HAL
open science

The Rao's distance between negative binomial distributions for Exploratory Analyses and Goodness-of-Fit Testing

C. Manté

▶ **To cite this version:**

C. Manté. The Rao's distance between negative binomial distributions for Exploratory Analyses and Goodness-of-Fit Testing. 61st World Statistics Congress - ISI2017, Jul 2017, Marrakech, Morocco. hal-01632444

HAL Id: hal-01632444

https://hal.science/hal-01632444v1

Submitted on 10 Nov 2017

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

The Rao's distance between negative binomial distributions for Exploratory Analyses and Goodness-of-Fit Testing

Manté Claude
Aix-Marseille Université, Université du Sud Toulon-Var, CNRS/INSU, IRD, MIO, UM 110, Marseille, France claude.mante@mio.osupytheas.fr

Abstract

The statistical analysis of counts of living organisms brings information about the collective behavior of species (schooling, habitat preference, etc), possibly depending on their biological characteristics (growth rate, reproductive power, survival rate, etc). The negative binomial distribution (NB) is widely used to model such data but the parametric approach is ill-suited from an exploratory point of view. Indeed, the "visual" distance between parameters is not relevant, because it depends on the chosen parametrization! On the contrary, considering the Riemannian manifold $NB(D_{\mathcal{R}})$ of negative binomial distributions equipped with the Fisher-Rao metrics, it is possible to compute intrinsic distances between species.
In this work, we focus on geometrical aspects of the χ^2 goodness-of-fit (GOF) test for distributions in $NB(D_{\mathcal{R}})$, in connection with the position of the reference distribution. We show that this position is critical for performances of this test, as Critchley & Marriott (2016) noticed in a different setting.

Keywords: Information geometry, Riemannian manifold, χ^2 test, overdispersed data.

1. Introduction

In a seminal paper, Rao (1945) noticed that, equipped with the Fisher information metrics denoted $\mathfrak{g}(\bullet)$, a family of probabilities depending on p parameters can be considered as a p-dimensional Riemannian manifold. The associated Riemannian (Rao's) distance between the distributions of parameters θ^1 and θ^2 is given by:

$$D_{\mathcal{R}}\left(\theta^1,\theta^2\right) := \int_0^1 \sqrt{\dot{\gamma}'(t).\mathfrak{g}(\gamma(t)).\dot{\gamma}(t)}dt \tag{1}$$

where γ is a **segment** (minimal length curve) connecting $\theta^1 = \gamma(0)$ to $\theta^2 = \gamma(1)$. As any Riemannian distance, $D_{\mathcal{R}}$ is **intrinsic** (*i.e.* does not depend on the parametrization used). Following this pioneer work, a number of authors used the Rao's distance to deal with various statistical topics: exploratory methods such as data visualization, clustering and classification, or hypothesis testing problems (Menendez *et al.* , 1995; Cubedo & Oller , 2002).
We will focus on the latest topic (GOF tests), in the setting of the Riemannian manifold $NB(D_{\mathcal{R}})$ of negative binomial distributions equipped with this distance. A probability distribution $\mathfrak{L}^i$ will be identified with its coordinates with respect to some chosen parametrization; for instance, we will write $\mathfrak{L}^i \equiv \left(\phi^i,\mu^i\right)$.

2. Essential elements of Riemannian geometry

Consider a Riemannian manifold $\mathfrak{M}$, and a parametric curve $\alpha : [a,b] \to \mathfrak{M}$. Its first derivative with respect to "time" will be denoted $\dot{\alpha}$. We will also consider for any $\theta \in \mathfrak{M}$ the local norm $\|V\|_g(\theta)$ associated with the metrics $\mathfrak{g}$ on the tangent space $T_\theta\mathfrak{M}$:

$$\forall V \in T_\theta\mathfrak{M},\ \|V\|_{\mathfrak{g}}(\theta) := \sqrt{V'.\mathfrak{g}(\theta).V}. \tag{2}$$

