Statistica Sinica **28** (2018), 2657-2670
doi:https://doi.org/10.5705/ss.202017.0060

APPLICATIONS OF PETER HALL'S MARTINGALE LIMIT THEORY TO ESTIMATING AND TESTING HIGH DIMENSIONAL COVARIANCE MATRICES

Danning Li, Lingzhou Xue and Hui Zou

Jilin University, Pennsylvania State University and University of Minnesota

Abstract: Martingale limit theory is increasingly important in modern probability theory and mathematical statistics. In this article, we give a selected overview of Peter Hall's contributions to both the theoretical foundations and the wide applicability of martingales. We highlight his celebrated coauthored book, Hall and Heyde (1980) and his ground-breaking paper, Hall (1984). To illustrate the power of his martingale limit theory, we present two contemporary applications to estimating and testing high dimensional covariance matrices. In the first, we use the martingale central limit theorem in Hall and Heyde (1980) to obtain the simultaneous risk optimality and consistency of Stein's unbiased risk estimation (SURE) information criterion for large covariance matrix estimation. In the second application, we use the central limit theorem for degenerate U-statistics in Hall (1984) to establish the consistent asymptotic size and power against more general alternatives when testing high-dimensional covariance matrices.

Key words and phrases: Degenerate U-statistics, hypothesis testing, large covariance matrix, martingale limit theory, Stein's unbiased risk estimation.

1. Introduction

The concept of *martingale* was first introduced by Paul Levy in probability theory, and its name was introduced later by Jean Ville, in 1939. The early development of martingale theory includes Levy's martingale characterization, Bernstein's inequality for weakly dependent random variables, and Doob's martingale convergence theorems. The interplay of theory and applications is evident in the history of probability and mathematical statistics. Statisticians have employed martingales as a technical tool in a wide range of applications since the 1970s. As a result, asymptotic properties of martingales were of increasing importance in studying complex probabilistic behaviors. Peter Hall became a world leader in the theory of martingales when he was working on his master and doctoral theses at Australian National University and Oxford University, advised by

Chris Heyde and John Kingman respectively. He was introduced as "Mr Martingale" when he visited the University of Cambridge in the mid-1970s (Delaigle and Speed (2016)). He made fundamental contributions to both the theoretical foundations and the wide applicability of martingales.

We first give a selected overview of Peter Hall's contributions to martingale limit theory. His main research interests focus on the martingale central limit theorems and invariance principles (Brown (1971); McLeish (1974)), which are the heart of the book by Hall and Heyde (1980). Hall (1977) derived the general martingale central limit theorems and invariance principles under relaxed conditions. Hall (1978) generalized Bernstein's discovery of the convergence of moments in the central limit theorem to the martingale case, and proved the convergence of moments in martingale central limit theorems. Hall and Heyde (1976) used the Skorokhod representation to obtain a unified approach to the law of the iterated logarithm for martingales, and Hall (1979a) worked out the powerful Skorokhod representation method to prove Martingale invariance principles under quite general conditions. Hall and Heyde (1981) obtained the nonuniform estimate of the rate of convergence in the martingale central limit theorem, which provides a martinagle analogue of Feller's generalization of the Berry-Esseen theorem.

Hall and Heyde (1980) is one of the most important reference books in martingales. It provides a comprehensive overview of the state-of-the-art martingale limit theory and wide applications to illustrate the power of martingale methods. The book bridged the gap between martingale theory and applications, and it has had a broad, significant and long-lasting impact on numerous areas of probability theory, mathematical statistics, and econometrics. In another ground-breaking paper, Hall (1984) used martingale theory to obtain a central limit theorem for degenerate U-statistics with applications to multivariate nonparametric density estimators. Consider the degenerate U-statistic $U_n = \sum\sum_{1\le i<j\le n} H_n(X_i, X_j)$ where $X_1, \ldots, X_n$ are independent and identically distributed random observations, and $E\{H_n(X_1, X_2)|X_1\} = 0$ almost surely. Hall (1984) assumed more practicable conditions to derive the central limit theorem of U_n. Let $G_n(x, y) = E\{H_n(X_1, x)H_n(X_1, y)\}$. More specifically, given that H_n is symmetric, $E[H_n^2(X_1, X_2)] < \infty$, and

$$\lim_{n\to\infty} \frac{E\{G_n^2(X_1, X_2)\} + (1/n)E\{H_n^4(X_1, X_2)\}}{[E\{H_n^2(X_1, X_2)\}]^2} = 0,$$

