very sparse matrix, especially well for identity matrix as seen in Figure 1.

A practical question is how to choose a proper threshold, as this is important to the performance of the thresholding estimator. In the above simulations, we chose $\tau = C\sqrt{\log p/n}$ with constant C slightly different for the four cases but all close to 1.5. In the next two applications to hypothesis testing, we employ the cross validation to choose a proper thresholding. The procedure consists of the following steps:

(1) The data is split into training data $D_{\mathcal{S}}^{(v)}$ of sample size n_1 and testing data $D_{\mathcal{S}^c}^{(v)}$ of sample size $n - n_1$ for m times, $v = 1, 2, \ldots, m$.

(2) The training data $D_{\mathcal{S}}^{(v)}$ is used to construct the thresholding estimator $Q(\tilde{\Sigma}_\tau^{(v)})$ under a sequence of thresholds while the testing data $D_{\mathcal{S}^c}^{(v)}$ constructs the nonthresholded ratio-consistent estimator $\hat{Q}^{(v)}$, for example, using CQ estimator of $\|\Sigma\|_{\mathsf{F}}^2$.

(3) The candidates of thresholds are $\tau_j = j\Delta\sqrt{\log(p)/n_1}$ for $j = 1, 2, \ldots, J$ where J is chosen to be a reasonably large number, say 50, and Δ is such that $J\Delta\sqrt{\log(p)/n_1} \le \hat{M} := \max_i \hat{\sigma}_{ii}$.

(4) The final j^* is taken to be the minimizer of the following problem:

$$\min_{j\in\{1,2,\ldots,J\}} \frac{1}{m}\sum_{v=1}^{m} |Q(\tilde{\Sigma}_{\tau_j}^{(v)}) - \hat{Q}^{(v)}|.$$

(5) The final estimator $Q(\tilde{\Sigma}_{\tau_{j^*}})$ is obtained by applying threshold $\tau_{j^*} = j^*\Delta\sqrt{\log(p)/n}$ to the empirical covariance matrix of the entire n data.

Bickel and Levina (2008a) suggested to use $n_1 = n/\log n$ for estimating covariance matrices. This is consistent with our experience for estimating functionals when no prior knowledge about the covariance matrix structure is provided. We will apply this splitting rule in the later simulation studies on high-dimensional hypothesis testing.

5.2. *Application to high-dimensional two-sample testing.* In this section, we apply the thresholding estimator of quadratic functionals to the high-dimensional two-sample testing problem. Two groups of data are simulated

from the Gaussian models:

$$X_{i,j} \sim \mathcal{N}(\mu_i, \Sigma) \qquad \text{for } i = 1, 2 \text{ and } j = 1, \ldots, n/2.$$

The dimensions considered for this problem are $(p, n) \in \{(500, 100), (1000, 150), (2000, 200)\}$. For simplicity, we choose Σ to be a correlation matrix and choose the sparse covariance structure to be 2 by 2 block diagonal matrices with 250 of them having correlations 0.3 and the rest having correlations 0. So the off-diagonal quadratic functional is always 45, which does not increase with p in our setting. The mean vectors μ_1 and μ_2 are chosen as follows. Let $\mu_1 = 0$ and the percentage of $\mu_{1,k} = \mu_{2,k}$ to be in $\{0\%, 50\%, 95\%, 100\%\}$. The 100% proportion corresponds to the case where the two groups are identical, thus gives information about accuracy of the size of the tests. The 95% proportion represents the situation where the alternative hypotheses are sparse. For those k such that $\mu_{1,k} \neq \mu_{2,k}$, we simply chose the value of each $\mu_{2,k}$ equally. To make the power comparable among different configurations, we use a constant signal-to-noise ratio $\eta = \|\mu_1 - \mu_2\|/\sqrt{\operatorname{tr}(\Sigma^2)} = 0.1$ across experiments.

Table 1 reports the empirical power and size of six testing methods based on 500 repetitions.

(BS) Bai and Saranadasa's original test.

(newBS) Bai and Saranadasa's modified test where $\operatorname{tr}(\Sigma^2)$ is estimated by thresholding the sample covariance matrix.

(CQ) Chen and Qin's original test.

TABLE 1
Empirical testing power and size of 6 *testing methods based on* 500 *simulations*

Prop. of equalities	**BS**	**newBS**	**CQ**	**newCQ**	**Bonf**	**BH**
			$p = 500, n = 100$			
0%	0.408	0.422	0.428	0.432	0.104	0.110
50%	0.396	0.422	0.418	0.428	0.110	0.116
95%	0.422	0.440	0.438	0.442	0.208	0.214
100% (size)	0.030	0.036	0.036	0.038	0.042	0.042
			$p = 1000, n = 150$			
0%	0.696	0.710	0.718	0.718	0.082	0.086
50%	0.698	0.712	0.712	0.714	0.106	0.112
95%	0.702	0.716	0.718	0.722	0.308	0.328
100% (size)	0.040	0.044	0.048	0.046	0.050	0.050
			$p = 2000, n = 200$			
0%	0.930	0.938	0.940	0.940	0.138	0.146
50%	0.918	0.922	0.924	0.928	0.104	0.106
95%	0.922	0.928	0.930	0.930	0.324	0.338
100% (size)	0.046	0.050	0.050	0.050	0.046	0.046

TABLE 2
Mean and SD of relative errors for estimating quadratic functionals (in percentage)

	$p = 500, n = 100$	$p = 1000, n = 150$	$p = 2000, n = 200$
BS	4.93 (2.48)	4.47 (1.56)	5.05 (1.10)
newBS	2.12 (1.43)	0.74 (0.56)	0.54 (0.40)
CQ	3.72 (1.97)	2.32 (1.24)	1.70 (0.91)
newCQ	2.77 (1.38)	1.27 (0.64)	0.62 (0.33)

(newCQ) Chen and Qin's modified test where $\mathrm{tr}(\Sigma_i^2)$ and $\mathrm{tr}(\Sigma_1\Sigma_2)$ are estimated by thresholding their empirical counterparts.

