Rank tests for PCA under weak identifiability

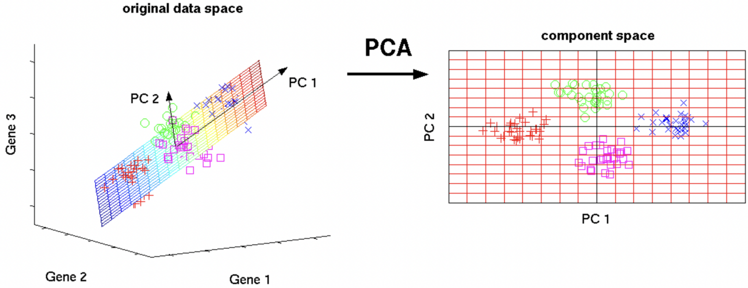
Davy Paindaveine, Laura Peralvo Maroto and Thomas Verdebout

Université libre de Bruxelles (ULB)

SPP : Mathematics PhD and Postdoc seminars, Bruxelles

Principal component analysis (PCA) is a classic multivariate statistical analysis technique.

↪ Objective: size reduction.



Let $\mathbf{X} = (X_1, \ldots, X_p)$ be an observed p-dimensional random vector from the p variate distribution P with finite second-order moments and covariance matrix $\mathbf{\Sigma}$.

We want to obtain ($\mathcal{S}^{p-1} := \{\mathbf{u} \in \mathbb{R}^p, \mathbf{u}'\mathbf{u} = 1\}$)

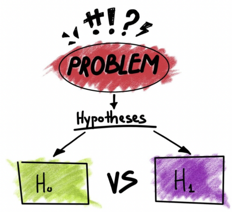
$$\boldsymbol{\beta}_1 := \mathrm{argmax}_{\boldsymbol{\beta} \in \mathcal{S}^{p-1}} \mathrm{Var}[\boldsymbol{\beta}'\mathbf{X}] = \mathrm{argmax}_{\boldsymbol{\beta} \in \mathcal{S}^{p-1}} \boldsymbol{\beta}'\mathbf{\Sigma}\boldsymbol{\beta}.$$

The first PC is then $Y_1 := \boldsymbol{\beta}_1'\mathbf{X}$. Then,

$$\boldsymbol{\beta}_2 := \mathrm{argmax}_{\boldsymbol{\beta} \in \mathcal{S}^{p-1}, \boldsymbol{\beta}'\boldsymbol{\beta}_1 = 0} \mathrm{Var}[\boldsymbol{\beta}'\mathbf{X}].$$

The second PC is then $Y_2 := \boldsymbol{\beta}_2'\mathbf{X}$.

It is well known that using the spectral decomposition of $\mathbf{\Sigma} = \sum_{j=1}^{p} \lambda_j \boldsymbol{\theta}_j \boldsymbol{\theta}_j'$ ($\lambda_1 > \lambda_2 > \ldots > \lambda_p$ are the eigenvalues and $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_p$ the associated eigenvectors of $\mathbf{\Sigma}$), then $\boldsymbol{\beta}_1 = \boldsymbol{\theta}_1, \boldsymbol{\beta}_2 = \boldsymbol{\theta}_2, \ldots$.



Let X be an observation described by a statistical model $\{\mathrm{P}_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \Theta\}$.

Consider a partition of Θ into $\Theta = \mathcal{H}_0 \oplus \mathcal{H}_1$ ($\oplus$ denotes a disjoint union) where $\mathcal{H}_0$ is called the "null hypothesis" and $\mathcal{H}_1$ is the "alternative".

A test problem is a decision problem in which only two decisions are possible.

(After observing X) → $R\mathcal{H}_0$ (rejection of $\mathcal{H}_0$)

or

→ $\bar{R}\mathcal{H}_0$ (non-rejection of $\mathcal{H}_0$)

Two mistakes can be made.

	$\boldsymbol{\theta} \in \mathcal{H}_0$	$\boldsymbol{\theta} \in \mathcal{H}_1$
$R\mathcal{H}_0$	Type I error (probability = α)	correct decision (probability = $1 - \beta$)
$\bar{R}\mathcal{H}_0$	correct decision probability = $1 - \alpha$	Type II error probability = β

$\longrightarrow$ The ideal test would be one that minimizes both first- and second-class risks.

