Universal asymptotics for high-dimensional sign tests

Davy Paindaveine[a] Thomas Verdebout[b]

[a] ECARES, Université libre de Bruxelles, Belgium
[b] EQUIPPE, Université Lille 3, France

Brussels, May 2014

Motivation

We consider high-dimensional directional data, that is, with data that live on

$$\mathcal{S}^{p-1} = \left\{ x \in \mathbb{R}^p : \|x\| = \sqrt{x'x} = 1 \right\}, \qquad p \text{ large}.$$

Motivation

We consider high-dimensional directional data, that is, with data that live on

$$\mathcal{S}^{p-1} = \left\{ x \in \mathbb{R}^p : \|x\| = \sqrt{x'x} = 1 \right\}, \qquad p \text{ large}.$$

Why?

Motivation

We consider high-dimensional directional data, that is, with data that live on

$$\mathcal{S}^{p-1} = \left\{ x \in \mathbb{R}^p : \|x\| = \sqrt{x'x} = 1 \right\}, \qquad p \text{ large}.$$

Why?

(i) Because many applications involve such data.

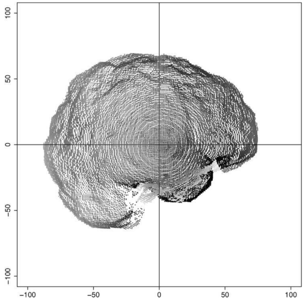
• text analysis: the data may consist of n texts, containing altogether p different words. The observations are then of the form

$$x_i = \begin{pmatrix} x_{i1} \\ \vdots \\ x_{ip} \end{pmatrix}, \quad i = 1, \ldots, n,$$

where x_{ij} is the frequency of the jth word in the ith text. When performing, e.g., clustering, one often replaces x_i with $x_i/\|x_i\|$ so that text length does not play a role; see, among others, Dhillon and Modha (2001), Banerjee et al. (2003).

Motivation

- Dryden (2005) considers an application in brain shape modelling.



Here, observations are of the form

$$x_i = \begin{pmatrix} x_{i1} \\ \vdots \\ x_{ip} \end{pmatrix}, \quad i = 1, \ldots, n,$$

where x_{ij} is the distance between the central landmark and the extremity of brain i in direction j.

- $p = 62{,}501$!
- One is interested in shape, and not in size $\Rightarrow$ one considers $x_i / \|x_i\|$

(ii) Because, in the HD setup, data sometimes automatically become of a directional nature

---

Consider $X \sim \mathcal{N}(0, \frac{1}{p} I_p)$.
Then $\|\sqrt{p}\, X\|^2 \sim \chi^2_p$, hence has mean p and variance $2p$.
Therefore, $\|X\|^2$ has mean 1 and variance $2/p$, so that, as $p \to \infty$,

$$\mathrm{E}[(\|X\|^2 - 1)^2] = \mathrm{Var}[\|X\|^2] \to 0.$$

We conclude that X "eventually belongs" to S^{p-1}.

(ii) Because, in the HD setup, data sometimes automatically become of a directional nature

---

Consider $X \sim \mathcal{N}(0, \frac{1}{p} I_p)$.
Then $\|\sqrt{p}\, X\|^2 \sim \chi^2_p$, hence has mean p and variance $2p$.
Therefore, $\|X\|^2$ has mean 1 and variance $2/p$, so that, as $p \to \infty$,

$$\mathrm{E}[(\|X\|^2 - 1)^2] = \mathrm{Var}[\|X\|^2] \to 0.$$

We conclude that X "eventually belongs" to S^{p-1}.

Remarks:
- This is in line with the fact that $\mathrm{E}[U] = 0$ and $\mathrm{Var}[U] = \frac{1}{p} I_p$ for $U \sim \mathrm{Unif}(\mathcal{S}^{p-1})$

(ii) Because, in the HD setup, data sometimes automatically become of a directional nature

---

Consider $X \sim \mathcal{N}(0, \frac{1}{p} I_p)$.
Then $\|\sqrt{p}\, X\|^2 \sim \chi^2_p$, hence has mean p and variance $2p$.
Therefore, $\|X\|^2$ has mean 1 and variance $2/p$, so that, as $p \to \infty$,

$$\mathrm{E}[(\|X\|^2 - 1)^2] = \mathrm{Var}[\|X\|^2] \to 0.$$

We conclude that X "eventually belongs" to S^{p-1}.

Remarks:

- This is in line with the fact that $\mathrm{E}[U] = 0$ and $\mathrm{Var}[U] = \frac{1}{p} I_p$ for $U \sim \mathrm{Unif}(\mathcal{S}^{p-1})$
- This naturally brings sign tests in the picture...

Two low-dimensional tests

Let $X_1, \ldots, X_n$ be i.i.d. with values in $\mathcal{S}^{p-1}$.
We consider the problem of testing for uniformity on $\mathcal{S}^{p-1}$.

