Part 3: Statistical Inference for HD Data

Topics in this chapter

a. One sample and two sample mean problems

b. Tests on covariance structures

c. Error variance estimation in HD regression

d. Tests on regression coefficients in HD regression

e. Confidence intervals for regression coefficients in HD regression

5.1 One sample and two sample mean problems

One sample mean problem:
Suppose that $\mathbf{x}_1, \cdots, \mathbf{x}_n$ is a random sample IID from a p-dimensional population with mean $\boldsymbol{\mu}$. We want to test

$$H_0 : \boldsymbol{\mu} = \boldsymbol{\mu}_0 \quad \text{versus} \quad H_1 : \boldsymbol{\mu} \neq \boldsymbol{\mu}_0$$

Two sample mean problem:
Suppose that $\mathbf{x}_{kj}$, $k = 1, 2$, and $j = 1, \cdots, n_k$ is a random sample IID from a population with mean $\boldsymbol{\mu}_k$. To test

$$H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 \quad \text{versus} \quad H_1 : \boldsymbol{\mu}_1 \neq \boldsymbol{\mu}_2$$

Hotelling T^2 test for one sample problem

Given a random sample $\mathbf{x}_1, \cdots, \mathbf{x}_n$ from $N_p(\boldsymbol{\mu}, \Sigma)$, define the sample mean $\bar{\mathbf{x}}$ as

$$\bar{\mathbf{x}} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i,$$

and the sample covariance matrix $\mathbf{S}$ as

$$\mathbf{S} = \frac{1}{n-1}\sum_{i=1}^{n}(\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^T.$$

Hotelling's T^2 test for one sample problem

Under $p < n$, the Hotelling's T^2 test for

$$H_0 : \boldsymbol{\mu} = \mathbf{0} \quad \text{vs} \quad H_1 : \boldsymbol{\mu} \neq \mathbf{0}$$

has its test statistics defined as

$$T^2 = n\bar{\mathbf{x}}'\mathbf{S}^{-1}\bar{\mathbf{x}}.$$

Under H_0,

$$T^2 \sim \frac{(n-1)p}{n-p} F_{p,n-p}.$$

For one sample normal mean problem, T^2 is equivalent to the LRT.

Hotelling's T^2 test for one sample problem

Under $p < n$, the Hotelling's T^2 test for

$$H_0 : \boldsymbol{\mu} = \mathbf{0} \quad \text{vs} \quad H_1 : \boldsymbol{\mu} \neq \mathbf{0}$$

has its test statistics defined as

$$T^2 = n\bar{\mathbf{x}}'\mathbf{S}^{-1}\bar{\mathbf{x}}.$$

Under H_0,

$$T^2 \sim \frac{(n-1)p}{n-p} F_{p,n-p}.$$

For one sample normal mean problem, T^2 is equivalent to the LRT.

Hotelling's T^2 test: two sample problem

Given two independent random samples,

- $\{\mathbf{x}_{11}, \ldots, \mathbf{x}_{1n_1}\}$ from $N_p(\boldsymbol{\mu}_1, \Sigma)$,
- $\{\mathbf{x}_{21}, \ldots, \mathbf{x}_{2n_2}\}$ from $N_p(\boldsymbol{\mu}_2, \Sigma)$.

Sample means: $\bar{\mathbf{x}}_1$ and $\bar{\mathbf{x}}_2$,
Sample covariance matrix: $\mathbf{S}_1$ and $\mathbf{S}_2$.
Define the pooled sample covariance matrix $\mathbf{S}_p$ as

$$\begin{aligned} \mathbf{S}_p &= \frac{1}{n}\sum_{k=1}^{2}\sum_{j=1}^{n_k}(\mathbf{x}_{kj} - \bar{\mathbf{x}}_k)(\mathbf{x}_{kj} - \bar{\mathbf{x}}_k)^T \qquad (1) \\ &= \frac{1}{n}((n_1 - 1)\mathbf{S}_1 + (n_2 - 1)\mathbf{S}_2), \end{aligned}$$

where $n = n_1 + n_2 - 2$.
This is not equivalent to an LRT.

Hotelling's T^2 test: two sample problem

Under $p < n$, the Hotelling's T^2 test for

$$H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 \quad \text{vs} \quad H_1 : \boldsymbol{\mu}_1 \neq \boldsymbol{\mu}_2$$

has its test statistics defined as

$$T^2 = \left(\frac{1}{n_1} + \frac{1}{n_2}\right)^{-1} (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)' \mathbf{S}_p^{-1} (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2). \tag{2}$$

Under H_0,

$$T^2 \sim \frac{np}{n-p+1} F_{p,n-p+1}. \tag{3}$$

Hotelling's T^2 test: two sample problem

Under $p < n$, the Hotelling's T^2 test for

$$H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 \quad \text{vs} \quad H_1 : \boldsymbol{\mu}_1 \neq \boldsymbol{\mu}_2$$

has its test statistics defined as

$$T^2 = \left(\frac{1}{n_1} + \frac{1}{n_2}\right)^{-1} (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)' \mathbf{S}_p^{-1} (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2). \tag{2}$$

Under H_0,

$$T^2 \sim \frac{np}{n - p + 1} F_{p,n-p+1}. \tag{3}$$

Limitations of T^2 under HD setting

- When $p > n$, the Hotelling's T^2 test is not well-defined due to the singularity of sample covariance matrix $\mathbf{S}$ and $\mathbf{S}_p$.
- When $p < n$, the Hotelling's T^2 test has poor power when p is close to n, where the sample covariance matrix is nearly singular (Bai and Saranadasa, 1996).

Modification for one sample problem

1. Replace $\mathbf{S}^{-1}$
   - Identity matrix: I
   - Diagonal matrix: D_S, the diagonal matrix of $\mathbf{S}$
   - Ridge-like estimator: $(\mathbf{S} + \lambda I)^{-1}$
2. Projection

Methods that replace $\mathbf{S}^{-1}$

Table 1: Overview of methods that replace $\mathbf{S}^{-1}$

Method	Test Statistic	Null Distribution	Normality
Dempster (1958, 1960)	$T_D^2 = \bar{\mathbf{x}}^T\bar{\mathbf{x}}/(\frac{1}{n}\mathrm{tr}\mathbf{S})$	$T_D^2 \sim F_{r,nr}$	Yes
Bai and Saranadasa (1996)	$T_{BS} = \bar{\mathbf{x}}^T\bar{\mathbf{x}} - \frac{1}{n}\mathrm{tr}\mathbf{S}$	$T_{BS}/\sqrt{\mathrm{Var}(T_{BS})} \to N(0,1)$	No
Chen and Qin (2010)	$T_{CQ} = \sum_{i\neq j}^{n} \mathbf{x}_i^T\mathbf{x}_j$	$T_{CQ}/\sqrt{\mathrm{Var}(T_{CQ})} \to N(0,1)$	No
Srivastava and Du (2008)	$T_{SD(adj)} = \bar{\mathbf{x}}^T(\frac{1}{n}D_S)^{-1}\bar{\mathbf{x}} - \frac{(n-1)p}{n-3}$	$T_{SD(adj)}/\sqrt{c_{p,n}\mathrm{Var}(T_{SD(adj)})} \to N(0,1)$	Yes
Srivastava (2009)	$T_{SD(noadj)} = \bar{\mathbf{x}}^T(\frac{1}{n}D_S)^{-1}\bar{\mathbf{x}} - \frac{(n-1)p}{n-3}$	$T_{SD(noadj)}/\sqrt{\mathrm{Var}(T_{SD(noadj)})} \to N(0,1)$	No
Chen et al. (2011)	$RHT(\lambda) = n\bar{\mathbf{x}}^T(\mathbf{S}+\lambda I)^{-1}\bar{\mathbf{x}}$	$\frac{\sqrt{p}(RHT(\lambda)/p-\Theta_1(\lambda,c))}{(2\Theta_2(\lambda,c))^{1/2}} \to N(0,1)$	Yes

Methods that replace $\mathbf{S}^{-1}$

Table 1: Overview of methods that replace $\mathbf{S}^{-1}$

Method	Test Statistic	Null Distribution	Normality
Dempster (1958, 1960)	$T_D^2 = \bar{\mathbf{x}}^T\bar{\mathbf{x}}/(\frac{1}{n}\mathrm{tr}\mathbf{S})$	$T_D^2 \sim F_{r,nr}$	Yes
Bai and Saranadasa (1996)	$T_{BS} = \bar{\mathbf{x}}^T\bar{\mathbf{x}} - \frac{1}{n}\mathrm{tr}\mathbf{S}$	$T_{BS}/\sqrt{\mathrm{Var}(T_{BS})} \to N(0,1)$	No
Chen and Qin (2010)	$T_{CQ} = \sum_{i\neq j}^n \mathbf{x}_i^T\mathbf{x}_j$	$T_{CQ}/\sqrt{\mathrm{Var}(T_{CQ})} \to N(0,1)$	No
Srivastava and Du (2008)	$T_{SD(adj)} = \bar{\mathbf{x}}^T(\frac{1}{n}D_S)^{-1}\bar{\mathbf{x}} - \frac{(n-1)p}{n-3}$	$T_{SD(adj)}/\sqrt{c_{p,n}\mathrm{Var}(T_{SD(adj)})} \to N(0,1)$	Yes
Srivastava (2009)	$T_{SD(noadj)} = \bar{\mathbf{x}}^T(\frac{1}{n}D_S)^{-1}\bar{\mathbf{x}} - \frac{(n-1)p}{n-3}$	$T_{SD(noadj)}/\sqrt{\mathrm{Var}(T_{SD(noadj)})} \to N(0,1)$	No
Chen et al. (2011)	$RHT(\lambda) = n\bar{\mathbf{x}}^T(\mathbf{S}+\lambda I)^{-1}\bar{\mathbf{x}}$	$\frac{\sqrt{p}(RHT(\lambda)/p-\Theta_1(\lambda,c))}{(2\Theta_2(\lambda,c))^{1/2}} \to N(0,1)$	Yes

$$\widehat{\mathrm{tr}(\Sigma^2)}_{\mathrm{BS}} = \frac{(n-1)^2}{n(n+3)}\left[\mathrm{tr}\mathbf{S}^2 - (\mathrm{tr}\mathbf{S})^2/(n-1)\right] \quad (4)$$

$$\widehat{\mathrm{tr}(\Sigma^2)}_{\mathrm{CQ}} = \frac{1}{(n-1)n}\mathrm{tr}\left(\sum_{j\neq k}^n(\mathbf{x}_j - \bar{\mathbf{x}}_{(j,k)})\mathbf{x}_j^T(\mathbf{x}_k - \bar{\mathbf{x}}_{(j,k)})\mathbf{x}_k^T\right) \quad (5)$$

Methods that replace $\mathbf{S}^{-1}$

Table 1: Overview of methods that replace $\mathbf{S}^{-1}$

Method	Test Statistic	Null Distribution	Normality
Dempster (1958, 1960)	$T_D^2 = \bar{\mathbf{x}}^T \bar{\mathbf{x}} / (\frac{1}{n}\mathrm{tr}\mathbf{S})$	$T_D^2 \sim F_{r,nr}$	Yes
Bai and Saranadasa (1996)	$T_{BS} = \bar{\mathbf{x}}^T \bar{\mathbf{x}} - \frac{1}{n}\mathrm{tr}\mathbf{S}$	$T_{BS}/\sqrt{\mathrm{Var}(T_{BS})} \to N(0,1)$	No
Chen and Qin (2010)	$T_{CQ} = \sum_{i \neq j}^{n} \mathbf{x}_i^T \mathbf{x}_j$	$T_{CQ}/\sqrt{\mathrm{Var}(T_{CQ})} \to N(0,1)$	No
Srivastava and Du (2008)	$T_{SD(adj)} = \bar{\mathbf{x}}^T (\frac{1}{n} D_S)^{-1} \bar{\mathbf{x}} - \frac{(n-1)p}{n-3}$	$T_{SD(adj)}/\sqrt{c_{p,n}\mathrm{Var}(T_{SD(adj)})} \to N(0,1)$	Yes
Srivastava (2009)	$T_{SD(noadj)} = \bar{\mathbf{x}}^T (\frac{1}{n} D_S)^{-1} \bar{\mathbf{x}} - \frac{(n-1)p}{n-3}$	$T_{SD(noadj)}/\sqrt{\mathrm{Var}(T_{SD(noadj)})} \to N(0,1)$	No
Chen et al. (2011)	$RHT(\lambda) = n\bar{\mathbf{x}}^T (\mathbf{S} + \lambda I)^{-1} \bar{\mathbf{x}}$	$\frac{\sqrt{p}(RHT(\lambda)/p - \Theta_1(\lambda,c))}{(2\Theta_2(\lambda,c))^{1/2}} \to N(0,1)$	Yes

Methods that replace $\mathbf{S}^{-1}$

Table 1: Overview of methods that replace $\mathbf{S}^{-1}$

Method	Test Statistic	Null Distribution	Normality
Dempster (1958, 1960)	$T_D^2 = \bar{\boldsymbol{x}}'\bar{\boldsymbol{x}}/(\frac{1}{n}\mathrm{tr}\mathbf{S})$	$T_D^2 \sim F_{r,nr}$	Yes
Bai and Saranadasa (1996)	$T_{BS} = \bar{\boldsymbol{x}}'\bar{\boldsymbol{x}} - \frac{1}{n}\mathrm{tr}S$	$T_{BS}/\sqrt{\mathrm{Var}(T_{BS})} \to N(0,1)$	No
Chen and Qin (2010)	$T_{CQ} = \sum_{i \neq j}^{n} \boldsymbol{x}_i'\boldsymbol{x}_j$	$T_{CQ}/\sqrt{\mathrm{Var}(T_{CQ})} \to N(0,1)$	No
Srivastava and Du (2008)	$T_{SD(adj)} = \bar{\boldsymbol{x}}'(\frac{1}{n}D_S)^{-1}\bar{\boldsymbol{x}} - \frac{(n-1)p}{n-3}$	$T_{SD(adj)}/\sqrt{c_{p,n}\mathrm{Var}(T_{SD(adj)})} \to N(0,1)$	Yes
Srivastava (2009)	$T_{SD(noadj)} = \bar{\boldsymbol{x}}'(\frac{1}{n}D_S)^{-1}\bar{\boldsymbol{x}} - \frac{(n-1)p}{n-3}$	$T_{SD(noadj)}/\sqrt{\mathrm{Var}(T_{SD(noadj)})} \to N(0,1)$	No
Chen et al. (2011)	$RHT(\lambda) = n\bar{\boldsymbol{x}}'(S+\lambda I)^{-1}\bar{\boldsymbol{x}}$	$\frac{\sqrt{p}(RHT(\lambda)/p - \Theta_1(\lambda,c))}{(2\Theta_2(\lambda,c))^{1/2}} \to N(0,1)$	Yes

Extensions of multinormal distribution

1. **Independent component model**:

$$\mathbf{x} = \boldsymbol{\mu} + \Gamma \mathbf{z},$$

where Γ is a $p \times m$ matrix for some $m \geq p$ such that $\Gamma\Gamma^T = \Sigma$, and $\mathbf{z} = (z_1, \cdots, z_m)^T$ consisting of m-variate independent and identically distributed random variables satisfying $E(z_j) = 0$, $\text{var}(z_j) = 1$, and $E(z_j^4) = \kappa < \infty$.