Power of a test = $\mathrm{P}[R\mathcal{H}_0] = 1 - \beta$ when $\boldsymbol{\theta} \in \mathcal{H}_1$.

$\hookrightarrow$ we want to maximize the power.

Let $f_{\boldsymbol{\theta}}$ be a density of $\mathrm{P}_{\boldsymbol{\theta}}$ with respect to some measure μ.

Definition

The model $(\mathrm{P}_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \Theta)$ is called *differentiable in quadratic mean at* $\boldsymbol{\theta}$ is there exists measurable functions $\dot{\boldsymbol{\ell}}_{\boldsymbol{\theta}}$ such that, as $\boldsymbol{h} \to 0$,

$$\frac{1}{\|\boldsymbol{h}\|^2} \int \left\{ f_{\boldsymbol{\theta}+\boldsymbol{h}}^{1/2} - f_{\boldsymbol{\theta}}^{1/2} - \frac{1}{2} \boldsymbol{h}' \dot{\boldsymbol{\ell}}_{\boldsymbol{\theta}} f_{\boldsymbol{\theta}}^{1/2} \right\}^2 d\mu = o(1).$$

Definition

The sequence of statistical models $(\mathrm{P}_{\boldsymbol{\theta}}^{(n)} : \boldsymbol{\theta} \in \Theta)$ is *locally asymptotically normal* (LAN) at $\boldsymbol{\theta}$ if there exist matrices $\boldsymbol{r}_n$ and $\boldsymbol{\Gamma}_{\boldsymbol{\theta}}$ and random vectors $\boldsymbol{\Delta}_{\boldsymbol{\theta}}^{(n)}$ such that, under $\mathrm{P}_{\boldsymbol{\theta}}^{(n)}$, for every converging sequence $\boldsymbol{h}_n \to \boldsymbol{h}$,

$$\log \frac{d\mathrm{P}^{(n)}_{\boldsymbol{\theta}+\boldsymbol{r}_n^{-1}\boldsymbol{h}_n}}{d\mathrm{P}^{(n)}_{\boldsymbol{\theta}}} = \boldsymbol{h}'\boldsymbol{\Delta}_{\boldsymbol{\theta}}^{(n)} - \frac{1}{2}\boldsymbol{h}'\boldsymbol{\Gamma}_{\boldsymbol{\theta}}\boldsymbol{h} + o_{\mathrm{P}}(1) \quad \text{and} \quad \boldsymbol{\Delta}_{\boldsymbol{\theta}}^{(n)} \xrightarrow{\mathcal{L}} \mathcal{N}(\mathbf{0}, \boldsymbol{\Gamma}_{\boldsymbol{\theta}}).$$

Note : if the experiment $(\mathrm{P}_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \Theta)$ is differentiable in quadratic mean, then the sequence of model $(\mathrm{P}_{\boldsymbol{\theta}}^{(n)} : \boldsymbol{\theta} \in \Theta)$ is LAN (with norming matrices $r_n = \sqrt{n}\mathbf{I}$).

Testing problem: throughout the presentation, we consider the following problem test :

$$\begin{cases} \mathcal{H}_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0 \\ \mathcal{H}_1 : \boldsymbol{\theta} \neq \boldsymbol{\theta}_0 \end{cases}$$

where $\boldsymbol{\theta}$ is the eigenvector associated with the largest eigenvalue λ_1 of the underlying covariance matrix and $\boldsymbol{\theta}_0$ is a given unit vector of $\mathbb{R}^p$.

We will consider situations where $\lambda_1 - \lambda_2$ is small, that is a situation of weak identifiability of $\boldsymbol{\theta}$.

Objective : Paindaveine, Remy and Verdebout (2020) studied purely Gaussian procedures for this problem in the Gaussian case.

$\hookrightarrow$ the objective of this work is to extend the results of Paindaveine, Remy and Verdebout (2020) to

(i) the elliptical case;

(ii) signed-rank procedures.