(Bonf) Bonferroni correction: This method regards the high-dimensional testing problem as p univariate testing problems. If there is a p-value that is less than $0.05/p$, the null hypothesis is rejected.

(BH) Benjamini–Hochberg method. The method is similar to the Bonferroni correction, but employs the Benjamini–Hochberg method in decision making.

For estimating quadratic functionals, the cross-validation is employed using $n/\log(n)$ splitting rule. The first four methods are evaluated at the 5% significance level while Bonferroni correction and Benjamini–Hochberg correction are evaluated at 5% family-wise error rate or FDR. We also list the average relative estimation errors for the quadratic functionals of the first four methods in Table 2. Here, the average is taken over four different proportions of equalities and the average for CQ and newCQ is also taken over errors in estimating $\mathrm{tr}(\Sigma_1^2)$ and $\mathrm{tr}(\Sigma_2^2)$.

Several comments are in order. First, the first four methods based on Wald-type of statistic with correlation ignored perform much better, in terms of the power, than the last two methods which combines individual tests. Even in the case that proportional of equalities is 95% where the individual difference is large for nonidentical means, aggregating the signals together in the Wald-type of statistic still outperforms. However, in the case of 0% identical means, the power of Bonferroni or FDR method is extremely small, due to small individual differences. Second, the method newCQ, which combines CQ and thresholding estimator of the quadratic functional, has the highest power and performs the best among all methods. The corrected BS method also improves the performance by estimating the quadratic functionals better compared with original BS. CQ indeed is more powerful than BS as claimed by Chen and Qin (2010), but we can even improve the performance of those two methods more by leveraging the sparsity structure of covariance matrices.

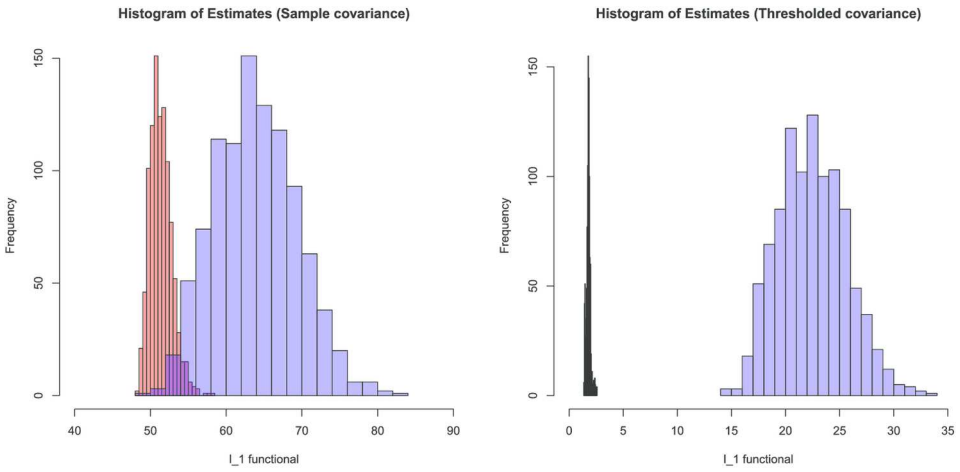

FIG. 3. *Histogram of* 1000 ℓ_1 *functional estimates for* H_0 *and* H_1 *by* $\ell_r(\hat{\Sigma})$ *(left) and optimal estimator* $\ell_r(\tilde{\Sigma}_\tau)$ *(right).*

5.3. *Estimation of* ℓ_r *functional and correlation detection.* In order to check the effectiveness of using ℓ_r norm of the thresholded sample matrix to detect correlation, let us take one simple matrix structure as an example and use $r = 1$. Under H_0, assume $X \sim \mathcal{N}(0, I_p)$; while under H_1, $X \sim \mathcal{N}(0, \Sigma)$, where $\Sigma_{ij} = 0.8$ if $i, j \in \mathcal{S}$ and $\mathcal{S}$ is a random subset of size $p/20$ in $\{1, 2, \ldots, n\}$. We chose to use $p = 500$ and generated $n = 100$ independent random vectors under both H_0 and H_1. The whole simulation was done for $N = 1000$ times.

We compare the ℓ_1 norm estimates based on empirical covariance matrix $\ell_1(\hat{\Sigma})$ and thresholded empirical covariance matrix $\ell_1(\tilde{\Sigma}_\tau)$. The threshold is decided by cross validation with $n/\log(n)$ splitting. The simulations yielded N estimates for both null and alternative hypotheses, which were plotted in Figure 3. The optimal estimator $\ell_1(\tilde{\Sigma}_\tau)$ perfectly discriminates the null and alternative hypotheses while $\ell_1(\hat{\Sigma})$ overestimates ℓ_1 functional and blurs the difference of the two hypotheses.