Test-specific:

- BS: $\boldsymbol{z}$ satisfies that $E(\prod_{k=1}^m z_k^{v_k})$ equals 0 when there is at least one $v_k = 1$ and equals 1 when there are two v_k's equal to 2, whenever $\sum_{k=1}^m v_k = 4$.
- CQ: $\boldsymbol{z}$ satisfies that $E(\prod_{k=1}^q z_{l_k}^{\alpha_k}) = \prod_{k=1}^q E(z_{l_k}^{\alpha_k})$ for a positive integer q such that $\sum_{l=1}^q \alpha_l \leq 8$ and $l_1 \neq \ldots \neq l_q$.

Extensions of multinormal distribution

2. **Elliptical distribution**: $\mathbf{x}$ has characteristic function

$$E \exp(i\mathbf{t}^T\mathbf{x}) = \exp(i\boldsymbol{\mu}^T\mathbf{t})\phi(\mathbf{t}^T\Sigma\mathbf{t})$$

If $\mathbf{x}$ has a density, its density is of the form

$$f(\mathbf{x}) = |\Sigma|^{-1/2}g\{(\mathbf{x}-\boldsymbol{\mu})^T\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})\},$$

and $E(\mathbf{x}) = \boldsymbol{\mu}$ and $\text{cov}(\mathbf{x}) \propto \Sigma$ if they exist. Thus, Σ is called the scatter matrix of $\mathbf{x}$.
Stochastic Representation:

$$\mathbf{x} = \boldsymbol{\mu} + R\Sigma^{1/2}\mathbf{u},$$

where R is a nonnegative random variable and is independent of $\mathbf{u}$, which follows a uniform distribution over the unit sphere $\{\mathbf{u} : \mathbf{u}^T\mathbf{u} = 1\}$.

Nonparametric/Robust test for HD data

Assume $\mathbf{x}_i$ follows an elliptical distribution:

$$\mathbf{x}_i = \boldsymbol{\mu} + \epsilon_i, \quad \text{and} \quad \epsilon_i = \Gamma R_i \mathbf{u}_i,$$

where Γ is a $p \times p$ matrix, $\mathbf{u}_i$ is a random vector uniformly distributed on the unit sphere in $\mathbb{R}^p$, and R_i is a nonnegative random variable independent of $\mathbf{u}_i$.

The distribution of $\mathbf{x}_i$ depends on Γ only through $\Gamma\Gamma^T \widehat{=} \Omega$.

In general, $\mathbf{x}_i$'s covariance matrix Σ is related to Ω by $\Sigma = p^{-1}E(R_i^2)\Omega$.

An important special case of elliptical distribution is $N(\boldsymbol{\mu}, \Sigma)$, for which $R_i^2 \sim \chi_p^2$ and $\Omega = \Sigma$.

Chen and Qin (2010)'s test

For the independent component models, the test statistic of Bai and Saranadasa (1996) is based on $\|\bar{\mathbf{x}}\|^2$.

The test statistic of Chen and Qin (2010) is based on $\sum_{i=1}^n \sum_{j=1, j\neq i}^n \mathbf{x}_i^T \mathbf{x}_j$.

By removing the diagonal elements in $\|\bar{\mathbf{x}}\|^2$, Chen and Qin (2010) was able to considerably relax the conditions on p and n.

New nonparametric test statistics

The **spatial sign function** of $\mathbf{x}_i$ is defined as

$$\mathbf{z}_i = \begin{cases} \frac{\mathbf{x}_i}{\|\mathbf{x}_i\|}, & \mathbf{x}_i \neq 0, \\ 0, & \mathbf{x}_i = 0, \end{cases}$$

where $\|\mathbf{x}_i\|^2 = \mathbf{x}_i^T \mathbf{x}_i$. The spatial sign vector is simply the unit vector in the direction of $\mathbf{x}_i$. In the univariate case, it reduces to the familiar sign function. Define

$$T_n = \sum_{i=1}^{n} \sum_{j=1}^{i-1} \mathbf{z}_i^T \mathbf{z}_j,$$

which indeed is a U-statistic.
Under H_0, $E(\mathbf{z}_i) = 0$ which implies $E(T_n) = 0$.

The new test statistic dismiss the diagonal elements in defining T_n, and can be deemed as a nonparametric extension of Chen and Qin (2010).

Connection to multivariate spatial sign test

The new test generalizes the multivariate spatial sign test (e.g., Brown, 1983; Chaudhuri, 1992; Möttönen and Oja, 1995) to the high-dimensional setting.

For fixed p, Möttönen, et al. (1997) derived the asymptotic relative efficiency (ARE) of the spatial sign test versus Hotelling's T^2 test and established its theoretical advantage for heavy-tailed distributions:

For example, when $\mathbf{x} \sim$ 10-dim t_ν, the ARE of the spatial sign test versus T^2 test is **2.42** when $\nu = 3$, and is **0.95** when $\nu = \infty$ (multivariate normality).

However, similarly as Hotelling's T^2 test, the multivariate spatial sign test is not defined when $p > n$.

Here we modify the sign test in a way such that its efficiency advantage can be preserved in the high-dimensional setting.

Theoretical properties

Conditions:

(C1) $\mathrm{Tr}(\Sigma^4) = o\{\mathrm{Tr}^2(\Sigma^2)\}$.

(C2) $\dfrac{\mathrm{Tr}^4(\Sigma)}{\mathrm{Tr}^2(\Sigma^2)} \exp\left\{ -\dfrac{\mathrm{Tr}^2(\Sigma)}{128p\lambda^2_{\max}(\Sigma)} \right\} = o(1)$.

Theorem

Assume Conditions (C1) and(C2) hold. Then under H_0,

$$\frac{T_n}{\sqrt{\frac{n(n-1)}{2} Tr(B^2)}} \to N(0,1)$$

in distribution, as $n, p \to \infty$, where $B = E\left(\frac{\epsilon_i \epsilon_i^T}{\|\epsilon_i\|^2}\right)$.

Comparison on dimensionalilty

Srivastava and Du (2008): $n = O(p^\delta)$ for some $1/2 < \delta \leq 1$.

Srivastava (2009): $n = O(p^\delta)$ for some $0 < \delta \leq 1$,

Lee, Limb, Li, Vannuccid and Petkova (2012): $p/n \to c > 0$

Srivastava, Katayama and Kano (2013): $n = O(p^\delta)$, $\delta > 1/2$,

Chen and Qin (2010): $\mathrm{Trace}(\Sigma^4) = \mathrm{o}(\mathrm{Trace}^2(\Sigma^2))$.

To apply T_n in practical data analysis, we need an estimator of $\mathrm{Tr}(B^2)$.

Following Chen and Qin (2010), we may estimate $\mathrm{Tr}(B^2)$ using the cross-validation approach as follows:

$$\widehat{\mathrm{Tr}(B^2)} = \{n(n-1)\}^{-1}\mathrm{Tr}\Big\{\sum_{1\le j\neq k\le n}(\mathbf{z}_j - \bar{\mathbf{z}}_{(j,k)})\mathbf{z}_j^T(\mathbf{z}_k - \bar{\mathbf{z}}_{(j,k)})\mathbf{z}_k^T\Big\},$$

where $\bar{\mathbf{z}}_{(j,k)}$ is the sample mean after excluding $\mathbf{z}_j$ and $\mathbf{z}_k$.

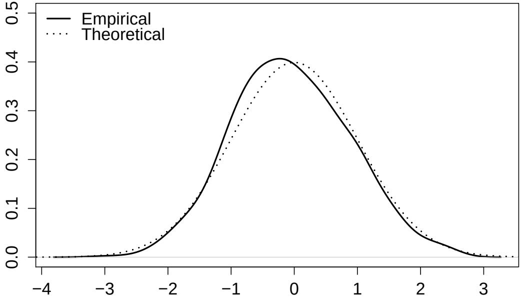


Figure: Density curves under the null hypothesis and $\mathbf{x} \sim N(0, (0.8^{|i-j|}))$. $(n = 50,\ p = 1000)$

Local power analysis

For the local power analysis, we impose the following additional conditions.

(C3) $\exp\left(-\frac{\mathrm{Tr}^2(\Sigma)}{256p\lambda_{\max}^2(\Sigma)}\right) = o\Big(\min\Big(\frac{\lambda_{\max}(\Sigma)}{\mathrm{Tr}(\Sigma)}, \frac{\lambda_{\min}(\Sigma)}{\lambda_{\max}(\Sigma)}\Big)\Big)$.

(C4) $\lambda_{\max}(\Sigma) = o(\mathrm{Tr}(\Sigma))$.

(C5) $\|\boldsymbol{\mu}\|^2 E(\|\epsilon\|^{-2}) = o\Big(\min\Big(n^{-1}\frac{\mathrm{Tr}(\Sigma^2)}{\lambda_{\max}(\Sigma)\mathrm{Tr}(\Sigma)}, n^{-1/2}\frac{\mathrm{Tr}^{1/2}(\Sigma^2)}{\mathrm{Tr}(\Sigma)}\Big)\Big)$.

(C6) For some $0 < \delta < 1$, $\|\boldsymbol{\mu}\|^{2\delta} E(\|\epsilon\|^{-2-2\delta}) = o\big(E^2(\|\epsilon\|^{-1})\big)$.

Theorem

Assume conditions (C1)-(C6) hold. Letting $A = E\Big\{\frac{1}{\|\epsilon_i\|}\big(I_p - \frac{\epsilon_i\epsilon_i^T}{\|\epsilon_i\|^2}\big)\Big\}$*. Then as* $n, p \to \infty$*,*

$$\frac{T_n - \frac{n(n-1)}{2}\boldsymbol{\mu}^T A^2 \boldsymbol{\mu}(1+o(1))}{\sqrt{\frac{n(n-1)}{2} Tr(B^2)}} \to N(0,1)$$

in distribution.

This theorem implies that under the local alternatives, the proposed test at level α has the local power

$$\beta_n = \Phi\Big(- z_\alpha + \sqrt{\frac{n(n-1)}{2}} \frac{\boldsymbol{\mu}^T A^2 \boldsymbol{\mu}(1+o(1))}{\sqrt{\mathrm{Tr}(B^2)}}\Big),$$

where $\Phi(\cdot)$ and z_α denote the cumulative distribution function and the upper α quantile of the $N(0,1)$ distribution, respectively.

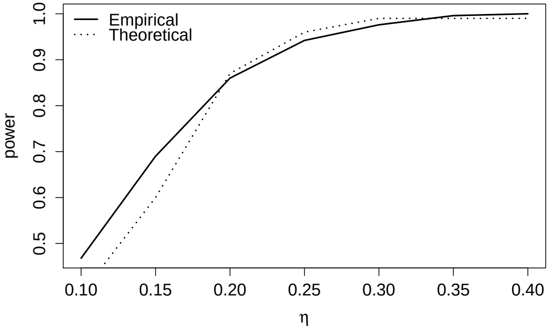


Figure: Comparing the empirical distribution of the new test with the theoretical distribution ($n = 50$, $p = 1000$) when $X \sim t_3(\Omega)$ with $\Omega = (0.8^{|i-j|})$. Here $\eta = \boldsymbol{\mu}^T A^2 \boldsymbol{\mu} / \sqrt{\mathrm{Tr}(B^2)}$.

Asymptotic Relative Efficiency (ARE)

For multivariate t-distribution $t_\nu(\Omega)$ with $\Omega = I_p$, when p is large, the ARE of T_n versus the Chen and Qin's test is approximately

$$\mathrm{ARE}_{T_n,\mathrm{CQ}} \approx \frac{2}{\nu - 2}\Big(\frac{\Gamma((\nu+1)/2)}{\Gamma(\nu/2)}\Big)^2.$$

ARE$\approx$ 2.54, 1.76, 1.51, 1.38 for $\nu = 3, 4, 5, 6$, respectively, and ARE$\approx$ 1 for $\nu = \infty$ (i.e., $N(0, I_p)$).

Recall that for $\nu = 3$, the ARE of the classical spatial sign test versus Hotelling's T^2 is 2.02 for $p = 3$ and 2.09 for $p = 10$ in the fixed dimensional case.