We consider triangular arrays of elliptically symmetric observations $\mathbf{X}_{ni}$, $i = 1, \ldots, n$, $n = 1, 2, \ldots$ where $\mathbf{X}_{n1}, \ldots, \mathbf{X}_{nn}$ form an observed n-tuple of mutually independent p-dimensional random vectors with probability density function of the form

$$f_{\sigma_n^2, \mathbf{V}_n, f_1}(\mathbf{x}) := c_{p,f_1} \frac{1}{\sigma_n^p |\mathbf{V}_n|^{1/2}} f_1\left(\frac{1}{\sigma_n}(\mathbf{x}'\mathbf{V}_n^{-1}\mathbf{x})^{1/2}\right), \quad \mathbf{x} \in \mathbb{R}^p, \tag{1}$$

where, $\boldsymbol{\Sigma}_n := \mathbf{I}_p + r_n v \boldsymbol{\theta}\boldsymbol{\theta}'$, $\sigma_n := |\boldsymbol{\Sigma}_n|^{1/(2p)} = (1 + r_n v)^{1/2p}$ is a *scale parameter* in $\mathbb{R}_0^+ := (0, \infty)$,

$$\mathbf{V}_n := \boldsymbol{\Sigma}_n/\sigma_n^2 := (1 + r_n v)^{-1/p}(\mathbf{I}_p + r_n v \boldsymbol{\theta}\boldsymbol{\theta}')$$

is a *shape parameter* with eigenvalues

$$\lambda_{n1,\mathbf{V}_n} = (1 + r_n v)^{(p-1)/p} \quad \text{et} \quad \lambda_{n2;\mathbf{V}_n} = \cdots = \lambda_{np;\mathbf{V}_n} = (1 + r_n v)^{-1/p},$$

où r_n et v are positive real numbers and $f_1 : \mathbb{R}_0^+ \to \mathbb{R}^+ := [0, \infty)$ is a *standardized radial density* ($f_1 \in \mathcal{F}_1$).

The resulting hypothesis will be denoted as $\mathrm{P}^{(n)}_{\boldsymbol{\theta},r_n,v,f_1}$.

Examples :

(i) the p-variate multinormal distribution, with radial density $f_1(r) = \phi_1(r) := \exp(-a_p r^2/2)$;

(ii) the p-variate Student distributions, with radial densities (for $\nu \in \mathbb{R}_0^+$ degrees of freedom) $f_1(r) = f^t_{1,\nu}(r) := (1 + a_{p,\nu} r^2/\nu)^{-(p+\nu)/2}$;

where the positive constants a_p and $a_{p,\nu}$ are such that $f_1 \in \mathcal{F}_1$.

We denote by

$$d_{ni} := d_{ni}(\mathbf{V}_n) := \|\mathbf{V}_n^{-1/2}\mathbf{X}_{ni}\| \quad \text{and} \quad \mathbf{U}_{ni} := \mathbf{U}_{ni}(\mathbf{V}_n) := \mathbf{V}_n^{-1/2}\mathbf{X}_{ni}/d_{ni}$$

respectively the *standardized elliptical distances* and the *multivariate signs*, $i = 1, \ldots, n$.

We studied likelihood ratio of the form

$$\Lambda_n := \log \frac{\mathrm{dP}^{(n)}_{\boldsymbol{\theta}_0+\nu_n\boldsymbol{\tau}_n, r_n, v, f_1}}{\mathrm{dP}^{(n)}_{\boldsymbol{\theta}_0, r_n, v, f_1}}$$

for some admissible perturbations $\boldsymbol{\tau}_n$, where (ν_n) is a positive real sequence.

Some sequences (r_n) does not provide LAN (locally and asymptotically normal) experiments!

We first obtain a general result to derive the asymptotic behavior of Λ_n. In the following result, $f_{\boldsymbol{\vartheta}_n}$ is a density associated with a certain distribution $\mathrm{P}_{\boldsymbol{\vartheta}_n, f}$ with respect to a dominated measure μ and $\mathbf{Z}_{n1}, \ldots, \mathbf{Z}_{nn}$ form a random sample from the distribution $\mathrm{P}_{\boldsymbol{\vartheta}_n, f}$.