Hall (1984) proved that U_n is asymptotically normally distributed with zero mean and covariance matrix $(1/2)n^2E\{H_n^2(X_1, X_2)\}$. Because of Hall and Heyde (1980) and Hall (1984), theoretical progress in martingales has led to a number of important research topics: weak convergence of U-statistics and empirical processes (Loynes (1978); Hall (1979b)), weak convergence of log-likelihood-ratio processes (Hall and Loynes (1977)), nonparametric function estimation and modeling (Hall (1984); Hardle, Marron and Park (1988); Hall, Marron and Park (1992); Racine and Li (2004)), sliced inverse regression (Hsing and Carroll (1992); Hall and Li (1993)), empirical likelihood estimation (Donald, Imbens and Newey (2003)), unit root tests in time series regression (Phillips and Perron (1988); Elliott, Rothenberg and Stock (1996)), structural change estimation in econometric models (Andrews (1993); Bai and Perron (1998)), autocorrelation matrix estimation (Andrews (1991)), and many others. In recent years, the martingale limit theory in Hall and Heyde (1980) and Hall (1984) has received considerable attention in the development of high-dimensional statistical inference such as high-dimensional mean tests (Chen and Qin (2010); Wang, Peng and Li (2015)), high-dimensional covariance tests (Schott (2007); Li and Xue (2015); He and Chen (2016)), and inference on conditional dependence (Wang et al. (2015)), among others.

In the rest of this paper, we present applications of Hall and Heyde (1980) and Hall (1984) to estimating and to testing high dimensional covariance matrices. Section 2 applies the martingale central limit theorem to obtain consistency for Stein's unbiased risk estimation (SURE) information criteria (Stein (1981); Efron (1986, 2004)) for large covariance matrix estimation. Section 3 applies the central limit theorem for degenerate U-statistics in Hall (1984) to establish the consistent asymptotic size and power for a new test statistic against more general alternatives when testing high-dimensional covariance matrices. Section 4 provides numerical studies to demonstrate the finite-sample performance. The complete proofs of main results are included in a separate supplementary file. consistency for Stein's unbiased risk estimation (SURE) information criteria (Stein (1981); Efron (1986, 2004)) for large covariance matrix estimation. Section 3 applies the central limit theorem for degenerate U-statistics in Hall (1984) to establish the consistent asymptotic size and power for a new test statistic against more general alternatives when testing high-dimensional covariance matrices. Section 4 provides numerical studies to demonstrate the finite-sample performance. The complete proofs of main results are included in a separate supplementary file.

2. Application to The SURE Information Criterion

Let $X_1, \ldots, X_n$ be independent and identically distributed p-dimensional

Gaussian observations with mean vector μ and covariance matrix $\Sigma_{p\times p} = \left(\sigma_{ij}\right)_{p\times p}$. We assume that $p \geq n$ and p is of a nearly exponential order of n (i.e., $\log(p) = o(n)$). The problem of estimating Σ is important to various multivariate statistical methods and theory. Let $\tilde{\Sigma}^s = \left(\tilde{\sigma}^s_{ij}\right)_{p\times p}$ be the sample covariance matrix. It is well-known that $\tilde{\Sigma}^s$ performs poorly when estimating Σ in high dimensions. Several regularized estimators of large covariance matrices have been proposed, including banding (Wu and Pourahmadi (2003); Bickel and Levina (2008a); Fan, Xue and Zou (2016)), tapering (Furrer and Bengtsson (2007); Cai, Zhang and Zhou (2010); Xue and Zou (2014)), and thresholding (Bickel and Levina (2008b); Rothman, Levina and Zhu (2009); Cai and Liu (2011); Xue, Ma and Zou (2012)). The minimax optimality was established for large covariance matrix estimation (Cai, Zhang and Zhou (2010); Cai and Zhou (2012); Xue and Zou (2013)).

Little is known about the model selection criterion when estimating large covariance matrices. Stein's unbiased risk estimation (SURE) information criterion (Stein (1981)) has shown appealing performances in adaptive wavelet thresholding (Donoho and Johnstone (1995)) and sparse linear regression (Efron et al. (2004); Zou, Hastie and Tibshirani (2007)). Based on martingale central limit theorems in Hall and Heyde (1980), we attempt to obtain model selection consistency of SURE information criterion for large covariance matrix estimation. To facilitate discussion, we focus on the estimation of large bandable covariance matrices, which have natural applications for modeling temporal and spatial dependence. Following Bickel and Levina (2008a) and Cai, Zhang and Zhou (2010), we assume that Σ is in

$$\mathcal{G}_\alpha = \{\Sigma : \ |\sigma_{ij}| \leq M_1 |i-j|^{-(\alpha+1)}, \forall\, i \neq j, \text{ and } \lambda_{\max}(\Sigma) \leq M_0\}, \tag{2.1}$$

where $\lambda_{\max}(\Sigma)$ is the largest eigenvalue of matrix Σ, and α, M_0, and M_1 are positive constants. The constant α controls the decay rate of the off-diagonal elements of Σ. Without loss of generality, we assume $\sigma_{ii} = 1$ for $1 \leq i \leq p$ in this section.

To estimate Σ in $\mathcal{G}_\alpha$, we consider the banded covariance matrix

$$\hat{\Sigma}^{(\tau)} = \left(\hat{\sigma}^{(\tau)}_{ij}\right)_{1\leq i,j\leq p}$$

where $\hat{\sigma}^{(\tau)}_{ij} = \omega^{(\tau)}_{ij}\tilde{\sigma}_{ij}$ and $\omega^{(\tau)}_{ij}$ is the banding weight satisfying: (i) $\omega^{(\tau)}_{ij} = 1$ for $|i-j| < \tau$; (ii) $\omega^{(\tau)}_{ij} = 0$ for $|i-j| \geq \tau$. We need to properly choose the banding parameter τ in practice.