This suggests that nonparametric test may have more substantial power gain in the high-dimensional case.

Beyond Elliptical Distribution

Symmetric independent component models (e.g., Ilmonen and Paindaveine, 2011):

$$\mathbf{x}_i = \mu + \Gamma \mathbf{z}_i,$$

where Γ is a full rank $p \times p$ positive definite matrix; $\mathbf{z}_i = (Z_{i1}, \ldots, Z_{ip})'$ has independent components Z_{ij} and Z_{ij} is symmetric about zero. Assume that Z_{ij} are standardized such that $Var(Z_{ij}) = 1$.

Theorem

Assume (C1) and (C2) hold for symmetric independent component models, $E(Z_{ij}^4)$ is finite, and $Tr(\Sigma^2) = o(Tr^2(\Sigma))$. Then $\frac{T_n}{\sqrt{\frac{n(n-1)}{2} Tr(B^2)}} \to N(0,1)$ in distribution under H_0, as $n, p \to \infty$.

Simulation Studies

Goal: to compare the performance of the new test with four alternatives:
(1) the test of Chen and Qin (CQ test, 2010),

(2) the test of Srivastava, Katayama and Kano (SKK test, 2013),

(3) the test based on multiple comparison with Bonferroni correction (BF test),

(4) the test based on multiple comparison with FDR control (FDR test, Benjamini and Hochberg, 1995).

The BF test controls Type I error rate at 0.05 and the FDR test controls the false discovery rate at 0.05. Both the BF test and FDR test are computed using the p-values from the t tests for the marginal hypotheses and reject H_0 if at least one marginal test is significant.

Simulation Studies

Sample size: $n = 20, 50$

Dimension: $p = 200, 1000$ and 2000. To save space, we report the results for $p = 1000$ and 2000 here.

The performance of the five tests are evaluated on 1000 simulation runs.

Example 1. $\mathbf{x} \sim N_p(\boldsymbol{\mu}, \Sigma)$

The three choices for $\boldsymbol{\mu}$ are:
(1) the null hypothesis $\boldsymbol{\mu}_0 = (0, \ldots, 0)^T$;
(2) the alternative $\boldsymbol{\mu}_1 = (0.25, 0.25, \ldots, 0.25)^T$;
(3) the alternative $\boldsymbol{\mu}_2 = (\boldsymbol{\mu}_{21}, \ldots, \boldsymbol{\mu}_{2p})^T$ with $\boldsymbol{\mu}_{21} = \ldots = \boldsymbol{\mu}_{2\frac{p}{3}} = 0$, $\boldsymbol{\mu}_{2(\frac{p}{3}+1)} = \ldots = \boldsymbol{\mu}_{2(\frac{2p}{3})} = 0.2$ and $\boldsymbol{\mu}_{2(\frac{2p}{3}+1)} = \ldots = \boldsymbol{\mu}_{2p} = -0.2$.

The three choices for Σ are:
(1) $\sigma_{ii} = 1$ and $\sigma_{ij} = 0.2$ $(i \neq j)$; (denoted by Σ_1 in tables)
(2) $\sigma_{ij} = 0.8^{|i-j|}$; (denoted by Σ_2)
(3) $\Sigma = DRD$, where $D = \text{diag}(d_1, \ldots, d_p)$ with $d_i = 2 + (p - i + 1)/p$, $R = (r_{ij})$ with $r_{ii} = 1$ and $r_{ij} = (-1)^{i+j}(0.2)^{|i-j|^{0.1}}$ for $i \neq j$. Denoted by Σ_3. This was considered in Srivastava, Katayama and Kano (2013).

Table 1a: Example 1 - multivariate normal distribution

Σ	μ	n	p	New	CQ	SKK	BF	FDR
Σ_1	μ_0	20	1000	0.066	0.069	0.061	0.046	0.046
		20	2000	0.073	0.070	0.046	0.052	0.053
		50	1000	0.059	0.060	0.043	0.035	0.038
		50	2000	0.058	0.061	0.029	0.043	0.047
	μ_1	20	1000	0.723	0.723	0.692	0.405	0.471
		20	2000	0.720	0.729	0.638	0.385	0.447
		50	1000	0.975	0.976	0.962	0.842	0.890
		50	2000	0.970	0.976	0.945	0.850	0.901
	μ_2	20	1000	0.951	0.826	0.650	0.382	0.443
		20	2000	0.954	0.821	0.567	0.404	0.464
		50	1000	1.000	1.000	1.000	0.964	0.997
		50	2000	1.000	1.000	1.000	0.973	0.998

Table 1b: Example 1 - multivariate normal distribution

Σ	μ	n	p	New	CQ	SKK	BF	FDR
Σ_2	μ_0	20	1000	0.055	0.055	0.067	0.052	0.052
		20	2000	0.059	0.061	0.048	0.042	0.044
		50	1000	0.052	0.052	0.061	0.042	0.044
		50	2000	0.045	0.045	0.052	0.048	0.048
	μ_1	20	1000	0.490	0.438	0.514	0.127	0.132
		20	2000	0.646	0.594	0.556	0.113	0.119
		50	1000	1.000	0.998	1.000	0.296	0.331
		50	2000	1.000	1.000	1.000	0.331	0.369
	μ_2	20	1000	0.242	0.225	0.335	0.100	0.104
		20	2000	0.342	0.310	0.318	0.084	0.092
		50	1000	0.932	0.862	0.982	0.239	0.260
		50	2000	0.991	0.987	1.000	0.269	0.296

Table 1c: Example 1 - multivariate normal distribution

Σ	μ	n	p	New	CQ	SKK	BF	FDR
Σ_3	μ_0	20	1000	0.052	0.051	0.047	0.038	0.041
		20	2000	0.058	0.059	0.023	0.047	0.047
		50	1000	0.048	0.050	0.051	0.043	0.048
		50	2000	0.060	0.060	0.052	0.054	0.061
	μ_1	20	1000	0.795	0.797	0.815	0.122	0.138
		20	2000	0.969	0.968	0.930	0.134	0.145
		50	1000	0.999	0.999	0.999	0.357	0.430
		50	2000	1.000	1.000	1.000	0.416	0.479
	μ_2	20	1000	0.540	0.549	0.579	0.092	0.102
		20	2000	0.790	0.788	0.695	0.102	0.112
		50	1000	0.992	0.991	0.994	0.265	0.289
		50	2000	1.000	1.000	1.000	0.343	0.381

Example 2: $\mathbf{x} \sim t_3(\Sigma)$

We simulate $\mathbf{x}_i$ from a p-variate t distribution with mean vector $\boldsymbol{\mu}$, covariance matrix Σ and 3 degrees of freedom.

The choices of $\boldsymbol{\mu}$ and Σ are set to be the same as those in Example 1. The distribution is heavy-tailed in this example.

Table 2a: Example 2 - multivariate t-distribution

Σ	μ	n	p	New	CQ	SKK	BF	FDR
Σ_1	μ_0	20	1000	0.083	0.088	0.012	0.011	0.013
		20	2000	0.064	0.072	0.007	0.010	0.010
		50	1000	0.053	0.063	0.015	0.011	0.012
		50	2000	0.069	0.076	0.008	0.010	0.010
	μ_1	20	1000	0.633	0.472	0.222	0.153	0.183
		20	2000	0.631	0.468	0.171	0.117	0.138
		50	1000	0.941	0.721	0.493	0.424	0.491
		50	2000	0.921	0.736	0.448	0.438	0.492
	μ_2	20	1000	0.815	0.371	0.076	0.129	0.150
		20	2000	0.830	0.363	0.040	0.107	0.122
		50	1000	1.000	0.803	0.333	0.427	0.485
		50	2000	1.000	0.825	0.288	0.493	0.571

Table 2b: Example 2 - multivariate t-distribution

Σ	μ	n	p	New	CQ	SKK	BF	FDR
Σ_2	μ_0	20	1000	0.054	0.058	0.001	0.023	0.023
		20	2000	0.061	0.057	0.000	0.011	0.011
		50	1000	0.056	0.057	0.002	0.015	0.015
		50	2000	0.048	0.042	0.001	0.015	0.015
	μ_1	20	1000	0.355	0.174	0.001	0.033	0.034
		20	2000	0.488	0.224	0.003	0.022	0.022
		50	1000	0.979	0.465	0.021	0.080	0.084
		50	2000	0.998	0.624	0.015	0.095	0.103
	μ_2	20	1000	0.198	0.113	0.001	0.034	0.034
		20	2000	0.251	0.141	0.000	0.018	0.019
		50	1000	0.766	0.249	0.010	0.059	0.064
		50	2000	0.922	0.349	0.005	0.070	0.073

Table 2c: Example 2 - multivariate t-distribution

Σ	μ	n	p	New	CQ	SKK	BF	FDR
Σ_3	μ_0	20	1000	0.052	0.053	0.000	0.011	0.012
		20	2000	0.058	0.060	0.000	0.013	0.013
		50	1000	0.052	0.053	0.000	0.015	0.015
		50	2000	0.059	0.060	0.000	0.020	0.021
	μ_1	20	1000	0.682	0.349	0.013	0.029	0.033
		20	2000	0.883	0.512	0.002	0.038	0.040
		50	1000	0.996	0.780	0.120	0.094	0.102
		50	2000	1.000	0.933	0.091	0.118	0.128
	μ_2	20	1000	0.441	0.228	0.001	0.027	0.029
		20	2000	0.654	0.339	0.000	0.027	0.028
		50	1000	0.942	0.570	0.037	0.067	0.071
		50	2000	0.998	0.785	0.018	0.077	0.086

Example 3: scale mixture of two normals

We simulate $\mathbf{x}_i$ from a scale mixture of two multivariate normal distributions $0.9 * N_p(\boldsymbol{\mu}, \Sigma) + 0.1 * N_p(\boldsymbol{\mu}, 9\Sigma)$.

We consider the same choices of $\boldsymbol{\mu}$ and Σ as in Example 1.

Table 3a: Example 3 - mixture of multivariate normal distributions

Σ	$\boldsymbol{\mu}$	n	p	New	CQ	SKK	BF	FDR
Σ_1	$\boldsymbol{\mu}_0$	20	1000	0.063	0.070	0.007	0.014	0.015
		20	2000	0.045	0.049	0.007	0.015	0.018
		50	1000	0.063	0.066	0.014	0.020	0.021
		50	2000	0.042	0.040	0.007	0.015	0.016
	$\boldsymbol{\mu}_1$	20	1000	0.649	0.548	0.277	0.209	0.259
		20	2000	0.627	0.542	0.229	0.193	0.231
		50	1000	0.941	0.859	0.730	0.600	0.682
		50	2000	0.964	0.867	0.701	0.619	0.687
	$\boldsymbol{\mu}_2$	20	1000	0.870	0.449	0.109	0.175	0.201
		20	2000	0.882	0.492	0.089	0.224	0.243
		50	1000	1.000	0.966	0.700	0.663	0.759
		50	2000	1.000	0.968	0.577	0.698	0.800

Table 3b: Example 3 - mixture of multivariate normal distributions

Σ	μ	n	p	New	CQ	SKK	BF	FDR
Σ_2	μ_0	20	1000	0.054	0.053	0.008	0.020	0.020
		20	2000	0.039	0.045	0.002	0.022	0.022
		50	1000	0.050	0.050	0.000	0.026	0.026
		50	2000	0.035	0.034	0.000	0.021	0.022
	μ_1	20	1000	0.342	0.207	0.045	0.058	0.061
		20	2000	0.493	0.307	0.072	0.046	0.049
		50	1000	0.995	0.730	0.050	0.124	0.129
		50	2000	1.000	0.879	0.035	0.111	0.119
	μ_2	20	1000	0.178	0.130	0.030	0.046	0.050
		20	2000	0.249	0.173	0.043	0.047	0.048
		50	1000	0.797	0.421	0.012	0.100	0.110
		50	2000	0.947	0.565	0.011	0.092	0.093

Table 3c: Example 3 - mixture of multivariate normal distributions

Σ	$\boldsymbol{\mu}$	n	p	New	CQ	SKK	BF	FDR
Σ_3	$\boldsymbol{\mu}_0$	20	1000	0.046	0.063	0.006	0.019	0.019
		20	2000	0.050	0.047	0.004	0.017	0.018
		50	1000	0.039	0.049	0.000	0.021	0.023
		50	2000	0.041	0.039	0.001	0.020	0.021
	$\boldsymbol{\mu}_1$	20	1000	0.678	0.485	0.093	0.059	0.067
		20	2000	0.914	0.700	0.126	0.069	0.070
		50	1000	0.998	0.943	0.230	0.157	0.177
		50	2000	1.000	0.995	0.167	0.190	0.211
	$\boldsymbol{\mu}_2$	20	1000	0.437	0.285	0.059	0.050	0.053
		20	2000	0.687	0.478	0.088	0.047	0.050
		50	1000	0.953	0.759	0.045	0.111	0.118
		50	2000	0.998	0.940	0.040	0.150	0.169

An application

Type 2 diabetes is one of the most common chronic diseases.

Insulin resistance in skeletal muscle is a prominent feature of Type 2 diabetes.

To study insulins ability to regulate gene expression, an experiment performed microarray analysis using the Affymetrix Hu95A chip of human skeletal muscle biopsies from 15 diabetic patients both before and after insulin treatment (Wu et al., 2007).

The gene expression alterations are promising to provide insights on new therapeutic targets for the treatment of this common disease.

Hence, we are interested in testing the hypothesis whether there is a change of the gene expression level due to the treatment.

The underlying genetics of Type 2 diabetes were recognized to be very complex.

It is believed that Type 2 diabetes is resulted from interactions between many genetic factors and the environment. The data were normalized by quantile normalization.

