On corrections of classical multivariate tests for high-dimensional data

Jian-feng YAO

UNIVERSITÉ DE RENNES 1

with

Zhidong BAI, Dandan JIANG, Shurong ZHENG

Overview

High dimensional data

High dimensional data $\neq$ high dimensional models

- Nonparametric regression: a very high-dimensional model (i.e. infinite dimensional model) but with one-dimensional data :

$$y_i = f(x_i) + \varepsilon_i, \qquad f : \mathbb{R} \mapsto \mathbb{R}, \qquad i = 1, \ldots, n$$

- High-dimensional data : observation vectors $y_i \in \mathbb{R}^p$, with p relatively high w.r.t. the sample size n

High dimensional data

Some typical data dimensions :

data ratio n/p

	data dimension p	sample size n	n/p
portfolio	~ 50	500	10
climate survey	320	600	1.9
speech analysis	$a \cdot 10^2$	$b \cdot 10^2$	~ 1
ORL face data base	1440	320	1.2
micro-arrays	2000	200	0.1

- Important: data ratio n/p not always large ; could be $\ll 1$
- Note: use of the Inverse data ratio: $y = p/n$

High-dimensional effect by an example

The two-sample problem:

- two independent samples:
$$\mathbf{x}_1, \ldots, \mathbf{x}_{n_1} \sim (\boldsymbol{\mu}_1, \Sigma), \qquad \mathbf{y}_1, \ldots, \mathbf{y}_{n_2} \sim (\boldsymbol{\mu}_2, \Sigma)$$
- want to test H_0: $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$ against H_1: $\boldsymbol{\mu}_1 \neq \boldsymbol{\mu}_2$.
- Classical approach: Hotelling's T^2 test

$$T^2 = \frac{n_1 n_2}{n} (\bar{\mathbf{x}} - \bar{\mathbf{y}})' S_n^{-1} (\bar{\mathbf{x}} - \bar{\mathbf{y}}),$$

where

$$\bar{\mathbf{x}} = \sum_{i=1}^{n_1} x_i, \qquad \bar{\mathbf{y}} = \sum_{j=1}^{n_2} y_j, \qquad n = n_1 + n_2,$$

$$S_n = \frac{1}{n-2} \left[\sum_{i=1}^{n_1} (x_i - \bar{\mathbf{x}}_i)(x_i - \bar{\mathbf{x}}_i)' + \sum_{j=1}^{n_2} (y_j - \bar{\mathbf{y}}_j)(y_j - \bar{\mathbf{y}}_j)' \right].$$

S_n: a sample covariance matrix

The two-sample problem:

Hotelling's T^2 test: nice properties

- invariance under linear transformations;
- finite-sample optimality if Gaussian; asymptotic optimality otherwise.

Hotelling's T^2 test: bad news

- low power even for moderate data dimensions;
- high instability in computing S_n^{-1} even for $p = 40$;
- very few is known for the non Gaussian case;
- fatal deficiency: when $p > n - 2$, S_n is not invertible.

Dempster's non-exact test (NET)

Dempster A.P., '58, '60

- A reasonable test must be based on $\bar{\mathbf{x}} - \bar{\mathbf{y}}$ even when $p > n-2$;
- choose a new basis in $\mathbb{R}^n$, project the data such that
  1. axis 1 $\|$ Ground mean: $(n_1\boldsymbol{\mu}_1 + n_2\boldsymbol{\mu}_2)/n$
  2. axis 2 $\|$ $(\bar{\mathbf{x}} - \bar{\mathbf{y}})$.
- let the data matrix $\underset{n\times p}{\mathbf{X}} = (\mathbf{x}_1, \ldots, \mathbf{x}_{n_1}, \mathbf{y}_1, \ldots, \mathbf{y}_{n_2})'$, and the (orthonormal) base change $\mathbf{H_n}$:

$$\underset{n\times p}{\mathbf{Z}} = \underset{n\times n}{\mathbf{H_n}}\mathbf{X} = \begin{pmatrix} h_1' \\ \vdots \\ h_n' \end{pmatrix}\mathbf{X} = \begin{pmatrix} \mathbf{z}_1' \\ \vdots \\ \mathbf{z}_n' \end{pmatrix}, \qquad h_1 = \frac{1}{\sqrt{n}}\mathbf{1}_n, \quad h_2 = \begin{pmatrix} \frac{n_2}{\sqrt{nn_1}}\mathbf{1}_{n_1} \\ -\frac{n_1}{\sqrt{nn_2}}\mathbf{1}_{n_2} \end{pmatrix}.$$