We introduce the SURE information criterion to select the banding parameter. Let $R(\tau) = \mathbb{E}\|\hat{\Sigma}^{(\tau)} - \Sigma\|_F^2$ be the Frobenius risk of $\hat{\Sigma}^{(\tau)}$. Here $R(\tau)$ satisfies the Stein's identity

$$R(\tau) = \mathbb{E}\|\hat{\Sigma}^{(\tau)} - \tilde{\Sigma}^s\|_F^2 - \sum_{i,j} \mathrm{var}(\tilde{\sigma}_{ij}^s) + 2\sum_{i,j} \mathrm{cov}\left(\hat{\sigma}_{ij}^{(\tau)}, \tilde{\sigma}_{ij}^s\right), \tag{2.2}$$

where we used the fact that $\tilde{\Sigma}^s$ is an unbiased estimate for Σ. The third term on the right-hand side is referred to as the covariance penalty (Efron (2004)). By definition, we obtain that $\mathrm{cov}(\hat{\sigma}_{ij}^{(\tau)}, \tilde{\sigma}_{ij}^s) = \{(n-1)/n\}\omega_{ij}^{(\tau)}\mathrm{var}(\tilde{\sigma}_{ij}^s)$. Let $\widehat{\mathrm{var}}(\tilde{\sigma}_{ij}^s)$ be an unbiased estimator of $\mathrm{var}(\tilde{\sigma}_{ij}^s)$. Then, we derive Stein's unbiased risk estimator of $R(\tau)$ as

$$\mathrm{SURE}(\tau) = \|\hat{\Sigma}^{(\tau)} - \tilde{\Sigma}^s\|_F^2 - \sum_{i,j} \widehat{\mathrm{var}}(\tilde{\sigma}_{ij}^s) + 2\frac{n-1}{n}\sum_{i,j} \omega_{ij}^{(\tau)}\widehat{\mathrm{var}}(\tilde{\sigma}_{ij}^s). \tag{2.3}$$

We find that $\mathbb{E}\{\mathrm{SURE}(\tau)\} = R(\tau)$. Following Yi and Zou (2013) and Li and Zou (2016), one sees that $\mathrm{SURE}(\tau)$ has an explicit expression as

$$\mathrm{SURE}(\tau) = \sum_{1\le i,j\le p}\left(\frac{n}{n-1} - \omega_{ij}^{(\tau)}\right)^2 \tilde{\sigma}_{ij}^2 + \sum_{1\le i,j\le p}\left(2\omega_{ij}^{(\tau)} - \frac{n}{n-1}\right)(a_n\tilde{\sigma}_{ij}^2 + b_n\tilde{\sigma}_{ii}\tilde{\sigma}_{jj}),$$

with $a_n = \{n(n-3)\}/\{(n-1)(n-2)(n+1)\}$ and $b_n = n/\{(n+1)(n-2)\}$.

Now, we can select the banding parameter by the SURE tuning

$$\hat{\tau}_n = \arg\min_{\tau}\ \mathrm{SURE}(\tau). \tag{2.4}$$

Efron (1986, 2004) showed that SURE is equivalent to AIC for regression models with an additive homoscedastic Gaussian noise. It is also known that AIC yields an asymptotic minimax optimal estimator (Yang (2005)). It was expected that $\mathrm{SURE}(\tau)$ might have the fundamental properties of AIC (Shao (1997); Yang (2005)) and result in a minimax optimal banded covariance matrix estimator. Li and Zou (2016) proved that by minimizing $\mathrm{SURE}(\tau)$ over all possible banded estimators, we obtain the minimax optimal rate of convergence and the resulting estimator $\hat{\Sigma}^{(\hat{\tau}_n)}$ is comparable to the oracle estimator $\hat{\Sigma}^{(k_0)}$ given the true banding parameter k_0,

$$\sup_{\Sigma\in\mathcal{G}_\alpha}\ \mathbb{E}\|\hat{\Sigma}^{(\hat{\tau}_n)} - \Sigma\|_F^2 \asymp \sup_{\Sigma\in\mathcal{G}_\alpha}\ \mathbb{E}\|\hat{\Sigma}^{(k_0)} - \Sigma\|_F^2.$$

Thus, we can regard $\mathrm{SURE}(\tau)$ as the analogue of AIC for large bandable covariance matrix estimation. Here we study the bandwidth selection property of SURE tuning. In applications, the SURE information criterion would be more appealing if it was consistent in identifying the true bandwidth. In traditional linear regression, AIC is risk optimal, and BIC is known for its selection consistency property (Shao (1997); Yang (2005)). Recently, certain AIC-type criteria have been shown to achieve the consistency property under a high-dimensional setting. For instance, Fujikoshi, Sakurai and Yanagihara (2014) and Yanagihara, Wakaki and Fujikoshi (2015) established the consistency of AIC-type criteria in high-dimensional multivariate linear regression, and Bai, Fujikoshi and Choi (2015) established the consistency of AIC-type criteria in high-dimensional principal component analysis. Here we use the martingale central limit theorem in Hall and Heyde (1980) to prove that when the true covariance matrix is banded, by minimizing SURE(τ) we select the true bandwidth with probability one.

