# Inférence basée sur les rangs pour l'ACP sous faible identifiabilité

Davy Paindaveine [1] & Laura Peralvo Maroto [2] & Thomas Verdebout [3]

[1] *ECARES et Département de Mathématique, Université Libre de Bruxelles, Belgique, davy.paindaveine@ulb.be*

[2] *ECARES et Département de Mathématique, Université Libre de Bruxelles, Belgique, laura.peralvo.maroto@ulb.be*

[3] *ECARES et Département de Mathématique, Université Libre de Bruxelles, Belgique, thomas.verdebout@ulb.be*

**Résumé.** Sur base d'un échantillon aléatoire elliptique $p$-dimensionnel de matrice de forme $\mathbf{V}_n$, on considère le problème testant l'hypothèse nulle $\mathcal{H}_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$ contre la contre-hypothèse $\mathcal{H}_1 : \boldsymbol{\theta} \neq \boldsymbol{\theta}_0$, où $\boldsymbol{\theta}$ est le « premier » vecteur propre de la matrice de forme $\mathbf{V}_n$ et $\boldsymbol{\theta}_0$ est un $p$-vecteur unitaire fixé. Nous montrons que le test des rangs signés, obtenu par Hallin, Paindaveine et Verdebout (2010) pour les vecteurs propres, s'avère plus performant que le test pseudo-gaussien classique, même dans des conditions d'identifiabilité faible de $\boldsymbol{\theta}$. Plus précisément, même lorsque le rapport des deux plus grandes valeurs propres $\lambda_{n1}/\lambda_{n2}$ de $\mathbf{V}_n$ converge vers un, le test basé sur les rangs de van der Waerden domine uniformément le test pseudo-gaussien sous l'hypothèse d'ellipticité. Pour obtenir nos résultats, nous dérivons des résultats généraux pour le comportement asymptotique du rapport de vraisemblances de tableaux triangulaires d'observations et fournissons une représentation asymptotique de la statistique du test des rangs signés sous faible identifiabilité. Nos résultats sont confirmés par des simulations de Monte-Carlo.

**Mots-clés.** Densités elliptiques, normalité locale asymptotique, rangs et signes multivariés, analyse en composantes principales, matrices de diffusion à pointes, matrices de forme à pointes, faible identifiabilité.

**Abstract.** On the basis of a $p$-dimensional elliptical random sample with shape matrix $\mathbf{V}_n$, we consider the problem of testing the null hypothesis $\mathcal{H}_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$ against the alternative $\mathcal{H}_1 : \boldsymbol{\theta} \neq \boldsymbol{\theta}_0$, where $\boldsymbol{\theta}$ is the "first" eigenvector of the shape matrix $\mathbf{V}_n$ and $\boldsymbol{\theta}_0$ is a fixed unit $p$-vector. We show that the signed-rank test, obtained by Hallin, Paindaveine and Verdebout (2010) for the eigenvectors, outperforms classical pseudo-Gaussian tests even under weak identifiability of $\boldsymbol{\theta}$. More precisely, even if the ratio of the two largest eigenvalues $\lambda_{n1}/\lambda_{n2}$ of $\mathbf{V}_n$ converges to one, the van der Waerden rank test uniformly dominates the pseudo-Gaussian test under the assumption of ellipticity. To obtain our results, we derive general results for the asymptotic behavior of likelihood ratio of triangular arrays of observations and provide an asymptotic representation of signed-rank test statistics under weak identifiability. Our results are confirmed by Monte-Carlo simulations.

**Keywords.** Elliptical densities, local asymptotic normality, multivariate ranks and signs, principal component analysis, spiked scatter matrices, spiked shape matrices, weak identifiability.

# 1 Objectifs et résultats de la communication

L'un des outils statistiques les plus largement utilisés dans l'analyse statistique multivariée est l'analyse en composantes principales (ACP) dont l'objectif est la réduction de la dimension d'un $p$-vecteur aléatoire $\mathbf{X}$ en le projetant dans un nouvel espace de dimension inférieure. Cette projection géométrique permet de représenter les données de manière plus compacte, ce qui est essentiel pour la visualisation et la compréhension. Cette réduction de la dimension de $\mathbf{X}$ s'effectue tout en préservant l'essentiel de sa variabilité totale, typiquement capturée dans la matrice de covariance. Soit $\mathbf{\Sigma}_{\text{cov}}$ la matrice de covariance de $\mathbf{X}$ admettant la décomposition spectrale $\mathbf{\Sigma}_{\text{cov}} = \sum_{j=1}^{p} \lambda_j \boldsymbol{\theta}_j \boldsymbol{\theta}_j'$ ($\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p$). La $j^{\text{e}}$ composante principale est $\boldsymbol{\theta}_j'\mathbf{X}$ et est de variance $\lambda_j$. Les composantes principales obtenues par l'ACP sont linéairement indépendantes et orthogonales les unes aux autres. Géométriquement, cela signifie qu'elles forment un ensemble d'axes principaux perpendiculaires dans l'espace des composantes principales. La matrice $\mathbf{\Sigma}_{\text{cov}}$ étant généralement inconnue dans la pratique, il est naturel d'effectuer de l'inférence sur celle-ci, et plus particulièrement sur ses vecteurs propres.