Under normality, we have:

- the z_i's are n independent $\mathcal{N}_p(*, \Sigma)$;
- $\mathbb{E}z_1 = \frac{1}{\sqrt{n}}(n_1\boldsymbol{\mu}_1 + n_2\boldsymbol{\mu}_2)$, $\qquad \mathbb{E}z_2 = \frac{n_1 n_2}{\sqrt{n}}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$,
  $\mathbb{E}z_3 = 0$, $\quad i = 3, \ldots, n.$

Dempster's non-exact test (NET)

Test statistic:

- $F_D = (n-2)\dfrac{\|z_2\|^2}{\|z_3\|^2 + \cdots + \|z_n\|^2}$
- Under H_0,

$$\|z_j\|^2 \sim Q := \sum_{k=1}^{r} \alpha_k \chi_1^2(k) \ ,$$

  where $\alpha_1 \geq \cdots \alpha_r > 0$ are the non null eigenvalues of Σ.
- The distribution of F_D is complicated
- Approximations - so the NET test : think as $\Sigma = I_p$,
  1. $Q \simeq m\chi_r^2$;
  2. next estimate r by $\hat{r}$;
- Finally, under H_0, $F_D \simeq F(\hat{r}, (n-2)\hat{r})$.

Dempster's non-exact test (NET)

Problems with the NET test:

- Difficult to construct the orthogonal transformation $\mathbf{H_n} = \{h_j\}$ for large n;
- even under Gaussianity, the exact power function depend on $\mathbf{H_n}$.

Bai and Saranadasa's test (ANT)

Bai & Saranadasa, '96

- Consider directly the statistic $M_n = \|\bar{\mathbf{x}} - \bar{\mathbf{y}}\|^2 - \frac{n}{n_1 n_2} \operatorname{tr} S_n$;
- generally under very mild conditions (here RMT comes!),

$$\frac{M_n}{\sigma_n^2} \Longrightarrow \mathcal{N}(0,1) , \qquad \sigma_n^2 := \operatorname{Var}(M_n) = \frac{n^2}{n_1^2 n_2^2} \frac{n-1}{n-2} \operatorname{tr} \Sigma^2 \quad .$$

- A ratio consistent estimator:

$$\widehat{\sigma}_n^2 = \frac{2n(n-1)(n-2)}{n_1 n_2 (n-3)} \left[\operatorname{tr} S_n^2 - \frac{1}{n-2} (\operatorname{tr} S_n)^2 \right] , \quad \widehat{\sigma}_n^2 / \sigma_n^2 \xrightarrow{P} 1.$$

- Finally, under H_0,

$$Z_n = \frac{M_n}{\widehat{\sigma}_n^2} \Longrightarrow \mathcal{N}(0,1)$$

This is the Bai-Saranadasa's asymptotic normal test (ANT).

Comparison between T^2, NET and ANT

Power functions:

- Assuming $p \to \infty$, $n \to \infty$, $p/n \to y \in (0,1)$, $n_1/n \to \kappa$;
- Hotelling's T^2, Dempster's NET and Bai-Saranadasa's ANT:

$$\beta_H(\boldsymbol{\mu}) = \Phi\left(-\xi_\alpha + \sqrt{\frac{n(1-y)}{2y}}\kappa(1-\kappa)\|\Sigma^{-1/2}\boldsymbol{\mu}\|^2\right) + o(1) ,$$

$$\beta_D(\boldsymbol{\mu}) = \Phi\left(-\xi_\alpha + \frac{n}{\sqrt{2\,\mathrm{tr}\,\Sigma^2}}\kappa(1-\kappa)\|\boldsymbol{\mu}\|^2\right) + o(1) = \beta_{BS}(\boldsymbol{\mu}) .$$

where α = test size, and

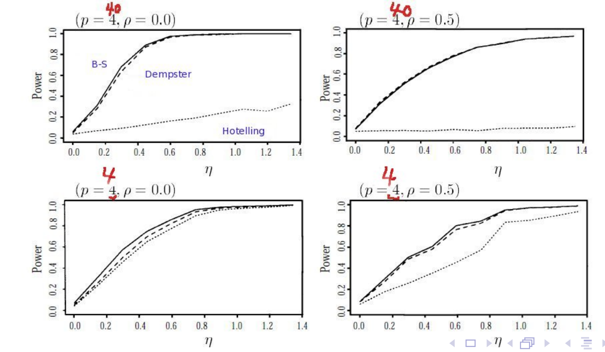
$$\boldsymbol{\mu} = \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2, \qquad \xi_\alpha = \Phi^{-1}(1-\alpha) .$$

- Important: because of the factor $(1-y)$, T^2 losses power when y increases, i.e. p increases relatively to n.