The celebrated Rayleigh test rejects the null at asymptotic level α if

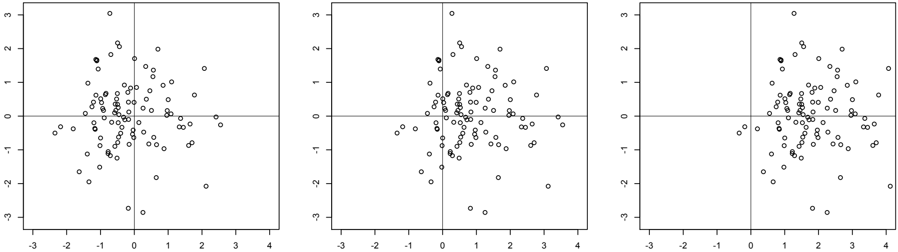
$$T_n = np\|\bar{X}\|^2 = \left(\frac{1}{\sqrt{n}}\sum_{i=1}^n X_i\right)'\left(\frac{1}{p}I_p\right)^{-1}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^n X_i\right) > \chi^2_{p,1-\alpha},$$

where $\chi^2_{d,1-\alpha}$ denotes the $(1-\alpha)$- quantile of the χ^2_d distribution.

In a non-directional framework, this test would be considered as a location test, rejecting the null $\mathcal{H}_0 : \mathrm{E}[X] = 0$ for large values of $\bar{X}$ (this test is valid under sphericity assumptions).

Two low-dimensional tests

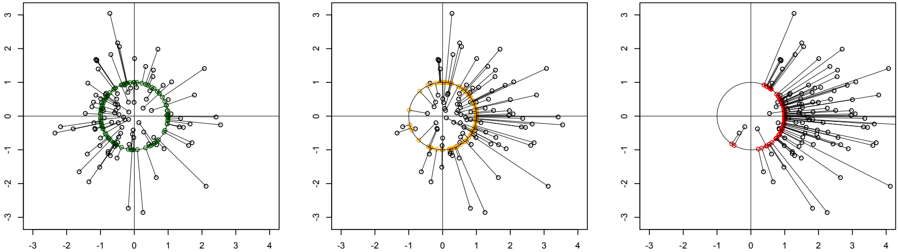
→ Increasingly severe alternatives →



$$X = \frac{Z + \theta}{\|Z + \theta\|}, \text{ with } Z \text{ spherical and } \theta \neq 0$$

Two low-dimensional tests

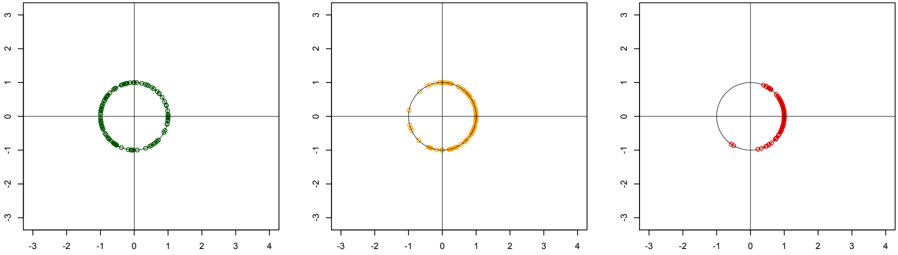
→ Increasingly severe alternatives →



$$X = \frac{Z + \theta}{\|Z + \theta\|}, \text{ with } Z \text{ spherical and } \theta \neq 0$$

Two low-dimensional tests

$\rightarrow$ Increasingly severe alternatives $\rightarrow$



$$X = \frac{Z+\theta}{\|Z+\theta\|}, \text{ with } Z \text{ spherical and } \theta \neq 0$$

Two low-dimensional tests

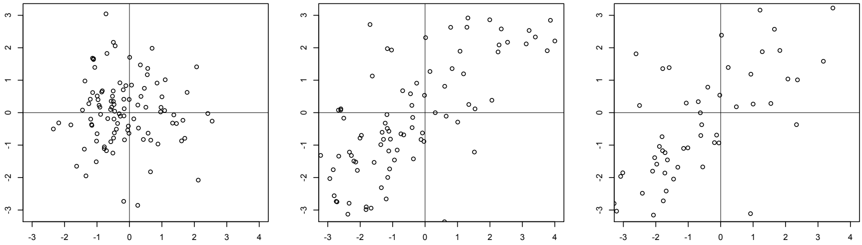
Still for the problem of testing uniformity on $\mathcal{S}^{p-1}$, an alternative test is the Hallin and Paindaveine (2006) test, that rejects the null if

$$\begin{aligned} T_n &= \frac{p(p+2)}{2n} \sum_{i,j=1}^{n} \Big((X_i' X_j)^2 - \frac{1}{p} \Big) > \chi^2_{d(p),1-\alpha} \\ &= \frac{np(p+2)}{2} \Big\| \frac{S}{\mathrm{tr}[S]} - \frac{1}{p} I_p \Big\|^2 \qquad \Big(\text{with } S = \frac{1}{n} \sum_{i=1}^{n} (X_i - 0)(X_i - 0)' \Big), \end{aligned}$$

with $\|A\|^2 = \mathrm{tr}[AA']$ and $d(p) = \frac{p(p+1)}{2} - 1$.

