# Robust morphometric analysis based on landmarks

**Alfonso García-Pérez**

Versión final publicada en *The Mathematics of the Uncertain*, Editores: **E. Gil, E. Gil, J. Gil y M.A. Gil, Editorial Springer-Verlag, 2018:165-174.**

https://link.springer.com/chapter/10.1007/978-3-319-73848-2_16

Departamento de Estadística, I.O. y C.N., Universidad Nacional de Educación a Distancia (UNED),

Paseo Senda del Rey 9, 28040 Madrid, Spain; agar-per@ccia.uned.es

Abstract: Procrustes Analysis is a Morphometric method based on Configurations of Landmarks that estimates the superimposition parameters by least-squares; for this reason, the procedure is very sensitive to outliers. There are classical results, based on the normality of the observations, to test whether there are significant differences between individuals. In this paper we determine a Von Mises plus Saddlepoint approximation for the tail probability (*p*-value) of this test for the Procrustes Statistic, when the observations come from a model close to the normal.

## 1 Introduction

This paper is about a robust classification problem of $n$ individuals based on their shapes, i.e., using their geometric information. The usual (classical or robust) methods based on Multivariate Analysis cannot extract all the geometric information from the individuals. For this reason, in recent years, morphometrics methods based on Configurations of landmarks have been developed. A landmark is a peculiar point whose position is common in all the individuals to classify. For instance, when we classify skulls, the landmarks could be the center of the supraorbital arch, the chin, etc.; or, if we classify projectile points found in an archaeological site, the landmarks could be the ends of the points.

In all the cases, the mathematical (geometric) information that we obtain from the individuals is the $k$ coordinates of their $p$ landmarks, $l_i = (ci1, \ldots, cik)$, $i = 1, \ldots, p$.

The matrix of landmarks coordinates is called a Configuration. For each individual with $p$ landmarks of dimension $k$ (where $k$ is equal to 2 or 3) we have a collection of landmark coordinates expressed in $p \times k$ matrix as

$$M = \begin{pmatrix} c_{11} & \cdots & c_{1k} \\ \cdots & \cdots & \cdots \\ c_{p1} & \cdots & c_{pk} \end{pmatrix}$$

There are many morphometric methods; see for instance [1] or [3]. In this paper we consider Superimposition Methods; namely, Procrustes Analysis, obtaining the Procrustes coordinates and adapting the Configurations to a common (local) reference system, matching them at the common center. For these reasons, a Local Coordinate Reference System is needed and a Geographical Information System very useful.

A common graphical representation of a Configuration is a scatter plot of its landmarks coordinates. Joining the resulting points with segments we obtaining a polygon where the landmarks coordinates define the vertices of the polygon.

Because we use the shape of the individuals in their classification and shape is a property of an object that is invariant under scaling, rotation and translation (otherwise, for instance, an object and itself with double size could be classified into two different groups), in order to classify them with a Procrustes Analysis, we have first to remove the effect of Size (scale), Location (translation) and Orientation (rotation) to standardize them and match them in a common center in order to make them comparable.

This means that we have to estimate by least-squares the superimposition parameters $\alpha$, $\beta$ and $\boldsymbol{\Gamma}$ (scale, translation and rotation) in order to minimize the full Procrustes distance $d_F$ between Configurations $M_1$ and $M_2$, i.e.,

$$\min d_F(M_1, M_2) = \min ||M_2 - \alpha M_1 \boldsymbol{\Gamma} - \mathbf{1}_p \beta'||$$

$$= \sqrt{trace[(M_2 - \alpha M_1\, \boldsymbol{\Gamma} - \mathbf{1}_p \beta')'(M_2 - \alpha M_1\, \boldsymbol{\Gamma} - \mathbf{1}_p \beta')]}$$

where $\alpha$ is a scalar representing the Size, $\beta$ is a vector of $k$ values corresponding to a Location parameter formed by the centroid coordinates, $\mathbf{1}_p$ is a column vector of dimension $p \times 1$ and $\boldsymbol{\Gamma}$ a $k \times k$ square rotation matrix.

The idea that we pursue with this transformation is to match both Configurations, i.e., a superimposition of $M_1$ onto $M_2$.

