Apprentissage non supervisé: Tests pour la position.

Hotelling test

Assume that $\mathbf{X}_1, \ldots, \mathbf{X}_n$ are i.i.d. $\mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. We want to test $\mathcal{H}_0 : \boldsymbol{\mu} = \mathbf{0}$ against $\mathcal{H}_1 : \boldsymbol{\mu} \neq \mathbf{0}$.

The classical asymptotic test for this problem is the so-called Hotelling test that rejects the null hypothesis when

$$\bar{\mathbf{X}}' \mathbf{S}^{-1} \bar{\mathbf{X}} > \chi^2_{p;1-\alpha},$$

where $\mathbf{S} := n^{-1} \sum_{i=1}^{n} (\mathbf{X}_i - \bar{\mathbf{X}})(\mathbf{X}_i - \bar{\mathbf{X}})'$ and $\chi^2_{q;\nu}$ is the quantile of order ν of the chi-square distribution with q degrees of freedom.

We focus here on hypothesis testing in high-dimensions.

We want to consider hypothesis testing in the high-dimensional framework where $p_n \to \infty$ as $n \to \infty$.

We first consider the Gaussian location problem. That is we consider the problem of testing $\mathcal{H}_0 : \boldsymbol{\mu} = \mathbf{0}$ against $\mathcal{H}_1 : \boldsymbol{\mu} \neq \mathbf{0}$.

The classical test for this problem is the Hotelling test that rejects the null (at the asymptotic level α) when

$$n\bar{\mathbf{X}}'\mathbf{S}^{-1}\bar{\mathbf{X}} > \chi^2_{p;1-\alpha}.$$

Same problem: when $p > n$, $\mathbf{S}$ is not invertible so that the Hotelling test is useless in practice

$\rightsquigarrow$ We consider first the case where $\mathbf{\Sigma}$ is known; we take for instance $\mathbf{\Sigma} = \mathbf{I}_p$ and consider the statistic $n\|\bar{\mathbf{X}}\|^2$.

HD

We consider the $n \to \infty$ and $p = p_n \to \infty$ framework.

$\rightsquigarrow$ We need to consider triangular arrays of the form

$$\begin{array}{llll} \mathbf{X}_{11} & & & \text{with values in } \mathbb{R}^{p_1-1} \\ \mathbf{X}_{21} & \mathbf{X}_{22} & & \text{with values in } \mathbb{R}^{p_2-1} \\ \vdots & & \ddots & \\ \mathbf{X}_{n1} & \mathbf{X}_{n2} & \dots \quad \mathbf{X}_{nn} & \text{with values in } \mathbb{R}^{p_n-1} \\ \vdots & & & \ddots \end{array}$$

where, under $\mathcal{H}_0$, observations in row n are i.i.d. $\mathcal{N}_{p_n}(\mathbf{0}, \mathbf{I}_{p_n})$

$$T_n = n\|\bar{\mathbf{X}}\|^2 \xrightarrow[n\to\infty]{\mathcal{D}} ???$$

Proposition

As p_n and $n \to \infty$, we have that

$$\frac{T_n - p_n}{\sqrt{2p_n}} \to \mathcal{N}(0,1).$$

Proof. We have that $\sqrt{n}\bar{\mathbf{X}} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_{p_n})$, so that

$$n\|\bar{\mathbf{X}}\|^2 =_{\mathcal{D}} \sum_{i=1}^{p_n} Z_i,$$

where the Z_i's are i.i.d χ^2_1. It follows from the CLT that

$$\frac{\sum_{i=1}^{p_n} Z_i - p_n}{\sqrt{2p_n}}$$

is asymptotically standard normal. □

As a result, a natural extension of the Hotteling test in high-dimension for the specified $\mathbf{\Sigma}$-case rejects the null at the asymptotic level α when

$$\frac{T_n - p_n}{\sqrt{2p_n}} > z_{1-\alpha},$$

where z_ν stand for the ν-quantile of a standard Gaussian random variable.

Now of course, the big challenge is the unspecified-$\mathbf{\Sigma}$ case, which is the “realistic case".