5.4. *Application to testing multifactor pricing model.* In this section, we test the validity of the Capital Asset Pricing Model (CAPM) and Fama–French models using Pesaran and Yamagata's method (2.2) for the securities in the Standard & Poor 500 (S&P 500) index. Following the literature, we used 60 monthly stock returns to construct test statistics since monthly returns are nearly independent. The composition of index keeps changing annually, so we selected only 276 large stocks. The monthly returns (adjusted for dividend) between January 1990 and December 2012 are downloaded from the Wharton Research Data Services (WRDS) database. The time

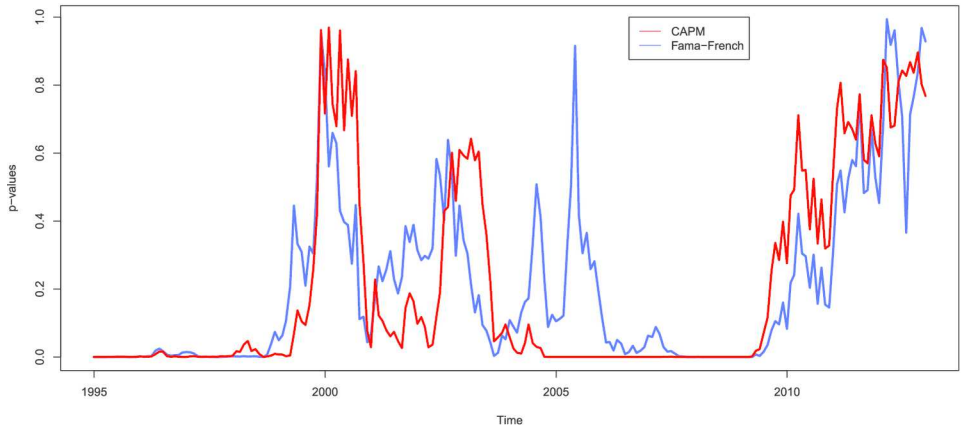

FIG. 4. *P-values of testing $H_0 : \alpha = 0$ in the CAMP and Fama–French 3 factor models based on S&P 500 monthly returns from January 1995 to December 2012.*

series on the risk-free rates and Fama–French three factors are obtained from Ken French's data library. If only the first factor, that is, the excessive return of the market portfolio is used, the Fama–French model reduces to the CAPM model. We tested the null hypothesis $H_0 : \alpha = 0$ for both models. The p-values of the tests are depicted in Figure 4, which are computed based on running windows of previous 60 months.

The results suggest that market efficiency is time dependent and the Fama–French model are rejected less frequently than the CAPM. Before 1998, the evidence that $\alpha \neq 0$ is very strong. After 1998, the Fama–French 3-factor model holds most of the time except the period 2007–2009 that contains the financial crisis. On the other hand, the CAPM is rejected for an extended period of time during this period.

APPENDIX A: A TECHNICAL LEMMA ON χ^2 DIVERGENCES

LEMMA A.1. *Consider a mixture of Gaussian product distributions*

$$\tilde{\mathbb{P}}^n = \frac{1}{m}\sum_{j=1}^{m} \mathbb{P}_j^n,$$

where $\mathbb{P}_j \sim \mathcal{N}_p(0, \Sigma_j)$ such that $\mathbb{P}_j \ll \mathbb{P}_0$. Then

$$\chi^2(\tilde{\mathbb{P}}^n, \mathbb{P}_0^n) = \frac{1}{m^2}\sum_{j,k=1}^{m} |I_p - (\Sigma_j - I)(\Sigma_k - I)|^{-n/2} - 1. \tag{A.1}$$

Furthermore, assume $2(\log p) \leq n$. Consider the mixture $\bar{\mathbb{P}}^n$ defined in the proof of Theorem 3.2 where k and a are defined as follows:

1. *If $R < 4(\frac{\log p}{n})^{q/2}$, then take $k = 1$ and $a = (R/4)^{1/q}$.*

2. If $R \geq 4(\frac{\log p}{n})^{q/2}$*, then take* k *to be the largest integer such that*

$$k \leq \frac{R}{2}\left(\frac{\log((p-1)/k^2+1)}{2n}\right)^{-q/2} \tag{A.2}$$

and

$$a = \left(\frac{\log((p-1)/k^2+1)}{2n}\right)^{1/2} \wedge (2k)^{-1/2}. \tag{A.3}$$

Such choices yield in both cases

$$\chi^2(\bar{\mathbb{P}}^n, \mathbb{P}_0^n) \leq e - 1. \tag{A.4}$$

Moreover, in case 2 we have that (i) $k \geq 2$ *and* (ii) *under the assumption that* $R^2 < (p-1)n^{-q}/2$*, we also have* $k^2 \leq R^2 n^q < (p-1)/2$.