Two low-dimensional tests

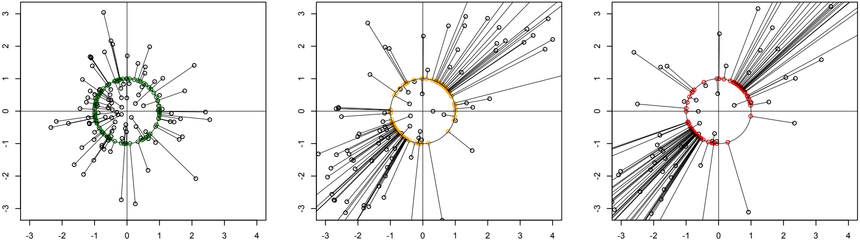
$\rightarrow$ Increasingly severe alternatives $\rightarrow$



$$X = \frac{AZ}{\|AZ\|}, \text{ with } Z \text{ spherical and } A = I_p + \lambda H$$

Two low-dimensional tests

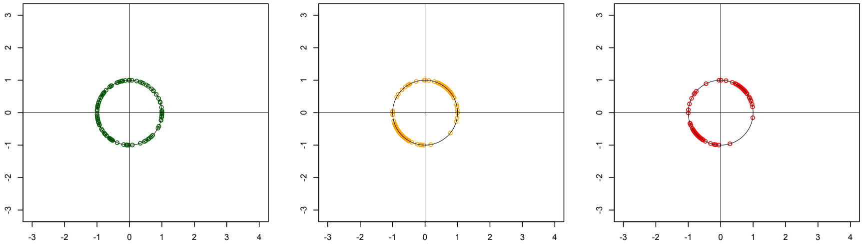
$\rightarrow$ Increasingly severe alternatives $\rightarrow$



$$X = \frac{AZ}{\|AZ\|}, \text{ with } Z \text{ spherical and } A = I_p + \lambda H$$

Two low-dimensional tests

$\rightarrow$ Increasingly severe alternatives $\rightarrow$



$$X = \frac{AZ}{\|AZ\|}, \text{ with } Z \text{ spherical and } A = I_p + \lambda H$$

Two low-dimensional tests

Both tests, and many more from multivariate analysis, rely on a null (fixed-p) asymptotic result of the form

$$T_n \xrightarrow[n\to\infty]{\mathcal{L}} \chi^2_{d(p)},$$

hence lead to rejection (at asymptotic level α) whenever

$$T_n > \chi^2_{d(p),1-\alpha}.$$

Of course, practical implementation of such tests requires $n >> p$.

Such tests therefore are not valid in the HD setup.

(Yet, their fixed-p optimality motivates studying their HD properties...)

The HD case

What would you expect in the HD setup where both $n, p \to \infty$?

Some heuristics...

The HD case

What would you expect in the HD setup where both $n, p \to \infty$?

Some heuristics... Assume that $d(p) \to \infty$ as $p \to \infty$.

- Ok for Rayleigh, $d(p) = p$
- Ok for Hallin and Paindaveine (2006), $d(p) = \frac{p(p+1)}{2} - 1$
- Ok for...

What would you expect in the HD setup where both $n, p \to \infty$?

Some heuristics... Assume that $d(p) \to \infty$ as $p \to \infty$.

Since

$$\frac{\chi^2_d - d}{\sqrt{2d}} = \frac{\chi^2_d - \mathrm{E}[\chi^2_d]}{\sqrt{\mathrm{Var}[\chi^2_d]}} \xrightarrow[d\to\infty]{} \mathcal{N}(0,1),$$

one may then expect that the fixed-p asymptotic result

$$T_n \xrightarrow[n\to\infty]{\mathcal{L}} \chi^2_{d(p)}$$

will lead to a double-asymptotic result of the form

$$T_n^{St} = \frac{T_n - d(p)}{\sqrt{2d(p)}} \xrightarrow[n,p\to\infty]{\mathcal{L}} \mathcal{N}(0,1).$$

Intuitively, this should hold if $p = p_n$ is going to ∞ sufficiently slowly.

The HD case

Natural questions are :

- Is this heuristics valid? That is, is there a $(p = p_n) \to \infty$ such that

$$T_n^{St} = \frac{T_n - d(p)}{\sqrt{2d(p)}} \xrightarrow[n,p\to\infty]{\mathcal{L}} \mathcal{N}(0,1) \text{ ?}$$

- How fast may p_n go to infinity?
- "For (p_n) such that the above convergence holds", which test should be favored for fixed (n, p)?

Test 1: reject at asymptotic level α if $\dfrac{T_n - d(p)}{\sqrt{2d(p)}} > \Phi^{-1}(1-\alpha)$.