Proposition

Let (ν_n) be a positive real sequence and suppose that for all sequence $(\boldsymbol{\vartheta}_n) \in \boldsymbol{\Theta} \subset \mathbb{R}^p$, there exists real valued functions $\dot{\ell}^{(n)}_{\boldsymbol{\vartheta}_n,\boldsymbol{\tau}_n}$, where $\boldsymbol{\tau}_n$ is a bounded sequence in $\mathbb{R}^p$, such that

$$\int \left(f^{1/2}_{\boldsymbol{\vartheta}_n+\nu_n\boldsymbol{\tau}_n}(\mathbf{z}) - f^{1/2}_{\boldsymbol{\vartheta}_n}(\mathbf{z}) - \frac{1}{2}\dot{\ell}^{(n)}_{\boldsymbol{\vartheta}_n,\boldsymbol{\tau}_n}(\mathbf{z}) f^{1/2}_{\boldsymbol{\vartheta}_n}(\mathbf{z}) \right)^2 d\mu(\mathbf{z}) = o(n^{-1})$$

as $n \to \infty$. Assume furthermore that

$$\mathrm{E}_{\mathrm{P}^{(n)}_{\boldsymbol{\vartheta}_n,f}} [n\dot{\ell}^{(n)}_{\boldsymbol{\vartheta}_n,\boldsymbol{\tau}_n}(\mathbf{Z}_{n1})] = o(1), \quad \Gamma^{(n)}_{\boldsymbol{\vartheta}_n,\boldsymbol{\tau}_n} := \mathrm{E}_{\mathrm{P}^{(n)}_{\boldsymbol{\vartheta}_n,f}} \left[n(\dot{\ell}^{(n)}_{\boldsymbol{\vartheta}_n,\boldsymbol{\tau}_n}(\mathbf{Z}_{n1}))^2 \right] = O(1),$$

$$\left(\sum_{i=1}^{n} (\dot{\ell}^{(n)}_{\boldsymbol{\vartheta}_n,\boldsymbol{\tau}_n}(\mathbf{Z}_{ni}))^2 \right) - \Gamma^{(n)}_{\boldsymbol{\vartheta}_n,\boldsymbol{\tau}_n} = o_{\mathrm{P}^{(n)}_{\boldsymbol{\vartheta}_n,f}}(1)$$

and

$$\mathrm{E}_{\mathrm{P}^{(n)}_{\boldsymbol{\vartheta}_n,f}} [n(\dot{\ell}^{(n)}_{\boldsymbol{\vartheta}_n,\boldsymbol{\tau}_n}(\mathbf{Z}_{n1}))^2 \mathbb{I}[n(\dot{\ell}^{(n)}_{\boldsymbol{\vartheta}_n,\boldsymbol{\tau}_n}(\mathbf{Z}_{n1}))^2 \geq n\epsilon^2]] = o(1)$$

as $n \to \infty$. Then, as $n \to \infty$ under $\mathrm{P}^{(n)}_{\boldsymbol{\vartheta}_n,f}$,

$$\log \frac{d\mathrm{P}^{(n)}_{\boldsymbol{\vartheta}_n+\nu_n\boldsymbol{\tau}_n,f}}{d\mathrm{P}^{(n)}_{\boldsymbol{\vartheta}_n,f}} = \sum_{i=1}^{n} \dot{\ell}^{(n)}_{\boldsymbol{\vartheta}_n,\boldsymbol{\tau}_n}(\mathbf{Z}_{ni}) - \frac{1}{2}\Gamma^{(n)}_{\boldsymbol{\vartheta}_n,\boldsymbol{\tau}_n} + o_{\mathrm{P}}(1).$$

We need some further mild regularity conditions on f_1 in the sequel.

Define $\varphi_{f_1} := -\dot{f}_1/f_1$ where $\dot{f}_1$ is the a.e. derivative of f_1.

In the next result, f_1 is supposed to be in the collection of all absolutely continuous radial standardized densities $\mathcal{F}_1^a$ for which

$$\mathcal{J}_p(f_1) := \mathrm{E}[\varphi_{f_1}^2(d_{ni}/\sigma)(d_{ni}/\sigma)^2] < \infty.$$

The following theorem describes the asymptotic behavior of Λ_n in the following four regimes:

(i) $r_n \equiv 1$: *away from contiguity*,

(ii) $r_n = o(1)$ with $\sqrt{n}r_n \to \infty$: *above contiguity*,

(iii) $r_n = 1/\sqrt{n}$: *under contiguity*,

(iv) $r_n = o(1/\sqrt{n})$: *under strict contiguity*.