It is possible to use a Classical Morphometric Analysis from a descriptive point of view. This is briefly exposed, together with its robustification by replacing the classical estimators with robust ones, in [6].

## 2 Classical Morphometric Analysis from an Inferential Point of View

Instead of considering a descriptive morphometric analysis it is more interesting to test if there are significant differences between two Configurations. From a classical point of view, we have the following result in [8, 11]: If $X_1$ and $X_2$ are two scaled

and centered Configurations of dimension $p \times k$, the *Residual Distance* between Configurations $X_1$ and $X_2$ is defined as

$$||X_2 - X_1||^2 = trace\left[(X_2 - X_1)'(X_2 - X_1)\right].$$

As saw before, the $k \times k$ square rotation matrix $\boldsymbol{\Gamma}$ is determined such that the Procrustes distance between these two Configurations $X_1$ and $X_2$ (i.e., between landmarks) is minimal

$$\min_{\boldsymbol{\Gamma}} ||X_2 - X_1\boldsymbol{\Gamma}||^2 = \min_{\boldsymbol{\Gamma}} trace\left[(X_2 - X_1\,\boldsymbol{\Gamma})'(X_2 - X_1\,\boldsymbol{\Gamma})\right].$$

This minimum obtained after matching (i.e., after translation, rotation and scaling) is called the *Procrustes statistic*:

$$G(X_1, X_2) = \min_{\boldsymbol{\Gamma}} ||X_2 - X_1\boldsymbol{\Gamma}||^2.$$

Under the null hypothesis $H_0$ that there is no systematic differences between Configurations $X_1$ and $X_2$, i.e., they belong to the same group, or more precisely, that if $\eta$ is a constant, they are of the form

$$X_2 = X_1 + \eta\,\boldsymbol{e}$$

where the $p \times k$ landmarks coordinates of Configuration $\boldsymbol{e}$ are univariate i.i.d. $N(0, 1)$, then

$$G(X_1, X_2) \approx \eta^2\,\chi^2_g$$

i.e., $G_s(X_1, X_2) = G(X_1, X_2)/\eta^2 \approx \chi^2_g$, where $g = kp - k(k+1)/2 - 1$. Hence, we can compute tail probabilities ($p$-values) for testing $H_0$. It must be $p > (k + 1)/2 + 1/k$ and obviously an integer.

## 3 Robust Morphometric Analysis from an Inferential Point of View

The standard normality of the landmarks coordinates is a very hard assumption. For this reason we shall use robust methods for testing $H_0$ assuming that the $p \times k$ landmarks coordinates of $e$ follow, not a standard normal distribution but a contaminated normal model:

$$\frac{X_2 - X_1}{\eta} \rightsquigarrow (1 - \varepsilon)N(0, 1) + \varepsilon N(0, \nu).$$

In this section we are going to compute the tail probabilities ($p$-values), assuming this contaminated model, using a VOM+SAD approximation.

We use this scale contaminated normal mixture model because the Configurations are matched at the common centroid that is the new origin and equal to 0, being the contamination in the scale the natural source of contamination in the observations.

## *3.1 Von Mises Approximations for the p-Value of the Procrustes Statistic*

In order to test the null hypothesis $H_0$ that there is no systematic differences between the standardized Configurations $X_1$ and $X_2$, using the Procrustes statistic $G_s(X_1, X_2)$ that follows a $\chi^2_g$ distribution under a normal model, we have the following result.

**Proposition 3.1** *Let $G_s(X_1, X_2)$ be the Procrustes statistic, that follows a $\chi^2_g$ distribution when the underlying model is a normal distribution, $\Phi_{\mu,\sigma}$. If the previous null hypothesis $H_0$ holds, the von Mises (VOM) approximation for the functional tail probability (if $F$ is close to the normal $\Phi_{\mu,\sigma}$) is*

$$P_F\{G_s(X_1, X_2) > t\} \simeq g \int_{-\infty}^{\infty} P\{\chi^2_{g-1} > t - (\tfrac{x-\mu}{\sigma})^2\}\, dF(x) - (g-1) P\{\chi^2_g > t\}.$$

*Proof* The von Mises (VOM) approximation for the functional tail probability is (if $F$ is close to the normal $\Phi_{\mu,\sigma}$)

$$p_g^F = P_F\{G_s(X_1, X_2) > t\} \simeq p_g^{\Phi} + \int \mathrm{TAIF}(x; t; \chi^2_g, \Phi_{\mu,\sigma})\, dF(x) \qquad (1)$$

where TAIF is the Tail Area Influence Function defined in [4].