Test 2: reject at asymptotic level α if

$$T_n > \chi^2_{d(p),1-\alpha}, \quad \text{or equivalently, if} \quad \frac{T_n - d(p)}{\sqrt{2d(p)}} > \frac{\chi^2_{d(p),1-\alpha} - d(p)}{\sqrt{2d(p)}}.$$

The HD case

To fix ideas, we restrict to Rayleigh's test, with test statistic

$$T_n = np_n \|\bar{X}\|^2 = \frac{p_n}{n} \sum_{i,j=1}^{n} X'_{ni} X_{nj},$$

which can be rewritten as

$$T_n = p_n + \frac{2p_n}{n} \sum_{1 \le i < j \le n}^{n} X'_{ni} X_{nj},$$

so that (recall $d(p) = p$ for Rayleigh's test)

$$T_n^{\mathrm{St}} = \frac{T_n - d(p_n)}{\sqrt{2d(p_n)}} = \frac{T_n - p_n}{\sqrt{2p_n}} = \frac{\sqrt{2p_n}}{n} \sum_{1 \le i < j \le n}^{n} X'_{ni} X_{nj}.$$

This is a U-statistic with an order-2 kernel that depends on $p = p_n$.

The HD case

To study the asymptotic behavior of this U-statistic

$$T_n^{\mathrm{St}} = \frac{\sqrt{2p_n}}{n} \sum_{1 \le i < j \le n}^{n} X'_{ni} X_{nj},$$

the key move is to decompose T_n^{St} into

$$T_n^{\mathrm{St}} = \sum_{\ell=1}^{n} D_{n\ell},$$

where the random variables

$$\begin{aligned} D_{n\ell} &= \mathrm{E}_{n\ell}\left[T_n^{\mathrm{St}}\right] - \mathrm{E}_{n,\ell-1}\left[T_n^{\mathrm{St}}\right] \qquad \ell = 1, \ldots, n \\ &= \frac{\sqrt{2p_n}}{n} \sum_{i=1}^{\ell-1} X'_{ni} X_{n\ell} \end{aligned}$$

form a martingale difference process; here, $\mathrm{E}_{n\ell}[\,\cdot\,]$ denotes expectation with respect to $\sigma(X_1, \ldots, X_\ell)$.

The HD case

We can then rely on a CLT for martingale differences, such as the following.

Theorem (Billingsley (1995), Theorem 35.12)

Let $D_{n\ell}$, $\ell = 1, \ldots, n$, $n = 1, 2, \ldots,$ be a triangular array of random variables such that, for any n, $D_{n1}, D_{n2}, \ldots, D_{nn}$ is a martingale difference sequence with respect to some filtration $\mathcal{F}_{n1}, \mathcal{F}_{n2}, \ldots, \mathcal{F}_{nn}$ (with $\mathcal{F}_{n0} := \{\emptyset, \Omega\}$). Assume that $\mathrm{E}[D_{n\ell}^2] < \infty$ for any n, ℓ ,and that

$$\sum_{\ell=1}^{n} \mathrm{E}\big[D_{n\ell}^2 \,|\, \mathcal{F}_{n,\ell-1}\big] \xrightarrow[n\to\infty]{\mathrm{P}} 1 \tag{1}$$

(where $\xrightarrow{\mathrm{P}}$ denotes convergence in probability), and

$$\sum_{\ell=1}^{n} \mathrm{E}\big[D_{n\ell}^2 \, \mathbb{I}[|D_{n\ell}| > \varepsilon]\big] \xrightarrow[n\to\infty]{} 0. \tag{2}$$

Then $\sum_{\ell=1}^{n} D_{n\ell}$ is asymptotically standard normal.

The HD case

After some work to establish (1)-(2), in the present context, we then obtain

Theorem 1

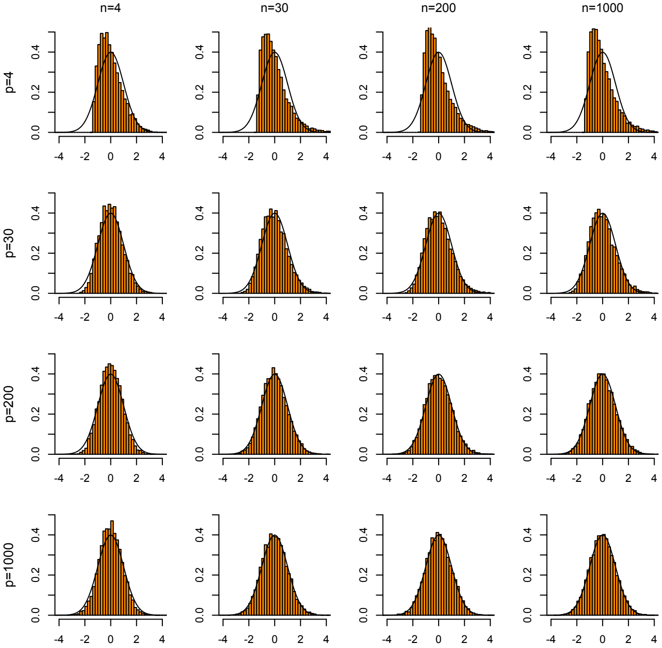
Let p_n be a sequence of positive integers converging to $+\infty$. Assume that X_{ni}, $i = 1, \ldots, n$, $n = 1, 2, \ldots$, is a triangular array such that for any n, the random p_n-vectors X_{ni}, $i = 1, \ldots, n$ are i.i.d. uniform on $\mathcal{S}^{p_n - 1}$. Then

$$T_n^{\mathrm{St}} = \frac{T_n - p_n}{\sqrt{2p_n}} = \frac{\sqrt{2p_n}}{n} \sum_{1 \leq i < j \leq n} X'_{ni} X_{nj} \xrightarrow[n \to \infty]{\mathcal{L}} \mathcal{N}(0, 1).$$