Theorem

(i) (away from contiguity) if $r_n \equiv 1$, then, with $\nu_n = 1/\sqrt{n}$,

$$\boldsymbol{\Delta}_{f_1}^{(n)} := \frac{v}{\sqrt{1+v}}\sqrt{n}(\mathbf{I}_p - \boldsymbol{\theta}_0\boldsymbol{\theta}_0')\Bigg(\frac{1}{n}\sum_{i=1}^{n}\varphi_{f_1}\left(\frac{d_{ni}}{\sigma_n}\right)\frac{d_{ni}}{\sigma_n}\mathbf{U}_{ni}\mathbf{U}_{ni}' - \mathbf{I}_p\Bigg)\boldsymbol{\theta}_0$$

and

$$\boldsymbol{\Gamma}_{f_1} = \frac{\mathcal{J}_p(f_1)v^2}{p(p+2)(1+v)}(\mathbf{I}_p - \boldsymbol{\theta}_0\boldsymbol{\theta}_0'),$$

we have that, under $\mathrm{P}^{(n)}_{\boldsymbol{\theta}_0,r_n,v,f_1}$,

$$\Lambda_n = \boldsymbol{\tau}_n'\boldsymbol{\Delta}_{f_1}^{(n)} - \frac{1}{2}\boldsymbol{\tau}_n'\boldsymbol{\Gamma}_{f_1}\boldsymbol{\tau}_n + o_{\mathrm{P}}(1)$$

and that $\boldsymbol{\Delta}_{f_1}^{(n)}$ is asymptotically normal with mean zero and covariance matrix $\boldsymbol{\Gamma}_{f_1}$;

Theorem

(ii) (above contiguity) if r_n is $o(1)$ with $\sqrt{n}r_n \to \infty$, then, with $\nu_n = 1/(\sqrt{n}r_n)$,

$$\boldsymbol{\Delta}_{f_1}^{(n)} := v\sqrt{n}(\mathbf{I}_p - \boldsymbol{\theta}_0\boldsymbol{\theta}_0')\left(\frac{1}{n}\sum_{i=1}^{n}\varphi_{f_1}\left(\frac{d_{ni}}{\sigma_n}\right)\frac{d_{ni}}{\sigma_n}\mathbf{U}_{ni}\mathbf{U}_{ni}' - \mathbf{I}_p\right)\boldsymbol{\theta}_0$$

and

$$\boldsymbol{\Gamma}_{f_1} = \frac{\mathcal{J}_p(f_1)v^2}{p(p+2)}(\mathbf{I}_p - \boldsymbol{\theta}_0\boldsymbol{\theta}_0'),$$

we have that, under $\mathrm{P}^{(n)}_{\boldsymbol{\theta}_0, r_n, v, f_1}$,

$$\Lambda_n = \boldsymbol{\tau}_n'\boldsymbol{\Delta}_{f_1}^{(n)} - \frac{1}{2}\boldsymbol{\tau}_n'\boldsymbol{\Gamma}_{f_1}\boldsymbol{\tau}_n + o_{\mathrm{P}}(1)$$

and that $\boldsymbol{\Delta}_{f_1}^{(n)}$ is asymptotically normal with mean zero and covariance matrix $\boldsymbol{\Gamma}_{f_1}$;