Definition 1. *(Berger , 2003) Let $\gamma : [0,1] \to \mathfrak{M}$ be a curve traced on $\mathfrak{M}$, and $\boldsymbol{D}$ be a connection on $\mathfrak{M}$. γ is a geodesic with respect to $\boldsymbol{D}$ if its acceleration $\boldsymbol{D}_{\dot{\gamma}(t)}\dot{\gamma}(t)$ is null $\forall t \in]0,1[$. In other words, **a geodesic has constant speed** in the local norm (2):*

$$\|\dot{\gamma}\|_{\mathfrak{g}} := \|\dot{\gamma}(\bullet)\|_{\mathfrak{g}}(\gamma(\bullet)) = \sqrt{\dot{\gamma}'(\bullet).\mathfrak{g}(\gamma(\bullet)).\dot{\gamma}(\bullet)}.$$

Geodesics on a p-dimensional Riemannian manifold with respect to the metric connection ∇ are solutions of the Euler-Lagrange equation (Berger , 2003; Burbea , 1986):

$$\forall\, 1 \leq k \leq p,\ \ddot{\gamma}_k(t) + \sum_{i,j=1}^{p} \Gamma_{i,j}^{k}\, \dot{\gamma}_i(t)\, \dot{\gamma}_j(t) = 0 \tag{3}$$

where each coefficient of ∇ (some "Christoffel symbol" $\Gamma_{i,j}^{k}$) only depends on $\mathfrak{g}$. To determine the shortest curve between two points of $\mathfrak{M}$, one applies the following result.

Lemma 1. *(Berger , 2003) Let $\mathfrak{M}$ be an abstract surface, and p, $q \in \mathfrak{M}$. Suppose that $\alpha : [a,b] \to \mathfrak{M}$ is a curve of minimal length connecting p to q. Then, α is a geodesic.*

Nevertheless, building the segment connecting $\mathfrak{L}^1$ to $\mathfrak{L}^2$ is not straightforward, since the lemma above only says that a segment is a geodesic. But a geodesic is not necessarily a segment, since it can include cut points; see (Berger , 2003; Manté & Kidé , 2016) for more details on cut points and their detection.

Definition 2. *(Berger , 2003) Let $\mathfrak{M}$ be a Riemann manifold and $x \in \mathfrak{M}$. The exponential map of $\mathfrak{M}$ at x is $\exp_x : W_x \to \mathfrak{M}$, defined on some neighborhood W_x of the origin of $T_x\mathfrak{M}$ by:*

$$\exp_x(V) := \alpha_{\mathcal{B}(V)}(\|V\|)$$

where $\mathcal{B}(V)$ is the projection of V onto the unit ball and $\alpha_{\mathcal{B}(V)}$ is the unique unit-speed geodesic in $\mathfrak{M}$ such that $\alpha_{\mathcal{B}(V)}(0) = x$ and $\dot{\alpha}_{\mathcal{B}(V)}(0) = \mathcal{B}(V)$.

Generally, $D_{\mathcal{R}}\left(\mathfrak{L}^1, \mathfrak{L}^2\right)$ cannot be obtained in a closed-form. It must be computed by finding the numerical solution of (3) completed by the boundary conditions

$$\left\{\gamma(0) = \theta^1,\ \gamma(1) = \theta^2\right\}. \tag{4}$$

But geodesics can be as well be computed by solving (3) under the alternative constraints

$$\left\{\gamma(0) = \theta^1,\ \dot{\gamma}(0) = V \in \mathbb{R}^2\right\} \tag{5}$$

where V is the initial velocity of the geodesic; this solution is associated with the exponential map at θ^1.

3. Tangent plane approximation and χ^2 GOF tests

Remember that $D_{\mathcal{R}}$ has been constructively defined by formula (1); the following proposition shows that it is really a metric, at least locally.

Proposition 1. *(Berger , 2003) For ϱ small enough, the exponential map at θ^0 is a local diffeomorphism, such that $\exp_{\theta^0}(\mathcal{B}(0,\varrho)) = B_{\mathcal{R}}\left(\theta^0, \varrho\right)$ (metric balls of radius ϱ of $T_{\theta^0}\mathfrak{M}$ and $\mathfrak{M}$, respectively).*