Comparison between T^2, NET and ANT

Simulation results 1: Gaussian case

- Choice of covariance: $\Sigma = (1-\rho)I_p + \rho J_p, \; J_p = \mathbf{1}_p\mathbf{1}_p'$
- noncentral parameter $\eta = \dfrac{\|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|^2}{\sqrt{\operatorname{tr}\Sigma^2}}$, $\quad (n_1, n_2) = (25, 20), \; n = 45$



A summary of the introduction

- High-dimensional effect need to be taken into account ;
- Surprisingly, asymptotic methods with RMT perform well even for small p (as low as $p = 4$) ;
- many of classical multivariate analysis methods have to be examined with respect to high-dimensional effects.

The Marčenko-Pastur distribution

Theorem. Assume : Marčenko & Pastur, 1967

- $\mathbf{X} = p \times n$ i.i.d. variables $(0, 1)$, $\Sigma = I_p$
- not necessarily Gaussian, but with finite 4-th moment
- $p \to \infty$, $n \to \infty$, $p/n \to y \in (0, 1]$

Then, the (empirical) distribution of the eigenvalues of $S_n = \frac{1}{n}\mathbf{X}\mathbf{X}^T$ converges to the distribution with density function

$$f(x) = \frac{1}{2\pi yx}\sqrt{(x-a)(b-x)}, \quad a \le x \le b,$$

where

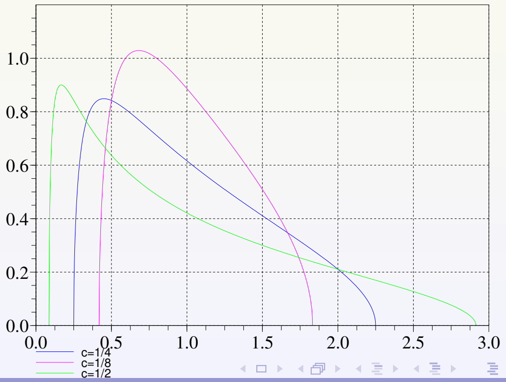
$$a = (1-\sqrt{y})^2, \quad b = (1+\sqrt{y})^2 \ .$$

The Marčenko-Pastur distribution

$$f(x) = \frac{1}{2\pi y x}\sqrt{(x-a)(b-x)}, \qquad (1-\sqrt{y})^2 = a \le x \le b = (1+\sqrt{y})^2.$$

$y \sim p/n$	a	b
1/8	0.42	1.83
1/4	0.25	2.25
1/2	0.09	2.91

Densités de la loi de Marcenko–Pastur



An explanation of the power deficiency of Hotelling's T^2

- when p increases, even in Gaussian case, S_n is different from its population counterpart Σ ;
- when $y = p/n \sim 1$, the left edge $a \sim 0$: small eigenvalues yield an instability of the T^2 statistic:

$$T^2 = \frac{n_1 n_2}{n} (\overline{\mathbf{x}} - \overline{\mathbf{y}})' S_n^{-1} (\overline{\mathbf{x}} - \overline{\mathbf{y}}) \ .$$

Bai and Silverstein's CLT for linear spectral statistics of S_n

Set

- the Empirical spectral distribution:

$$F_n = \frac{1}{p} \sum_{j=1}^{p} \delta_{\lambda_j},$$

  where λ_j's are p eigenvalues of S_n;
- $y_n = \frac{p}{n}$;
- $[a, b] \subset \mathcal{U} \text{ open } \subset \mathbb{C}$.
- for any g analytic on $\mathcal{U}$

$$G_n(g) = p \left[F_n(g) - \mu^{y_n}(g) \right]$$

  where μ^{α} is the MP distribution of index $\alpha \in (0, 1)$.