What is interesting is what is *not* there, namely a restriction on how fast p_n should go to infinity with n.

In other words, the result holds as soon as $\min(n, p) \to \infty$ ($\rightsquigarrow$ "universal asymptotics")

This extends to the second test considered
(and actually to various *sign* tests from multivariate analysis).


n=4
n=30
n=200
n=1000
p=4
p=30
p=200
p=1000

The HD case

To the best of our knowledge, this is the first universal (n, p)-asymptotic result. Typically, in previous works,

- one restricts the way p may go to infinity with n. It is standard to have

$$\frac{p_n}{n} \to c \in C$$

  for some convex $C \subset (0, \infty)$ (e.g., $C = (0, 1)$, $[1, \infty)$, etc.)

or

- no such restrictions are imposed, but...
  different (n, p)-regimes lead to different asymptotic distributions;
  see, e.g., Cai and Jiang (2012), Cai, Fan, and Jiang (2013).

Theorem 6 (Extreme Law: Sub-Exponential Case) *Let $p = p_n \to \infty$ satisfy $\frac{\log n}{p} \to 0$ as $n \to \infty$. Then*

(i). $\max_{1 \le i < j \le n} |\Theta_{ij} - \frac{\pi}{2}| \to 0$ *in probability as* $n \to \infty$;

(ii). *As* $n \to \infty$, $2p \log \sin \Theta_{\min} + 4 \log n - \log \log n$ *converges weakly to the extreme value distribution with the distribution function* $F(y) = 1 - e^{-Ke^{y/2}}$, $y \in \mathbb{R}$ *and* $K = 1/(4\sqrt{2\pi})$. *The conclusion still holds if* $\Theta_{\min}$ *is replaced by* $\Theta_{\max}$.

Theorem 8 (Extreme Law: Exponential Case) *Let $p = p_n$ satisfy $\frac{\log n}{p} \to \beta \in (0, \infty)$ as $n \to \infty$, then*

(i). $\Theta_{\min} \to \cos^{-1} \sqrt{1 - e^{-4\beta}}$ *and* $\Theta_{\max} \to \pi - \cos^{-1} \sqrt{1 - e^{-4\beta}}$ *in probability as* $n \to \infty$;

(ii). *As* $n \to \infty$, $2p \log \sin \Theta_{\min} + 4 \log n - \log \log n$ *converges weakly to a distribution with the distribution function*

$$F(y) = 1 - \exp\left\{-K(\beta)e^{(y+8\beta)/2}\right\},\ y \in \mathbb{R},\ \text{where } K(\beta) = \left(\frac{\beta}{8\pi(1 - e^{-4\beta})}\right)^{1/2},$$

and the conclusion still holds if $\Theta_{\min}$ *is replaced by* $\Theta_{\max}$.

Theorem 9 (Extreme Law: Super-Exponential Case) *Let $p = p_n$ satisfy $\frac{\log n}{p} \to \infty$ as $n \to \infty$. Then,*

(i). $\Theta_{\min} \to 0$ *and* $\Theta_{\max} \to \pi$ *in probability as* $n \to \infty$;

(ii). *As* $n \to \infty$, $2p \log \sin \Theta_{\min} + \frac{4p}{p-1} \log n - \log p$ *converges weakly to the extreme value distribution with the distribution function* $F(y) = 1 - e^{-Ke^{y/2}}$, $y \in \mathbb{R}$ *with* $K = 1/(2\sqrt{2\pi})$. *The conclusion still holds if* $\Theta_{\min}$ *is replaced by* $\Theta_{\max}$.

Theorem 6 (Extreme Law: Sub-Exponential Case) *Let $p = p_n \to \infty$ satisfy $\boxed{\frac{\log n}{p} \to}$ 0 as $n \to \infty$. Then*

(i). $\max_{1 \le i < j \le n} |\Theta_{ij} - \frac{\pi}{2}| \to 0$ *in probability as* $n \to \infty$;

(ii). *As* $n \to \infty$, $2p \log \sin \Theta_{\min} + 4 \log n - \log \log n$ *converges weakly to the extreme value distribution with the distribution function* $F(y) = 1 - e^{-Ke^{y/2}}$, $y \in \mathbb{R}$ *and* $K = 1/(4\sqrt{2\pi})$. *The conclusion still holds if* $\Theta_{\min}$ *is replaced by* $\Theta_{\max}$.