One idea is to replace $\mathbf{S}$ in $n\bar{\mathbf{X}}'\mathbf{S}^{-1}\bar{\mathbf{X}}$ by a regularized version of $\mathbf{S}$, that is to consider a test statistic of the form

$$T_{\text{reg}}(\lambda) := n\bar{\mathbf{X}}'(\mathbf{S} + \lambda \mathbf{I}_p)^{-1}\bar{\mathbf{X}}$$

What is the limiting behaviour of this quantity under the null hypothesis?

First note that

$$\begin{aligned} T_{\text{reg}}(\lambda) &= \sqrt{n}(\boldsymbol{\Sigma}^{-1/2}\bar{\mathbf{X}})'\boldsymbol{\Sigma}^{1/2}(\mathbf{S} + \lambda \mathbf{I}_p)^{-1}\boldsymbol{\Sigma}^{1/2}(\sqrt{n}\boldsymbol{\Sigma}^{-1/2}\bar{\mathbf{X}}) \\ &= \mathbf{Z}'\boldsymbol{\Sigma}^{1/2}(\mathbf{S} + \lambda \mathbf{I}_p)^{-1}\boldsymbol{\Sigma}^{1/2}\mathbf{Z}, \end{aligned}$$

where $\mathbf{Z} := \sqrt{n}\boldsymbol{\Sigma}^{-1/2}\bar{\mathbf{X}} \sim \mathcal{N}_{p_n}(\mathbf{0}, \mathbf{I}_{p_n})$ when the $\mathbf{X}_{ni}$'s are i.i.d. $\mathcal{N}_{p_n}(\mathbf{0}, \boldsymbol{\Sigma})$

Note that $\mathbf{Z}$ and $\mathbf{S}$ are independent. We have that

$$\begin{aligned} \mathrm{E}[T_{\mathrm{reg}}(\lambda)] &= \mathrm{E}[\mathrm{tr}(\mathbf{Z}'\mathbf{\Sigma}^{1/2}(\mathbf{S}+\lambda\mathbf{I}_p)^{-1}\mathbf{\Sigma}^{1/2}\mathbf{Z})] \\ &= \mathrm{E}[\mathrm{tr}(\mathbf{Z}\mathbf{Z}'\mathbf{\Sigma}^{1/2}(\mathbf{S}+\lambda\mathbf{I}_p)^{-1}\mathbf{\Sigma}^{1/2})] \\ &= \mathrm{E}[\mathrm{tr}(\mathbf{\Sigma}^{1/2}(\mathbf{S}+\lambda\mathbf{I}_p)^{-1}\mathbf{\Sigma}^{1/2})]. \end{aligned}$$

As a result, the quantity

$$\mathrm{tr}(\mathbf{\Sigma}^{1/2}(\mathbf{S}+\lambda\mathbf{I}_p)^{-1}\mathbf{\Sigma}^{1/2})$$

plays a really important role. Its asymptotic behavior clearly depends on the asymptotic behavior of the eigenvalues of $\mathbf{S}$.

Histogram of the eigenvalues of **S** computed from 5000 Standard Gaussian observations with $p = 4$