A CLT for linear spectral statistics

Bai and Silverstein, '04

Theorem

Assume that

- $g_1, \cdots, g_k$ *are* k *analytic functions on* $\mathcal{U}$;
- *the matrix entries* x_{ij} *are i.i.d. real-valued random variables such that* $Ex_{ij} = 0$, $Ex_{ij}^2 = 1$, $E{x_{ij}}^4 = 3$.
- *as* $n, p \to \infty$, $y_n = \dfrac{p}{n} \to y \in (0, 1)$;

Then,

$$(G_n(g_1), \cdots, G_n(g_k)) \Rightarrow \mathcal{N}_k(m, V),$$

with a given mean vector $m = m(g_1, \ldots, g_k)$ *and asymptotic covariance matrix* $V = V(g_1, \ldots, g_k)$.

Other versions exist: Lytova & Pastur '09; Bai & Wang '09

Random Fisher matrices

- two independent samples:
$$\mathbf{x}_1, \ldots, \mathbf{x}_{n_1} \sim (0, I_p), \qquad \mathbf{y}_1, \ldots, \mathbf{y}_{n_2} \sim (0, I_p)$$
with i.i.d coordinates of mean 0 and variance 1
- Associated sample covariance matrices:
$$S_1 = \frac{1}{n_1} \sum_{i=1}^{n_1} \mathbf{x}_i \mathbf{x}_i^*, \qquad S_2 = \frac{1}{n_2} \sum_{j=1}^{n_2} \mathbf{y}_j \mathbf{y}_j^*.$$
- Fisher matrix: $V_n = S_1 S_2^{-1}$ where $n_2 > p$.

Random Fisher matrices

- Assume

$$y_{n_1} = \frac{p}{n_1} \to y_1 \in (0,1), \qquad y_{n_2} = \frac{p}{n_2} \to y_2 \in (0,1) \ .$$

- Under mild moment conditions, the ESD $F_n^{V_n}$ of V_n has a LSD F_{y_1,y_2} with density:

$$\ell(x) = \begin{cases} \frac{(1-y_2)\sqrt{(b-x)(x-a)}}{2\pi x(y_1+y_2x)}, & a \le x \le b, \\ 0, & \text{otherwise} \end{cases}$$

where

$$a = (1-y_2)^{-2}\left(1-\sqrt{y_1+y_2-y_1y_2}\right)^2 , \quad b = (1-y_2)^{-2}\left(1+\sqrt{y_1+y_2-y_1y_2}\right)^2 .$$

CLT for LSS of random Fisher matrices

- let

$$\left[I_{(0,1)}(y_1)\frac{(1-\sqrt{y_1})^2}{(1+\sqrt{y_2})^2}, \quad \frac{(1+\sqrt{y_1})^2}{(1-\sqrt{y_2})^2} \right] \subset \widetilde{\mathcal{U}} \ \text{ open } \ \subset \mathbb{C},$$

- for an analytic function f on $\widetilde{\mathcal{U}}$, define

$$\widetilde{G_n}(f) = p \cdot \int_{-\infty}^{+\infty} f(x) \left[F_n^{V_n} - F_{y_{n_1}, y_{n_2}} \right] (dx) \ ,$$

where $F_{y_{n_1}, y_{n_2}}$ is the LSD with indexes y_{n_k}, $k = 1, 2$.

CLT for LSS of random Fisher matrices

Zheng, '08

Theorem

Assume $\mathbb{E}\mathbf{x}_{11}^4 < \infty$, $\mathbb{E}\mathbf{y}_{11}^4 < \infty$ *and let*

$$\beta_x = E|\mathbf{x}_{11}|^4 - 3, \quad \beta_y = E|\mathbf{y}_{11}|^4 - 3.$$

Then for any analytic functions $f_1, \cdots, f_k$ *defined on* $\widetilde{\mathcal{U}}$,

$$\left[\widetilde{G_n}(f_1), \cdots, \widetilde{G_n}(f_k)\right] \Longrightarrow \mathcal{N}_k(m, \upsilon)$$

with suitable asymptotic mean and covariance functions m *and* υ.

CLT for LSS of random Fisher matrices

Zheng, '08

Limiting mean function m

$$m(f_j) = \lim_{r \to 1_+} [(1) + (2) + (3)]$$

$$\frac{1}{4\pi i} \oint_{|\zeta|=1} f_j(z(\zeta)) \left[\frac{1}{\zeta - \frac{1}{r}} + \frac{1}{\zeta + \frac{1}{r}} - \frac{2}{\zeta + \frac{y_2}{hr}} \right] d\zeta \qquad (1)$$

$$+ \frac{\beta_x \cdot y_1 (1-y_2)^2}{2\pi i \cdot h^2} \oint_{|\zeta|=1} f_j(z(\zeta)) \frac{1}{(\zeta + \frac{y_2}{hr})^3} d\zeta \qquad (2)$$

$$+ \frac{\beta_y \cdot y_2 (1-y_2)}{2\pi i \cdot h} \oint_{|\zeta|=1} f_j(z(\zeta)) \frac{\zeta + \frac{1}{hr}}{(\zeta + \frac{y_2}{hr})^3} d\zeta, \qquad (3)$$

where

$$z(\zeta) = (1-y_2)^{-2} \left[1 + h^2 + 2h\mathcal{R}(\zeta) \right], \qquad h = \sqrt{y_1 + y_2 - y_1 y_2}.$$