Theorem 1. *Let $\Sigma_0 \in \mathcal{G}_\alpha$ be the true banded matrix with bandwidth k_0, where $\sigma_{ij} = 0$ if $|i-j| \geq k_0$. In $(1/p)\min_{h \leq k_0 - 1} \sum_{|i-j|=h} \sigma_{ij}^2 \gg \log n/n$, then, with probability one, SURE achieves the bandwidth selection consistency that $\hat{\tau}_n = k_0$.*

3. Application to Testing the Covariance Structure

Let $X_1, \ldots, X_n$ be independent and identically distributed p-dimensional Gaussian observations with mean vector μ and covariance matrix Σ. We assume that $p \gg n$ and $\lambda_{max}(\Sigma) < M_0$ for some constant M_0. Testing the covariance structure in Σ is of importance in a wide range of research fields. In Section 3, we consider testing the hypothesis that Σ is banded with some given bandwidth $k_0 \geq 1$,

$$\mathbf{H}_0 : \sigma_{ij} = 0, \quad \forall (i,j) \text{ such that } |i-j| \geq k_0. \tag{3.1}$$

When $k_0 = 1$, $\mathbf{H}_0$ corresponds to testing the mutual independence of Gaussian random variables. In the literature, $\mathbf{H}_0$ has been considered in Cai and Jiang (2011), Qiu and Chen (2012, 2015), among others. We introduce two parameter spaces for Σ:

$$\mathcal{G}_1 = \left\{ \Sigma = (\sigma_{ij})_{p \times p} : \ \sigma_{ii} = \sigma_{ji} \text{ and } \max_{|i-j| \geq k_0} |\sigma_{ij}| > C\sqrt{\frac{\log p}{n}} \right\};$$
$$\mathcal{G}_2 = \left\{ \Sigma = (\sigma_{ij})_{p \times p} : \ \sigma_{ii} = \sigma_{ji} \text{ and } \frac{n}{p} \sum_{|i-j| \geq k_0} \sigma_{ij}^2 \gg \log p \right\}.$$

In this $\mathcal{G}_1$ represents the parameter space in which the covariance has a few relatively large entries with $|i-j| \geq k_0$, and $\mathcal{G}_2$ denotes the parameter space in which the covariance contains a lot of small nonzero entries with $|i-j| \geq k_0$. In current literature, extreme-value type statistics test against the sparse alternative $\mathcal{G}_1$ (Cai and Jiang (2011)), and sum-of-squares type statistics test against the dense alternative $\mathcal{G}_2$ (Qiu and Chen (2012, 2015)). As we do not have the prior knowledge of the sparse or dense alternative in practice, it is important to effectively test against general alternatives. Here we are interested in testing procedure that boosts power against the more general alternative that

$$\mathbf{H}_1 : \Sigma \in \mathcal{G}_1 \cup \mathcal{G}_2. \tag{3.2}$$

Let $\Gamma = (\rho_{ij})_{p\times p}$ be the corresponding correlation matrix, and $\tilde{\Gamma} = (\tilde{\rho}_{ij})$ its sample estimate where $\bar{x}_k = (1/n)\sum_{i=1}^n x_{ik}$ and

$$\tilde{\rho}_{ij} = \frac{(x_i - \bar{x}_i)^T (x_j - \bar{x}_j)}{\|x_i - \bar{x}_i\| \cdot \|x_j - \bar{x}_j\|}, \quad 1 \le i, j \le p \tag{3.3}$$

Cai and Jiang (2011) proposed the maximum test statistic

$$L_n = \max_{|i-j|\ge k_0} |\tilde{\rho}_{ij}|, \tag{3.4}$$

Let $\Gamma_{p,\delta} = \{1 \le i \le p\,;\ |\rho_{ij}| > 1-\delta$ for some $1 \le j \le p$ with $j \ne i\}$ for any $0 < \delta < 1$. When $p \to \infty$ with $\log p = o(n^{1/3})$ and $|\Gamma_{p,\delta}| = o(p)$, Cai and Jiang (2011) proved that $nL_n^2 - 4\log p + \log\log p$ converges weakly to an extreme distribution of type I with the distribution function $F(y) = e^{-1/\sqrt{8\pi}e^{-y/2}}$, $\forall y \in \mathbb{R}$ under $\mathbf{H}_0$. However, Hall (1979c) and Li and Xue (2015) point out that the extreme-value form statistic L_n may suffer from low power against dense alternatives with $\Sigma \in \mathcal{G}_2$.