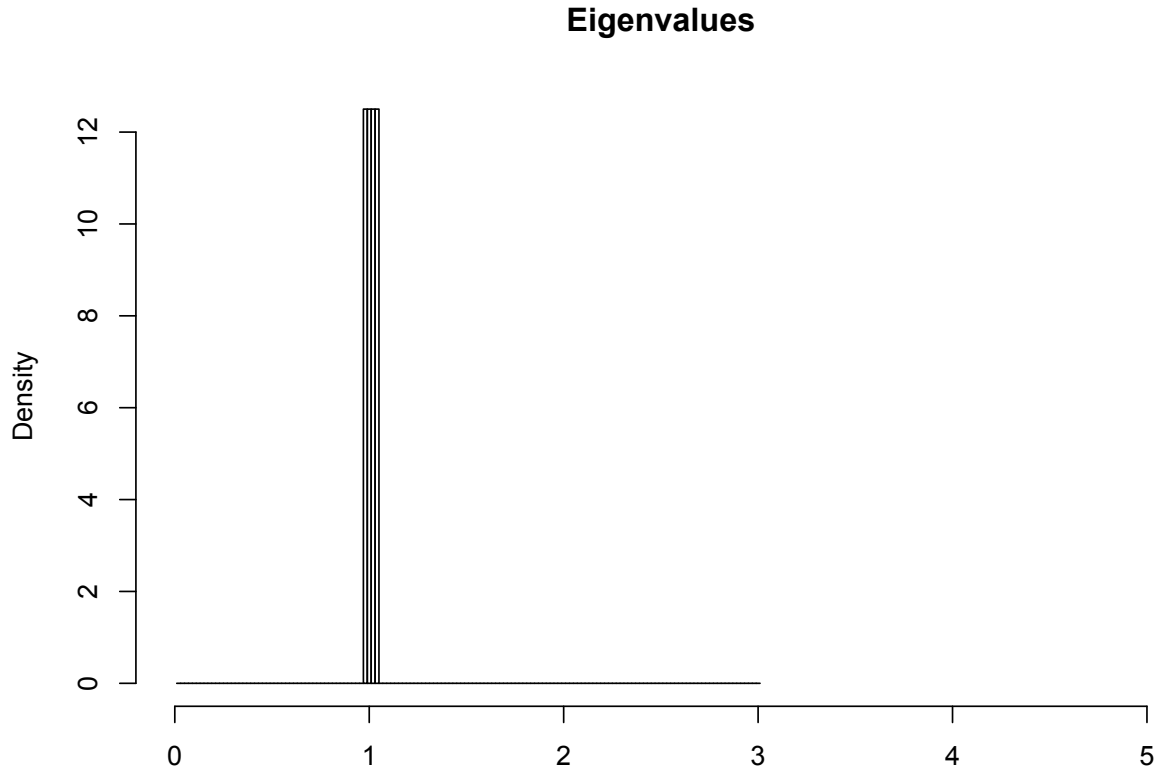


Figure: Histogram of eigenvalues

Histogram of the eigenvalues of **S** computed from 5000 Standard Gaussian observations with $p = 2000$