Theorem 8 (Extreme Law: Exponential Case) *Let $p = p_n$ satisfy $\boxed{\frac{\log n}{p} \to \beta \in (0, \infty)}$ as $n \to \infty$, then*

(i). $\Theta_{\min} \to \cos^{-1} \sqrt{1 - e^{-4\beta}}$ *and* $\Theta_{\max} \to \pi - \cos^{-1} \sqrt{1 - e^{-4\beta}}$ *in probability as* $n \to \infty$;

(ii). *As* $n \to \infty$, $2p \log \sin \Theta_{\min} + 4 \log n - \log \log n$ *converges weakly to a distribution with the distribution function*

$$F(y) = 1 - \exp\left\{-K(\beta)e^{(y+8\beta)/2}\right\}, \ y \in \mathbb{R}, \ \text{where } K(\beta) = \left(\frac{\beta}{8\pi(1 - e^{-4\beta})}\right)^{1/2},$$

and the conclusion still holds if $\Theta_{\min}$ *is replaced by* $\Theta_{\max}$.

Theorem 9 (Extreme Law: Super-Exponential Case) *Let $p = p_n$ satisfy $\boxed{\frac{\log n}{p} \to \infty}$ as $n \to \infty$. Then,*

(i). $\Theta_{\min} \to 0$ *and* $\Theta_{\max} \to \pi$ *in probability as* $n \to \infty$;

(ii). *As* $n \to \infty$, $2p \log \sin \Theta_{\min} + \frac{4p}{p-1} \log n - \log p$ *converges weakly to the extreme value distribution with the distribution function* $F(y) = 1 - e^{-Ke^{y/2}}$, $y \in \mathbb{R}$ *with* $K = 1/(2\sqrt{2\pi})$. *The conclusion still holds if* $\Theta_{\min}$ *is replaced by* $\Theta_{\max}$.

Theorem 6 (Extreme Law: Sub-Exponential Case) *Let* $p = p_n \to \infty$ *satisfy* $\frac{\log n}{p} \to 0$ *as* $n \to \infty$. *Then*

(i). $\max_{1 \le i < j \le n} |\Theta_{ij} - \frac{\pi}{2}| \to 0$ *in probability as* $n \to \infty$;

(ii). *As* $n \to \infty$, $2p \log \sin \Theta_{\min} + 4 \log n - \log \log n$ *converges weakly to the extreme value distribution with the distribution function* $F(y) = 1 - e^{-Ke^{y/2}}$, $y \in \mathbb{R}$ *and* $K = 1/(4\sqrt{2\pi})$. *The conclusion still holds if* $\Theta_{\min}$ *is replaced by* $\Theta_{\max}$.

Theorem 8 (Extreme Law: Exponential Case) *Let* $p = p_n$ *satisfy* $\frac{\log n}{p} \to \beta \in (0, \infty)$ *as* $n \to \infty$, *then*

(i). $\Theta_{\min} \to \cos^{-1} \sqrt{1 - e^{-4\beta}}$ *and* $\Theta_{\max} \to \pi - \cos^{-1} \sqrt{1 - e^{-4\beta}}$ *in probability as* $n \to \infty$;

(ii). *As* $n \to \infty$, $2p \log \sin \Theta_{\min} + 4 \log n - \log \log n$ *converges weakly to a distribution with the distribution function*

$$F(y) = 1 - \exp\left\{-K(\beta)e^{(y+8\beta)/2}\right\}, \quad y \in \mathbb{R}, \quad \text{where } K(\beta) = \left(\frac{\beta}{8\pi(1 - e^{-4\beta})}\right)^{1/2},$$

and the conclusion still holds if $\Theta_{\min}$ *is replaced by* $\Theta_{\max}$.

Theorem 9 (Extreme Law: Super-Exponential Case) *Let* $p = p_n$ *satisfy* $\frac{\log n}{p} \to \infty$ *as* $n \to \infty$. *Then,*

(i). $\Theta_{\min} \to 0$ *and* $\Theta_{\max} \to \pi$ *in probability as* $n \to \infty$;

(ii). *As* $n \to \infty$, $2p \log \sin \Theta_{\min} + \frac{4p}{p-1} \log n - \log p$ *converges weakly to the extreme value distribution with the distribution function* $F(y) = 1 - e^{-Ke^{y/2}}$, $y \in \mathbb{R}$ *with* $K = 1/(2\sqrt{2\pi})$. *The conclusion still holds if* $\Theta_{\min}$ *is replaced by* $\Theta_{\max}$.