Replacing the normal model by the contaminated normal model $\Phi^{\varepsilon} = (1-\varepsilon)\,\Phi_{\mu,\sigma} + \varepsilon\,\delta_x$ and computing the derivative at $\varepsilon = 0$ we obtain that

$$\begin{aligned}\mathrm{TAIF}(x; t; \chi^2_g, \Phi_{\mu,\sigma}) &= \frac{\partial}{\partial \varepsilon} P_{\Phi^{\varepsilon}}\{G_s(X_1, X_2) > t\}\Big|_{\varepsilon=0} \\ &= g P\{\chi^2_{g-1} > t - (x-\mu)^2/\sigma^2\} - g P\{\chi^2_g > t\}\end{aligned}$$

integrating now, we obtain the result. □

Considering a scale contaminated normal (SCN) model

$$(1-\varepsilon)N(0, 1) + \varepsilon N(0, \nu)$$

**Table 1** *Exact* and approximate $p$-values with $g = 3$

| $t$ | "Exact" | Approximate |
|---|---|---|
| 6 | 0′149 | 0′148 |
| 8 | 0′077 | 0′076 |
| 10 | 0′042 | 0′042 |
| 12 | 0′024 | 0′025 |
| 14 | 0′016 | 0′016 |
| 16 | 0′011 | 0′011 |
| 18 | 0′007 | 0′008 |

the VOM approximation is

$$p_g^F \simeq (1 - g\,\varepsilon) P\{\chi_g^2 > t\} + g\,\varepsilon \int_{-\infty}^{\infty} P\{\chi_{g-1}^2 > t - x^2\}\, d\Phi_{0,\nu}(x).$$

In Table 1 appear, [10], the *Exact* values (obtained through a simulation of 100.000 samples) and the VOM approximations when $\varepsilon = 0'05$, $\nu = 2$ and $g = 3$.

To obtain the previous numerical results we had to deal with numerical integration. Sometimes, we would like to have analytic expressions of $p_g^F$ to value the effect of contamination $\varepsilon$, etc. For this reason, and for controlling the relative error of the approximation, in the next section we shall compute the Saddlepoint approximation for the $p$-value of the Procrustes Statistic.

## *3.2 Saddlepoint Approximations for the p-Value of the Procrustes Statistic*

Using Lugannani and Rice formula, [9], for the sample mean of $g$ independent square normal variables, we obtain the VOM+SAD approximation given in the next result.

**Proposition 3.2** *Let $G_s(X_1, X_2)$ be the Procrustes statistic, that follows a $\chi_g^2$ distribution when the underlying model is a normal distribution, $\Phi_{\mu,\sigma}$. If the null hypothesis $H_0$ holds, the saddlepoint approximation of the von Mises expansion, VOM+SAD approximation, for the functional tail probability (if $F$ is close to the normal $\Phi_{\mu,\sigma}$) is*

$$P_F\{G_s(X_1, X_2) > t\} \simeq P\left\{\chi_g^2 > t\right\} - B + B \int_{-\infty}^{\infty} \frac{\sqrt{g}}{\sqrt{t}}\, e^{\frac{(t-g)(x-\mu)^2}{2t\sigma^2}}\, dF(x) \qquad (2)$$

*where* $B = \frac{g\sqrt{g}}{\sqrt{\pi}\,(t-g)}\, e^{-(t-g-g\cdot\log(t/g))/2}$.

*Proof* If $G_s(X_1, X_2)$ follows a $\chi^2_g$ distribution, and $Y_1, \ldots, Y_g$ are $g$ independent gamma distributions $\gamma(1/2, 1/2)$ with moment generating function $M$ and cumulant generating function $K = \log M$, it is, following [2, 5, 9] or [7],

$$P_\Phi\left\{\frac{G_s(X_1, X_2)}{g} > t\right\} = P\left\{\frac{1}{g}\sum_{i=1}^{g} Y_i > t\right\}$$

$$= 1 - \Phi_s(w) + \phi_s(w)\left\{\frac{1}{r} - \frac{1}{w} + O(g^{-3/2})\right\} \qquad (3)$$

where $\Phi_s$ and $\phi_s$ are the cumulative distribution and density functions of the standard normal distribution.