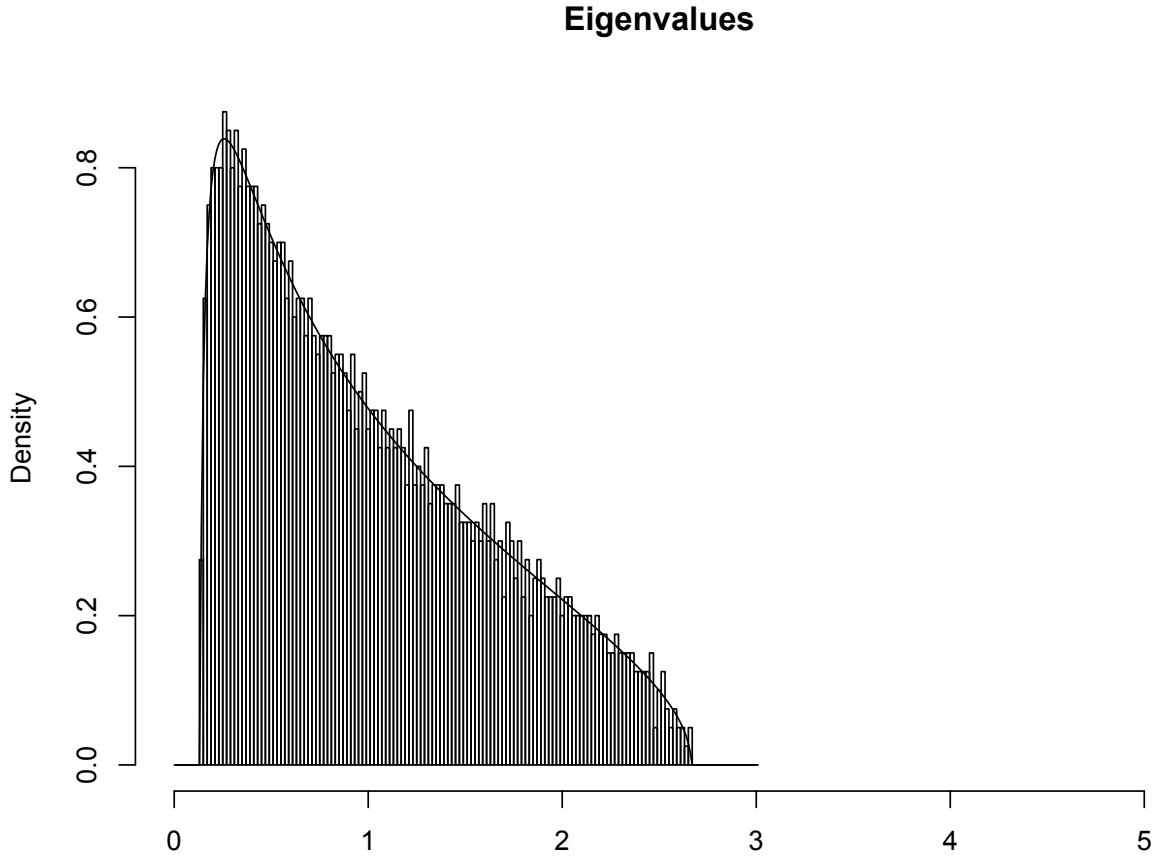


Figure: Histogram of eigenvalues

Consider the empirical spectral distribution of **S** defined as

$$F_{n,p}(t) := \frac{1}{p}\#\{\hat{\lambda}_i, \hat{\lambda}_i < t\}.$$

A milestone result by Marcenko-Pastur (1967) shows that when $p_n/n \to \gamma \in (0, 1]$, $F_{n,p}(t)$ converges weakly to the *Marcenko-Pastur* distribution $F_{\mathrm{MP}}(t)$ with density

$$u \to f(u) = \frac{\sqrt{(b_\gamma - x)(x - a_\gamma)}}{2\pi x\gamma},$$

where

$$a_\gamma := (1 - \sqrt{\gamma})^2 \quad \text{and} \quad b_\gamma := (1 + \sqrt{\gamma})^2$$

The so-called *Stieltjes transform* play an important role in the proof of such results. Let $z = u + iv \in \mathbb{C}$ with $v > 0$. The Stieltjes transform of a probability distribution F is defined as

$$s_F(z) := \int_{\mathcal{X}} \frac{1}{x - z} dF(x)$$

To show that $F_{n,p}$ converges weakly to F_{MP}, it is enough to show that the empirical Stieltjes tranform

$$s_{F_{n,p}}(z) = \frac{1}{p} \sum_{j=1}^{p} \frac{1}{\hat{\lambda}_j - z} = \frac{1}{p} \mathrm{tr}((\mathbf{S} - z\mathbf{I}_p)^{-1})$$

converges pointwise to $s_{F_{\mathrm{MP}}}(z)$, which is the Stieltjes transform of the Marcenko-Pastur distribution. Actually, this holds.

Remember that in the Hotteling test statistic we are interested in $\mathrm{tr}(\mathbf{\Sigma}^{1/2}(\mathbf{S} + \lambda\mathbf{I}_p)^{-1}\mathbf{\Sigma}^{1/2})$.

Using the fact that

$$\frac{1}{p_n}\mathrm{tr}((\mathbf{S}+\lambda\mathbf{I}_{p_n})^{-1}) - s_{F_{\mathrm{MP}}}(-\lambda) = o_P(1)$$

as n with $p_n/n \to \gamma \in (0,1]$,

$$\frac{1}{p_n}\mathrm{tr}(\mathbf{\Sigma}^{1/2}(\mathbf{S}+\lambda\mathbf{I}_p)^{-1}\mathbf{\Sigma}^{1/2}) - \theta_1(\lambda,\gamma) = o_P(1)$$

as n with $p_n/n \to \gamma \in (0,1]$, where

$$\theta_1(\lambda,\gamma) := \frac{1-\lambda s_{F_{\mathrm{MP}}}(-\lambda)}{1-\gamma(1-\lambda s_{F_{\mathrm{MP}}}(-\lambda))}.$$