To boost the power of L_n against $\mathbf{H}_1$, we introduce a quadratic form statistic. To this end, take $Z_i = \{1/\sqrt{i(i+1)}\}(X_1 + \cdots + X_i) - \{i/\sqrt{i(i+1)}\}X_{i+1}$ for $1 \le i \le n-1$ and $Z_n = 1/\sqrt{n}\sum_{i=1}^n X_i$. Note that $Z_1, \ldots, Z_{n-1}$ are *i.i.d.* $N_p(0,\Sigma)$ random vectors. Using Theorem 3.1.2 from Muirhead (1982), $\tilde{\Sigma} = (1/n)\sum_{k=1}^n (X_k - \bar{X})(X_k - \bar{X})^T$ is equal to $\hat{\Sigma} = \left(\hat{\sigma}_{ij}\right)_{1\le i,j\le p} = (1/n)\sum_{k=1}^{n-1} Z_k Z_k^T$. Now we define the quadratic form statistic as follows:

$$Q_n^2 = \frac{S_n^2(k_0)}{S}, \tag{3.5}$$

where

$$\begin{aligned} S_n^2(k_0) &= \sum_{1\le i,j\le p} \omega_{ij}^{(k_0)} \left\{ \hat{\sigma}_{ij}^2 - \sum_{m=1}^{n-1} \frac{(z_{mi}z_{mj})^2}{n^2} \right\}, \\ \text{and} \quad S^2 &= \sum_{1\le l<m\le n} \left\{ \frac{1}{n^2} \sum_{1\le i,j\le p} 2\omega_{ij}^{(k_0)} z_{mi}z_{mj}z_{li}z_{lj} \right\}^2. \end{aligned} \tag{3.6}$$

We follow Hall (1984) to derive the central limit theorem for $S_n^2(k_0)$. Let $H_n(Z_m, Z_l) = (1/n^2)\sum_{1\le i,j\le p} 2\omega_{ij}^{(k_0)}(z_{mi}z_{mj} - \sigma_{ij})(z_{li}z_{lj} - \sigma_{ij})$ and

$$Y_m = \frac{2(n-2)}{n^2} \sum_{1\le i,j\le p} \omega_{ij}^{(k_0)} \sigma_{ij}(z_{mi}z_{mj} - \sigma_{ij}),$$

where $\omega_{ij}^{(k_0)}$s are the same banding weights defined in Section 2. We can rewrite the difference $S_n^2(k_0) - ES_n^2(k_0)$ as

$$S_n^2(k_0) - ES_n^2(k_0) = \sum_{1\le i,j\le p} \omega_{ij}^{(k_0)} \left\{ \hat{\sigma}_{ij}^2 - \sum_{m=1}^{n-1} \frac{(z_{mi}z_{mj})^2}{n^2} - \frac{n(n-1)}{n^2}\sigma_{ij}^2 \right\}$$
$$= \sum_{m=2}^{n-1}\sum_{l=1}^{m-1} H_n(Z_m, Z_l) + \sum_{m=2}^{n-1} Y_m, \tag{3.7}$$

where we used the fact that $E\hat{\Sigma} = \Sigma$. Under $\mathbf{H}_0$, $Y_m = 0$ and $ES_n^2(k_0) = 0$. Then as shown in (3.7), $S_n^2(k_0) - ES_n^2(k_0)$ is a degenerate U statistic of the form of U_n in Hall (1984).

We follow Theorem 1 of Hall (1984) to show a central limit theorem for $S_n^2(k_0)$.

Theorem 2. *Let* $\mathrm{Var}_n(k_0) = \{n(n-1)\}/2E\{H_n(Z_1, Z_2)^2\}$*. Under* $\mathbf{H}_0$*,*

$$\mathrm{Var}_n(k_0)^{-1/2}\{S_n^2(k_0) - ES_n^2(k_0)\} \to N(0,1)$$

in distribution as $n \to \infty$*. Further,*

$$\sup_t \left| P\left(\frac{S_n^2(k_0) - ES_n^2(k_0)}{\sqrt{\mathrm{Var}_n(k_0)}} \le t \right) - \Phi(t) \right| \le Cn^{-1/5}.$$

As well, we have the convergency of S^2 in probability to $\mathrm{Var}_n(k_0)$.

Theorem 3. *Under* $\mathbf{H}_0$*,* $S^2/\{\mathrm{Var}_n(k_0)\} \to 1$ *in probability as* $n \to \infty$*.*

Combining Theorems 3–4 and Slutsky's theorem, we obtain a central limit theorem for Q_n^2.

Theorem 4. *Under* $\mathbf{H}_0$*,* Q_n^2 *converges weakly to* $N(0,1)$ *as* $n \to \infty$*.*

Now, we combine the strengths of both Q_n^2 and L_n and propose a new testing procedure:

$$TS = I_{\{Q_n^2 + (nL_n^2 - 4\log p + \log\log p) \ge c_\alpha\}}$$

where the threshold c_α is defined as the α upper quantile of the convolution distribution $\Phi \star F$. Here $TS = 1$ leads to the rejection of $\mathbf{H}_0$. In what follows, we provide the theoretical guarantee of its asymptotic size and power. To this end, we define the marginal distribution functions of Q_n and L_n as

$$P_{Q_n}(z) = P\big(Q_n^2 \le z\big),\ P_{L_n}(y) = P\big(nL_n^2 - 4\log p + \log\log p \le y\big),$$

as well as their joint distribution function as

$$P_{Q_n,L_n}(z,y) = P\big(\{Q_n^2 \le z\} \cap (nL_n^2 - 4\log p + \log\log p \ge y)\big).$$

We derive the explicit joint limiting law of Q_n and L_n, that shares the spirit of Li and Xue (2015).