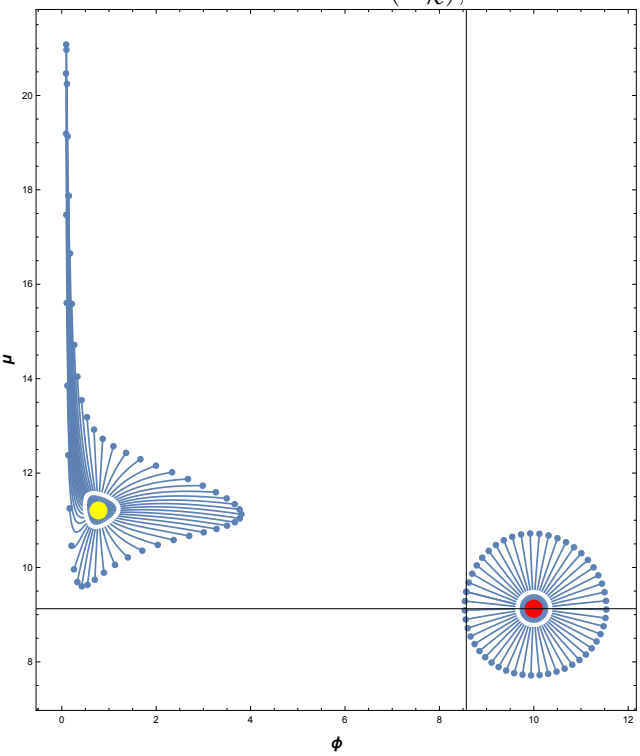
The maximal admissible value of ϱ (named injectivity radius of $\mathfrak{M}$ at θ^0) is denoted $\bar{\varrho}\left(\theta^0\right)$. Considering the first-order approximation of $D_{\mathcal{R}}$ we can write, if θ is close enough to θ^0:

$$D_{\mathcal{R}}\left(\mathfrak{L}^{\theta^0}, \mathfrak{L}^{\theta}\right) \equiv D_{\mathcal{R}}\left(\theta^0, \theta\right) \approx \sqrt{{}^t\left(\theta^0 - \theta\right).\mathfrak{g}\left(\theta^0\right).\left(\theta^0 - \theta\right)} \equiv \left\|\theta^0 - \theta\right\|_{\mathfrak{g}}\left(\theta^0\right). \tag{6}$$

Thus, if $D_{\mathcal{R}}\left(\mathfrak{L}^{\theta^0}, \mathfrak{L}^{\theta}\right) = \varrho \leq \bar{\varrho}\left(\theta^0\right)$ is small enough, the metric sphere $S_{\mathcal{R}}\left(\theta^0, \varrho\right) := \left\{\theta \in \mathfrak{M} :\ D_{\mathcal{R}}\left(\theta^0, \theta\right) = \varrho\right\}$ can be isometrically identified with the centered ellipsoid $\mathcal{E}_{\theta^0}(\varrho)$ drawn on $T_{\theta^0}\mathfrak{M}$. This is illustrated on Figure 1, where we displayed $\exp_{\theta}(\mathcal{E}_{\theta}(\varrho)) = S_{\mathcal{R}}(\theta, \varrho)$ for NB distributions of parameters $\theta^1 = \{0.7767, 11.2078\}$ (yellow point, a rather aggregative distribution) and $\theta^2 = \{10., 9.12624\}$ (red point, a bell-shaped distribution). We can see on Figure 1 that for $\varrho = 0.3$ and $\varrho = 1.5$, the geodesics emanating from θ^2 draw circles, while the situation is very different for θ^1 : for $\varrho = 1.5$, the ball is dramatically anisotropic, very far from an ellipse while for $\varrho = 0.3$, the tangent approximation (6) seems reasonable.

Consider now the application $\mathfrak{P} : \Theta \to \mathfrak{M}$ associating to θ the probability $\mathfrak{L}^{\theta}$. Suppose that the classical Fréchet- Darmois- Cramer- Rao (FDCR) assumptions (Rao , 1945, 1965, chapter 5) are fulfilled by the family

Figure 1: Sampled metric balls (50 geodesics) of radius 1.5 (blue curves) and 0.35 (white contours) for two distributions from $NB(D_{\mathcal{R}})$; see comments in the text.