Theorem

(iii) (under contiguity) if $r_n = 1/\sqrt{n}$, then, letting $\nu_n \equiv 1$,

$$\Lambda_n = \boldsymbol{\tau}_n' \Big[v\sqrt{n}\Big(\frac{1}{n}\sum_{i=1}^n \varphi_{f_1}\Big(\frac{d_{ni}}{\sigma_n}\Big)\frac{d_{ni}}{\sigma_n}\mathbf{U}_{ni}\mathbf{U}_{ni}' - \mathbf{I}_p\Big)\Big(\boldsymbol{\theta}_0 + \frac{1}{2}\boldsymbol{\tau}_n\Big)\Big] - \frac{\mathcal{J}_p(f_1)v^2}{p(p+2)}\Big(\frac{\|\boldsymbol{\tau}_n\|^2}{2} - \frac{\|\boldsymbol{\tau}_n\|^4}{8}\Big) + o_{\mathrm{P}}(1),$$

under $\mathrm{P}^{(n)}_{\boldsymbol{\theta}_0, r_n, v, f_1}$, where, if $\boldsymbol{\tau}_n \to \boldsymbol{\tau}$, then

$$\boldsymbol{\tau}_n'\sqrt{n}\Big(\frac{1}{n}\sum_{i=1}^n \varphi_{f_1}\Big(\frac{d_{ni}}{\sigma_n}\Big)\frac{d_{ni}}{\sigma_n}\mathbf{U}_{ni}\mathbf{U}_{ni}' - \mathbf{I}_p\Big)\Big(\boldsymbol{\theta}_0 + \frac{1}{2}\boldsymbol{\tau}_n\Big)$$

is asymptotically normal with mean zero and covariance matrix

$$\frac{\mathcal{J}_p(f_1)}{p(p+2)}\Big(\|\boldsymbol{\tau}\|^2 - \frac{\|\boldsymbol{\tau}\|^4}{4}\Big);$$

Theorem

(iv) (under strict contiguity) if $r_n = o(1/\sqrt{n})$, then, even with $\nu_n \equiv 1$, we have that $\Lambda_n = o_{\mathrm{P}}(1)$ under $\mathrm{P}^{(n)}_{\boldsymbol{\theta}_0, r_n, v, f_1}$.

It follows that, for any fixed $v > 0$ and for any fixed sequence (r_n) associated with regime (i) or regime (ii), the sequence of models is LAN with central sequence

$$\boldsymbol{\Delta}^{(n)}_{\delta, f_1} := \frac{\sqrt{n} v}{\sqrt{1+\delta v}} (\mathbf{I}_p - \boldsymbol{\theta}_0 \boldsymbol{\theta}_0') \left(\frac{1}{n} \sum_{i=1}^{n} \varphi_{f_1}\left(\frac{d_{ni}}{\sigma_n} \right) \frac{d_{ni}}{\sigma_n} \mathbf{U}_{ni} \mathbf{U}_{ni}' - \mathbf{I}_p \right) \boldsymbol{\theta}_0$$

and Fisher information matrix

$$\boldsymbol{\Gamma}_{\delta, f_1} = \frac{\mathcal{J}_p(f_1) v^2}{p(p+2)(1+\delta v)} (\mathbf{I}_p - \boldsymbol{\theta}_0 \boldsymbol{\theta}_0')$$

where $\delta := 1$ if regime (i) is considered and $\delta := 0$ otherwise.

Below, we denote by $R_{ni}(\mathbf{V}_n)$ the rank of $d_{ni}(\mathbf{V}_n)$ among $d_{n1}(\mathbf{V}_n), \ldots, d_{nn}(\mathbf{V}_n)$.

The rank-based test $\phi_K = \phi_K^{(n)}$ of Hallin, Paindaveine and Verdebout (2010) rejects the null hypothesis (at asymptotic level α) when

$$Q_K := \frac{np(p+2)}{\mathcal{J}_p(K)} \sum_{j=2}^{p} \left(\tilde{\boldsymbol{\theta}}_j' \mathbf{S}_K^{(n)} \boldsymbol{\theta}_0\right)^2 > \chi^2_{p-1;1-\alpha}$$

where $\mathcal{J}_p(K)$ is a constant, $\tilde{\boldsymbol{\theta}}_j$ stands for a constrained estimator of $\hat{\mathbf{V}}_{\mathrm{Tyler}}$'s jth eigenvector for the shape estimator $\hat{\mathbf{V}}_{\mathrm{Tyler}}$ of Tyler (1987) and the signed-rank covariance matrix is of the form

$$\mathbf{S}_K^{(n)} := \frac{1}{n} \sum_{i=1}^{n} K\left(\frac{R_{ni}}{n+1}\right) \mathbf{U}_{ni} \mathbf{U}_{ni}',$$