Based on this, they obtained that

$$\frac{\sqrt{p_n}(p_n^{-1} T_{\text{reg}}(\lambda) - \theta_1(\lambda, \gamma))}{(2\theta_2(\lambda, \gamma))^{1/2}}$$

is asymptotically standard normal under the null hypothesis where $\theta_2(\lambda, \gamma)$ is defined as

$$\begin{aligned} \theta_2(\lambda, \gamma) &= \frac{1 - \lambda s_{F_{\text{MP}}}(-\lambda)}{(1 - \gamma(1 - \lambda s_{F_{\text{MP}}}(-\lambda)))^3} \\ &\quad -\lambda \frac{s_{F_{\text{MP}}}(-\lambda) - \lambda s'_{F_{\text{MP}}}(-\lambda)}{(1 - \gamma(1 - \lambda s_{F_{\text{MP}}}(-\lambda)))^4} \end{aligned}$$

A test can be constructed based on estimated versions of $\theta_1(\lambda, \gamma)$ and $\theta_2(\lambda, \gamma)$.

Another interesting test for the problem is the so-called *sign-test*.

Consider for a moment the fixed-p case. A test for the same problem can be based on the multivariate signs

$$\mathbf{U}_{n1} := \frac{\mathbf{X}_{n1}}{\|\mathbf{X}_{n1}\|}, \ldots, \mathbf{U}_{nn} := \frac{\mathbf{X}_{n1}}{\|\mathbf{X}_{nn}\|}$$

of the observations.

The signs are taking values on the unit sphere
$\mathcal{S}^{p_n-1} := \{\mathbf{x} \in \mathbb{R}^{p_n}, \mathbf{x}'\mathbf{x} = 1\}$

HD

Assume that the $\mathbf{X}_{ni}$'s are i.i.d. $\mathcal{N}_{p_n}(\mathbf{0}, \mathbf{I}_{p_n})$ under the null hypothesis. The signs are uniformly distributed on $\mathcal{S}^{p_n-1} := \{\mathbf{x} \in \mathbb{R}^{p_n}, \mathbf{x}'\mathbf{x} = 1\}$.

Let $\mathbf{U} \sim \mathrm{unif}(\mathcal{S}^{p-1})$, then $\mathbf{OU} =_{\mathcal{D}} \mathbf{U}$ for any rotation $\mathbf{O}$. In particular $\mathbf{U} =_{\mathcal{D}} -\mathbf{U}$, so that

$$\mathrm{E}[\mathbf{U}] = \mathbf{0}.$$

Moreover

$$\mathrm{Var}[\mathbf{U}] := \mathrm{E}[\mathbf{U}\mathbf{U}'] = \frac{1}{p}\mathbf{I}_p.$$

Therefore, letting $\bar{\mathbf{U}} := n^{-1}\sum_{i=1}^{n}\mathbf{U}_{ni}$, the central limit theorem entails that in the fixed-p_n case,

$$np\|\bar{\mathbf{U}}\| \to_{\mathcal{D}} \chi^2_p$$

as $n \to \infty$ when the $\mathbf{U}_{ni}$'s are uniformly distributed on $\mathcal{S}^{p_n-1}$

We need to consider triangular arrays of observations of the form

$$
\begin{array}{llllll}
\mathbf{U}_{11} & & & & & \text{with values in } \mathcal{S}^{p_1-1} \\
\mathbf{U}_{21} & \mathbf{U}_{22} & & & & \text{with values in } \mathcal{S}^{p_2-1} \\
\vdots & & \ddots & & & \\
\mathbf{U}_{n1} & \mathbf{U}_{n2} & \dots & \mathbf{U}_{nn} & & \text{with values in } \mathcal{S}^{p_n-1} \\
\vdots & & & & \ddots &
\end{array}
$$

We assume that $\mathbf{U}_{n1}, \mathbf{U}_{n2}, \dots, \mathbf{U}_{nn}$ are mutually independent from the uniform distribution on $\mathcal{S}^{p_n-1}$.