CLT for LSS of random Fisher matrices

Zheng, '08

Limiting covariance function υ

$$\upsilon(f_j, f_\ell) = \lim_{1<r_1<r_2\to 1_+} [(4) + (5))]$$

$$-\frac{1}{2\pi^2}\oint_{|\zeta_2|=1}\oint_{|\zeta_1|=1} \frac{f_j(z(r_1\zeta_1))f_\ell(z(r_2\zeta_2))r_1r_2}{(r_2\zeta_2 - r_1\zeta_1)^2} d\zeta_1 d\zeta_2, \quad (4)$$

$$-\frac{(\beta_x y_1 + \beta_y y_2)(1-y_2)^2}{4\pi^2 h^2}\oint_{|\zeta_1|=1} \frac{f_j\left(z(\zeta_1)\right)}{(\zeta_1 + \frac{y_2}{hr_1})^2} d\zeta_1 \oint_{|\zeta_2|=1} \frac{f_\ell\left(z(\zeta_2)\right)}{(\zeta_2 + \frac{y_2}{hr_2})^2} d\zeta_2 \quad (5)$$

$$j, \ell \in \{1, \cdots, k\}.$$

One-sample test on covariance matrices

- a sample $\mathbf{x}_1, \ldots, \mathbf{x}_n) \sim \mathcal{N}_p(\mu, \Sigma)$
- want to test $H_0 : \ \Sigma = I_p$
- in high-dimensional case, several previous work exist:
  Ledoit & Wolf '02; Schott '07; Srivastava '05 ...
- we focus on the LR statistic:

$$T_n = n\left[trS_n - \log|S_n| - p\right], \quad S_n = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})',$$

Classical LRT:

- Data dimension p is fixed, and when $n \to \infty$, $T_n \Longrightarrow \chi^2_{p(p+1)/2}$.
- Will see: rapidly deficient when p is not "small".

RMT Corrected LRT:

Bai, Jiang, Y and Zheng '09

Theorem

Assume $p/n \to y \in (0,1)$ *and let* $g(x) = x - \log x - 1$. *Then, under* H_0 *and when* $n \to \infty$

$$\left[\frac{T_n}{n} - p \cdot F^{y_n}(g)\right] \Rightarrow \mathcal{N}(m(g), v(g)),$$

where F^{y_n} *is the Marčenko-Pastur law of index* y_n *and*

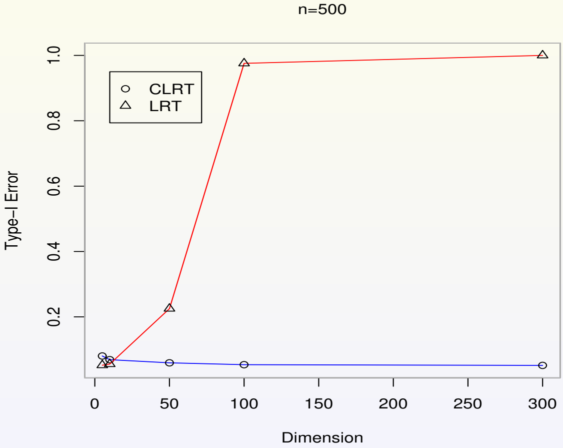
$$m(g) = -\frac{\log(1-y)}{2},$$
$$v(g) = -2\log(1-y) - 2y.$$

Comparison of LRT and CLRT by simulation

- nominal test level $\alpha = 0.05$;
- for each (p, n), 10,000 independent replications with real Gaussian variables.
- Powers are estimated under the alternative H_1: $\Sigma = \mathrm{diag}(1, 0.05, 0.05, 0.05, \ldots, 0.05)$.