The HD case

Our universal asymptotic results validate the use of two different (asymptotically equivalent) tests, namely

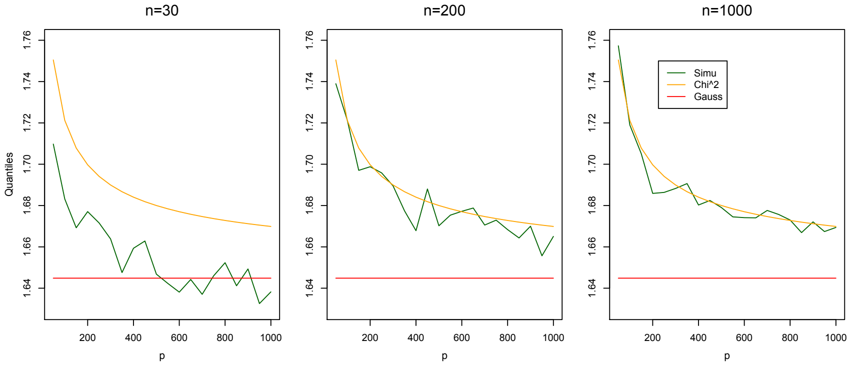
- Test 1: reject at asymptotic level α if

$$T_n^{\mathrm{St}} = \frac{T_n - d(p)}{\sqrt{2d(p)}} > \Phi^{-1}(1-\alpha).$$

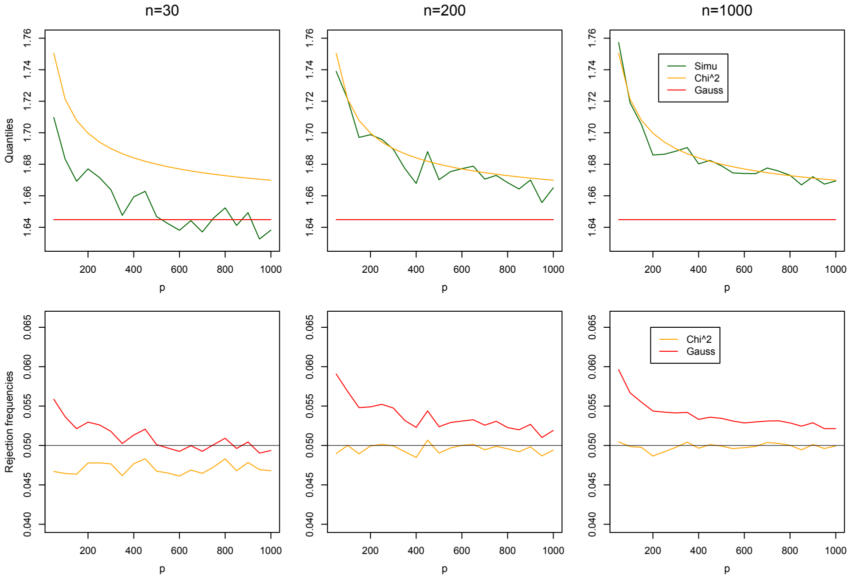
- Test 2: reject at asymptotic level α if $T_n > \chi^2_{d(p),1-\alpha}$, or equivalently, if

$$T_n^{\mathrm{St}} = \frac{T_n - d(p)}{\sqrt{2d(p)}} > \frac{\chi^2_{d(p),1-\alpha} - d(p)}{\sqrt{2d(p)}}.$$

What test should be favored for fixed (n, p)?



From 100,000 replications


n=30
n=200
n=1000
Quantiles
Rejection frequencies
p
Simu
Chi^2
Gauss

Another problem

Ley, Swan, Thiam, and Verdebout (2013) discussed R-estimation in the spherical location problem, that is in the problem of estimating θ from a random sample $X_1, \ldots, X_n$ with common *rotationally symmetric* density

$$x \mapsto c_{p,f}\, f(x'\theta),$$

on $\mathcal{S}^{p-1}$; see Saw (1978).

Paindaveine and Verdebout (2014) recently proposed signed-rank tests for $\mathcal{H}_0 : \theta = \theta_0$, including a sign test that rejects the null whenever

$$T_n = \frac{p-1}{n} \sum_{i,j=1}^{n} U_i'(\theta_0) U_j(\theta_0) > \chi^2_{p-1,1-\alpha},$$

where

$$U_i(\theta_0) = \frac{(I_p - \theta_0\theta_0')X_i}{\|(I_p - \theta_0\theta_0')X_i\|}, \quad i = 1, \ldots, n$$

is the "sign" of the projection of X_i onto the tangent space to $\mathcal{S}^{p-1}$ at θ_0.

Mutatis mutandis, one can establish

Theorem 2

Let p_n be a sequence of positive integers converging to $+\infty$. Assume that X_{ni}, $i = 1, \ldots, n$, $n = 1, 2, \ldots,$ is a triangular array such that for any n, the random p_n-vectors X_{ni}, $i = 1, \ldots, n$ are i.i.d. rotationally symmetric about $\theta_0 \in \mathcal{S}^{p_n-1}$ (with $X_{n1} \neq \theta_0$ a.s.) Then

$$T_n^{\mathrm{St}} = \frac{T_n - (p_n - 1)}{\sqrt{2(p_n - 1)}} \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}(0, 1).$$

Another problem

For the more classical Watson (1983) test, that rejects the null whenever

$$W_n = \frac{n(p-1)\bar{X}'(I_k - \theta_0\theta_0')\bar{X}}{1 - \frac{1}{n}\sum_{i=1}^{n}(X_i'\theta_0)^2} > \chi^2_{p-1,1-\alpha},$$

we can prove the following (where we let $u_{ni} = \sqrt{1 - (X_{ni}'\theta_0)^2}$).