Theorem 5. *If* $|\Gamma_{p,\delta}| = o(p)$ *for* $\delta \in (0,1)$ *and* $p \to \infty$ *with* $\log p = o(n^{1/5})$, *then, under* $\mathbf{H}_0$, *for any* z *and* y *we have*

$$P_{Q_n,L_n}(z,y) \to \Phi(z)\left(1 - e^{(-1/\sqrt{8\pi})e^{-(y/2)}}\right). \tag{3.8}$$

Let $P_{\mathbf{H}_0}(\cdot)$ be the probability given the null hypothesis $\mathbf{H}_0$, and $P_{\mathbf{H}_1}(\cdot)$ be the probability given the alternative hypothesis $\mathbf{H}_1$. $P_{\mathbf{H}_0}(TS = 1)$ is the conditional probability of rejecting $\mathbf{H}_0$ given that $\mathbf{H}_0$ is true, and $P_{\mathbf{H}_1}(TS = 1)$ is the conditional probability of correctly rejecting $\mathbf{H}_0$. In the sequel, we prove that TS does control the significance level and also achieves consistent power.

Theorem 6. *Under the conditions of Theorem* 5, *we have*

$$P_{\mathbf{H}_0}(TS = 1) \to \alpha \quad \text{as } n \to \infty.$$

Otherwise, if $p/n \to \infty$ *and* $\Sigma \in \mathcal{G}_1 \cup \mathcal{G}_2$, *we have*

$$\inf_{\Sigma \in \mathcal{G}_1 \cup \mathcal{G}_2} P_{\mathbf{H}_1}(TS = 1) \to 1 \quad \text{as } n \to \infty.$$

4. Numerical Properties

In this section, we demonstrate the numerical performance of our proposed SURE information criterion and our proposed new testing procedure. We consider three different models to simulate the independent observations $X_1, \ldots, X_n$ that are $N_p(0,\Sigma)$, and $\Sigma = (\sigma_{ij})_{p\times p}$ specifies the covariance structure:

- `Model 1.` $\sigma_{ij} = I(i=j) + (1/4)I(|i-j| \le 4)$ for $1 \le i,j \le p$.
- `Model 2.` $\sigma_{ij} = I(i=j) + (1/4)I(|i-j| \le 4) + 0.45I(i=7, j=1) + 0.45I(i=1, j=7)$ for $1 \le i,j \le p$.
- `Model 3.` $\sigma_{ij} = I(i=j) + (1/4)I(|i-j| \le 4) + 2.5\sqrt{\log p/n}\,I(|i-j| \ge 5)$ for $1 \le i,j \le p$.

Model 1 specifies a banded covariance matrix with bandwidth 5 to evaluate the proposed SURE information criterion. Model 1 mimics the null hypothesis $\mathbf{H}_0$ in Section 3 to examine the size. Model 2 corresponds to a covariance matrix in $\mathcal{G}_1$ with only two relatively large entries (i.e., σ_{17} and σ_{71}) with $|i-j| > 4$. Model 3 corresponds to a covariance matrix in $\mathcal{G}_2$ with many small disturbances. For each simulation model, we let $n = 200$ and $p = 50, 100, 200, 400, 800$, and generated 1,000 independent repetitions.

Table 1. Selection performance of SURE information criterion in Model 1.

Selected bandwidth	4	5	6
$p = 200$	0/1,000	1,000/1,000	0/1,000
$p = 400$	0/1,000	1,000/1,000	0/1,000
$p = 800$	0/1,000	1,000/1,000	0/1,000

Table 2. Performance of different testing procedures in Model 1.

p	Q_n^2	L_n	TS
50	0.0476	0.0266	0.034
100	0.044	0.029	0.0348
200	0.0408	0.026	0.0272
400	0.045	0.0226	0.0234
800	0.0446	0.0218	0.0204

Table 3. Performance of different testing procedures in Model 2.

p	Q_n^2	L_n	TS
50	0.1416	0.996	0.996
100	0.078	0.99	0.9902
200	0.0546	0.9788	0.9756
400	0.053	0.953	0.9502
800	0.0514	0.914	0.8504

Table 4. Performance of different testing procedures in Model 3.

p	Q_n^2	L_n	TS
50	0.0972	0.8256	0.8238
100	0.0716	0.8612	0.8582
200	0.0562	0.89	0.887
400	0.0492	0.913	0.9108
800	0.0442	0.9344	0.9284

To check the finite-sample performance of our proposed SURE selection in Model 1, we report the frequencies of selecting the corresponding bandwidth among 1,000 repetitions in Table 1. Our proposed SURE achieves the desired selection consistency, which is consistent with Theorem 1 of Section 2.