with $K : (0,1) \to \mathbb{R}$ stands for some *score function*, $\mathbf{U}_{ni} = \mathbf{U}_{ni}(\tilde{\boldsymbol{\theta}}_0 \hat{\Lambda}_{\mathrm{Tyler}} \tilde{\boldsymbol{\theta}}_0')$ et $R_{ni} = R_{ni}(\tilde{\boldsymbol{\theta}}_0 \hat{\Lambda}_{\mathrm{Tyler}} \tilde{\boldsymbol{\theta}}_0')$ with $\tilde{\boldsymbol{\theta}}_0 := (\boldsymbol{\theta}_0, \tilde{\boldsymbol{\theta}}_2, \ldots, \tilde{\boldsymbol{\theta}}_p)$.

Proposition

Fix a unit p-vector $\boldsymbol{\theta}_0$, $v > 0$ *and* $g_1 \in \mathcal{F}_1$. *Then, for any sequence* r_n, Q_K *converges weakly to a chi-square random variable with* $p - 1$ *degrees of freedom under* $\mathrm{P}^{(n)}_{\boldsymbol{\theta}_0, r_n, v, g_1}$.

$\hookrightarrow$ ϕ_K is robust to weak identifiability.

$\hookrightarrow$ Impact on asymptotic efficiency properties?

Proposition

Fix a unit p-vector $\boldsymbol{\theta}_0$, $v > 0$, $g_1 \in \mathcal{F}_1$. *Then,*

(a) *when (i)* $r_n \equiv 1$ $(\delta = 1)$ *or (ii)* $r_n = o(1)$ *with* $\sqrt{n}r_n \to \infty$ $(\delta = 0)$, *we have that under* $\mathrm{P}^{(n)}_{\boldsymbol{\theta}_0+\boldsymbol{\tau}_n/(\sqrt{n}r_n),r_n,v,g_1}$, *the test statistic* Q_K *converges weakly to a chi-square random variable with* $p-1$ *degrees of freedom and non-centrality parameter*

$$\frac{\mathcal{J}_p^2(K,g_1)v^2}{\mathcal{J}_p(K)p(p+2)(1+\delta v)}\|\boldsymbol{\tau}\|^2,$$

where $\boldsymbol{\tau} := \lim_{n\to\infty}\boldsymbol{\tau}_n$;

Proposition

(b) *under* $\mathrm{P}^{(n)}_{\boldsymbol{\theta}_0+\boldsymbol{\tau}_n,1/\sqrt{n},v,g_1}$, Q_K *converges weakly to a chi-square random variable with* $p-1$ *degrees of freedom and with non-centrality parameter*

$$\frac{v^2\mathcal{J}_p^2(K,g_1)}{16\mathcal{J}_p(K)p(p+2)}\|\boldsymbol{\tau}\|^2\left(4-\|\boldsymbol{\tau}\|^2\right)\left(2-\|\boldsymbol{\tau}\|^2\right)^2; \tag{2}$$

(c) *when* $r_n = o(1)$ *with* $\sqrt{n}r_n \to 0$, *we have that under* $\mathrm{P}^{(n)}_{\boldsymbol{\theta}_0+\boldsymbol{\tau}_n,r_n,v,g_1}$, Q_K *converges weakly to a chi-square random variable with* $p-1$ *degrees of freedom.*

Therefore :

$$\mathrm{ARE}_{\boldsymbol{\theta}_0,r_n,v,g_1}(\phi_K/\phi_G) := \frac{(1+\kappa_p(g_1))\mathcal{J}_p^2(K,g_1)}{p(p+2)\mathcal{J}_p(K)}$$

$\hookrightarrow$ not affected by weak identifiability.

Simulations

First simulation exercise : for any $b = 0, 1, \ldots, 5$, we generate $M = 2500$ mutually independent random samples $\mathbf{X}_i^{(b,r)}$, $i = 1, \ldots, n = 200$, from the $p = 3$-variate t_1 ($r = 1$), t_5 ($r = 2$) and normal ($r = 3$) distributions, with mean zero and scatter matrix

$$\mathbf{\Sigma}_n^{(b)} := \mathbf{I}_p + n^{-b/6}\boldsymbol{\theta}_0\boldsymbol{\theta}_0'$$

where $\boldsymbol{\theta}_0 = (1, 0, 0)'$. This covers regimes (i) ($b = 0$), (ii) ($b = 1, 2$), (iii) ($b = 3$) and (iv) ($b = 4, 5$).