Denote the corresponding sequence of hypotheses as $P_0^{(n)}$.
What is the asymptotic distribution of $R_n = np\|\bar{\mathbf{U}}\|^2$ under $P_0^{(n)}$ if $p_n \to \infty$?

HD

- The fixed-p asymptotic result

$$R_n \xrightarrow[n\to\infty]{\mathcal{L}} \chi^2_p$$

leads to

$$\frac{R_n - p}{\sqrt{2p}} \xrightarrow[n\to\infty]{\mathcal{L}} \frac{\chi^2_p - p}{\sqrt{2p}} = \frac{\chi^2_p - \mathrm{E}[\chi^2_p]}{\sqrt{\mathrm{Var}[\chi^2_p]}} \xrightarrow[p\to\infty]{} \mathcal{N}(0,1).$$

- This suggests the (n,p)-asymptotic result

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

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

$$R^{St}_n = \frac{R_n - p_n}{\sqrt{2p_n}} \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}(0,1) \ ?$$

Yes (Paindaveine, D., and Verdebout, T. (2016). On high-dimensional sign tests. ***Bernoulli***)

- Rewrite the Rayleigh statistic as

$$R_n = np_n\|\bar{\mathbf{X}}_n\|^2 = \frac{p_n}{n}\sum_{i,j=1}^{n}\mathbf{U}'_{ni}\mathbf{U}_{nj} = p_n + \frac{2p_n}{n}\sum_{1\le i<j\le n}\mathbf{U}'_{ni}\mathbf{U}_{nj},$$

so that

$$R_n^{\mathrm{St}} = \frac{R_n - p_n}{\sqrt{2p_n}} = \frac{\sqrt{2p_n}}{n}\sum_{1\le i<j\le n}\mathbf{U}'_{ni}\mathbf{U}_{nj}.$$

- To study this U-statistic with an order-2 kernel depending on $p = p_n$, write

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

where the random variables $(\mathrm{E}_{n\ell}[\,\cdot\,] = \mathrm{E}_{n\ell}[\,\cdot\,|\mathbf{U}_1,\ldots,\mathbf{U}_\ell])$

$$D_{n\ell} = \mathrm{E}_{n\ell}\big[R_n^{\mathrm{St}}\big] - \mathrm{E}_{n,\ell-1}\big[R_n^{\mathrm{St}}\big] = \frac{\sqrt{2p_n}}{n}\sum_{i=1}^{\ell-1}\mathbf{U}'_{ni}\mathbf{U}_{n\ell}, \qquad \ell = 1,\ldots,n,$$

form a martingale difference process.

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

$$\sum_{\ell=1}^{n} \mathrm{E}\left[D_{n\ell}^2 \,|\, \mathcal{F}_{n,\ell-1}\right] \xrightarrow[n\to\infty]{\mathrm{P}} 1 \tag{1.1}$$

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

$$\sum_{\ell=1}^{n} \mathrm{E}\left[D_{n\ell}^2 \, \mathbb{I}[|D_{n\ell}| > \varepsilon]\right] \xrightarrow[n\to\infty]{} 0. \tag{1.2}$$

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

HD

We first consider alternatives associated with triangular arrays of the form

$$\begin{array}{llllll}
\mathbf{X}_{11} & & & & & \text{with values in } \mathcal{S}^{p_1-1} \\
\mathbf{X}_{21} & \mathbf{X}_{22} & & & & \text{with values in } \mathcal{S}^{p_2-1} \\
\vdots & & \ddots & & & \\
\mathbf{X}_{n1} & \mathbf{X}_{n2} & \dots & \mathbf{X}_{nn} & & \text{with values in } \mathcal{S}^{p_n-1} \\
\vdots & & & & \ddots &
\end{array}$$

where $\mathbf{X}_{n1}, \mathbf{X}_{n2}, \dots, \mathbf{X}_{nn}$ are mutually independent from the rotationally symmetric distribution $P_{\boldsymbol{\theta}_n, \kappa_n, f}$ with density

$$\mathbf{x} \mapsto f\left(\kappa_n \mathbf{x}' \boldsymbol{\theta}_n\right);$$

here, the sequence $(\boldsymbol{\theta}_n)$ is such that $\boldsymbol{\theta}_n \in \mathcal{S}^{p_n-1}$ for any n, (κ_n) is a positive sequence, and $f : \mathbb{R} \mapsto \mathbb{R}^+$ is fixed.