	CLRT			LRT	
(p, n)	Size	Difference with 5%	Power	Size	Power
(5, 500)	0.0803	0.0303	0.6013	0.0521	0.5233
(10, 500)	0.0690	0.0190	0.9517	0.0555	0.9417
(50, 500)	0.0594	0.0094	1	0.2252	1
(100, 500)	0.0537	0.0037	1	0.9757	1
(300, 500)	0.0515	0.0015	1	1	1

On a plot


n=500
CLRT
LRT
Type−I Error
1.0
0.8
0.6
0.4
0.2
0
50
100
150
200
250
300
Dimension

Two-samples test on covariance matrices

- two samples
  $\mathbf{x}_1, \ldots, \mathbf{x}_{n_1} \sim \mathcal{N}_p(\mu_1, \Sigma_1)$, $\mathbf{y}_1, \ldots, \mathbf{y}_{n_2} \sim \mathcal{N}_p(\mu_2, \Sigma_2)$
- want to test $H_0 : \Sigma_1 = \Sigma_2$
- The associated sample covariance matrices are

$$S_1 = \frac{1}{n_1} \sum_{i=1}^{n_1} (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})', \quad S_2 = \frac{1}{n_2} \sum_{i=1}^{n_2} (\mathbf{y}_i - \bar{\mathbf{y}})(\mathbf{y}_i - \bar{\mathbf{y}})',$$

- Let the LR statistic

$$L_1 = \frac{\left|S_1 S_2^{-1}\right|^{\frac{n_1}{2}}}{\left|c_1 S_1 S_2^{-1} + c_2 I_p\right|^{\frac{n}{2}}},$$

where $n = n_1 + n_2$ and $c_k = \frac{n_k}{n}, k = 1, 2.$

Two-samples test on covariance matrices

Classical LRT:

- Data dimension p is fixed, and when $n_1, n_2 \to \infty$ and under H_0 ,

$$T_n = -2 \log L_1 \Rightarrow \chi^2_{p(p+1)/2} \ .$$

- Will see: rapidly deficient when p is not "small".

RMT Corrected LRT:

Bai, Jiang, Y and Zheng '09

Theorem

Assuming that the conditions of CLT for LSS of Fisher matrices hold and let

$$f(x) = \log(y_1 + y_2 x) - \frac{y_2}{y_1 + y_2} \log x - \log(y_1 + y_2).$$

Then under H_0 *and as* $n_1 \wedge n_2 \to \infty$,

$$\left[-\frac{2 \log L_1}{n} - p \cdot F_{y_{n_1}, y_{n_2}}(f) \right] \Rightarrow \mathcal{N}\left(m(f), \upsilon(f)\right),$$

with

$$m(f) = \frac{1}{2} \left[\log\left(\frac{y_1 + y_2 - y_1 y_2}{y_1 + y_2} \right) - \frac{y_1}{y_1 + y_2} \log(1 - y_2) - \frac{y_2}{y_1 + y_2} \log(1 - y_1) \right],$$

$$\upsilon(f) = -\frac{2y_2^2}{(y_1 + y_2)^2} \log(1 - y_1) - \frac{2y_1^2}{(y_1 + y_2)^2} \log(1 - y_2) - 2 \log \frac{y_1 + y_2}{y_1 + y_2 - y_1 y_2}.$$

Comparison of LRT and CLRT by simulation

- nominal test level $\alpha = 0.05$;
- for each (p, n_1, n_2), 10,000 independent replications with real Gaussian variables.
- Powers are estimated under the alternative H_1: $\Sigma_1 \Sigma_2^{-1} = \mathrm{diag}(3, 1, 1, \cdots ,)$.

Comparison of LRT and CLRT by simulation

with $(y_1, y_2) = (0.05, 0.05)$:

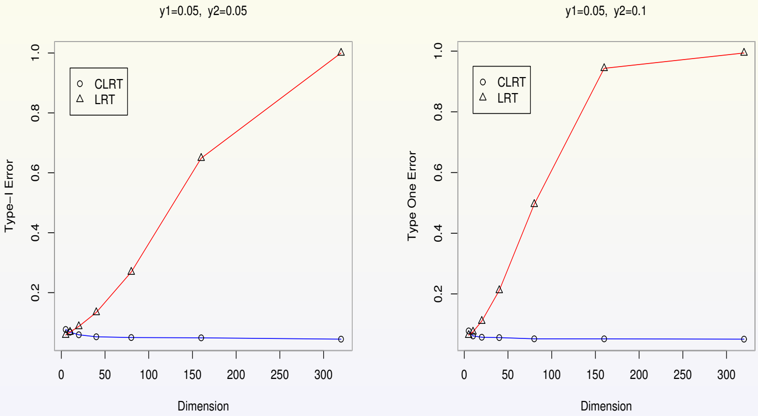
	CLRT			LRT	
(p, n_1, n_2)	Size	Difference with 5%	Power	Size	Power
(5, 100, 100)	0.0770	0.0270	1	0.0582	1
(10, 200, 200)	0.0680	0.0180	1	0.0684	1
(20, 400, 400)	0.0593	0.0093	1	0.0872	1
(40, 800, 800)	0.0526	0.0026	1	0.1339	1
(80, 1600, 1600)	0.0501	0.0001	1	0.2687	1
(160, 3200, 3200)	0.0491	-0.0009	1	0.6488	1
(320, 6400, 6400)	0.0447	-0.0053	0.9671	1	1

Comparison of LRT and CLRT by simulation

with $(y_1, y_2) = (0.05, 0.1)$:

	CLRT			LRT	
(p, n_1, n_2)	Size	Difference with 5%	Power	Size	Power
(5, 100, 50)	0.0781	0.0281	0.9925	0.0640	0.9849
(10, 200, 100)	0.0617	0.0117	0.9847	0.0752	0.9904
(20, 400, 200)	0.0573	0.0073	0.9775	0.1104	0.9938
(40, 800, 400)	0.0561	0.0061	0.9765	0.2115	0.9975
(80, 1600, 800)	0.0521	0.0021	0.9702	0.4954	0.9998
(160, 3200, 1600)	0.0520	0.0020	0.9702	0.9433	1
(320, 6400, 3200)	0.0510	0.0010	1	0.9939	1

Comparisons of LRT and CLRT



A general linear hypothesis in a multivariate regression

A p-th dimensional regression model:

$$\mathbf{x}_i = \mathbf{B}\mathbf{z}_i + \varepsilon_i, \quad i = 1, \ldots, n$$

where

$$\varepsilon_i \sim \mathcal{N}_p(0, \mathbf{\Sigma}), \quad \mathbf{x} \in \mathbb{R}^p, \quad \mathbf{z}_i \in \mathbb{R}^q, \quad n \geq p + q.$$

A general linear hypothesis:

- Write a bloc decomposition $\mathbf{B} = (\mathbf{B}_1, \mathbf{B}_2)$ with q_1 and q_2 columns
- To test

$$H_0 : \ \mathbf{B}_1 = \mathbf{B}_1^* \ ,$$

with a given $\mathbf{B}_1^*$.

Wilk's $\mathbf{\Lambda}$

- Let $\widehat{\mathbf{\Sigma}}_0$ and $\widehat{\mathbf{\Sigma}}_1$ be the likelihood "estimator" of $\mathbf{\Sigma}$ under H_0 and the alternative, respectively
- LRT statistic equals

$$\mathscr{L}_0/\mathscr{L}_1 = (\mathbf{\Lambda}_n)^{n/2}, \quad \mathbf{\Lambda}_n = \frac{|\widehat{\mathbf{\Sigma}}|}{|\widehat{\mathbf{\Sigma}}_0|} ,$$

  where $\mathbf{\Lambda}_n$ is the celebrated Wilk's $\mathbf{\Lambda}$: Wilks '32, '34 ; Bartlett '34.
- Classic (low dimensional) approximation of LRT: for fixed p and q, $n \to \infty$ and under H_0:

$$U_n = -n \log \mathbf{\Lambda}_n \Rightarrow \chi^2_{pq_1} .$$

- Less biased Bartlett's correction:

$$\tilde{U}_n = -k \log \mathbf{\Lambda}_n, \quad k = n - q - \frac{1}{2}(p - q_1 + 1) .$$

High-dimensional correction of Wilk's Λ

Bai, Jiang, Y and Zheng, '10

Theorem

Let $p \to \infty,\ q_1 \to \infty,\ n - q \to \infty$ and

$$y_{n_1} = \frac{p}{q_1} \to y_1 \in (0,1), \quad y_{n_2} = \frac{p}{n-q} \to y_2 \in (0,1).$$

