Statistica Sinica Preprint No: SS-2017-0060.R2	
Title	Applications of Peter Hall's martingale limit theory to estimating and testing high dimensional covariance matrices
Manuscript ID	SS-2017-0060.R2
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202017.0060
Complete List of Authors	Danning Li Lingzhou Xue and Hui Zou
Corresponding Author	Hui Zou
E-mail	zouxx019@umn.edu
Notice: Accepted version subject to English editing.	

Statistica Sinica

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 has been increasingly important in modern probability theory and mathematical statistics. In this article, we first give a selected overview of Peter Hall's contributions to both theoretical foundations and wide applicability of martingales. We highlight his celebrated coauthored book by Hall and Heyde (1980) and his ground-breaking paper by 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 application, 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

Danning Li and Lingzhou Xue contributed equally to this paper. Address for correspondence: Hui Zou, School of Statistics, University of Minnesota. Email: zouxx019@umn.edu

covariance matrices.

Key words and phrases: Large covariance matrix, Martingale limit theory, Degenerate U-statistics, Stein's unbiased risk estimation, Hypothesis testing.

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 began to adopt 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 to study complex probabilistic behaviors. Peter Hall became the world leader in the theory of martingales when he was working on his master and doctoral theses at the Australian National University and the 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 theoretical foundations and wide applicability of martingales.

In the sequel, we first give a selected overview of Peter Hall's contributions to the 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 celebrated 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.

Peter Hall's celebrated coauthored book (i.e., Hall and Heyde (1980)) becomes one of the most important reference books in martingales. It provides a comprehensive overview of the state of the art martingale limit theory and also wide applications to illustrate the power of martingale methods. Hall and Heyde (1980) 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 the 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 $X_1, \cdots, X_n$ 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

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

Hall (1984) proved that U_n is asymptotically normally distributed with zero mean and covariance matrix $\frac{1}{2}n^2E[H_n^2(X_1, X_2)]$. Because of Hall and Heyde (1980) and Hall (1984), the theoretical progress in martingales has lead to a number of important research topics in the past decades: 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 et al., 1988; Hall et al., 1992; Racine and Li, 2004), sliced inverse regression (Hsing and Carroll, 1992; Hall and Li, 1993), empirical likelihood estimation (Donald et al., 2003), unit root tests in time series regression (Phillips and Perron, 1988; Elliott et al., 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 et al., 2015), high-dimensional covariance tests (Schott 2007; Lan et al., 2015; Li and Xue, 2015; He and Chen, 2016), inference on conditional dependence (Wang et al., 2015), among others.

In the rest of this paper, we present two contemporary applications of Hall and Heyde (1980) and Hall (1984) to estimating and testing high dimensional covariance matrices respectively. 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.

2. Application to SURE Information Criterion

Let $X_1, \cdots, X_n$ 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 on 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 were proposed in recent literature, including banding (Wu and Pourahmadi, 2003; Bickel and Levina, 2008a; Fan et al., 2016), tapering (Furrer and Bengtsson, 2007; Cai et al., 2010; Xue and Zou, 2014), and thresholding (Bickel and Levina, 2008b; Rothman et al., 2009; Cai and Liu, 2011; Xue et al., 2012). The minimax optimality was established for large covariance matrix estimation (Cai et al., 2010; Cai and Zhou, 2012; Xue and Zou, 2013).

Yet little is known about the model selection criterion when estimating large covariance matrices. Note that 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 et al., 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 et al. (2010), we assume that Σ comes from the following parameter space:

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

where $\lambda_{\max}(\Sigma)$ is the largest eigenvalue of matrix Σ, and $\alpha, M_0, \&\ M_1$ 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

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

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

In what follows, 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 $\hat{\Sigma}^{(\tau)}$. Note that $R(\tau)$ has the following Stein's identity

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

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

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

Note that $\mathbb{E}[\mathrm{SURE}(\tau)] = R(\tau)$. Following Yi and Zou (2013) and Li and Zou (2016), it is easy to see that $\mathrm{SURE}(\tau)$ has an explicit expression as follows:

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

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

Now, we can select the banding parameter by the following 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 is expected that SURE(τ) might have the fundamental properties of AIC (Shao, 1997; Yang 2005) and it results in a minimax optimal banded covariance matrix estimator. Li and Zou (2016) proved that by minimizing SURE(τ) over all possible banded estimators, we obtained 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. Namely,

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

Thus, we may regard SURE(τ) as the analogue of AIC for large bandable covariance matrix estimation. Now, we shall study the bandwidth selection property of SURE tuning. In real-world applications, SURE information criterion would be more appealing if it could also be consistent in identifying the true bandwidth. Recall that 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 the high-dimensional setting. For instance, Fujikoshi et al. (2014) and Yanagihara et al. (2015) established the consistency of AIC-type criteria in high-dimensional multivariate linear regression, and Bai et al. (2015) established the consistency of AIC-type criteria in high-dimensional principal component analysis. In the following theorem, we used 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 the bandwidth k_0, where $\sigma_{ij} = 0$ if $|i - j| \geq k_0$. Assume that $\frac{1}{p} \min_{h \leq k_0 - 1} \sum_{|i-j|=h} \sigma_{ij}^2 \gg \log n/n$. With probability one, SURE achieves the bandwidth selection consistency that $\hat{\tau}_n = k_0$.*

Therefore, the proposed SURE information criterion achieves the simultaneous risk optimality and consistency when estimating large bandable covariance matrices.