When multiple probes are associated with the same gene, their expression values are consolidated by taking the average.

In our analysis, we considered 2519 curated gene sets. The gene sets we used are from the C2 collection of the GSEA online pathway databases (http://www.broadinstitute.org/gsea/msigdb/collection_details.jsp#C2). The largest gene set contains 1607 genes, which makes the hypothesis testing problem a high-dimensional one.

We applied both the new test and the CQ test at 5% significance level with the Bonferroni correction to control the family-wise error rate at 0.05 level.

For the CQ method, 520 gene sets (20.64% of all candidates) are identified as significant

For the new method, 954 gene sets (37.87% of all candidates) are selected as significant.

We observe that the significant gene sets selected by the new test include those identified by the CQ test with only one exception (HASLINGER_B_CLL_WITH_CHROMOSOME_12_TRISOMY).

Table: The top 10 significant gene sets selected by the new test and the CQ test

Gene set	New test	CQ test
ZWANG_CLASS_2_TRANSIENTLY_INDUCED_BY_EGF	24.34	20.11
NAGASHIMA_EGF_SIGNALING_UP	22.44	17.01
SHIPP_DLBCL_CURED_VS_FATAL_DN	22.34	18.24
WILLERT_WNT_SIGNALING	19.66	
UZONYI_RESPONSE_TO_LEUKOTRIENE_AND_THROMBIN	19.63	18.46
PID_HIF2PATHWAY	19.46	15.65
PHONG_TNF_TARGETS_UP	19.21	18.90
AMIT_EGF_RESPONSE_60_HELA	18.64	16.38
MCCLUNG_CREB1_TARGETS_DN	18.43	
SEMENZA_HIF1_TARGETS	18.34	
AMIT_SERUM_RESPONSE_40_MCF10A		15.98
AMIT_SERUM_RESPONSE_60_MCF10A		15.43
PLASARI_TGFB1_TARGETS_1HR_UP		15.00

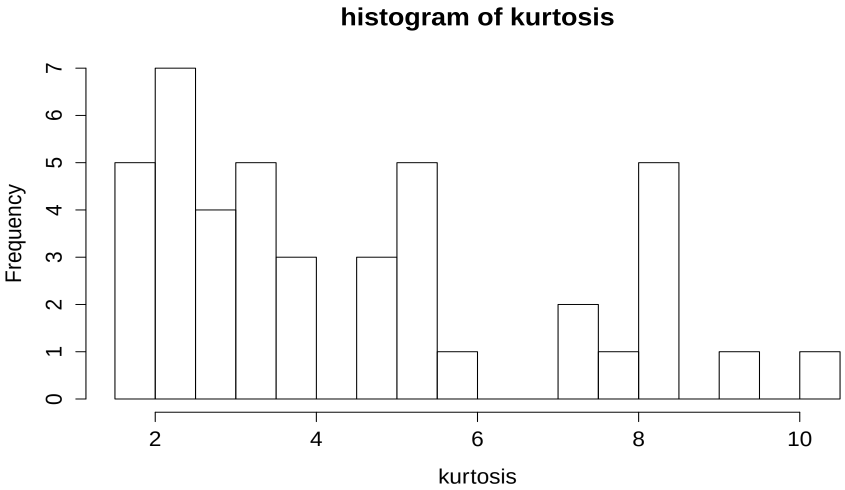


Figure: The histogram of marginal kurtosises for all genes in MCCLUNG_CREB1_TARGETS_DN gene set.

5.2 Test of covariance structures

Test of covariance structure is of great interest in classical multivariate data analysis:

◦ Chapters 8 and 11 of Muirhead (1982)

◦ Chapters 9 and 10 of Anderson (2003)

Earlier work on this topic

Ledoit and Wolf (2002, AoS) studied two tests on covariance

$$H_{10} : \Sigma = \sigma^2 I \quad \text{versus} \quad H_{11} : \Sigma \neq \sigma^2 I$$

and

$$H_{10a} : \Sigma = I \quad \text{versus} \quad H_{11a} : \Sigma \neq I$$

when we have an IID sample from $N_p(\boldsymbol{\mu}, \Sigma)$, where I is the $p \times p$ identity matrix.

The idea can be directly applicable for $H_{10} : \Sigma = \sigma^2 \Sigma_0$ with known Σ_0.

$\bar{\mathbf{x}}$: the sample mean, $\mathbf{S}$: the sample covariance matrix

The LRTs for H_{10} and H_{10a} are

$$U = \frac{1}{p}\mathrm{tr}\left[\left\{\frac{\mathbf{S}}{p^{-1}\mathrm{tr}(\mathbf{S})} - I\right\}^2\right] \quad \text{and} \quad V = \frac{1}{p}\mathrm{tr}\{(\mathbf{S} - I)^2\}$$

A quick take-home message

Dimensionality: $p/n \to y \in (0, \infty)$. p and n have the same order.

U is robust against high dimensionality, but V is not.

Denote by $\lambda_1, \cdots, \lambda_p$ the eigenvalues of Σ.

Let $\bar{\lambda} = p^{-1} \sum_{j=1}^{p} \lambda_j = p^{-1}\text{tr}(\Sigma)$ and $s_\lambda^2 = p^{-1} \sum_{j=1}^{p} (\lambda_j - \bar{\lambda})^2$.

This paper shows that under some conditions (the sample 3rd and 4th moments of λ_js have finite limites

$$p^{-1}\text{tr}(S) \to \bar{\lambda} \quad \text{and} \quad p^{-1}\text{tr}(S^2) \to (1+y)\bar{\lambda}^2 + s_\lambda^2.$$

in probability.

Impact of dimensionality?

$s_\lambda^2 = 0$ if $\Sigma = I$.

$$U = \frac{p^{-1}\mathrm{tr}(\mathbf{S}^2)}{[p^{-1}\mathrm{tr}(S)]^2} - 1 \to y + \frac{s_\lambda^2}{\bar{\lambda}^2}$$

Corrected LRT is to take y into consideration by approximating y by p/n. This constitutes an (n, p)-consistent test.

The authors proposed to modify V by

$$W = \frac{1}{p}\mathrm{tr}\{(\mathbf{S} - I)^2\} - \frac{p}{n}\{p^{-1}\mathrm{tr}(\mathbf{S})\}^2 + p/n$$

Adding p/n makes W have a limiting distribution of $\chi^2_{p(p+1)/2}$.

Srivastava (2005, J. Japan Statist. Soc)

This paper assumes that (1) $n = O(p^\delta)$ with $0 < \delta \leq 1$, and (2) normality Let $a_j = p^{-1}\mathrm{tr}(\Sigma^j)$. The author proved that

$$\widehat{a}_1 = p^{-1}\mathrm{tr}\mathbf{S}$$

and

$$\widehat{a}_2 = \frac{n^2}{(n-1)(n+2)}\frac{1}{p}\{\mathrm{tr}\mathbf{S}^2 - \frac{1}{n}(\mathrm{tr}\mathbf{S})^2\}$$

are unbiased and consistent estimate for a_1 and a_2.
Apply these results for testing H_{10}, H_{10a} and test of Σ being diagonal (test of independence under multinormality).

Some more works

Birke and Detter (2005, Statist & Probab. Letter) considered H_{10} and H_{10a} under normality assumption with $p/n \to y$, and $y = 0$ or ∞.

Schott (2007, CSDA) consider $H_0 : \Sigma_1 = \cdots = \Sigma_K$ under normality assumption.
LRT for this null hypothesis is

$$M = n \log |\mathbf{S}| - \sum_{k=1}^{K} n_k \log |\mathbf{S}_k|$$

where $\mathbf{S}$ is the pooled sample covariance matrix, and $\mathbf{S}_k$ is the sample covariance matrix from the k-the population.
Problem: $\mathbf{S}_k$ becomes singular if $n_k < p$.

He proposed using Wald statistic:

$$W = \frac{n}{2}\{\sum_{k=1}^{K} \frac{n_k}{n}\text{tr}(\mathbf{S}_k\mathbf{S}^{-1}\mathbf{S}_k\mathbf{S}^{-1}) - \sum_{k=1}^{K}\sum_{l=1}^{K} \frac{n_k n_l}{n^2}\text{tr}(\mathbf{S}_k\mathbf{S}^{-1}\mathbf{S}_k\mathbf{S}^{-1})\}$$

if $\sum_{k=1}^{K} n_k \geq p$.

LRT under normality assumption

Under multinormality assumption, Jiang and Yang (2013) obtained the limiting distribution of LRT tests studied in Chapter 8 of Muirhead (1982) for $p/n \to y \in (0,1)$. and Jiang and Qi (2015) further obtained the limiting distribution when $y = 1$.

Cai and Ma (2013) proposed a further corrected LRT that allows $p/n \to \infty$.

Test of covariance for independent component models

$$\mathbf{x} = \boldsymbol{\mu} + \Gamma \mathbf{w},$$

where $\mathbf{w}$ has iid component with $E(w_j) = 0$, $\text{var}(w_j) = 1$, and $E(w_j^4) = \kappa < \infty$.

Major probability tools: random matrix theory.

For $H_{10a} : \Sigma = I$:
Bai, Jiang, Yao and Zheng (2009, AoS) proposed correcting LRT when $p/n \to y \in (0, 1)$.

Wang, Yang, Miao and Cao (2013) also studied LRT

For H_{10}**, test of sphericity**:
Chen, Zhang and Zhong (2010, JASA) demonstrated the classical LRT may become invalid for HD data and proposed a test based on U-statistics.

Wang and Yao (2013) proposed the corrected LRT with $y \in (0, 1)$ and corrected John's test with $y \in (0, \infty)$.

Test of linear structures

Develop two tests on linear covariance structure for HD data.

For a set of pre-specified symmetric $p \times p$ matrices $(\mathbf{A}_1, \ldots, \mathbf{A}_K)$ with fixed and finite K,

$$H_0 : \Sigma = \theta_1 \mathbf{A}_1 + \theta_2 \mathbf{A}_2 + \ldots + \theta_K \mathbf{A}_K,$$

where $\{\theta_j, j = 1, \ldots, K\}$ are unknown parameters

Statistical Setting

- *Assumption A.* (**ICM**) Assume that the population $\mathbf{X}$ can be represented as $\mathbf{x} = \boldsymbol{\mu} + \Gamma \mathbf{w}$, where $\mathbf{w} = (w_1, \cdots, w_p)^T$, and $w_1, \cdots, w_p$ being IID and $E(w_j) = 0$, $E(w_j^2) = 1$ and $E(w_j^4) = \kappa < \infty$.
- *Assumption B.* Denote $y_n = p/n$. Assume that $y_n \to y \in (0, \infty)$.

Parameter estimation

Without a likelihood, we propose estimating $\boldsymbol{\theta}$ by minimizing the following squared loss function

$$\min_{\boldsymbol{\theta}} \operatorname{tr}(\mathbf{S}_n - \theta_1 \mathbf{A}_1 - \ldots - \theta_K \mathbf{A}_K)^2, \tag{1}$$

where $\mathbf{S}_n$ is the sample covariance matrix.
Let $\mathbf{C}$ be a $K \times K$ matrix with (i, j)-element being $\operatorname{tr}\mathbf{A}_i\mathbf{A}_j$ and $\mathbf{a}$ be a $K \times 1$ vector with j-th element being $\operatorname{tr}\mathbf{S}_n\mathbf{A}_j$. Minimizing (1) yields a least squares type estimate for $\boldsymbol{\theta}$:

$$\widehat{\boldsymbol{\theta}} = \mathbf{C}^{-1}\mathbf{a}. \tag{2}$$

We can show that under Assumptions A and B, $\widehat{\theta}_k = \theta_k + O_p(n^{-1})$, $k = 1, \cdots, K$.

Denote $\Sigma_0 = \theta_1 \mathbf{A}_1 + \theta_2 \mathbf{A}_2 + \ldots + \theta_K \mathbf{A}_K$.

$$H_0 : \Sigma = \Sigma_0.$$

Under H_0, we estimate $\boldsymbol{\theta}$ by $\widehat{\boldsymbol{\theta}}$ given in (2), and then an estimator of Σ is $\widehat{\Sigma}_0 = \widehat{\theta}_1 \mathbf{A}_1 + \ldots + \widehat{\theta}_K \mathbf{A}_K$.

Without the linear structure assumption, a natural estimator for Σ is the sample covariance matrix $\mathbf{S}_n$.

Entropy loss based test

Motivated by the entropy loss (EL) used for covariance matrix estimation (James and Stein, 1961; Muirhead, 1982), we propose our first test for H_0. For $p < n-1$,

$$T_{n1} = \text{tr}\mathbf{S}_n\widehat{\Sigma}_0^{-1} - \log(|\mathbf{S}_n\widehat{\Sigma}_0^{-1}|).$$

Denote $\{\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p\}$ to be the eigenvalues of $\widehat{\Sigma}_0^{-1/2}\mathbf{S}_n\widehat{\Sigma}_0^{-1/2}$. Then we can write T_{n1} as

$$T_{n1} = p\left(p^{-1}\sum_{j=1}^{p}\lambda_j - p^{-1}\sum_{j=1}^{p}\log\lambda_j\right).$$

This motivates us to further extend the test to the situation with $p \geq n-1$ by defining

$$T_{n1} = (n-1)\left(p^{-1}\sum_{j=1}^{n-1}\lambda_j - (n-1)^{-1}\sum_{j=1}^{n-1}\log\lambda_j\right).$$

Entropy loss based test

Define $q = \min\{p, n-1\}$. T_{n1} can be written in a unified form for $p < n-1$ and $p \geq n-1$:

$$T_{n1} = q\left(p^{-1}\sum_{j=1}^{q}\lambda_j - q^{-1}\sum_{j=1}^{q}\log\lambda_j\right). \tag{3}$$

Since this test is motivated by the entropy loss, we refer this test to be as **EL-test**.