PROOF. To unify the notation, we will work directly with $\mathbb{P}_S, S \in \mathcal{S}$ rather than $\mathbb{P}_j, j \in [m]$. However, in the first part of the proof, we will not use the specific form Σ_S nor that of $\mathcal{S}$. For now, we simply assume that Σ_S is invertible (we will check this later on). Recall that

$$\chi^2(\bar{\mathbb{P}}^n, \mathbb{P}_0^n) = \mathbb{E}_0\left[\left(\frac{\mathrm{d}\bar{\mathbb{P}}^n}{\mathrm{d}\mathbb{P}_0^n} - 1\right)^2\right] = \frac{1}{|\mathcal{S}|^2}\sum_{S,T\in\mathcal{S}}\left(\mathbb{E}_0\left[\frac{\mathrm{d}\mathbb{P}_S}{\mathrm{d}\mathbb{P}_0}\frac{\mathrm{d}\mathbb{P}_T}{\mathrm{d}\mathbb{P}_0}\right]\right)^n - 1,$$

where $\mathbb{E}_0$ denotes the expectation with respect to $\mathbb{P}_0$. Furthermore,

$$\mathbb{E}_0\left[\frac{\mathrm{d}\mathbb{P}_S}{\mathrm{d}\mathbb{P}_0}\frac{\mathrm{d}\mathbb{P}_T}{\mathrm{d}\mathbb{P}_0}\right] = \frac{1}{(|\Sigma_S||\Sigma_T|)^{1/2}}\mathbb{E}_0\left[\exp\left(-\frac{1}{2}X^\top(\Sigma_S^{-1}+\Sigma_T^{-1}-2I_p)X\right)\right].$$

Consider the spectral decomposition of $\Sigma_S^{-1}+\Sigma_T^{-1}-2I_p = U\Lambda U^\top$, where U is an orthogonal matrix and Λ is a diagonal matrix with eigenvalues $\lambda_1, \ldots, \lambda_p$ on its diagonal. Then, by rotational invariance of the Gaussian distribution, it holds

$$\begin{aligned}
&\mathbb{E}_0\left[\exp\left(-\frac{1}{2}X^\top(\Sigma_S^{-1}+\Sigma_T^{-1}-2I_p)X\right)\right] \\
&\quad= \mathbb{E}_0\left[\exp\left(-\frac{1}{2}X^\top\Lambda X\right)\right] \\
&\quad= \prod_{j=1}^p \mathbb{E}_0\left[\exp\left(-\frac{1}{2}\lambda_j X_j^2\right)\right] \\
&\quad= \begin{cases} \displaystyle\prod_{j=1}^p (1+\lambda_j)^{-1/2} = |I+\Lambda|^{-1/2}, & \text{if } \max_j \lambda_j < 1, \\ \infty, & \text{otherwise.} \end{cases}
\end{aligned}$$

To ensure that the above expression is finite, note that the Cauchy–Schwarz inequality yields

$$\left(\mathbb{E}_0\left[\frac{\mathrm{d}\mathbb{P}_S}{\mathrm{d}\mathbb{P}_0}\frac{\mathrm{d}\mathbb{P}_T}{\mathrm{d}\mathbb{P}_0}\right]\right)^2 \le \mathbb{E}_0\left[\left(\frac{\mathrm{d}\mathbb{P}_S}{\mathrm{d}\mathbb{P}_0}\right)^2\right]\mathbb{E}_0\left[\left(\frac{\mathrm{d}\mathbb{P}_T}{\mathrm{d}\mathbb{P}_0}\right)^2\right]$$
$$= (\chi^2(\mathbb{P}_S,\mathbb{P}_0)+1)(\chi^2(\mathbb{P}_T,\mathbb{P}_0)+1) < \infty,$$

where the two χ^2 divergences are finite because $\mathbb{P}_S \ll \mathbb{P}_0$ for any $S \in \mathcal{S}$. Therefore,

$$\mathbb{E}_0\left[\frac{\mathrm{d}\mathbb{P}_S}{\mathrm{d}\mathbb{P}_0}\frac{\mathrm{d}\mathbb{P}_T}{\mathrm{d}\mathbb{P}_0}\right] = \frac{|I+\Lambda|^{-1/2}}{(|\Sigma_S||\Sigma_T|)^{1/2}} = \frac{|\Sigma_S^{-1}+\Sigma_T^{-1}-I_p|^{-1/2}}{(|\Sigma_S||\Sigma_T|)^{1/2}}.$$

Next, observe that

$$(|\Sigma_S||\Sigma_T||\Sigma_S^{-1}+\Sigma_T^{-1}-I_p|)^{-1/2} = ((|\Sigma_S(\Sigma_S^{-1}+\Sigma_T^{-1}-I_p)\Sigma_T)|)^{-1/2}$$
$$= |I-(\Sigma_S-I)(\Sigma_T-I)|^{-1/2}.$$

Since we have not used the specific form of Σ_S, $S \in \mathcal{S}$, this bound is valid for any mixture and completes the proof of (A.1).