We perform, at nominal level 5%, the Wilcoxon test ϕ_{K_1} and the van der Waerden test ϕ_{K_Φ} for $\mathcal{H}_0^{(n)} : \boldsymbol{\theta} = \boldsymbol{\theta}_0$.

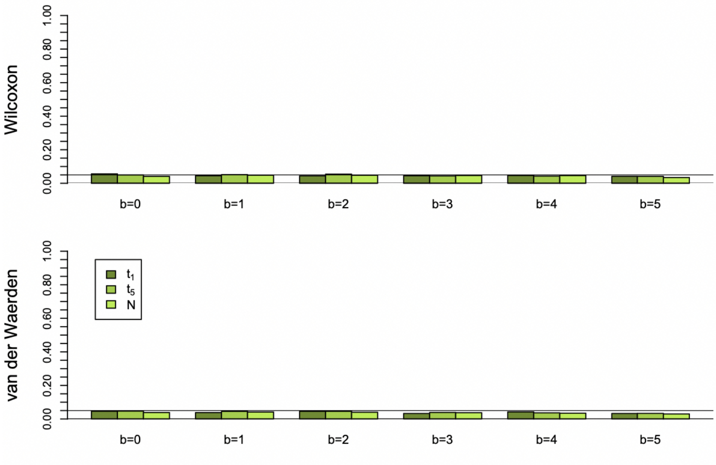


Figure: Empirical rejection frequencies, under the null hypothesis, of the Wilcoxon test ϕ_{K_1} and the van der Waerden test ϕ_{K_Φ}, performed at nominal level 5%.

Simulations

Second simulation exercice : we generate $M = 100\,000$ mutually independent random samples $\mathbf{X}_i^{(\ell)}$, $i = 1, \ldots, n = 10\,000$, $\ell = 0, 1, \ldots, 20$, from the $p = 2$-variate t_1 distribution with mean zero and scatter matrix

$$\boldsymbol{\Sigma}_n^{(\ell)} := \mathbf{I}_p + n^{-1/2}(\boldsymbol{\theta}_0 + \boldsymbol{\tau}_\ell)(\boldsymbol{\theta}_0 + \boldsymbol{\tau}_\ell)',$$

where $\boldsymbol{\theta}_0 = (1, 0)'$ et $\boldsymbol{\theta}_0 + \boldsymbol{\tau}_\ell = (\cos(\ell\pi/40), \sin(\ell\pi/40))'$.

The value $\ell = 0$ is associated to the null hypothesis, while values $\ell = 1, \ldots, 20$ provide increasingly severe alternatives.

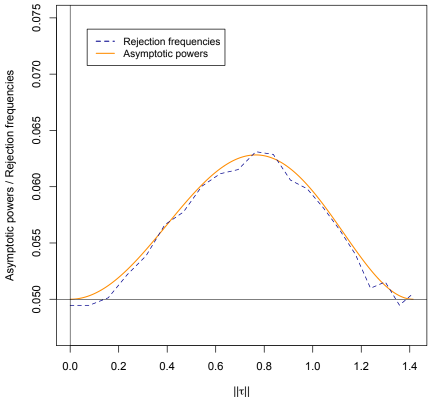


Figure: Empirical rejection frequencies, under the null hypothesis and local alternatives, of the van der Waerden test (ϕ_{K_ϕ}), performed at nominal level 5%.

References

- Hallin, M., Paindaveine, D. and Verdebout, Th. (2010). Optimal rank-based testing for principal components. *Annals of Statistics*, 38, 3245–3299.
- Paindaveine, D., Remy, J. and Verdebout, Th. (2020). Testing for principal component directions under weak identifiability. *Annals of Statistics*, 48, 324-345.
- Tyler, D. E. (1987). A distribution-free M-estimator of multivariate component vectors. *Annals of Statistics*, 11, 1243-1250.