Quadratic loss based test

Motivated by the quadratic loss (QL), another popular loss function in covariance matrix estimation (Olkin and Selliah, 1977; Haff, 1980; Muirhead, 1982), we propose the second test

$$T_{n2} = \mathrm{tr}(\mathbf{S}_n \widehat{\Sigma}_0^{-1} - \mathbf{I}_p)^2, \tag{4}$$

and refer this test to be as **QL-test**.

Example 1. Test of sphericity: $H_{10} : \Sigma = \sigma^2 I_p$

Chen, Zhang & Zhong (2010) demonstrated that the classical LRT may become invalid for HD data and proposed a test based on U-statistics with $p, n \to \infty$.

Jiang & Yang (2013) derived the asymptotic distribution of the LRT with $p/n \to (0, 1]$ under normality assumption on $\mathbf{w}$.

Wang & Yao (2013) proposed the corrected LRT with $p/n \to (0, 1)$ and the corrected John's (CJ) test with $p/n \to y \in (0, \infty)$

The CJ test behaves similarly with the proposal by Chen, Zhang & Zhong (2010) on powers as $p/n \to y \in (0, \infty)$ and the corrected LRT had greater powers than the CJ test and Chen, Zhang & Zhong (2010)'s test when the dimension p is not large relative to the sample size n. But when p is large relative to n ($p < n$), the corrected LRT had less powers than the CJ test and the test proposed in Chen, Zhang & Zhong (2010).

Example 1. Test of sphericity: $H_{10}: \Sigma = \sigma^2 I_p$

Under H_{10}, σ^2 can be estimated by $\widehat{\sigma}^2 = p^{-1}\mathrm{tr}\mathbf{S}_n$. It is easy to see that $\widehat{\sigma}^2 = \sigma^2 + O_p(n^{-1})$ under the condition that $p/n \to y \in (0, \infty)$. Let $\{\lambda_1 \geq \lambda_2 \ldots \geq \lambda_p\}$ be the eigenvalues of $\mathbf{S}_n/(p^{-1}\mathrm{tr}\mathbf{S}_n)$. The EL and QL tests have the following expressive forms:

$$T_{n1} = q\left(p^{-1}\sum_{j=1}^{q}\lambda_j - q^{-1}\sum_{j=1}^{q}\log\lambda_j\right), \tag{5}$$

where $q = \min\{p, n-1\}$, and

$$T_{n2} = \mathrm{tr}[\mathbf{S}_n/(p^{-1}\mathrm{tr}\mathbf{S}_n) - \mathbf{I}_p]^2. \tag{6}$$

T_{n1} is equivalent to the LRT under normality assumption and $p < n-1$.
T_{n2} in (6) coincides with the CJ test proposed by Wang & Yao (2013).

Example 2: Compound symmetric structure

Let $\mathbf{A}_1 = \mathbf{I}_p$ and $\mathbf{A}_2 = \mathbf{1}_p \mathbf{1}_p^T$, where $\mathbf{1}_p$ stands for a p-dimensional column vectors with all elements being 1. Testing compound symmetric structure is to test

$$H_{20} : \Sigma = \theta_1 \mathbf{A}_1 + \theta_2 \mathbf{A}_2,$$

where $\theta_1 > 0$ and $-1/(p-1) < \theta_2/(\theta_1 + \theta_2) < 1$.

Under normality assumption, Kato, Yamada and Fujikoshi (2010) studied the asymptotic behavior of the corresponding likelihood ratio test when $p < n$, and Srivastava and Reid (2012) proposed a new test statistic for H_{20} when $p \geq n$.

Under H_{20},

$$\widehat{\theta}_1 = p^{-1}(p-1)^{-1}(p\,\mathrm{tr}\mathbf{S}_n - \mathbf{1}_p^T\mathbf{S}_n\mathbf{1}_p), \quad \widehat{\theta}_2 = p^{-1}(p-1)^{-1}(\mathbf{1}_p^T\mathbf{S}_n\mathbf{1}_p - \mathrm{tr}\mathbf{S}_n).$$

respectively. It can be shown that $\widehat{\theta}_k = \theta_k + O_p(n^{-1})$, $k = 1, 2$

T_{n1} is equivalent to the LRT in Kato, Yamada and Fujikoshi (2010) under normality assumption when $p < n - 1$.

Srivastava and Reid (2012) recast this test problem to test of independence under normality. So it is different from the LRT, T_{n1} and T_{n2}.

Example 3: Testing bandedness structure

Let $\mathbf{A}_1 = \mathbf{I}_p$ and for $2 \le k \le K$, $\mathbf{A}_k$ to be a $p \times p$ matrix with (i, j)-element being 1 if $|i - j| = k - 1$ and 0 otherwise. Testing the $(K-1)$-banded covariance matrix is to test

$$H_{30} : \Sigma = \theta_1 \mathbf{A}_1 + \cdots + \theta_K \mathbf{A}_K,$$

where θ_k's are unknown. By (2), we have

$$\begin{aligned} \widehat{\theta}_1 &= p^{-1} \mathrm{tr} \mathbf{S}_n \\ \widehat{\theta}_k &= \frac{1}{2}(p - k + 1)^{-1} \mathrm{tr} \mathbf{S}_n \mathbf{A}_k, \quad \text{for} \quad 2 \le k \le K. \end{aligned}$$

Under Assumptions A and C, $\widehat{\theta}_k = \theta_k + O_p(n^{-1})$ if $p/n \to y \in (0, \infty)$ when K is a finite.

Qiu and Chen (2012) proposed a U-statistic test for the banded structure, and is different from T_{n1} and T_{n2}.

A factor model assumes $\mathbf{x} = v_1\mathbf{U}_1 + \cdots + v_{K-1}\mathbf{U}_{K-1} + \epsilon$, where $v_1, \cdots, v_{K-1}$ are random variables and $\mathbf{U}_k$, $k = 1, \cdots, K-1$ are random vectors.

Suppose that $v_1, \cdots, v_{K-1}$, $\mathbf{U}_1, \ldots, \mathbf{U}_{K-1}$ and ϵ are mutually independent. Suppose that $\mathrm{Cov}(\epsilon) = \theta_1 \mathbf{I}_p$ and conditioning on $\mathbf{U}_k$, $k = 1, \cdots, K-1$, the factor model has a covariance structure $\Sigma = \theta_1 \mathbf{I}_p + \theta_2 \mathbf{U}_1\mathbf{U}_1^T + \cdots + \theta_K \mathbf{U}_{K-1}\mathbf{U}_{K-1}^T$, where $\theta_{k+1} = \mathrm{Var}(v_k)$.

Let $\mathbf{A}_1 = I_p$, and $\mathbf{A}_{k+1} = \mathbf{U}_k\mathbf{U}_k^T$ for $k = 1, \cdots, K-1$. Thus, it is of interest to test

$$H_{40} : \Sigma = \theta_1 \mathbf{A}_1 + \cdots + \theta_K \mathbf{A}_K,$$

where θ_k's are unknown parameters.

Example 4: Factor model

In practice one typically sets $\mathbf{U}_k$ to be orthogonal to each other and have been standardized so that $\mathbf{U}_s^T \mathbf{U}_t = p$ for $s = t$ and 0 for $s \neq t$. The parameters θ_k can be estimated by

$$\widehat{\theta}_1 = (p - K + 1)^{-1} \left(\mathrm{tr}\mathbf{S}_n - p^{-1} \sum_{k=1}^{K-1} \mathbf{U}_k^T \mathbf{S}_n \mathbf{U}_k \right)$$

$$\widehat{\theta}_{k+1} = p^{-2} (\mathbf{U}_k^T \mathbf{S}_n \mathbf{U}_k - p\widehat{\theta}_1)$$

for $k = 1, \cdots, K-1$. Thus, when K is finite and p/n has a finite positive limit, then $\widehat{\theta}_k = \theta_k + O_p(n^{-1})$ under H_{40}. Then testing H_{40} can be carried out by using the EL and QL tests.

Example 5: Test a particular pattern (McDonald, 1974)

For even p which is finite and fixed, McDonald (1974) considered

$$H_{50} : \Sigma = \begin{pmatrix} \theta_1 \mathbf{I}_{p/2} + \theta_2 \mathbf{1}_{p/2} \mathbf{1}_{p/2}^T & \theta_3 \mathbf{I}_{p/2} \\ \theta_3 \mathbf{I}_{p/2} & \theta_1 \mathbf{I}_{p/2} + \theta_2 \mathbf{1}_{p/2} \mathbf{1}_{p/2}^T \end{pmatrix}.$$

Let $\mathbf{A}_1 = \mathbf{I}_p$,

$$\mathbf{A}_2 = \begin{pmatrix} \mathbf{1}_{p/2} \mathbf{1}_{p/2}^T & \mathbf{0} \\ \mathbf{0} & \mathbf{1}_{p/2} \mathbf{1}_{p/2}^T \end{pmatrix} \quad \text{and} \quad \mathbf{A}_3 = \begin{pmatrix} \mathbf{0} & \mathbf{I}_{p/2} \\ \mathbf{I}_{p/2} & \mathbf{0} \end{pmatrix}.$$

Then H_{50} can be written as $H_{50} : \Sigma = \theta_1 \mathbf{A}_1 + \theta_2 \mathbf{A}_2 + \theta_3 \mathbf{A}_3$. Thus, the proposed EL and QL test procedure can be used to test H_{50} with high-dimensional data.

Some results on random matrix

To derive the limiting distribution of T_{n1} and T_{n2}, we need the limiting distributions of $\log(|\mathbf{S}_n\widehat{\Sigma}_0^{-1}|)$, $\mathrm{tr}(\mathbf{S}_n\widehat{\Sigma}_0^{-1})$ and $\mathrm{tr}\{(\mathbf{S}_n\widehat{\Sigma}_0^{-1})^2\}$.

All these quantities are functional of eigenvalues of $\widehat{\Sigma}_0^{-1/2}\mathbf{S}_n\widehat{\Sigma}_0^{-1/2}$. This motivates us to study the limiting distribution of $\Sigma^{-1/2}\mathbf{S}_n\Sigma^{-1/2}\widehat{=}\mathbf{F}$, the sample covariance matrix of $\mathbf{w}_1, \cdots, \mathbf{w}_n$.

Zheng, Chen, Cui and Li (2017) derived the limiting distribution of $(\mathrm{tr}\mathbf{F}, \sum_{j=1}^{q}\log\{\lambda_j(\mathbf{F})\})$.

Theorem 1.

Suppose that $\mathbf{C}_0$ is a $p \times p$ symmetric constant matrix, and $p^{-1}\mathrm{tr}\mathbf{C}_0$ has finite limit. Under Assumptions A and B, it follows that

$$p^{-1}\mathrm{tr}\mathbf{F}\mathbf{C}_0 - p^{-1}\mathrm{tr}\mathbf{C}_0 = o_p(1), \quad (7)$$

$$p^{-1}\mathrm{tr}\mathbf{F}^2\mathbf{C}_0 - (1 + y_{n-1})p^{-1}\mathrm{tr}\mathbf{C}_0 = o_p(1). \quad (8)$$

Joint limiting distribution

Theorem 2.

Suppose that $\mathbf{C}_1$ and $\mathbf{C}_2$ are symmetric constant matrices, and their ESDs weakly converge as $p \to \infty$. Assume that $\Sigma_n^{(1)} \to \Sigma^{(1)}$ exists. Under Assumptions A and B, it follows that

$$\begin{pmatrix} \mathrm{tr}\mathbf{F}\mathbf{C}_1\mathbf{F}\mathbf{C}_1 \\ \mathrm{tr}\mathbf{F}\mathbf{C}_2 \end{pmatrix} - \boldsymbol{\mu}_n^{(1)} \to N(\mathbf{0}, \Sigma^{(1)})$$

in distribution as $n \to \infty$, where $\boldsymbol{\mu}_n^{(1)}$, $\Sigma_n^{(1)}$ and $\Sigma^{(1)}$ are given in Zheng, Chen, Cui and Li (2017).

Theorem 3.

Under H_0 and Assumptions A and B, it follows that

(a) For $p < n-1$,

$$\frac{T_{n1} + (p-n+1)\log(1-y_{n-1}) - 2p + \alpha_1(y_{n-1})}{\sqrt{\alpha_2(y_{n-1})}} \to N(0,1), \qquad (9)$$

where $\alpha_1(y) = 0.5\log(1-y) - 0.5(\kappa-3)y$ and $\alpha_2(y) = -2y - 2\log(1-y)$;

Theorem 3

(b) For $p \geq (n-1)$,

$$\frac{T_{n1} - p[y_{n-1}^{-1} - \alpha_3(y_{n-1})] - y_{n-1}^{-1} m_1(y_{n-1}) + m_2(y_{n-1})}{\sqrt{y^{-2} v_{11}(y_{n-1}) + v_{22}(y_{n-1}) - 2y_{n-1}^{-1} v_{12}(y_{n-1})}} \to N(0,1) \quad (10)$$

where $\alpha_3(y)$, $m_1(y)$, $m_2(y)$, $v_{11}(y)$, $v_{22}(y)$, $v_{12}(y)$ are given in the paper.
(c) $\frac{1}{2}[T_{n2} - py_{n-1} - (\kappa - 2)y]/\sigma \to N(0,1)$, where

$$\sigma^2 = y_{n-1}^2 + 2y_{n-1}^3 p^{-1} \mathrm{tr}(\Sigma_0 \mathbf{B})^2 + (\kappa - 3) y_{n-1}^3 p^{-1} \sum_{i=1}^{p} (\mathbf{e}_i^T \Gamma^T \mathbf{B} \Gamma \mathbf{e}_i)^2 - (\kappa - 1) y_{n-1}^3$$

with $\mathbf{B} = \sum_{k=1}^{K} d_k \mathbf{A}_k$, $\mathbf{d} = (d_1, \cdots, d_K)^T = \mathbf{C}^{-1}\mathbf{c}$ and $\mathbf{c}$ being a K-dimensional column vector with k-th element being $\mathrm{tr}\mathbf{A}_k \Sigma_0^{-1}$.