Next, we apply this bound to the specific choice for Σ_S of (3.10). Note that the minimal eigenvalue of the matrices $\Sigma_S, S \in \mathcal{S}$ is $1-\sqrt{ka^2}$. Later we will show $2a^2k \le 1$, which implies that Σ_S is always positive definite. In particular, this implies that $\mathbb{P}_S \ll \mathbb{P}_0$ for any $S \in \mathcal{S}$. Moreover, it follows from definition (3.10) that

$$I-(\Sigma_S-I)(\Sigma_T-I) = \begin{pmatrix} 1-a^2\mathbf{1}_S^\top\mathbf{1}_T & 0 \\ 0 & I-a^2\mathbf{1}_S\mathbf{1}_T^\top \end{pmatrix},$$

where 0 is a generic symbol to indicate space filled by zeros. Expanding the determinant along the first row (or column), we get

$$|I-(\Sigma_S-I)(\Sigma_T-I)|^{-1/2} = (1-a^2\mathbf{1}_S^\top\mathbf{1}_T)^{-1/2}|I-a^2\mathbf{1}_S\mathbf{1}_T^\top|^{-1/2}$$
$$= (1-a^2\mathbf{1}_S^\top\mathbf{1}_T)^{-1},$$

where in the second equality, we used Sylvester's determinant theorem. By (A.1), we have

$$\chi^2(\bar{\mathbb{P}}^n,\mathbb{P}_0^n) = \frac{1}{|\mathcal{S}|^2}\sum_{S,T\in\mathcal{S}}(1-a^2\mathbf{1}_S^\top\mathbf{1}_T)^{-n}-1.$$

As to be verified later, $2a^2k \le 1$. Using the fact that $(1-x)^{-1} \le \exp(2x)$ for $x \in [0,1/2]$ and the symmetry, we have

$$\chi^2(\bar{\mathbb{P}}^n,\mathbb{P}_0^n) \le \frac{1}{|\mathcal{S}|}\sum_{S\in\mathcal{S}}\exp(2na^2|S\cap[k]|)-1 = \mathcal{E}[\exp(2na^2|S\cap[k]|)-1],$$

where $\mathcal{E}$ denotes the expectation with respect to the distribution of S randomly chosen from $\mathcal{S}$. In particular, $|S \cap [k]| = \sum_{i=1}^{k} \mathbb{1}(i \in S)$ is the sum of k negatively associated random variables. Therefore, using negative association, the above expectation is further bounded by

$$\prod_{i=1}^{k} \mathcal{E}[e^{2na^2\mathbb{1}(i\in S)}].$$

Next, for a given by (A.3), we have

$$\prod_{i=1}^{k} \mathcal{E}[e^{2na^2\mathbb{1}(i\in S)}] = \left[(e^{2na^2}-1)\frac{k}{p-1}+1\right]^k \le \left[1+\frac{1}{k}\right]^k \le e.$$

We now show for both cases of the lemma, we have $2a^2k \le 1$. Indeed for case 1, we get $2a^2k = 2(R/4)^{2/q} < 2(\log p)/n \le 1$. For case 2, $2a^2k \le 1$ follows trivially from the definition of a. Also observe that $k \ge 2$ since

$$\frac{R}{2}\left(\frac{\log((p-1)/4+1)}{n}\right)^{-q/2} \ge 2\left(\frac{\log p}{n}\right)^{q/2}\left(\frac{\log((p-1)/4+1)}{n}\right)^{-q/2} \ge 2.$$

This proves part (i) of the statement on k. To prove part (ii), observe that $R^2 < (p-1)n^{-q}/2$ implies that

$$2^{1-2/q} < 1 < \log\left(\frac{p-1}{R^2n^q}+1\right),$$

which is equivalent to

$$Rn^{q/2} > \frac{R}{2}\left(\frac{\log((p-1)/(R^2n^q)+1)}{2n}\right)^{-q/2}.$$

Therefore, $k^2 \le R^2n^q < (p-1)/2$. $\square$

APPENDIX B: PROOF OF THEOREM 4.1

The proof follows a similar idea to that of Theorem 3.2. For the second part of the lower bound, we use exactly the same construction of two hypotheses as in Lemma A.1. Then it follows that for the ℓ_r functional,

$$\inf_{\hat{L}} \max_{\Sigma\in\mathcal{H}} \mathbb{E}_\Sigma[(\hat{L}-\ell_r(\Sigma))^2] \ge Ck^2a^{2r},$$

for some positive constant C, where the infimum is taken over all estimators $\hat{L}$ of $\ell_r(\Sigma)$ based on n observations. In case 1, $k^2a^{2r} = CR^{2r/q}$ while in case 2, following the same arguments as before,

$$k^2a^{2r} \ge \frac{R^2}{16}\left(\frac{\log((p-1)/(R^2n^q)+1)}{2n}\right)^{r-q} \wedge \frac{1}{2}.$$

This completes the second part of the lower bound.

The first part of the result is a little bit more complicated than the construction of $A^{(k)}$ and $B^{(k)}$ in the proof of Theorem 3.2 due to the extra $\log p$ term in the lower bound. We need to consider a mixture of measures in order to capture the complexity of the problem. With a slight abuse of notation, we redefine $(2k) \times (2k)$ matrices $A^{(k)}, B^{(k)}$ as follows:

$$A^{(k)} = \begin{pmatrix} \mathbf{1}\mathbf{1}^\top & a\mathbf{1}\mathbf{1}^\top \\ a\mathbf{1}\mathbf{1}^\top & \mathbf{1}\mathbf{1}^\top \end{pmatrix}, \qquad B^{(k)} = \begin{pmatrix} \mathbf{1}\mathbf{1}^\top & b\mathbf{1}\mathbf{1}^\top \\ b\mathbf{1}\mathbf{1}^\top & \mathbf{1}\mathbf{1}^\top \end{pmatrix},$$