If $K$ is the cumulant generating function, that is the functional of $\Phi_{\mu,\sigma}$,

$$K(\theta) = \log \int_{-\infty}^{\infty} e^{\theta\,(u-\mu)^2/\sigma^2}\, d\Phi_{\mu,\sigma}(u)$$

and $z_0$ is the (functional) saddlepoint, i.e., it is the solution of the equation $K'(z_0) = t$, the functionals that appear in (3) are

$$w = sign(z_0)\sqrt{2\,g\cdot(z_0\,t - K(z_0))} = \sqrt{g}\,sign(z_0)\sqrt{2\,(z_0\,t - K(z_0))} := \sqrt{g}\,w_1$$

$$r = z_0\sqrt{g\cdot K''(z_0)} = \sqrt{g}\,z_0\sqrt{K''(z_0)} := \sqrt{g}\,r_1.$$

As we saw before, the VOM approximation for the tail probability depends on the TAIF. To obtain the TAIF of $G_s(X_1, X_2)/g$ at $\Phi_{\mu,\sigma}$ we have to replace the model $\Phi_{\mu,\sigma}$ by the contaminated model $\Phi^\varepsilon = (1-\varepsilon)\Phi_{\mu,\sigma} + \varepsilon\,\delta_x$ in all the functionals in the right side of (3) that depend on $\Phi_{\mu,\sigma}$, and then to obtain the derivative at $\varepsilon = 0$; this process is represented with a dot over the functional. Since $\phi_s'(w) = -\phi_s(w)\,w$ and $\phi_s(w) \le 1$ , we obtain that

$$\mathrm{TAIF}\left(x; t; \frac{G_s(X_1, X_2)}{g}, \Phi_{\mu,\sigma}\right) = \frac{\partial}{\partial\varepsilon} P_{\Phi^\varepsilon}\left\{\frac{G_s(X_1, X_2)}{g} > t\right\}\Big|_{\varepsilon=0}$$

$$= -\phi_s(w)\,\dot{w} + \phi_s'(w)\,\dot{w}\left\{\frac{1}{r} - \frac{1}{w} + O(g^{-3/2})\right\} + \phi_s(w)\left\{-\frac{\dot{r}}{r^2} + \frac{\dot{w}}{w^2} + O(g^{-3/2})\right\}$$

$$= \phi_s(w)\left[-\frac{w\,\dot{w}}{r} - \frac{\dot{r}}{r^2} + \frac{\dot{w}}{w^2}\right] + O(g^{-1})$$

$$= \phi_s(w)\left[-\frac{\sqrt{g}\,w_1\,\sqrt{g}\;\overset{\bullet}{w}_1}{\sqrt{g}\,r_1} - \frac{\sqrt{g}\;\overset{\bullet}{r}_1}{g\,r_1^2} + \frac{\sqrt{g}\;\overset{\bullet}{w}_1}{g\,w_1^2}\right] + O(g^{-1})$$

$$= \frac{\phi_s(w)}{r_1}\left[-\sqrt{g}\cdot w_1\;\overset{\bullet}{w}_1\right] + O(g^{-1/2})$$

because the functionals $w_1$, $\overset{\bullet}{w}_1$, $r_1$ and $\overset{\bullet}{r}_1$ do not depend on $g$. Since

$$\overset{\bullet}{w}_1 = sign(z_0)\,\frac{2(\overset{\bullet}{z}_0\;t - \overset{\bullet}{K}\;(z_0))}{2\sqrt{2(z_0\,t - K(z_0))}} = \frac{\overset{\bullet}{z}_0\;t - \overset{\bullet}{K}\;(z_0))}{w_1}$$

it will be

$$\text{TAIF}\left(x; t; \frac{G_s(X_1, X_2)}{g}, \Phi_{\mu,\sigma}\right) = \frac{\phi_s(w)}{r_1}\sqrt{g}\left[\overset{\bullet}{K}\;(z_0) - \overset{\bullet}{z}_0\;t\right] + O(g^{-1/2}). \quad (4)$$