$\left\{\mathfrak{L}^{\theta} : \theta \in \Theta\right\}$ and that $\hat{\theta}$ is an unbiased first-order efficient estimator of θ. Then, $\mathfrak{g}\left(\theta^0\right) \approx V\left(\hat{\theta}\right)^{-1}$ and, if $\varrho \leq \bar{\varrho}\left(\theta^0\right)$, because of the Mahalanobis-like relationship (6), $\mathfrak{P}\left(\mathcal{E}_{\theta^0}\left(\varrho\right)\right)$ can be confused with the metric sphere of radius ϱ centered on $\mathfrak{L}^{\theta^0}$. But $\bar{\varrho}\left(\theta^0\right)$ is unknown in general; so, what if $\left\|\theta^0 - \theta\right\|_{\mathfrak{g}}\left(\theta^0\right)$ is too large $\left(\gg \bar{\varrho}\left(\theta^0\right)\right)$? This is an important issue because of the χ^2 GOF test (see for instance Rao , 1945; Menendez *et al.* , 1995; Cubedo & Oller , 2002) associated with formula (6): if $\widehat{\theta}_N$ is the estimation obtained from some N-sample of $\mathfrak{L}^{\theta^0}$, $N\, D^2_{\mathcal{R}}\left(\theta^0, \widehat{\theta}_N\right)$ should obey $\chi^2_{(p)}$, asymptotically.

4. The geometry of $NB(D_{\mathcal{R}})$

There is a number of parametrizations for the negative binomial distribution; because of its orthogonality, we chose the one used by Chua & Ong (2013):

$$P\left(X = j; (\phi, \mu)\right) = \left(\begin{array}{c} \phi + j - 1 \\ j \end{array}\right) \left(\frac{\mu}{\mu + \phi}\right)^{j} \left(1 - \frac{\mu}{\mu + \phi}\right)^{\phi}, \; j \geq 0 \tag{7}$$

$(\phi, \mu) \in \mathbb{R}^+ \times \mathbb{R}^+$; here, μ is the mean of the distribution. $D_{\mathcal{R}}\left(\mathfrak{L}^1, \mathfrak{L}^2\right)$ cannot be obtained in a closed-form but must be computed by finding the numerical solution of a the Euler-Lagrange equation (3), completed in the parametrization (7) by the boundary conditions

$$\left\{\gamma\left(0\right) = \left(\phi^1, \mu^1\right), \; \gamma\left(1\right) = \left(\phi^2, \mu^2\right)\right\}. \tag{8}$$

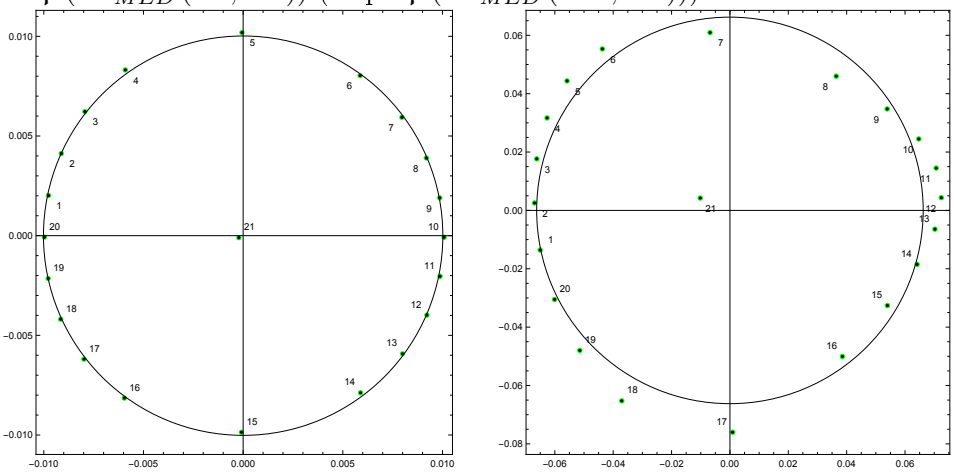
Geodesics can be as well be computed by solving (3) under the alternative constraints

$$\left\{\gamma\left(0\right) = \left(\phi^1, \mu^1\right), \; \dot{\gamma}\left(0\right) = V \in \mathbb{R}^2\right\} \tag{9}$$

where V is the initial velocity of the geodesic; this solution is associated with the exponential map at $\left(\phi^1, \mu^1\right)$.

5. Plotting χ^2 metric spheres around counts distributions

Figure 2: The case of the MED distribution. Left (resp. right) panel: representation through MDS of $\mathfrak{P}\left(\mathcal{SE}_{MED}\left(0.1,100\right)\right)$ (resp. $\mathfrak{P}\left(\mathcal{SE}_{MED}\left(0.99,100\right)\right)$).