where $a, b \in (0, 1/2)$ and $\mathbf{1}$ denotes a vector of ones of length k. Since $R^2 < p$, we now construct the block diagonal covariance matrices

$$\Sigma_m^{(k)} = \operatorname{diag}(C_1, C_2, \ldots, C_M, I_{p-2kM}), \qquad m = 1, 2, \ldots, M,$$

where the mth diagonal block is chosen to be $C_m = B^{(k)}$ while others are $C_i = A^{(k)}$ for $i \neq m$ and $M = \lfloor p/R \rfloor$. Also define $\Sigma_0^{(k)}$ to be of the same structure with $C_i = A^{(k)}$ for all i. Then we have $\Sigma_m^{(R/2)} \in \mathcal{G}_q(R)$ for $m = 0, 1, \ldots, M$, since each row of $\Sigma_m^{(R/2)}$ only contains at most R nonzero elements that are bounded by 1.

Let $\mathbb{P}_0$ denote the distribution of $X \sim \mathcal{N}_p(0, \Sigma_0^{(R/2)})$ and $\mathbb{P}_m$ denote the distribution of $X \sim \mathcal{N}_p(0, \Sigma_m^{(R/2)})$. Let $\mathbb{P}_0^n$ (resp., $\mathbb{P}_m^n$) denote the distribution of $\mathbf{X} = (X_1, \ldots, X_n)$ of n i.i.d. random variables drawn from $\mathbb{P}_0$ (resp., $\mathbb{P}_m$). Moreover, let $\bar{\mathbb{P}}^n$ denote the uniform mixture of $\mathbb{P}_m^n$ over $m \in [M]$. Consider the testing problem

$$H_0 : \quad \mathbf{X} \sim \mathbb{P}_0^n \quad \text{vs.} \quad H_1 : \quad \mathbf{X} \sim \bar{\mathbb{P}}^n.$$

Using Theorem 2.2, part (iii) of Tsybakov (2009) as before, we need to show χ^2-divergence can be bounded by a constant. By the same calculation as in Lemma A.1, we have

$$\chi^2(\bar{\mathbb{P}}^n, \mathbb{P}_0^n) = \frac{1}{M^2} \sum_{1 \le i,j \le M} |I - [(\Sigma_0^{(1)})^{-1}\Sigma_i^{(1)} - I][(\Sigma_0^{(1)})^{-1}\Sigma_j^{(1)} - I]|^{-n/2} - 1.$$

Note that χ^2-divergence here depends on $\Sigma_m^{(1)}$ instead of $\Sigma_m^{(R/2)}$ since perfectly correlated random variables do not add additional information and hence do not affect χ^2-divergence (see the proof of Theorem 3.2). Using the definition of $\Sigma_m^{(1)}$'s, we obtain

$$\begin{aligned} &|I - [(\Sigma_0^{(1)})^{-1}\Sigma_i^{(1)} - I][(\Sigma_0^{(1)})^{-1}\Sigma_j^{(1)} - I]| \\ &\quad = \begin{cases} 1 - 2(1+a^2)\left(\dfrac{a-b}{1-a^2}\right)^2 + \dfrac{(a-b)^4}{(1-a^2)^2}, & \text{if } i = j, \\ 1, & \text{otherwise.} \end{cases} \end{aligned}$$

Therefore,

$$\chi^2(\bar{\mathbb{P}}^n, \mathbb{P}_0^n) = \frac{1}{M}\left\{\left(1 - 2(1+a^2)\left(\frac{a-b}{1-a^2}\right)^2 + \frac{(a-b)^4}{(1-a^2)^2}\right)^{-n/2} - 1\right\},$$

which is bounded by $((1-5(a-b)^2)^{-n/2}-1)/M$ due to the fact $2(1+a^2)/(1-a^2)^2 \le 5$ for $a \le 1/2$. Now choose

$$a = \frac{1}{4}, \qquad b = a + \frac{1}{4}\sqrt{\frac{\log p}{n}}.$$

By assumption, there exists a constant $c_0 > 1$ such that $R^2 \le c_0 p$. Thus,

$$\chi^2(\bar{\mathbb{P}}^n, \mathbb{P}_0^n) \le \frac{1}{M}\left\{\left(1 - \frac{\log p}{2n}\right)^{-n/2} - 1\right\} \le \frac{R}{p} e^{\log p/2} \le \sqrt{c_0}.$$

Using standard techniques to reduce estimation problems to testing problems as before, we find

$$\inf_{\hat{L}} \max_{\Sigma \in \{\Sigma_m^{(R/2)}: m=0,\dots,M\}} \mathbb{E}_\Sigma[(\hat{L} - \ell_r(\Sigma))^2] \ge C(\ell_r(\Sigma_0^{(R/2)}) - \ell_r(\Sigma_1^{(R/2)}))^2.$$

For the above choice of $\Sigma_m^{(k)}$, we have

$$(\ell_r(\Sigma_0^{(R/2)}) - \ell_r(\Sigma_1^{(R/2)}))^2 = \frac{R^2}{4}(b^r - a^r)^2 \ge CR^2\frac{\log p}{n}.$$

Since $\Sigma_m^{(R/2)} \in \mathcal{G}_q(R)$, the above two displays imply that

$$\inf_{\hat{L}} \max_{\Sigma \in \mathcal{G}_q(R)} \mathbb{E}_\Sigma[(\hat{L} - \ell_r(\Sigma))^2] \ge C\frac{R^2 \log p}{n},$$

which together with the other part of the lower bound, completes the proof of the theorem.