3. Application to Testing the Covariance Structure

Let $X_1, \cdots, X_n$ 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 significant importance in a wide range of research fields. In Section 3, we consider testing the hypothesis that Σ is banded with some given bandwidth $k_0 \geq 1$, namely,

$$\mathbf{H}_0 : \sigma_{ij} = 0, \quad \forall (i,j) \text{ such that } |i-j| \geq k_0. \tag{3.5}$$

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. Next, we introduce two parameter spaces for Σ:

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

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

$$\mathbf{H}_1 : \Sigma \in \mathcal{G}_1 \cup \mathcal{G}_2. \tag{3.6}$$

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

$$\tilde{\rho}_{ij} = \frac{(x_i - \bar{x}_i)^T(x_j - \bar{x}_j)}{\|x_i - \bar{x}_i\| \cdot \|x_j - \bar{x}_j\|}, \quad 1 \le i, j \le p \tag{3.7}$$

Cai and Jiang (2011) proposed the following maximum test statistic

$$L_n = \max_{|i-j|\ge k_0} |\tilde{\rho}_{ij}|, \tag{3.8}$$

Define $\Gamma_{p,\delta} = \{1 \le i \le p;\ |\rho_{ij}| > 1-\delta$ for some $1 \le j \le p$ with $j \ne i\}$ for any $0 < \delta < 1$. Assuming that $p \to \infty$ 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 $F(y) = e^{-\frac{1}{\sqrt{8\pi}}e^{-y/2}}$, $\forall y \in \mathbb{R}$ under $\mathbf{H}_0$. However, Hall (1979c) and Li and Xue (2015) pointed 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 in the sequel. To this end, we define $Z_i = \frac{1}{\sqrt{i(i+1)}}(X_1 + \ldots + X_i) - \frac{i}{\sqrt{i(i+1)}}X_{i+1}$ for $1 \le i \le n-1$ and $Z_n = \frac{1}{\sqrt{n}}\sum_{i=1}^n X_i$. Note that $Z_1, \cdots, Z_{n-1}$

be *i.i.d.* $N_p(0, \Sigma)$ random vectors. Using Theorem 3.1.2 from Murihead (1983), $\tilde{\Sigma} = \frac{1}{n}\sum_{k=1}^{n}(X_k - \bar{X})(X_k - \bar{X})^T$ is equal to $\hat{\Sigma} = \big(\hat{\sigma}_{ij}\big)_{1\le i,j\le p} = \frac{1}{n}\sum_{k=1}^{n-1} Z_k Z_k^T$. Now we define the quadratic form statistic as follows:

$$Q_n^2 = \frac{S_n^2(k_0)}{S}, \tag{3.9}$$

where

$$\begin{aligned} S_n^2(k_0) &= \sum_{1\le i,j\le p} \omega_{ij}^{(k_0)}\{\hat{\sigma}_{ij}^2 - \sum_{m=1}^{n-1}\frac{(z_{mi}z_{mj})^2}{n^2}\}, \\ \text{and} \quad S^2 &= \sum_{1\le l<m\le n}\{\frac{1}{n^2}\sum_{1\le i,j\le p} 2\omega_{ij}^{(k_0)} z_{mi}z_{mj}z_{li}z_{lj}\}^2. \end{aligned} \tag{3.10}$$

First, we follow Hall (1984) to derive the central limit theorem of $S_n^2(k_0)$. Let $H_n(Z_m, Z_l) = \frac{1}{n^2}\sum\limits_{1\le i,j\le p} 2\omega_{ij}^{(k_0)}(z_{mi}z_{mj} - \sigma_{ij})(z_{li}z_{lj} - \sigma_{ij})$ 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 follow:

$$\begin{aligned} S_n^2(k_0) - ES_n^2(k_0) &= \sum_{1\le i,j\le p}\omega_{ij}^{(k_0)}(\hat{\sigma}_{ij}^2 - \sum_{m=1}^{n-1}\frac{(z_{mi}z_{mj})^2}{n^2} - \frac{n(n-1)}{n^2}\sigma_{ij}^2) \\ &= \sum_{m=2}^{n-1}\sum_{l=1}^{m-1} H_n(Z_m, Z_l) + \sum_{m=2}^{n-1} Y_m, \end{aligned} \tag{3.11}$$

where we used the simple fact that $E\hat{\Sigma} = \Sigma$. Under $\mathbf{H}_0$, $Y_m = 0$ and $ES_n^2(k_0) = 0$. Then as shown in (3.11), $S_n^2(k_0) - ES_n^2(k_0)$ becomes a degenerate U statistic of the same form as U_n in Hall (1984).