We examine the proposed new testing procedure together with the maximum form test statistic L_n in (3.4) and the quadratic form test statistic Q_n^2 in (3.5). Simulation results are summarized in Tables 2-4. As shown in Table 2, all three testing procedures achieve the reasonably good size in Model 1. As to power, L_n clearly suffers from low power against dense alternatives, and Q_n^2 suffers from low power against sparse alternatives. However, TS retains good power against the sparse alternative in Model 2 and the dense alternative in Model 3.

Supplementary Materials

In the online supplement, we provide the complete proofs of Theorems 1, 2, 3, 5 and 6.

Acknowledgment

Li and Xue are joint first-authors. The authors thank the Co-Editor, an associate editor, and two referees for their constructive comments and suggestions. Xue's research is partially supported by the National Science Foundation grant DMS-1505256. Zou's research is partially supported by the National Science Foundation grant DMS-1505111.

References

Andrews, D. W. (1991). Heteroskedasticity and autocorrelation consistent covariance matrix estimation. *Econometrica* **59**, 817-858.

Andrews, D. W. (1993). Tests for parameter instability and structural change with unknown change point. *Econometrica* **61**, 821-856.

Bai, J. and Perron, P. (1998). Estimating and testing linear models with multiple structural changes. *Econometrica* **66**, 47-78.

Bai, Z., Fujikoshi, Y. and Choi, K. P. (2015) High-dimensional consistency of AIC and BIC for estimating the number of significant components in principal component analysis. Technical Report. Hiroshima University.

Brown, B. M. (1971). Martingale central limit theorems. *Ann. Math. Statist.* **42**, 59-66.

Bickel, P. J. and Levina, E. (2008a). Regularized estimation of large covariance matrices. *Ann. Statist.* **36**, 199-227.

Bickel, P. J. and Levina, E. (2008b). Covariance regularization by thresholding. *Ann. Statist.* **36**, 2577-2604.

Cai, T. and Jiang, T. (2011). Limiting laws of coherence of random matrices with applications to testing covariance structure and construction of compressed sensing matrices. *Ann. Statist.* **39**, 1496–1525.

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

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

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

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.

Delaigle, A. and Speed, T. (2016). A stats boffin the world knew as Mr Martingale. Available at http://www.smh.com.au/comment/obituaries/a-stats-boffin-the-world-knew-as-mr-martingale-20160321-gnn8g1.html

Donald, S. G., Imbens, G. W. and Newey, W. K. (2003). Empirical likelihood estimation and consistent tests with conditional moment restrictions. *J. Econometrics* **117**, 55-93.

Donoho, D. and Johnstone, I. (1995). Adapting to unknown smoothness via wavelet shrinkage. *J. Amer. Statist. Assoc.* **90**, 1200-1224.

Efron, B. (1986). How biased is the apparent error rate of a prediction rule. *J. Amer. Statist. Assoc.* **81**, 461-470.

Efron, B. (2004). The estimation of prediction error: covariance penalties and cross-validation. *J. Amer. Statist. Assoc.* **99**, 619-632.

Efron, B., Hastie, T., Johnstone, I. and Tibshirani, R. (2004). Least angle regression (with discussion). *Ann. Statist.* **32**, 407-499.

Elliott, G., Rothenberg, T. J. and Stock, J. H. (1996). Efficient tests for an autoregressive unit root. *Econometrica* **64**, 813-836.

Fan, J., Xue, L. and Zou, H. (2016). Multi-task quantile regression under the transnormal model. *J. Amer. Statist. Assoc.* **111**, 1726-1735.

Fujikoshi, Y., Sakurai, T. and Yanagihara, H. (2014). Consistency of high-dimensional AIC-type and Cp-type criteria in multivariate linear regression. *J. Multivariate Anal.* **123**, 184-200.

Furrer, R. and Bengtsson, T. (2007). Estimation of high-dimensional prior and posterior covariance matrices in Kalman filter variants. *J. Multivariate Anal.* **98**, 227-255.

Hall, P. (1977). Martingale invariance principles. *Ann. Probab.* **5**, 875-887.

Hall, P. (1978). The convergence of moments in the martingale central limit theorem. *Zeitschrift fur Wahrscheinlichkeitstheorie und Verwandte Gebiete* **44**, 253-260.

Hall, P. (1979a). On the skorokhod representation approach to martingale invariance principles. *Ann. Probab.* 371-376.

Hall, P. (1979b). On the invariance principle for U-statistics. *Stochastic Process. Appl.* **9**, 163-174.

Hall, P. (1979c). On the rate of convergence of normal extremes. *J. Appl. Probab.* **16**, 433-439.

Hall, P. (1984). Central limit theorem for integrated square error of multivariate nonparametric density estimators. *J. Multivariate Anal.* **14**, 1-16.

Hall, P. G. and Heyde, C. C. (1976). On a unified approach to the law of the iterated logarithm for martingales. *Bull. Aust. Math. Soc.* **14**, 435-447.

Hall, P. and Heyde, C. C. (1980). *Martingale Limit Theory and Its Application*, Academic Press.

Hall, P. and Heyde, C. C. (1981). Rates of convergence in the martingale central limit theorem. *Ann. Probab.* **9**, 395-404.