Then, under H_0,

$$T_n = \upsilon(f)^{-\frac{1}{2}} \left[-\log \Lambda_n - p \cdot F_{y_{n_1}, y_{n_2}}(f) - m(f) \right] \Rightarrow \mathcal{N}(0,1),$$

where $m(f), \upsilon(f)$ and $F_{y_{n_1}, y_{n_2}}(f)$ are suitable constants computed from

$$f(x) = \log(1 + \frac{y_{n_2}}{y_{n_1}} x) .$$

The centering term:

$$
\begin{aligned}
F_{y_{n_1},y_{n_2}}(f) &= \frac{y_{n_2}-1}{y_{n_2}}\log c_n + \frac{y_{n_1}-1}{y_{n_1}}\log(c_n - d_n h_n) \\
&= +\frac{y_{n_1}+y_{n_2}}{y_{n_1}y_{n_2}}\log\left(\frac{c_n h_n - d_n y_{n_2}}{h_n}\right),
\end{aligned}
$$

where

$$
\begin{aligned}
h_n &= \sqrt{y_{n_1}+y_{n_2}-y_{n_1}y_{n_2}} \\
a_n, b_n &= \frac{(1 \mp h_n)^2}{(1-y_{n_2})^2} \\
c_n, d_n &= \frac{1}{2}\left[\sqrt{1+\frac{y_{n_2}}{y_{n_1}}b_n} \pm \sqrt{1+\frac{y_{n_2}}{y_{n_1}}a_n}\right], c_n > d_n,
\end{aligned}
$$

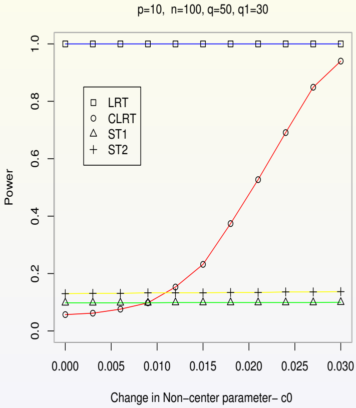
The limiting parameters:

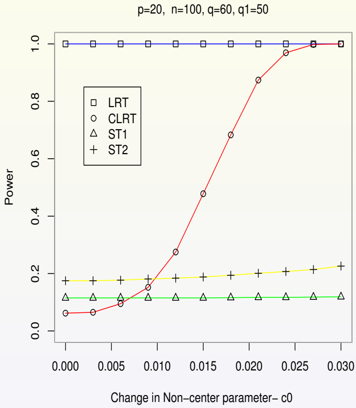
$$m(f) = \frac{1}{2}\log\frac{(c^2-d^2)h^2}{(ch-y_2d)^2},$$
$$\upsilon(f) = 2\log\left(\frac{c^2}{c^2-d^2}\right),$$

where

$$h = \sqrt{y_1+y_2-y_1y_2}$$
$$a_0, b_0 = \frac{(1\mp h)^2}{(1-y_2)^2}$$
$$c, d = \frac{1}{2}\left[\sqrt{1+\frac{y_2}{y_1}b_0} \pm \sqrt{1+\frac{y_2}{y_1}a_0}\right], c > d.$$

A simulation experiment





- Gaussian entries,
- non central parameter $c_0 \sim d(H, H_0)$.

Some conclusions:

- High dimensional effects should be taken into account;
- RMT for sample covariance matrices is a powerful tool to correct classical multivariate procedures ;
- Each time some $\boldsymbol{\Sigma}$ is to be estimated, one should take care of the "natural" estimator S_n : for high-dimensional data,

$$S_n \neq \boldsymbol{\Sigma}.$$

- Yet the RMT is not suffciently developped for statistics:
  1. dependent observations: time series ;
  2. not identically distributed variables.

Some references

- Bai, Z. D. and Saranadasa, H. (1996). Effect of high dimension comparison of significance tests for a high dimensional two sample problem. *Statistica Sinica.* **6**, 311-329.
- Bai, Z. D. and Silverstein, J. W. (2004). CLT for linear spectral statistics of large-dimensional sample covariance matrices. *Ann.Probab.* **32**, 553-605.
- Z. D. Bai, D. Jiang, J. Yao and S. Zheng, 2009. Corrections to LRT on Large Dimensional Covariance Matrix by RMT. *Annals of Statistics* **37**, 3822–3840
- Z. D. Bai, D. Jiang, J. Yao and S. Zheng, 2010. Large regression analysis. *submitted*
- Dempster, A. P. (1958). A high dimensional two sample significance test. *Ann. Math. Statist.* **29**, 995-1010.
- Zheng, S. (2008). Central Limit Theorem for Linear Spectral Statistics of Large Dimensional F Matrix. *Preprint, Northern-Est Normal University*