The limiting null distribution can be used to construct the rejection regions for T_{n1} and T_{n2}.

We next study the asymptotic power of these two tests. Define

$$\begin{pmatrix} \theta_1^* \\ \theta_2^* \\ \dots \\ \theta_K^* \end{pmatrix} = \begin{pmatrix} \mathrm{tr}\mathbf{A}_1^2 & \mathrm{tr}\mathbf{A}_2\mathbf{A}_1 & \dots & \mathrm{tr}\mathbf{A}_K\mathbf{A}_1 \\ \mathrm{tr}\mathbf{A}_1\mathbf{A}_2 & \mathrm{tr}\mathbf{A}_2^2 & \dots & \mathrm{tr}\mathbf{A}_K\mathbf{A}_2 \\ \dots & \dots & \dots & \dots \\ \mathrm{tr}\mathbf{A}_1\mathbf{A}_K & \mathrm{tr}\mathbf{A}_2\mathbf{A}_K & \dots & \mathrm{tr}\mathbf{A}_K^2 \end{pmatrix}^{-1} \begin{pmatrix} \mathrm{tr}\Sigma\mathbf{A}_1 \\ \mathrm{tr}\Sigma\mathbf{A}_2 \\ \dots \\ \mathrm{tr}\Sigma\mathbf{A}_K \end{pmatrix}$$

and $\Sigma_{0*} = \theta_1^*\mathbf{A}_1 + \ldots + \theta_K^*\mathbf{A}_K$. Under $H_0 : \Sigma = \Sigma_0$, it follows that $\Sigma_{0*} = \Sigma_0$. Under Assumptions A and B, it can be shown that $\widehat{\theta}_k = \theta_k^* + O_p(n^{-1}), k = 1, \ldots, K$.

Let $G_p(t) = p^{-1}\sum_{j=1}^p I(\lambda_j \leq t)$ be the ESD of $\Sigma_{0*}^{-1/2}\Sigma\Sigma_{0*}^{-1/2}$.

If $G_p(t) \to G(t)$, in which $G(t)$ is assumed not to be degenerated to a single point distribution. That is, $\Sigma \neq \tau\Sigma_{0*}$ for some constant $\tau > 0$. Under this condition, we have the limiting distributions of T_{n1} and T_{n2}.

Limiting distributions under alternative

Theorem 4

Under Assumptions A and B, and under $H_1 : G_p(t) \to G(t)$, a non-degenerated distribution, it follows that

$$\frac{T_{n1} - pF_2^{y_{n-1},G} - \mu_2^{(1)}}{\sigma_{2n}^{(1)}} \to N(0,1), \quad \text{for} \quad p < n-1$$

$$\frac{T_{n1} - pF_3^{y_{n-1},G} - \mu_3^{(1)}}{\sigma_{3n}^{(1)}} \to N(0,1), \quad \text{for} \quad p \geq n-1$$

$$\frac{T_{n2} - \mu_1^{(1)}}{\sigma_{1n}^{(1)}} \to N(0,1),$$

where $\mu_j^{(1)}, j = 1,2,3$, $F_j^{y_{n-1},G}, j = 2,3$ and $\sigma_{jn}^{(1)}, j = 1,2,3$ are given in the proof of Theorem .

Power function of T_{n2} and unbiased test

For a level α, the power function of T_{n2} is
$1-\Phi((\mu_0-\mu_1^{(1)})/\sigma_{1n}^{(1)}-2q_{\alpha/2}\sigma/\sigma_{1n}^{(1)})+\Phi((\mu_0-\mu_1^{(1)})/\sigma_{1n}^{(1)}+2q_{\alpha/2}\sigma/\sigma_{1n}^{(1)})$,
where $q_{\alpha/2}$ is the $\alpha/2$ quantile of $N(0,1)$ and $\mu_0 = py_{n-1}+(\kappa-2)y$.

Theorem 5

Suppose that Assumptions A and B are satisfied and the limit of $\sigma_{1n}^{(1)}$ exists. If $\Gamma^T\Sigma_{0*}^{-1}\Gamma = \mathbf{I}_p+\mathbf{A}$ whose ESD weakly converges and $\mathrm{tr}\mathbf{A}^2 > \delta > 0$, then we have

$$\beta_{T_{n2}} > \alpha$$

when n is large enough and δ is any given small constant.

Thus, T_{n2} is an asymptotically unbiased test. Furthermore, if $p^{-1}\mathrm{tr}\mathbf{A} \to c_1 \neq 0$, then $\beta_{T_{n2}} \to 1$ when $n \to \infty$.

Issue in Practical Implementation

Estimation of κ: Under Assumption A, we construct an consistent estimate for κ by using

$$E\{\mathbf{w}^T\Sigma\mathbf{w} - \mathrm{tr}(\Sigma)\}^2 = 2\mathrm{tr}(\Sigma^2) + (\kappa - 3)\sum_{j=1}^{p}\sigma_{jj}^2.$$

Our simulation shows that we may get more stable estimate for κ by using the linear structure under H_0.

Numerical Examples

In our simulation, $\mathbf{x} = \Sigma^{1/2}\mathbf{w}$, where Σ will be set according to the hypothesis to be tested.

In order to examine the performances of the proposed tests under different distributions, we consider the elements of $\mathbf{w}$ being independent and identically distributed as (a) $N(0, 1)$ or (b) Gamma(4,2)-2. The two distributions are both normalized so that their means 0 and variances 1.

For each setting, we conduct 1000 Monte Carlo simulations. The Monte Carlo simulation error rate is $1.96\sqrt{0.05 \times 0.95/1000} \approx 0.0135$ at level 0.05.

Test of compound symmetry $H_{20} : \Sigma = \theta_1 \mathbf{I}_p + \theta_2 \mathbf{1}_p \mathbf{1}_p^T$

We set $\Sigma = \theta_1 \mathbf{I}_p + \theta_2 \mathbf{1}_p \mathbf{1}_p^T + \theta_3 \mathbf{u}_p \mathbf{u}_p^T$, where $\mathbf{u}_p$ is a p-dimensional random vector following uniform distribution over $[-1, 1]$.

The third term is to examine the empirical powers when $\theta_3 \neq 0$. In our simulation, we set $(\theta_1, \theta_2) = (6, 1)$ and $\theta_3 = 0, 0.5, 1$, respectively.

We set $\theta_3 = 0$ to examine Type I error rate and $\theta_3 = 0.5, 1$ to study the empirical power of the proposed tests. The sample size is set as $n = 100, 200$ and the dimension is taken to be $p = 50, 100, 500, 1000$.

Compare the performances of proposed testing procedures and the test proposed in Srivastava and Reid (2012)

Table: Empirical power for H_{20} (in percentage) with $n = 100$

		$w_j \sim N(0,1)$				$w_j \sim$ Gamma(4,2)−2			
θ_3	Test	p =50	100	500	1000	50	100	500	1000
0	QL	5.23	5.40	5.12	5.19	6.48	5.99	5.64	5.41
	EL	5.32	6.51	5.12	5.18	5.77	6.35	5.46	5.54
	SR	4.90	5.01	4.91	4.98	9.60	8.96	8.15	8.04
0.5	QL	40.25	80.58	100.00	100.00	41.01	80.70	100.00	100.00
	EL	13.42	11.38	99.78	100.00	13.74	11.42	99.74	100.00
	SR	24.46	59.04	99.99	100.00	41.22	73.86	100.00	100.00
1.0	QL	95.88	99.97	100.00	100.00	95.98	99.97	100.00	100.00
	EL	53.53	29.97	100.00	100.00	53.71	30.18	100.00	100.00
	SR	87.90	99.67	100.00	100.00	93.61	99.87	100.00	100.00

Table: Empirical power for H_{20} (in percentage) with $n = 200$

		$w_j \sim N(0,1)$				$w_j \sim$ Gamma(4,2)−2			
θ_3	Test	p =50	100	500	1000	50	100	500	1000
0	QL	5.22	5.14	5.12	5.19	6.32	5.78	5.31	5.34
	EL	5.18	5.12	5.05	5.13	5.94	5.42	5.23	5.31
	SR	4.98	4.93	4.93	5.03	9.95	9.23	8.44	8.43
0.5	QL	79.86	99.32	100.00	100.00	78.56	99.28	100.00	100.00
	EL	42.00	58.62	100.00	100.00	41.22	58.62	100.00	100.00
	SR	61.74	95.79	100.00	100.00	75.78	98.24	100.00	100.00
1.0	QL	99.98	100.00	100.00	100.00	99.96	100.00	100.00	100.00
	EL	97.23	99.53	100.00	100.00	96.86	99.55	100.00	100.00
	SR	99.81	100.00	100.00	100.00	99.91	100.00	100.00	100.00

Test covariance matrix structure in H_{30}

We construct a banded matrix defined in Example 3 with width of band $K = 3$. Therefore, the null hypothesis H_{30} has the linear decomposition $\Sigma = \theta_1 \mathbf{I}_p + \theta_2 \mathbf{A}_2 + \theta_3 \mathbf{A}_3 + \theta_4 \mathbf{u}_p \mathbf{u_p}^T$, where $\mathbf{A}_2$ and $\mathbf{A}_3$ are defined in Example 3 and $\mathbf{u}_p$ is generated by the same way in the last example.

We take $(\theta_1, \theta_2, \theta_3) = (6, 1, 0.5)$ and $\theta_4 = 0$ to examine Type I error rates and take $\theta_4 = 0.5, 1$ to examine powers. In the simulation studies, we still set sample size $n = 100, 200$ and dimension $p = 50, 100, 500, 1000$.

In this example, we compare the test proposed in Qiu and Chen (2012) for the banded covariance matrix with our proposed tests, and referred their test as "QC" test hereinafter.

Table: Simulation Results for H_{30} (in percentage)

			$w_j \sim N(0,1)$				$w_j \sim$ Gamma(4,2)$-$2			
θ_4	n	Test	p =50	100	500	1000	50	100	500	1000
0	100	QL	5.30	5.19	5.29	5.34	6.61	6.19	5.83	5.89
		EL	5.31	6.34	5.20	5.36	5.83	6.32	5.53	5.81
		QC	5.00	5.50	5.50	5.90	5.00	5.80	6.10	5.20
0.5	100	QL	46.01	84.91	100.00	100.00	45.13	83.62	100.00	100.00
		EL	15.20	12.23	99.90	100.00	15.35	12.13	99.90	100.00
		QC	17.00	70.00	100.00	100.00	24.00	85.30	100.00	100.00
1.0	100	QL	97.49	99.98	100.00	100.00	96.96	99.98	100.00	100.00
		EL	60.45	33.68	100.00	100.00	59.80	33.71	100.00	100.00
		QC	83.00	100.00	100.00	100.00	78.00	99.90	100.00	100.00

Table: Simulation Results for H_{30} (in percentage)

			$w_j \sim N(0,1)$				$w_j \sim$ Gamma(4,2)−2			
θ_4	n	Test	p =50	100	500	1000	50	100	500	1000
0	200	QL	5.25	5.19	5.16	5.10	6.44	5.84	5.56	5.55
		EL	5.24	5.15	5.04	4.94	6.01	5.46	5.25	5.26
		QC	6.00	5.40	5.50	4.50	4.00	4.50	5.80	5.10
0.5	200	QL	85.30	99.66	100.00	100.00	84.09	99.64	100.00	100.00
		EL	48.46	65.39	100.00	100.00	47.62	65.20	100.00	100.00
		QC	49.00	100.00	100.00	100.00	49.00	99.70	100.00	100.00
1.0	200	QL	99.99	100.00	100.00	100.00	99.98	100.00	100.00	100.00
		EL	98.27	99.80	100.00	100.00	98.00	99.76	100.00	100.00
		QC	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00

Test of factor model H_{40}

In this example, we examine Type I error rate and empirical power of the proposed tests for factor model.

We first generate several mutually orthogonal factors. Suppose that $\mathbf{u}_k^*$, $k = 1, \cdots, K$ are IID random vectors following $N_p(\mathbf{0}, \mathbf{I}_p)$. Let $\mathbf{u}_1 = \mathbf{u}_1^*$ and $\mathbf{u}_k = (\mathbf{I}_p - \mathbf{P}_k)\mathbf{u}_k^*$, where $\mathbf{P}_k$ is the projection matrix on $\mathbf{u}_1, \cdots, \mathbf{u}_{k-1}$ for $k = 2, \cdots, K$.

Given $\mathbf{u}_k$'s, we have the covariance matrix structure

$$\Sigma = \theta_0 \, \mathbf{I}_p + \sum_{k=1}^{K} \theta_k \mathbf{u}_k \mathbf{u}_k^T$$

for the factor model.

We set $K = 4$ and the coefficient vector $(\theta_0, \theta_1, \theta_2, \theta_3)^T = (4, 3, 2, 1)^T$.

Similarly, $\theta_4 = 0$ is for Type I error rates and $\theta_4 = 0.5, 1$ is for powers.