Hence, we have to compute the influence functions $\overset{\bullet}{K}\;(z_0)$ and $\overset{\bullet}{z}_0$. To do this, because

$$K'(\theta) = \frac{\displaystyle\int_{-\infty}^{\infty} e^{\theta\,(u-\mu)^2/\sigma^2}\;\left(\frac{u-\mu}{\sigma}\right)^2 d\Phi_{\mu,\sigma}(u)}{\displaystyle\int_{-\infty}^{\infty} e^{\theta\,(u-\mu)^2/\sigma^2}\;\;d\Phi_{\mu,\sigma}(u)}$$

from the saddlepoint equation, $K'(z_0) = t$ , we obtain

$$\int_{-\infty}^{\infty} e^{z_0\,(u-\mu)^2/\sigma^2}\;\left[\left(\frac{u-\mu}{\sigma}\right)^2 - t\right]\,d\Phi_{\mu,\sigma}(u) = 0.$$

Replacing again the model by the contaminated model $\Phi^{\varepsilon} = (1-\varepsilon)\,\Phi_{\mu,\sigma} + \varepsilon\,\delta_x$ before obtaining the derivative at $\varepsilon = 0$, and making the change of variable $(u - \mu)/\sigma = y$, we obtain

$$\overset{\bullet}{z}_0\left[\int_{-\infty}^{\infty} e^{z_0\,y^2}\;y^4\,d\Phi_s(y) - t\int_{-\infty}^{\infty} e^{z_0\,y^2}\;y^2\,d\Phi_s(y)\right] + e^{z_0\,(x-\mu)^2/\sigma^2}\left[\left(\frac{x-\mu}{\sigma}\right)^2 - t\right] = 0$$

i.e.,

$$\overset{\bullet}{z}_0 = \frac{1}{2}\,t^{-5/2}\,e^{\frac{(t-1)(x-\mu)^2}{2t\sigma^2}}\;\left[t - \left(\frac{x-\mu}{\sigma}\right)^2\right].$$

In a similar way, we obtain that

$$\dot{K}\,(z_0) = \frac{3}{2}\,t^{-1/2}\,e^{z_0\,(x-\mu)^2/\sigma^2} - \frac{1}{2}\,t^{-3/2}\,e^{z_0\,(x-\mu)^2/\sigma^2}\,\left(\frac{x-\mu}{\sigma}\right)^2 - 1.$$

Also it is

$$r_1 = z_0\,\sqrt{K''(z_0)} = \frac{t-1}{\sqrt{2}} \quad \text{and} \quad \phi_s(w) = \frac{1}{\sqrt{2\pi}}\,e^{-g\cdot(t-1-\log t)/2}.$$

Therefore, from (4), it will be

$$\text{TAIF}\left(x;\,t;\,\frac{G_s(X_1,\,X_2)}{g},\,\Phi_{\mu,\sigma}\right) = A\left(\frac{1}{\sqrt{t}}\,e^{\frac{(t-1)(x-\mu)^2}{2t\sigma^2}} - 1\right) + O(g^{-1/2})$$

where

$$A = \frac{\sqrt{g}}{\sqrt{\pi}\,(t-1)}\,e^{-g\cdot(t-1-\log t)/2}.$$

From (1), we obtain now the VOM+SAD approximation for the $p$-value of the test statistic $G_s(X_1,\,X_2)/g$,

$$P_F\left\{\frac{G_s(X_1,\,X_2)}{g} > t\right\} \simeq P\left\{\chi^2_g > g\,t\right\} - A + A\int_{-\infty}^{\infty}\frac{1}{\sqrt{t}}\,e^{\frac{(t-1)(x-\mu)^2}{2t\sigma^2}}\,dF(x)$$

and from this, we obtain the approximation (2) for the test statistic $G_s(X_1,\,X_2)$. □

If $F$ is the location contaminated normal mixture (LCN),

$$F = (1-\varepsilon)\,N(0,\,1) + \varepsilon\,N(\theta,\,1)$$

the VOM+SAD approximation is

$$P_F\left\{G_s(X_1,\,X_2) > t\right\} \simeq P\left\{\chi^2_g > t\right\} + \varepsilon\,B\left[e^{-(1-t/g)\theta^2/2} - 1\right].$$

In Table 2 appear the *Exact* values (obtained through simulation of 100.000 samples), the VOM and the VOM+SAD approximations when $\varepsilon = 0'01$, $\theta = 1$ and $g = 5$.