In the following theorem, we follow Theorem 1 of Hall (1984) to show the central limit theorem of $S_n^2(k_0)$.

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

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

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

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

Moreover, we show the convergency of S^2 in probability to $\mathrm{Var}_n(k_0)$ as follows.

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

Combining Theorems 3–4 and Slutsky's theorem, we immediately obtain the following 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 shall combine both strengths of Q_n^2 and L_n to propose the following 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 $\Phi \star F$. Note that $TS = 1$ leads to the rejection of $\mathbf{H}_0$. In what follows, we shall 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:

$$P_{Q_n}(z) = P\big(Q_n^2 \leq z\big),$$

and

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

as well as their joint distribution function:

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

In the following theorem, we derive the explicit joint limiting law of Q_n and L_n, which shares the similar spirit with Li and Xue (2015).

Theorem 5. *Assuming that* $|\Gamma_{p,\delta}| = o(p)$ *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*

$$P_{Q_n,L_n}(z,y) \to \Phi(z)\left(1 - e^{\frac{-1}{\sqrt{8\pi}}e^{\frac{-y}{2}}}\right). \tag{3.12}$$

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 $\mathbf{H}_1$. $P_{\mathbf{H}_0}(TS = 1)$ denotes 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 shall prove that TS does well control the significance level and also achieve the consistent power.

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

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

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

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

4. Numerical Properties

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

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

Model 1 specifies a banded covariance matrix with the given bandwidth 5 to evaluate the proposed SURE information criterion. Also note that

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 generate 1000 independent repetitions respectively.

First, we check the finite-sample performance of our proposed SURE selection in Model 1. We report the frequencies of selecting the corresponding bandwidth among 1000 repetitions in Table 1. Our proposed SURE achieves the desired selection consistency, which is consistent with Theorem 1 of Section 2.

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

Selected bandwidth	4	5	6
$p = 200$	0/1000	1000/1000	0/1000
$p = 400$	0/1000	1000/1000	0/1000
$p = 800$	0/1000	1000/1000	0/1000

Now, we examine the proposed new testing procedure together with the maximum form test statistic L_n in (3.8) and the quadratic form test statistic Q_n^2 in (3.9). 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. Next, we compare their 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 the nice power against not only the sparse alternative in Model 2 but also the dense alternative in Model 3.

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

Supplementary Materials

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

Acknowledgements

The authors thank the Co-Editor, an associate editor, and two referees for their constructive comments and suggestions. Lingzhou Xue's

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

research is partially supported by the National Science Foundation grant DMS-1505256. Hui 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 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. *Statistica 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.

De la Peña, V.H. (1999). A general class of exponential inequalities for martingales and ratios. *Ann. Probab.* 27, 537-564.

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

Dharmadhikari, S. W., Fabian, V., Jogdeo, K. (1968). Bounds on the moments of martingales. *The Annals of Mathematical Statistics* 1719-1723.

Donald, S. G., Imbens, G. W. and Newey, W. K. (2003). Empirical likelihood estimation and consistent tests with conditional moment restrictions. *Journal of 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(2), 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 Processes and their Applications*, 9, 163-174.

Hall, P. (1979c). On the rate of convergence of normal extremes. *Journal of Applied Probability*, 16(02), 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. *Bulletin of the Australian Mathematical Society*, 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. *Probability Theory and 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. *Journal of Econometrics*, 194, 283-297.

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

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 Transactions on Information 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* (Vol. 83, No. 02, pp. 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(1), 99-130.

Rothman, A. J., Levina, E. and Zhu, J. (2009). Generalized thresholding of large covariance matrices. *J. Amer. Statist. Assoc.* 104(485), 177-186.

Schott, J. R. (2005). Testing for complete independence in high dimensions. *Biometrika*, 92, 951-956.

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). *Statistica Sinica*, 7, 221-242.

Stein, C. (1981). Estimation of the mean of a multivariate normal distribution. *Ann. Statist.* 9(6), 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(512), 1726-1734.

Withers, C.S. (1985) The moments of the multivariate normal, *Bulletin of the Australian Mathematical Society* 32, 103-107.

Wu, W. B. and Pourahmadi, M. (2003). Nonparametric estimation of large covariance matrices of longitudinal data. *Biometrika* 90(4), 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. *Statistica 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. *Electronic Journal of Statistics*, 9(1), 869-897.

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

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

Danning Li

School of Mathematics, Jilin University

E-mail: danningli@jlu.edu.cn

Lingzhou Xue

Department of Statistics, Pennsylvania State University

E-mail: lzxue@psu.edu

Hui Zou

School of Statistics, University of Minnesota

E-mail: zouxx019@umn.edu