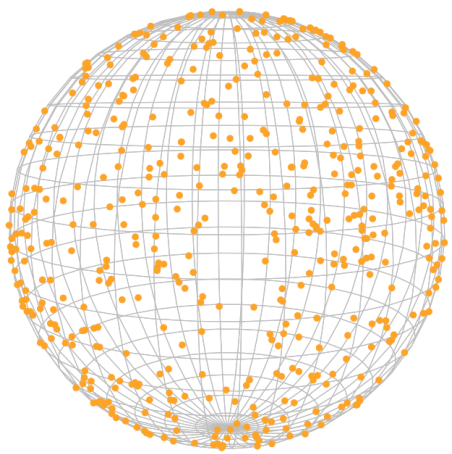
We denote the corresponding sequence of hypotheses as $P^{(n)}_{\boldsymbol{\theta}_n, \kappa_n, f}$.

HD

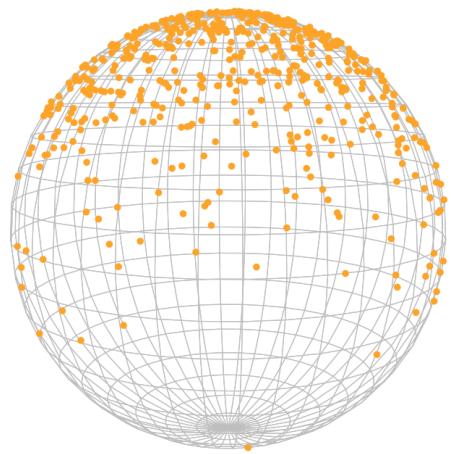
The Rayleigh test is the likelihood ratio test for testing uniformity against absolutely continuous alternatives with densities

$$\mathbf{u} \mapsto c_{p,\kappa} \exp(\kappa \mathbf{u}'\boldsymbol{\theta}),$$

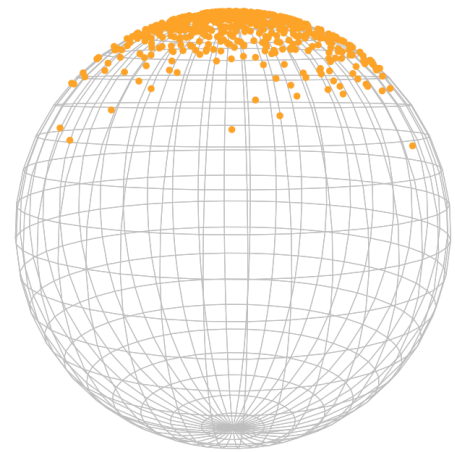
with $\kappa > 0$



$\kappa \to 0$



$\kappa = 3$



$\kappa = 10$

The larger κ, the more concentrated the distribution is about θ

$\kappa \to 0$ corresponds to $\mathrm{Unif}(S^{p-1})$

Proposition (Cutting, Paindaveine and Verdebout (2017), *Annals of Statistics*)

Let (p_n) be a sequence of positive integers diverging to ∞.
Let (θ_n) be a sequence such that $\theta_n \in \mathcal{S}^{p_n-1}$ for all n.
Then,
(i) if $\kappa_n = \tau p_n^{3/4}/\sqrt{n}$ ($\tau > 0$), the asymptotic power of Rayleigh, under $\mathrm{P}^{(n)}_{\theta_n,\kappa_n,f}$, is

$$1 - \Phi\left(\Phi^{-1}(1-\alpha) - \frac{\tau^2}{\sqrt{2}}\right).$$

(ii) if $\kappa_n = o(p_n^{3/4}/\sqrt{n})$, its asymptotic power is α.

HD

We furthemore obtain that the Rayleigh is locally and asymptotically most powerfull within the class of rotation-invariant tests in high-dimension.

The first optimality result in high-dimension to the best of our knowledge!