**Corollary 3.1** *To test the null hypothesis $H_0$ that there is no systematic differences between the standardized Configurations $X_1$ and $X_2$ with p landmarks of dimension k (i.e., $X_1$ and $X_2$ belong to the same classification group) using the Procrustes statistic $G_s(X_1,\,X_2)$ and assuming that the error difference between Configurations*

**Table 2** *Exact* and approximate $p$-values with $g = 5$

| $t$ | "Exact" | VOM appr. | VOM+SAD appr. |
|---|---|---|---|
| 9 | 0′1125 | 0′1129 | 0′1136 |
| 11 | 0′0538 | 0′0539 | 0′0545 |
| 13 | 0′0251 | 0′0249 | 0′0253 |
| 15 | 0′0114 | 0′0112 | 0′0115 |
| 17 | 0′0050 | 0′0049 | 0′0051 |
| 19 | 0′0022 | 0′0022 | 0′0023 |

$$\frac{X_2 - X_1}{\eta}$$

*follows a scale contamination normal model* $(1-\varepsilon)N(0,1) + \varepsilon N(0,\nu)$, *the VOM+SAD approximation for the tail probability (p-value) is*

$$P\{G_s(X_1, X_2) > t\} \approx P\{\chi^2_g > t\} + \varepsilon \frac{g^{3/2}}{\sqrt{\pi}(t-g)} \left[ \frac{\sqrt{g}}{\sqrt{t - \nu^2(t-g)}} - 1 \right]$$

$$\cdot \exp\left\{ -\frac{1}{2} \left( t - g - g \cdot \log \frac{t}{g} \right) \right\} \tag{5}$$

*where* $g = kp - k(k+1)/2 - 1$. *It must be* $p > (k+1)/2 + 1/k$ *and obviously an integer.*

*Then, if* $k = 2$, *it is* $g = 2p - 4$ *and* $p > 2$. *And if* $k = 3$, *it is* $g = 3p - 7$ *and* $p \geq 3$.

There are some applications of this approximation in [6]. There we test if there are significance differences between dots of *Notch tips and bay leaves*, of Solutrense period, that were found in caves of Asturias (Spain). We do this analysis using a photo of the "Museo Arqueológico de Asturias" (Oviedo), including this photo in QGIS as a raster layer.

## 4 Conclusions

Classical Morphometric Analysis based on Landmarks is not robust because it is based on sample means and least-squares estimation using a Normal distribution as model.

In this paper we consider a Contaminated Normal Model to make robust inferences. Namely, for this mixture model we obtain an von Mises approximation of the

$p$-value of a test for the null hypothesis of no significance differences between two individuals based on their shapes.

We also obtain a very accurate saddlepoint approximation of this von Mises approximation.

**Acknowledgements** This work is partially supported by Grant MTM2015-67057-P from Ministerio de Economía, Industria y Competitividad (Spain).

## References

1. Claude J (2008) Morphometrics with R. Springer, New York
2. Daniels HE (1983) Saddlepoint approximations for estimating equations. Biometrika 70:89–96
3. Dryden IL, Mardia KV (2016) Statistical shape analysis (with applications in R). Wiley, Chichester
4. Field CA, Ronchetti E (1985) A tail area influence function and its application to testing. Commun Stat 4:19–41
5. García-Pérez A (2006) Chi-square tests under models close to the normal distribution. Metrika 63:343–354
6. García-Pérez A, Cabrero-Ortega MY (2017) Robust morphometric analysis based on landmarks. Applications. arXiv:1703.04642
7. Jensen JL (1995) Saddlepoint approximations. Clarendon Press, Oxford
8. Langron SP, Collins AJ (1985) Perturbation theory for procrustes analysis. J R Stat Soc Ser B 47:277–284
9. Lugannani R, Rice S (1980) Saddle point approximation for the distribution of the sum of independent random variables. Adv Appl Probab 12:475–490
10. R Development Core Team (2016) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Viena. http://www.R-project.org
11. Sibson R (1979) Studies in the robustness of multidimensional scaling: perturbational analysis of classical scaling. J R Stat Soc Ser B Stat Methodol 41:217–229