Table: Empirical powers for H_{40} (in percentage)

			$w_j \sim N(0,1)$				$w_j \sim$ Gamma(4,2)−2			
θ_5	n	Test	p =50	100	500	1000	50	100	500	1000
0	100	QL	5.46	5.46	6.02	6.27	6.89	6.53	6.58	6.94
		EL	5.40	6.40	5.79	6.20	6.03	6.42	6.30	6.52
0.5	100	QL	99.99	100.00	100.00	100.00	99.96	100.00	100.00	100.00
		EL	97.58	86.91	100.00	100.00	97.38	87.25	100.00	100.00
1.0	100	QL	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
		EL	100.00	99.93	100.00	100.00	99.99	99.93	100.00	100.00
0	200	QL	5.32	5.28	5.42	5.52	6.57	6.05	5.89	6.05
		EL	5.30	5.22	5.33	5.65	6.15	5.61	5.60	5.65
0.5	200	QL	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
		EL	99.99	100.00	100.00	100.00	100.00	100.00	100.00	100.00
1.0	200	QL	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
		EL	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00

Test for special pattern H_{50}

To investigate the performance of QL and EL tests for H_{50}, represent Σ as a linear combination

$$\Sigma = \sum_{k=1}^{4} \theta_k \mathbf{A}_k,$$

where $\mathbf{A}_1$, $\mathbf{A}_2$ and $\mathbf{A}_3$ are defined in Example 5 and $\mathbf{A}_4 = \mathbf{u}_p \mathbf{u}_p^T$ with $\mathbf{u}_p \sim N_p(\mathbf{0}, \mathbf{I}_p)$. We set the first three coefficients $(\theta_1, \theta_2, \theta_3) = (6, 0.5, 0.1)$ and $\theta_4 = 0$, 0.5 and 1 for examining Type I error rates and powers, respectively.

Table: Empirical power for H_{50} (in percentage)

			$w_j \sim N(0,1)$				$w_j \sim$ Gamma(4,2)−2			
θ_5	n	Test	p =50	100	500	1000	50	100	500	1000
0	100	QL	5.28	5.19	5.23	5.41	6.59	6.15	5.84	6.16
		EL	5.27	6.33	5.17	5.48	5.85	6.35	5.61	5.73
0.5	100	QL	99.64	100.00	100.00	100.00	99.61	100.00	100.00	100.00
		EL	85.90	57.86	100.00	100.00	85.89	59.17	100.00	100.00
1.0	100	QL	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
		EL	99.87	97.67	100.00	100.00	99.81	97.86	100.00	100.00
0	200	QL	5.25	5.11	5.08	5.18	6.40	5.84	5.58	5.62
		EL	5.25	5.15	5.06	5.05	6.01	5.44	5.29	5.30
0.5	200	QL	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
		EL	99.75	99.99	100.00	100.00	99.72	99.99	100.00	100.00
1.0	200	QL	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
		EL	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00

5.3 Error variance estimation

Consider a linear regression model

$$Y_i = \mathbf{x}_i^T \boldsymbol{\beta} + \varepsilon,$$

where $E(\varepsilon|\mathbf{x}) = 0$ and $\mathrm{Var}(\varepsilon|\mathbf{x}) = \sigma^2$.

If $d = \dim(\mathbf{x}_i)$ finite and fixed, the MSE is a good estimate of σ^2.

If $d > n$,
(a) impose sparsity assumption, and select significant variables $\Rightarrow$ spurious correlation

(b) not impose sparsity assumption, leads to overfit model and all residuals equal to zero, and MSE does not perform well in this case.

Illustration of spurious correlation

To explain more clearly the concept about spurious correlation, we simulate $n = 50$ data points iid $N(0,1)$ $\{X_j\}_{j=1}^p$ (with $p = 1000$) and iid response $Y \sim N(0,1)$.

In this **null** model: $Y = \varepsilon$, all covariates $\{X_j\}_{j=1}^p$ and the response Y are independent and follow $N(0,1)$.

Estimate the maximum "linear" spurious correlation:

$$\zeta_n = \max_{1 \le j \le p} |\widehat{\mathrm{corr}}(X_j, Y)|$$

In other words, one variable ($s = 1$) is selected to predict the realized noise vector best.

Idea can be extended to s variables: the correlation between the response and fitted values using the "best subset" of s-variables.

Spurious correlation and its impact on estimation of σ^2

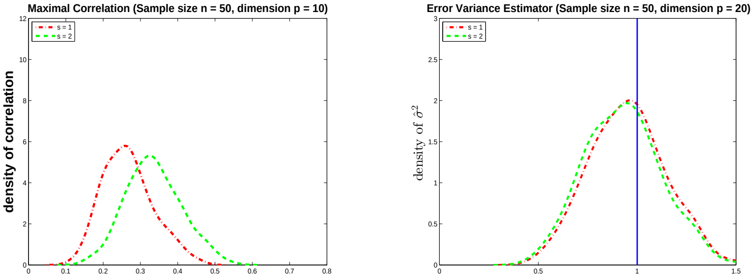
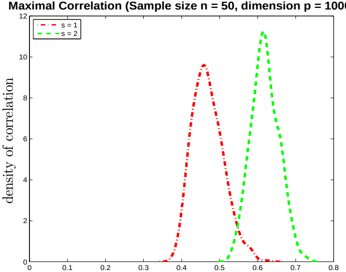


Figure: Distributions of the maximum spurious correlations between Y and X_j for linear model when y and $\mathbf{x}$ are independent, and their impact on the estimation of σ^2 based on 500 simulations. ($s = 1, 2$, $(n, p) = (50, 10)$)

Spurious correlation and its impact on estimation of σ^2



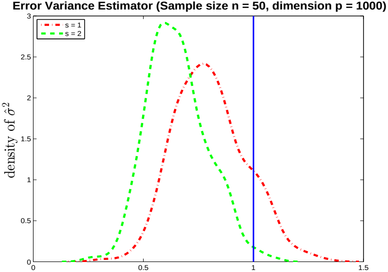


Figure: Distributions of the maximum spurious correlations between Y and X_j for linear model when y and $\mathbf{x}$ are independent, and their impact on the estimation of σ^2 based on 500 simulations. $(s = 1, 2, (n, p) = (50, 1000))$

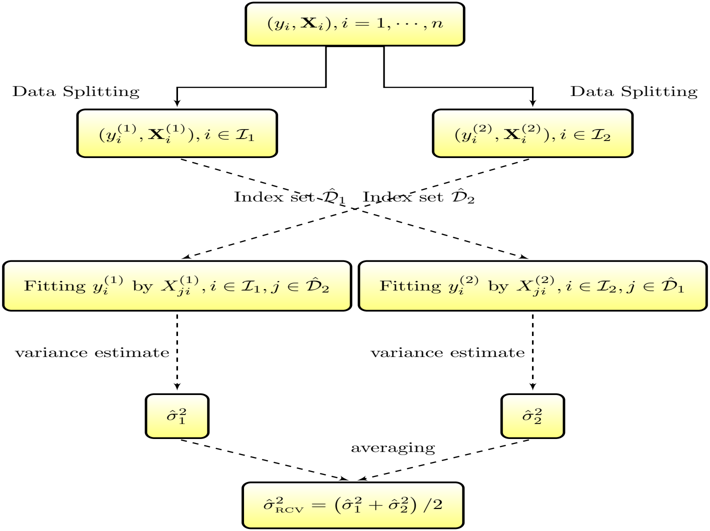
How to get rid of spurious correlation

Data splitting:

Splitting data into two parts: one for model selection and for model fitting and error variance estimation.

This leads to refitted cross-validation (RCV) method for error variance estimation (Fan, Guo and Hao, 2012, JRSSB)

Flow Chart of RCV Algorithm



RCV for linear models

Ref: Fan, Guo and Hao (2012, JRSSB)

Linear regression model:

$$\mathbf{y} = \mathbf{X}\beta + \varepsilon$$

where $E(\varepsilon|\mathbf{X}) = 0$ and $\mathrm{Var}(\varepsilon|\mathbf{X}) = \sigma^2 I$.

Naive estimation procedure:

Step 1. Use feature screening/variable selection procedure to choose a model $\widehat{M}$

Step 2. Let $\mathbf{P}_M = \mathbf{X}_M(\mathbf{X}_M^T\mathbf{X}_M)^{-1}\mathbf{X}_M$, the projection matrix of $\mathbf{X}_M$.

$$\widehat{\sigma}^2_{\widehat{M}} = (n - \widehat{s})^{-1}\mathbf{y}^T(I - \mathbf{P}_{\widehat{M}})\mathbf{y} = (n - \widehat{s})^{-1}\varepsilon^T(I - \mathbf{P}_{\widehat{M}})\varepsilon$$

where $\widehat{s} = |\widehat{M}|$.

Asymptotic behavior of naive estimate

Denote $\widehat{\gamma}_n = \varepsilon^T \mathbf{P}_{\widehat{M}} \varepsilon / \varepsilon^T \varepsilon$

$$\widehat{\sigma}_n^2 = (n - \widehat{s})^{-1}(1 - \widehat{\gamma}_n^2)\varepsilon^T \varepsilon$$

Assumptions:
1. Errors are iid with mean zero and finite variance σ^2 and independent of **x**.
2. Lower bounded of minimal eigenvalue of design matrix $\frac{1}{n}\mathbf{X}_M^T \mathbf{X}_M$
3. **x** is bounded
4. Error ε is sub-Gaussian.

Fan, Guo and Hao (2012) proved that
(1) If a procedure satisfies the sure screening property with $\widehat{s} \leq b_n$ where $b_n = o(n)$, then $\widehat{\sigma}_n^2/(1-\widehat{\gamma}_n^2) \to \sigma^2$ in probability and

$$\sqrt{n}\{\widehat{\sigma}_n^2/(1-\widehat{\gamma}_n^2) - \sigma^2\} \to N(0, E(\varepsilon_1^4) - \sigma^4)$$

in distribution.

(b) If, in addition, $\log(p)/n = O(1)$, then $\widehat{\gamma} = O_P[\sqrt{\widehat{s}\log(p)/n}]$.

RCV procedure:

1. Split data into two subsets $(\mathbf{X}^{(1)}, y^{(1)})$ and $(\mathbf{X}^{(2)}, y^{(2)})$
2. A variable selection/feature tool is performed on $(\mathbf{X}^{(1)}, y^{(1)})$, and let $\widehat{M}_1$ denote the set of variable selected. The variance σ^2 is then estimated on the second data set $(\mathbf{X}^{(2)}, y^{(2)})$, namely

$$\widehat{\sigma}_1^2 = (n/2 - |\widehat{M}_1|)^{-1}\mathbf{y}^{(2)T}(I - \mathbf{P}^{(2)}_{\widehat{M}_1})\mathbf{y}^{(2)}$$

   Similarly, we obtain

$$\widehat{\sigma}_2^2 = (n/2 - |\widehat{M}_2|)^{-1}\mathbf{y}^{(1)T}(I - \mathbf{P}^{(1)}_{\widehat{M}_2})\mathbf{y}^{(1)}$$

3. Define the final estimator as

$$\widehat{\sigma}_{RCV}^2 = (\widehat{\sigma}_1^2 + \widehat{\sigma}_2^2)/2$$

Under Assumption 1 and 2, $E(\varepsilon_1^4) < \infty$. If a procedure satisfies the sure screening property with $\widehat{s}_1 \leq b_n$ and $\widehat{s}_2 \leq b_n$, and $b_n = o(n)$, then

$$\sqrt{n}(\widehat{\sigma}_{RCV} - \sigma^2) \to N(0, E(\varepsilon_1^4) - \sigma^4)$$

Let $\boldsymbol{\beta}^*$ is the true value of $\boldsymbol{\beta}$, define

$$\widehat{\sigma}_O^2 = n^{-1} \sum_{i=1}^{n} (Y_i - \mathbf{x}_i^T \boldsymbol{\beta}^*)^2$$

$\widehat{\sigma}_O^2$ and $\widehat{\sigma}_{RCV}^2$ have the same asymptotic variance.

Discussions

1. Generality

2. Conditions

Scaled sparse linear regression

Ref: Sun and Zhang (2012, Biometrika)

Model: $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ with $\mathrm{Var}(\varepsilon_1) = \sigma^2$

Strategy: Estimate $\boldsymbol{\beta}$ and σ^2 jointly.

Penalized least squares:

$$L_\lambda(\boldsymbol{\beta}) = \frac{1}{2n}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \lambda^2 \sum_{j=1}^{p} \rho(|\beta_j|/\lambda)$$

where $\rho(\cdot)$ is a penalty function.

Key observation: Scale-invariance considerations and existing theory suggest that $\lambda \propto \sigma$.

Scaled Lasso

$$L_{\lambda_0}(\boldsymbol{\beta},\sigma) = \frac{\|\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\|^2}{2n\sigma} + \frac{(1-a)\sigma}{2} + \lambda_0 \sum_{j=1}^{p} |\beta_j|,$$

where λ_0 is a pre-specified value. The paper uses $\lambda_0 = 2^{j-1}\sqrt{\log(p)/n}$ with $j = 1, 2$ and 3.

a is a parameter to adjust degrees of freedom. In general, we may set $a = 0$.

Alternating minimization $\boldsymbol{\beta}$ and σ^2 leads to

1. Given $\widehat{\sigma}$, we solve a Lasso problem with $\lambda = \lambda_0\widehat{\sigma}$

$$L_{\lambda_0}(\boldsymbol{\beta},\widehat{\sigma}) = \frac{\|\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\|^2}{2n} + (\lambda_0\sigma)\sum_{j=1}^{p} |\beta_j|.$$

2. Given $\widehat{\boldsymbol{\beta}}$, we have

$$\widehat{\sigma} = [\|\mathbf{y}-\mathbf{X}\widehat{\boldsymbol{\beta}}\|^2/\{(1-a)n\}]^{1/2}$$

Asymptotic properties

1. $\widehat{\sigma}/\sigma \to 1$ in probability

2. $\sqrt{n}(\widehat{\sigma}/\sigma - 1) \to N(0, 1/2)$ in distribution.