APPENDIX C: PROOF OF THEOREM 4.4

The proof is similar to that of Theorem 3.2, but simpler since $r \le 1$ where no elbow effect exists. Consider hypothesis construction (3.10) with $\Sigma_S = I_p + \kappa\bar{\Sigma}$ and

$$a = \kappa k^{-1/r} \quad \text{and} \quad k = \left\lceil R\left(\frac{\log(\nu p/(R^2 n^q))}{2n}\right)^{-q/2} \right\rceil.$$

Choose ν sufficiently small so that $R(\frac{\log(\nu p/(R^2 n^q))}{2n})^{1-q/2} \le 1/2$, which implies $2ka^2 \le 1$ and guarantees the positive semi-definiteness of Σ_S. Furthermore, $ka^q \le R$ holds, so $\Sigma_S \in \mathcal{G}_q(R)$. By the same derivation as in Theorem 3.2, we are able to show

$$\chi^2(\bar{\mathbb{P}}^n, \mathbb{P}_0^n) \le e^\nu - 1,$$

which by Theorem 2.2(iii) of Tsybakov (2009) leads to the final conclusion.

SUPPLEMENTARY MATERIAL

Technical proofs Fan, Rigollet and Wang (2015)
(DOI: 10.1214/15-AOS1357SUPP; .pdf). This supplementary material contains the introduction to two-sample high-dimensional testing methods and the proofs of upper bounds that were omitted from the paper.

REFERENCES

Amini, A. A. and Wainwright, M. J. (2009). High-dimensional analysis of semidefinite relaxations for sparse principal components. *Ann. Statist.* **37** 2877–2921. MR2541450

Arias-Castro, E., Bubeck, S. and Lugosi, G. (2015). Detecting positive correlations in a multivariate sample. *Bernoulli* **21** 209–241. MR3322317

Bai, Z. and Saranadasa, H. (1996). Effect of high-dimension: By an example of a two sample problem. *Statist. Sinica* **6** 311–329. MR1399305

Berthet, Q. and Rigollet, P. (2013a). Complexity theoretic lower bounds for sparse principal component detection. *J. Mach. Learn. Res.* **30** 1046–1066.

Berthet, Q. and Rigollet, P. (2013b). Optimal detection of sparse principal components in high-dimension. *Ann. Statist.* **41** 1780–1815. MR3127849

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

Bickel, P. J. and Ritov, Y. (1988). Estimating integrated squared density derivatives: Sharp best order of convergence estimates. *Sankhyā Ser. A* **50** 381–393. MR1065550

Birnbaum, A., Johnstone, I. M., Nadler, B. and Paul, D. (2013). Minimax bounds for sparse PCA with noisy high-dimensional data. *Ann. Statist.* **41** 1055–1084. MR3113803

Butucea, C. (2007). Goodness-of-fit testing and quadratic functional estimation from indirect observations. *Ann. Statist.* **35** 1907–1930. MR2363957

Butucea, C. and Meziani, K. (2011). Quadratic functional estimation in inverse problems. *Stat. Methodol.* **8** 31–41. MR2741507

Cai, T. and Liu, W. (2011). Adaptive thresholding for sparse covariance matrix estimation. *J. Amer. Statist. Assoc.* **106** 672–684. MR2847949

Cai, T. T. and Low, M. G. (2005). Nonquadratic estimators of a quadratic functional. *Ann. Statist.* **33** 2930–2956. MR2253108

Cai, T. T. and Low, M. G. (2006). Optimal adaptive estimation of a quadratic functional. *Ann. Statist.* **34** 2298–2325. MR2291501

Cai, T. T., Ma, Z. and Wu, Y. (2013). Sparse PCA: Optimal rates and adaptive estimation. *Ann. Statist.* **41** 3074–3110. MR3161458

Cai, T., Ma, Z. and Wu, Y. (2015). Optimal estimation and rank detection for sparse spiked covariance matrices. *Probab. Theory Related Fields* **161** 781–815. MR3334281

Cai, T. T., Ren, Z. and Zhou, H. H. (2013). Optimal rates of convergence for estimating Toeplitz covariance matrices. *Probab. Theory Related Fields* **156** 101–143. MR3055254

Cai, T. T. and Yuan, M. (2012). Adaptive covariance matrix estimation through block thresholding. *Ann. Statist.* **40** 2014–2042. MR3059075

Cai, T. T., Zhang, C.-H. and Zhou, H. H. (2010). Optimal rates of convergence for covariance matrix estimation. *Ann. Statist.* **38** 2118–2144. MR2676885

Cai, T. T. and Zhou, H. H. (2012). Minimax estimation of large covariance matrices under ℓ_1-norm. *Statist. Sinica* **22** 1319–1349. MR3027084

Chen, S. X. and Qin, Y.-L. (2010). A two-sample test for high-dimensional data with applications to gene-set testing. *Ann. Statist.* **38** 808–835. MR2604697

Donoho, D. L. and Nussbaum, M. (1990). Minimax quadratic estimation of a quadratic functional. *J. Complexity* **6** 290–323. MR1081043

Efromovich, S. and Low, M. (1996). On optimal adaptive estimation of a quadratic functional. *Ann. Statist.* **24** 1106–1125. MR1401840

El Karoui, N. (2008). Operator norm consistent estimation of large-dimensional sparse covariance matrices. *Ann. Statist.* **36** 2717–2756. MR2485011

Fama, E. F. and French, K. R. (1993). Common risk factors in the returns on stocks and bonds. *Journal of Financial Economics* **33** 3–56.