Bien que les problèmes d'inférence traditionnels s'expriment à partir de la matrice de covariance $\mathbf{\Sigma}_{\text{cov}}$, ceux-ci peuvent s'étendre naturellement à des modèles elliptiques arbitraires caractérisés par un vecteur de *position* $\boldsymbol{\mu} \in \mathbb{R}^p$, un paramètre d'*échelle* $\sigma \in \mathbb{R}_0^+$, une matrice réelle de taille $p \times p$ symétrique définie positive $\mathbf{V}$ appelée matrice de *forme*, et une *densité radiale standardisée* $f_1$. Lorsque cette densité elliptique admet des moments finis d'ordre deux, la matrice de forme $\mathbf{V}$ et la matrice de covariance $\mathbf{\Sigma}_{\text{cov}}$ sont proportionnelles et partagent donc la même collection de vecteurs propres ainsi, qu'à un facteur positif près, la même collection de valeurs propres.

Notre intérêt réside plus précisément ici dans le problème de tester l'hypothèse nulle $\mathcal{H}_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$ contre la contre-hypothèse $\mathcal{H}_1 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$ , où $\boldsymbol{\theta}_0$ est un vecteur unitaire fixé de $\mathbb{R}^p$. Ce problème a été traité par Hallin, Paindaveine et Verdebout (2010) sur base d'un $n$-uple $\mathbf{X}_1, \ldots, \mathbf{X}_n$ de $p$-vecteurs elliptiques de position $\boldsymbol{\mu}$ et de matrice de forme $\mathbf{V}$. Plus précisément, en notant $\mathbf{Z}_i := \mathbf{V}^{-1/2}(\mathbf{X}_i - \boldsymbol{\mu})$ la version sphérique de $\mathbf{X}_i$ ($\mathbf{A}^{1/2}$ représente la racine symétrique définie positive de la matrice symétrique définie positive $\mathbf{A}$), par $\mathbf{U}_i = \mathbf{U}_i(\boldsymbol{\mu}, \mathbf{V}) := \mathbf{Z}_i / \|\mathbf{Z}_i\|$ les signes multivariés correspondants et par $R_i^{(n)} = R_i^{(n)}(\boldsymbol{\mu}, \mathbf{V})$ les rangs des normes $\|\mathbf{Z}_i\|$, $i = 1, \ldots, n$, le test basé sur les rangs de Hallin, Paindaveine et Verdebout (2010) $\underset{\sim}{\phi}{}_K^{(n)}$ rejette l'hypothèse nulle (au niveau asymptotique $\alpha$) lorsque

$$\underset{\sim}{Q}{}_K^{(n)} := \frac{np(p+2)}{\mathcal{J}_p(K)} \sum_{j=2}^{p} \left( \tilde{\boldsymbol{\theta}}_j' \, \underset{\sim}{\mathbf{S}}{}_K^{(n)} \boldsymbol{\theta}_0 \right)^2 \tag{1}$$

dépasse le quantile supérieur d'ordre $\alpha$ de la distribution chi-carré à $(p-1)$ degrés de liberté, où $\mathcal{J}_p(K)$ est une constante de normalisation, $\tilde{\boldsymbol{\theta}}_j$ représente un estimateur contraint du $j^{\text{e}}$ vecteur propre de $\mathbf{V}$ et la matrice de covariance de rangs signés est de la forme

$$\underset{\sim}{\mathbf{S}}{}_K^{(n)} := \frac{1}{n} \sum_{i=1}^{n} K\left( \frac{R_i^{(n)}}{n+1} \right) \mathbf{U}_i \mathbf{U}_i',$$

avec $K : (0,1) \to \mathbb{R}$ une *fonction de score*, et $\mathbf{U}_i = \mathbf{U}_i(\hat{\boldsymbol{\mu}}, \hat{\mathbf{V}})$ et $R_i^{(n)} = R_i^{(n)}(\hat{\boldsymbol{\mu}}, \hat{\mathbf{V}})$ sont calculés à partir d'estimateurs appropriés $\hat{\boldsymbol{\mu}}$ et $\hat{\mathbf{V}}$ de $\boldsymbol{\mu}$ et $\mathbf{V}$. En plus d'être robuste en termes de validité, Hallin, Paindaveine et Verdebout (2010) ont montré que ce test est, pour une fonction de score bien particulière $K = K_{f_1}$, *localement et asymptotiquement optimal* au sens de Le Cam) sous toute densité radiale régulière $f_1$.