Discussion

Conditions are hard to verify.

Seems to be specific for linear regression model. Difficulty to be extended for other models.

Moment estimator

Ref: Dicker (2014, Biometrika)

Model: $\mathbf{y} = \mathbf{X}\beta + \varepsilon$

Assumption: $\mathbf{x}$ and ε are normal, $E(\mathbf{x}) = 0$, $\mathrm{Cov}(\mathbf{x}) = \Sigma$ and $p/n^2 \to 0$ for consistency and $p/n \to y \in [0, \infty)$ for normality.

Key observation:

$$\frac{1}{n}E(\|\mathbf{y}\|^2) = \sigma^2 + \tau_1^2 \tag{1}$$
$$\frac{1}{n^2}E(\|\mathbf{X}^T\mathbf{y}\|^2) = \frac{p}{n}m_1\sigma^2 + \frac{p}{n}m_1\tau_1^2 + (1 + \frac{1}{n})\tau_2^2 \tag{2}$$

where $\tau_k^2 = \beta^T\Sigma^k\beta$, $m_k = p^{-1}\mathrm{tr}(\Sigma^k)$
(2) needs normality assumption.

When $\Sigma = I$, then $\tau_1 = \tau_2 = \|\beta\|^2$ and $m_1 = 1$. Then

$$\frac{1}{n}E(\|\mathbf{y}\|^2) = \sigma^2 + \tau_1^2 \tag{3}$$

$$\frac{1}{n^2}E(\|\mathbf{X}^T\mathbf{y}\|^2) = \frac{p}{n}\sigma^2 + \frac{n+p+1}{n}\tau_1^2 \tag{4}$$

From these two equations, we can construct an unbiased moment estimator for (σ^2, τ_1).

Estimable Σ

If Σ is unknown, but a norm-consistent estimator $\widehat{\Sigma}$ for Σ is available such that $\|\widehat{\Sigma} - \Sigma\| \to 0$. Then we rewrite

$$\mathbf{y} = \mathbf{X}\Sigma^{-1/2}\Sigma^{1/2}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \widehat{=} \mathbf{Z}\boldsymbol{\gamma} + \boldsymbol{\varepsilon}$$

Substituting Σ in $\mathbf{Z}$ by $\widehat{\Sigma}$ and approximately reduce this case to $\Sigma = I$.

Nonestimable Σ

The author claims that

$$\boldsymbol{\beta}^T \Sigma^k \boldsymbol{\beta} \approx \|\boldsymbol{\beta}\|^2 p^{-1} \mathrm{tr}(\Sigma^k)$$

for positive integer k.

Numerical comparison

Reid, Tibshirani and Friedman (2016, Statistica Sinica) compare 11 methods for error estimation methods including three methods introduced in this section.

See this paper for more details.

5.4 Tests on regression coefficients in HD regression

Consider HD linear regression model

$$\mathbf{y} = \mathbf{X}\beta + \varepsilon$$

in matrix notation.

Of interest is to test

$$H_0 : \boldsymbol{\beta} = \boldsymbol{\beta}_0 \quad \text{versus} \quad H_1 : \boldsymbol{\beta} \neq \boldsymbol{\beta}_0$$

For low-dimensional case, the LSE $\widehat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$ and $\mathrm{Cov}(\widehat{\beta}) = \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}$.
A natural χ^2-type test is

$$T_1 = (\widehat{\beta} - \beta_0)^T(\mathbf{X}^T\mathbf{X})(\widehat{\beta} - \beta_0)/\widehat{\sigma}^2$$

Under normality assumption, it follows a F-distribution.

Challenges in HD cases

1. $(\mathbf{X}^T\mathbf{X})$ is not invertible, and LSE is not well-defined.

2. Estimation of σ^2.

Some references

1. Zhong, P.S. and Chen, S. X. (2011). Tests in high dimensional regression coefficients with factorial designs. *JASA*, **106**, 260 - 274.
2. Wang, S. and Cui, H. (2013). Generalized F test for high dimensional linear regression coefficients. *Journal of Multivariate Analysis*, **117**, 134 - 149.
3. Wang, S. and Cui, H. (2015). A new test for part of high dimensional regression coefficients. *Journal of Multivariate Analysis*, **137**, 187 - 203.

These works extended Chen and Qin (2010) and applied theory of U-statistics to derive the limiting distribution under H_0 and the local power under H_1.

Connection with HD one-sample mean problem

Note that

$$y = \mathbf{x}^T \boldsymbol{\beta} + \varepsilon \Rightarrow \mathbf{x}y = \mathbf{x}\mathbf{x}^T \boldsymbol{\beta} + \mathbf{x}\varepsilon$$

This implies that $E(\mathbf{x}y) = \Sigma\boldsymbol{\beta}$ by assuming $E(\varepsilon|\mathbf{x}) = 0$, where $\Sigma = E(\mathbf{x}\mathbf{x}^T)$.

Note that

$$H_0 : \boldsymbol{\beta} = \boldsymbol{\beta}_0 \Leftrightarrow H_0 : \Sigma(\boldsymbol{\beta} - \boldsymbol{\beta}_0) = 0$$

if Σ is full rank (i.e. positive definite).

WLOG, assume $\boldsymbol{\beta}_0 = \mathbf{0}$. Define $\mathbf{z} = \mathbf{x}y$ and $\boldsymbol{\mu}_z = E(\mathbf{z})$.

$$H_0 : \boldsymbol{\beta} = \boldsymbol{\beta}_0 \Leftrightarrow H_0 : \boldsymbol{\mu}_z = 0.$$

This is the one sample mean problem. We are able to deal with it.

Connections to test on covariance structure

Write $\mathbf{w} = (\mathbf{x}^T, \mathbf{y})^T$.

$$\text{Cov}(\mathbf{w}) = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}$$

$\Sigma_{12} = \text{Cov}(\mathbf{x}, \mathbf{y}) = E(\mathbf{x}\mathbf{y}^T)$ if we assume $E(\mathbf{x}) = \mathbf{0}$.

In particular, when $\dim(y)$ =1, this is equivalent to test $E(\mathbf{x}y) = \mathbf{0}$.

Under normality assumption, test $H_0 : \Sigma_{12} = \mathbf{0}$ is test of independence between $\mathbf{x}$ and $\mathbf{y}$. This can be done by test of one sample mean problem too.

Let $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^T, \boldsymbol{\beta}_2^T)^T$. $\dim(\boldsymbol{\beta}_1)$ is finite and small. Of interest is to test

$$H_0 : \boldsymbol{\beta}_2 = \mathbf{0} \quad \text{versus} \quad H_1 : \boldsymbol{\beta}_2 \neq \mathbf{0}$$

Observe

$$\begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix} \begin{pmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \end{pmatrix} = \boldsymbol{\mu}_z$$

$\Rightarrow$

$$\begin{pmatrix} I & \mathbf{0} \\ -\Sigma_{21}\Sigma_{11}^{-1} & I \end{pmatrix} \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix} \begin{pmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \end{pmatrix} = \begin{pmatrix} I & \mathbf{0} \\ -\Sigma_{21}\Sigma_{11}^{-1} & I \end{pmatrix} \boldsymbol{\mu}_z$$

$\Rightarrow$

$$\begin{pmatrix} \Sigma_{11} & * \\ \mathbf{0} & \Sigma_{22.1} \end{pmatrix} \begin{pmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \end{pmatrix} = \begin{pmatrix} \boldsymbol{\mu}_{z1} \\ \boldsymbol{\mu}_{z2} - \Sigma_{21}\Sigma_{11}^{-1}\boldsymbol{\mu}_{z1} \end{pmatrix}$$

$\Rightarrow$

$$H_0 : \boldsymbol{\beta}_2 = \mathbf{0} \Leftrightarrow H_0 : \Sigma_{22.1}\boldsymbol{\beta}_2 = \mathbf{0} \Leftrightarrow H_0 : \boldsymbol{\mu}_{z2} - \Sigma_{21}\Sigma_{11}^{-1}\boldsymbol{\mu}_{z1} = \mathbf{0}$$

This strategy is called "decorrelation".

When $\dim(\boldsymbol{\beta}_1)$ is finite and small, we may can consistent estimate for Σ_{21} and Σ_{11}^{-1}. This problem can be cast into "one-sample mean problem".

• When $\dim(\boldsymbol{\beta}_1)$ is large, while $\dim(\boldsymbol{\beta}_2)$ is finite and fixed. To get motivation, let us consider $\dim(\boldsymbol{\beta}_2) = 1$. Thus, $\Sigma_{11}^{-1}\Sigma_{12}$ is a column vector, denoted by $\boldsymbol{\alpha}$

• **Challenge**: Σ_{11}^{-1} may not be estimated well.

• **Key observation**: $\boldsymbol{\alpha}$ is the regression coefficient of x_p on $(x_1, \cdots, x_{p-1})^T$.

• Instead of estimating Σ_{11}^{-1} and Σ_{12} separately, we estimate $\boldsymbol{\alpha}$ directly.

How to estimate $\boldsymbol{\alpha}$?

You are expert now!

HD regression beyond linear models

Objective function: $\ell(\boldsymbol{\beta})$ such as negative log-likelihood or loss function

M-estimator:

$$\widehat{\boldsymbol{\beta}} = \mathrm{argmin}_{\boldsymbol{\beta}} \ell(\boldsymbol{\beta})$$

Of interest is to test

$$H_0 : \boldsymbol{\beta} = \boldsymbol{\beta}_0 \quad \text{versus} \quad H_1 : \boldsymbol{\beta} \neq \boldsymbol{\beta}_0$$

Consider score function $\mathbf{s}(\boldsymbol{\beta}) = \partial \ell(\boldsymbol{\beta})/\partial \boldsymbol{\beta}$.

$$\mathbf{s}(\boldsymbol{\beta}) = \mathbf{s}(\boldsymbol{\beta}_0) + [\partial^2 \ell(\boldsymbol{\beta}^*)/\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T](\boldsymbol{\beta} - \boldsymbol{\beta}_0)$$

where $\boldsymbol{\beta}^*$ lies between $\boldsymbol{\beta}$ and $\boldsymbol{\beta}_0$.
Under $H_0 : \boldsymbol{\beta} = \boldsymbol{\beta}_0$, $E\mathbf{s}(\boldsymbol{\beta}_0) = 0$. Thus, if $E[\partial^2 \ell(\boldsymbol{\beta}^*)/\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T]$ is positive definite, we can cast $H_0 : \boldsymbol{\beta} = \boldsymbol{\beta}_0$ into $H_0 : E\mathbf{s}(\boldsymbol{\beta}) = \mathbf{0}$.

Denote $\mathbf{z}_i = \mathbf{s}_i(\boldsymbol{\beta}_0)$, where $\mathbf{s}(\boldsymbol{\beta}_0) = \sum_{i=1}^n \mathbf{s}_i(\boldsymbol{\beta}_0)$. We are testing $H_0 : \boldsymbol{\mu}_z = E(\mathbf{z}) = \mathbf{0}$. Again this is one-sample mean problem.

Let $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^T, \boldsymbol{\beta}_2^T)^T$. How to test $H_0 : \boldsymbol{\beta}_2 = \mathbf{0}$?

We may extend the ideas for HD linear models.

Ning and Liu (2017, AOS) consider dim($\boldsymbol{\beta}_2$)=1 and used Dantzig selector to estimate $\boldsymbol{\alpha}$, and call their method as decorrected score test.

5.5 Confidence intervals for HD regression

Let us start with HD linear model

$$\ell(\boldsymbol{\beta}) = \frac{1}{2}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2$$

Score function

$$\mathbf{s}(\boldsymbol{\beta}) = \mathbf{X}^T(\mathbf{X}\boldsymbol{\beta} - \mathbf{y}) = \sum_{i=1}^{n} \mathbf{x}_i(\mathbf{x}_i^T\boldsymbol{\beta} - y_i).$$

and Hessian matrix is $\mathbf{X}^T\mathbf{X}$. What is the asymptotical distribution of $\mathbf{s}(\boldsymbol{\beta})/\sqrt{n}$ if the true value of $\boldsymbol{\beta}$ is $\boldsymbol{\beta}_0$.

$$\frac{1}{\sqrt{n}}\mathbf{s}(\boldsymbol{\beta}) = \frac{1}{n}\mathbf{X}^T\mathbf{X}\{\sqrt{n}(\boldsymbol{\beta} - \boldsymbol{\beta}_0)\} + \frac{1}{\sqrt{n}}\mathbf{X}\varepsilon$$

So its asymptotic mean is $\sqrt{n}\Sigma(\boldsymbol{\beta} - \boldsymbol{\beta}_0)$

How to construct a confidence interval for β_p? That is, $\boldsymbol{\beta}_2$ when its dim equals 1.

Using decorrelation strategy, we have

$$s_p(\boldsymbol{\beta}) - \Sigma_{21}\Sigma_{11}^{-1}\mathbf{s}_1(\boldsymbol{\beta})$$

follow an asymptotic normal distribution with mean $\sqrt{n}\sigma_{22.1}(\beta_p - \beta_{p0})$

In practice, we replace $\boldsymbol{\beta}$ by a penalized estimator, and $\boldsymbol{\alpha} = \Sigma_{11}^{-1}\Sigma_{12}$ by an penalized estimate of regression coefficient of x_p on $x_1, \cdots, x_{p-1}$.

Ning and Liu (2017, AoS): $\widehat{\boldsymbol{\beta}}$ is the SCAD or MCP estimator and $\widehat{\boldsymbol{\alpha}}$ is the Dantzig type estimate.

Zhang and Zhang (2014, JRSSB) used similar strategy but focus on LASSO.