Fan, J. (1991). On the estimation of quadratic functionals. *Ann. Statist.* **19** 1273–1294. MR1126325

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

Fan, J., Rigollet, P. and Wang, W. (2015). Supplement to "Estimation of functionals of sparse covariance matrices." DOI:10.1214/15-AOS1357SUPP.

Foucart, S. and Rauhut, H. (2013). *A Mathematical Introduction to Compressive Sensing*. Birkhäuser/Springer, New York. MR3100033

Hall, P. and Marron, J. S. (1987). Estimation of integrated squared density derivatives. *Statist. Probab. Lett.* **6** 109–115. MR0907270

Ibragimov, I. A., Nemirovskiĭ, A. S. and Khas'minskiĭ, R. Z. (1987). Some problems of nonparametric estimation in Gaussian white noise. *Theory Probab. Appl.* **31** 391–406.

Johnstone, I. M. and Lu, A. Y. (2009). On consistency and sparsity for principal components analysis in high-dimensions. *J. Amer. Statist. Assoc.* **104** 682–693. MR2751448

Jung, S. and Marron, J. S. (2009). PCA consistency in high-dimension, low sample size context. *Ann. Statist.* **37** 4104–4130. MR2572454

Klemelä, J. (2006). Sharp adaptive estimation of quadratic functionals. *Probab. Theory Related Fields* **134** 539–564. MR2214904

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

Levina, E. and Vershynin, R. (2012). Partial estimation of covariance matrices. *Probab. Theory Related Fields* **153** 405–419. MR2948681

Lintner, J. (1965). The valuation of risk assets and the selection of risky investments in stock portfolios and capital budgets. *The Review of Economics and Statistics* **47** 13–37.

Ma, Z. (2013). Sparse principal component analysis and iterative thresholding. *Ann. Statist.* **41** 772–801. MR3099121

Mossin, J. (1966). Equilibrium in a capital asset market. *Econometrica* **34** 768–783.

Nemirovski, A. (2000). Topics in nonparametric statistics. In *Lectures on Probability Theory and Statistics (Saint-Flour, 1998). Lecture Notes in Math.* **1738** 85–277. Springer, Berlin. MR1775640

Nemirovskiĭ, A. S. and Khas'minskiĭ, R. Z. (1987). Nonparametric estimation of the functionals of the products of a signal observed in white noise. *Problemy Peredachi Informatsii* **23** 27–38. MR0914348

Onatski, A., Moreira, M. J. and Hallin, M. (2013). Asymptotic power of sphericity tests for high-dimensional data. *Ann. Statist.* **41** 1204–1231. MR3113808

Paul, D. and Johnstone, I. M. (2012). Augmented sparse principal component analysis for high-dimensional data. Available at arXiv:1202.1242v1.

Pesaran, M. H. and Yamagata, T. (2012). Testing capm with a large number of assets. IZA Discussion Papers 6469, Institute for the Study of Labor.

Ravikumar, P., Wainwright, M. J., Raskutti, G. and Yu, B. (2011). High-dimensional covariance estimation by minimizing ℓ_1-penalized log-determinant divergence. *Electron. J. Stat.* **5** 935–980. MR2836766

Rigollet, P. and Tsybakov, A. B. (2012). Comment: "Minimax estimation of large covariance matrices under ℓ_1-norm" [MR3027084]. *Statist. Sinica* **22** 1358–1367. MR3027087

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

Sharpe, W. F. (1964). Capital asset prices: A theory of market equilibrium under conditions of risk. *J. Finance* **19** 425–442.

Srivastava, M. S. and Du, M. (2008). A test for the mean vector with fewer observations than the dimension. *J. Multivariate Anal.* **99** 386–402. MR2396970

Tsybakov, A. B. (2009). *Introduction to Nonparametric Estimation*. Springer, New York. MR2724359

Verzelen, N. (2012). Minimax risks for sparse regressions: Ultra-high-dimensional phenomenons. *Electron. J. Stat.* **6** 38–90. MR2879672

Vu, V. and Lei, J. (2012). Minimax rates of estimation for sparse PCA in high-dimensions. In *Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics April 21–23, 2012, JMLR W&CP* **22** 1278–1286.

Zou, H., Hastie, T. and Tibshirani, R. (2006). Sparse principal component analysis. *J. Comput. Graph. Statist.* **15** 265–286. MR2252527

J. Fan
W. Wang
Department of Operations Research
and Financial Engineering
Princeton University
Princeton, New Jersey 08544
USA
E-mail: jqfan@princeton.edu
weichenw@princeton.edu

P. Rigollet
Department of Mathematics
Massachusetts Institute of Technology
77 Massachusetts Avenue
Cambridge, Massachusetts 02139-4307
USA
E-mail: rigollet@math.mit.edu