Theorem 3

Let X_{ni}, $i = 1, \ldots, n$, $n = 1, 2, \ldots,$ form a triangular array of random vectors satisfying the following conditions : (i) for any n, $X_{n1}, X_{n2}, \ldots, X_{nn}$ are mutually independent and share a common rotationally symmetric distribution on $\mathcal{S}^{p_n-1}$ with location θ_0; (ii) $p_n \to \infty$ as $n \to \infty$; (iii) $\mathrm{E}[u_{n1}^2] > 0$ for any n; (iv) $\mathrm{E}[u_{n1}^4]/(\mathrm{E}[u_{n1}^2])^2 = o(n)$ as $n \to \infty$. Then

$$W_n^{\mathrm{St}} = \frac{W_n - (p_n - 1)}{\sqrt{2(p_n - 1)}} \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}(0, 1).$$

See Ley, Paindaveine and Verdebout (2014).

Another problem

For the more classical Watson (1983) test, that rejects the null whenever

$$W_n = \frac{n(p-1)\bar{X}'(I_k - \theta_0\theta_0')\bar{X}}{1 - \frac{1}{n}\sum_{i=1}^{n}(X_i'\theta_0)^2} > \chi^2_{p-1,1-\alpha},$$

we can prove the following (where we let $u_{ni} = \sqrt{1-(X_{ni}'\theta_0)^2}$).

Theorem 3

Let X_{ni}, $i = 1, \ldots, n$, $n = 1, 2, \ldots,$ form a triangular array of random vectors satisfying the following conditions : (i) for any n, $X_{n1}, X_{n2}, \ldots, X_{nn}$ are mutually independent and share a common rotationally symmetric distribution on $\mathcal{S}^{p_n-1}$ with location θ_0; (ii) $p_n \to \infty$ as $n \to \infty$; (iii) $\mathrm{E}[u_{n1}^2] > 0$ for any n; (iv) $\mathrm{E}[u_{n1}^4]/(\mathrm{E}[u_{n1}^2])^2 = o(n)$ as $n \to \infty$. Then

$$W_n^{\mathrm{St}} = \frac{W_n - (p_n - 1)}{\sqrt{2(p_n-1)}} \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}(0,1).$$

See Ley, Paindaveine and Verdebout (2014).

Another problem

No universal consistency

Still...

• Imposing (iii) only excludes the degenerate case for which $X_{n1} = \theta_0$ a.s., which would imply that W_n — hence also W_n^{St} — is not well-defined.

• If (iv) does not hold, we must then have that, for some constant $C > 0$,

$$\mathrm{E}[(X'_{n1}\theta_0)^2] \geq 1 - \frac{C}{\sqrt{n}}$$

for infinitely many n. In the high-dimensional setup considered, this is extremely pathological, since it corresponds to the distribution of X_{n1} concentrating in *one* particular direction — namely, the direction θ_0 — in the expanding Euclidean space $\mathbb{R}^{p_n}$.

References

Banerjee, A., Dhillon, I., Ghosh, J., and Sra, S. (2003). Generative model-based clustering of directional data. Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining, 19–28.

Cai, T., Fan, J., and Jiang, T. (2013). Distributions of angles in random packing on spheres. Journal of Machine Learning Research 14, 1801–1828.

Cai, T., and Jiang, T. (2012). Phase transition in limiting distributions of coherence of high-dimensional random matrices. Journal of Multivariate Analysis 107, 24–39.

Dhillon, I., and Modha, D. (2001). Concept decompositions for large sparse text data using clustering. Machine Learning 42, 143–175.

Dryden, I. (2005). Statistical analysis on high-dimenional spheres and shape spaces. Annals of Statistics 33, 1643–1665.

Hallin, M., and Paindaveine, D. (2006). Semiparametrically efficient rank-based inference for shape. I. Optimal rank-based tests for sphericity. Annals of Statistics 34, 2707–2756.

Ley, C., Paindaveine, D., and Verdebout, T. (2014). High-dimensional tests for spherical location and spiked covariance. Submitted.

Ley, C., Swan, Y., Thiam, B., and Verdebout, T. (2013), Optimal R-estimation of a spherical location. Statistica Sinica 23, 305–333.

Paindaveine, D., and Verdebout, T. (2014). Optimal rank-based tests for the location parameter of a rotationally symmetric distribution on the hypersphere (with Th. Verdebout). In M. Hallin, D. Mason, D. Pfeifer, and J. Steinebach Eds, Mathematical Statistics and Limit Theorems: Festschrift in Honor of Paul Deheuvels. Springer, to appear.

Paindaveine, D., and Verdebout, T. (2014). High-dimensional sign tests: from universal asymptotic theory to practice. Submitted.

Papers with P. are downloadable from http://homepages.ulb.ac.be/˜dpaindav