Hall, P. and Li, K. C. (1993). On almost linearity of low dimensional projections from high dimensional data. *Ann. Statist.* **21**, 867-889.

Hall, W. J. and Loynes, R. M. (1977). Weak convergence of processes related to likelihood ratios. *Ann. Statist.* **5**, 330-341.

Hall, P., Marron, J. S. and Park, B. U. (1992). Smoothed cross-validation. *Probab. Theory Related Fields* **92**, 1-20.

Hardle, W., Hall, P. and Marron, J. S. (1988). How far are automatically chosen regression smoothing parameters from their optimum? *J. Amer. Statist. Assoc.* **83**, 86-95.

He, J. and Chen, S. X. (2016). Testing super-diagonal structure in high dimensional covariance matrices. *J. Econometrics* **194**, 283-297.

Hsing, T. and Carroll, R. J. (1992). An asymptotic theory for sliced inverse regression. *Ann. Statist.* **20**, 1040-1061.

Lan, W., Luo, R., Tsai, C. L., Wang, H. and Yang, Y. (2015). Testing the diagonality of a large covariance matrix in a regression setting. *J. Bus. Econom. Statist.* **33**, 76-86.

Li, D. and Xue, L. (2015). Joint limiting laws for high-dimensional independence tests. *arXiv preprint* arXiv:1512.08819.

Li, D. and Zou, H. (2016). SURE information criteria for large covariance matrix estimation and their asymptotic properties. *IEEE Trans. Inf. Theory* **62**, 2153-2169.

Loynes, R. M. (1978). On the weak convergence of U-statistic processes, and of the empirical process. In *Mathematical Proceedings of the Cambridge Philosophical Society* **83**, 269-272. Cambridge University Press.

McLeish, D. L. (1974). Dependent central limit theorems and invariance principles. *Ann. Probab.* **2**, 620-628.

Muirhead, R. J. (1982). *Aspects of Multivariate Statistical Theory.* Wiley.

Phillips, P. C. and Perron, P. (1988). Testing for a unit root in time series regression. *Biometrika* **75**, 335-346.

Qiu, Y. and Chen, S. X. (2012). Test for bandedness of high-dimensional covariance matrices and bandwidth estimation. *Ann. Statist.* **40**, 1285-1314.

Qiu, Y. and Chen, S. X. (2015). Bandwidth selection for high-dimensional covariance matrix estimation. *J. Amer. Statist. Assoc.* **110**, 1160-1174.

Racine, J. and Li, Q. (2004). Nonparametric estimation of regression functions with both categorical and continuous data. *Journal of Econometrics* **119**, 99-130.

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

Schott, J. R. (2007). Some high-dimensional tests for a one-way MANOVA. *J. Multivariate Anal.* **98**, 1825-1839.

Shao, J. (1997). An asymptotic theory for linear model selection (with discussion). *Statist. Sinica* **7**, 221-242.

Stein, C. (1981). Estimation of the mean of a multivariate normal distribution. *Ann. Statist.* **9**, 1135-1151.

Wang, L., Peng, B. and Li, R. (2015). A high-dimensional nonparametric multivariate test for mean vector. *J. Amer. Statist. Assoc.* **110**, 1658-1669.

Wang, X., Pan, W., Hu, W., Tian, Y. and Zhang, H. (2015). Conditional distance correlation. *J. Amer. Statist. Assoc.* **110**, 1726-1734.

Wu, W. B. and Pourahmadi, M. (2003). Nonparametric estimation of large covariance matrices of longitudinal data. *Biometrika* **90**, 831-844.

Xue, L., Ma, S. and Zou, H. (2012). Positive-definite ℓ_1-penalized estimation of large covariance matrices. *J. Amer. Statist. Assoc.* **107**, 1480-1491.

Xue, L. and Zou, H. (2013). Minimax optimal estimation of general bandable covariance matrices. *J. Multivariate Anal.* **116**, 45-51.

Xue, L. and Zou, H. (2014). Rank-based tapering estimation of bandable correlation matrices. *Statist. Sinica* **24**, 83-100.

Yang, Y. (2005). Can the strengths of AIC and BIC be shared? A conflict between model indentification and regression estimation. *Biometrika* **92**, 937-950.

Yanagihara, H., Wakaki, H. and Fujikoshi, Y. (2015). A consistency property of the AIC for multivariate linear models when the dimension and the sample size are large. *Electron. J. Stat.* **9**, 869-897.

Yi, F. and Zou, H. (2013). SURE-tuned tapering estimation of large covariance matrices. *Comput. Statist. Data Anal.* **58**, 339-351.

Zou, H., Hastie, T. and Tibshirani, R. (2007). On the degrees of freedom of the Lasso. *Ann. Statist.* **35**, 2173-2192.

Mathematics School and Institute, Jilin University, Changchun City, 130012 China.

E-mail: danningli@jlu.edu.cn

Department of Statistics, Pennsylvania State University, State College, PA 16801, USA.

E-mail: lzxue@psu.edu

School of Statistics, University of Minnesota, Minneapolis, MN 55455, USA.

E-mail: zouxx019@umn.edu

(Received January 2017; accepted October 2017)