Notre objectif est l'étude des propriétés asymptotiques de ce test dans un contexte d'*identifiabilité faible* introduit par Paindaveine, Remy et Verdebout (2020), en ce sens que le premier vecteur propre $\boldsymbol{\theta}$ n'est pas correctement identifié à la limite. En d'autres termes, nous considérons un cadre asymptotique non standard dans lequel les valeurs propres peuvent dépendre de $n$ et où $\lambda_{n1}/\lambda_{n2}$ converge vers 1 lorsque $n$ diverge vers l'infini. Nous nous limitons aux spectres de la forme $\lambda_{n1} > \lambda_{n2} = \ldots = \lambda_{np}$ en considérant des tableaux triangulaires d'observations à symétrie elliptique $\mathbf{X}_{ni}$, $i = 1, \ldots, n$, $n = 1, 2, \ldots$, où $\mathbf{X}_{n1}, \ldots, \mathbf{X}_{nn}$ forment un $n$-uple de $p$-vecteurs aléatoires mutuellement indépendants issu d'une distribution elliptique caractérisée par un paramètre de position $\boldsymbol{\mu}$, une matrice de forme définie positive

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

où $v$ est un nombre réel positif, $(r_n)$ est une suite réelle positive bornée, $\mathbf{I}_\ell$ désigne la matrice identité de dimension $\ell$ et pour une certaine fonction $f_1$. Les valeurs propres de la matrice de forme $\mathbf{V}_n$ sont alors $\lambda_{n1} = (1 + r_n v)^{(p-1)/p}$ (de vecteur propre correspondant $\boldsymbol{\theta}$) et $\lambda_{n2} = \cdots = \lambda_{np,\mathbf{V}_n} = (1 + r_n v)^{-1/p}$ (l'espace propre correspondant étant le complément orthogonal de $\boldsymbol{\theta}$ dans $\mathbb{R}^p$). Lorsque $r_n \equiv 1$, le rapport des deux premières valeurs propres valant $1 + v$, le premier vecteur propre reste bien identifié à la limite. Il en est de même dans le cas plus général où $r_n$ reste éloigné de 0 quand $n$ diverge vers l'infini. Par contre, si la suite $r_n$ converge vers zéro lorsque $n$ diverge vers l'infini, le premier vecteur propre $\boldsymbol{\theta}$ n'est plus correctement identifié à la limite. Ce concept d'identifiabilité faible rend donc le problème de test consistant à tester $\mathcal{H}_0^{(n)} : \boldsymbol{\theta} = \boldsymbol{\theta}_0$ contre $\mathcal{H}_1^{(n)} : \boldsymbol{\theta} = \boldsymbol{\theta}_0$ de plus en plus difficile lorsque $n$ diverge vers l'infini.

Dans un premier temps, nous construisons un résultat de représentation asymptotique de la statistique du test des rangs signés $\underset{\sim}{\phi}_K^{(n)}$ de Hallin, Paindaveine et Verdebout (2010), valable sous faible identifiabilité du paramètre $\boldsymbol{\theta}$ qui simplifie grandement l'étude du comportement asymptotique de ce test sous faible identifiabilité. A partir de ce résultat, nous montrons que le test des rangs signés $\underset{\sim}{\phi}_K^{(n)}$ est robuste en termes de validité à l'identifiabilité faible en ce sens qu'il respectera asymptotiquement la contrainte de niveau nominal dans des scénarios arbitrairement proches du cas sphérique, sans aucune hypothèse requise sur la densité sous-jacente au-delà de l'ellipticité.

Dans un second temps, nous souhaitons étudier des propriétés d'optimalité sur le test basé sur les rangs $\underset{\sim}{\phi}_K^{(n)}$ de Hallin, Paindaveine et Verdebout (2010). Dans ce sens, nous construisons un résultat général sur le comportement asymptotique du rapport de vraisemblances dans une configuration de tableaux triangulaires. En particulier, nous généralisons un résultat bien connu dans Van der Vaart (1998) aux tableaux triangulaires d'observations à savoir que la différentiabilité en moyenne quadratique d'une suite d'expériences implique que cette même suite est localement et asymptotiquement normale. Cette généralisation aux tableaux triangulaires nous permet par suite de dériver le comportement asymptotique du rapport de

vraisemblances elliptiques (avec une potentielle identifiabilité faible) et d'obtenir le comportement asymptotique du test basé sur les rangs sous des alternatives locales. On montre ainsi que ce test bénéficie encore de fortes propriétés d'optimalité sous faible identifiabilité.

Finalement, nous comparons le test basé sur les rangs $\underset{\sim}{\phi}{}_K^{(n)}$ au test pseudo-gaussien classique obtenu par Hallin, Paindaveine et Verdebout (2010) en calculant l'efficacité relative asymptotique du test basé sur les rangs par rapport au test pseudo-gaussien. On montre que celle-ci n'est pas impactée par l'identifibialité faible de $\boldsymbol{\theta}$.

Les différents résultats obtenus sont finalement illustrés à travers plusieurs exercices de Monte-Carlo.

# Bibliographie


Hallin, M., Paindaveine, D. et Verdebout, T. (2010), Optimal rank-based testing for principal components, *Annals of Statistics*, 38, pp. 3245-3299.

Paindaveine, D., Remy, J. et Verdebout, T. (2020), Testing for principal component directions under weak identifiability, *Annals of Statistics*, 48, pp. 324-345.

Van der Vaart, A. W. (1998), Asymptotic Statistics. Cambridge University Press, United States of America.