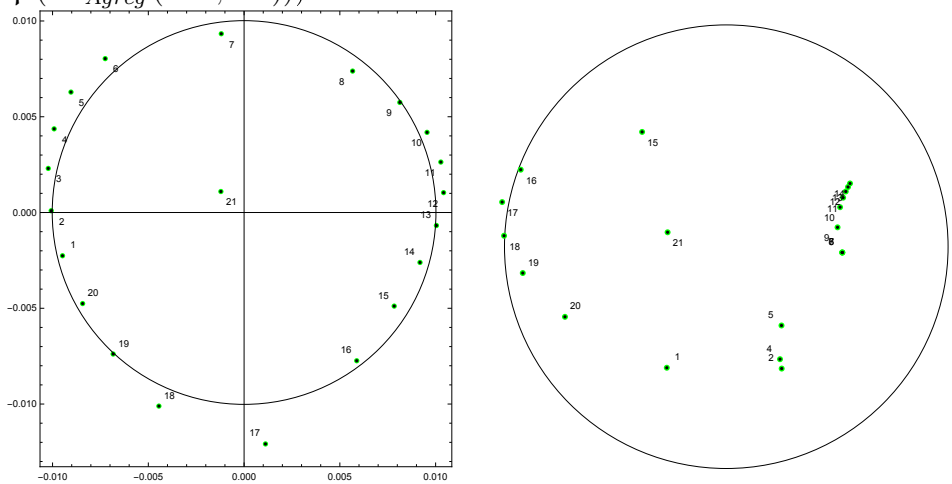


The manifold $NB(D_{\mathcal{R}})$ was considered by Manté & Kidé (2016) for computing coordinates-free distances between marine species characterized by count distributions. In that work, we mainly focused on numerical problems met in approximating $D_{\mathcal{R}}\left(\mathfrak{L}^1,\mathfrak{L}^2\right)$: this is not an easy task, because of the possible presence of cut points on geodesics (Manté & Kidé , 2016). A "visual" distance between species was afterwards obtained through Multidimensional Scaling (MDS) of the Rao's distance table.
In this work, our purpose is more geometrical, in connection with GOF tests. Consider some $\theta^0=\left(\phi^0,\mu^0\right)\in\Theta:=\mathbb{R}^+\times\mathbb{R}^+$ and, for $\alpha<1$, the ellipsoid $\mathcal{E}_{\theta^0}\left(\alpha,N\right):=\mathcal{E}_{\theta^0}\left(\chi^2_{(2)}\left(\alpha\right)/N\right)$, where N is the sample size and $\chi^2_{(2)}\left(\alpha\right)$ denotes the quantile of order α of $\chi^2_{(2)}$ (see for instance Figure 4). Is $\mathfrak{P}\left(\mathcal{E}_{\theta^0}\left(\alpha,N\right)\right)$ yet a metric sphere for usual values of α (0.1, 0.5, 0.95,...) and any $\mathfrak{L}^{\theta^0}$? If the answer is negative for some $\left(\theta^0,\alpha\right)\in\Theta\times]0,1[$, the GOF test will be "anisotropic", *i.e.* there will be a pair of probabilities $\left(\mathfrak{L}^{\underline{\theta}},\mathfrak{L}^{\overline{\theta}}\right)$ with identical risk, such that $D_{\mathcal{R}}\left(\mathfrak{L}^{\theta^0},\mathfrak{L}^{\underline{\theta}}\right)<D_{\mathcal{R}}\left(\mathfrak{L}^{\theta^0},\mathfrak{L}^{\overline{\theta}}\right)$. We put in limelight five very different test distributions:

- "BSh" (for "bell-shaped") is $NB\left(10,144.3\right)$
- "Moy" $=NB\left(1.193,87.268\right)$ is the mean of a bivariate distribution fitting the parameters of non-aggregative species found in a specific habitat
- "MED"$=NB\left(0.7767,11.2078\right)$ is the spatial median (Serfling , 2004) of the same sample of parameters
- "Agreg"$=NB\left(0.01,0.1443\right)$ corresponds to a theoretical aggregative species (borderline case: $\phi\to0^+$)
- "Boundary"$=NB\left(6,0.05\right)$, designed for investigating the case $\mu\to0^+$ (second borderline case).

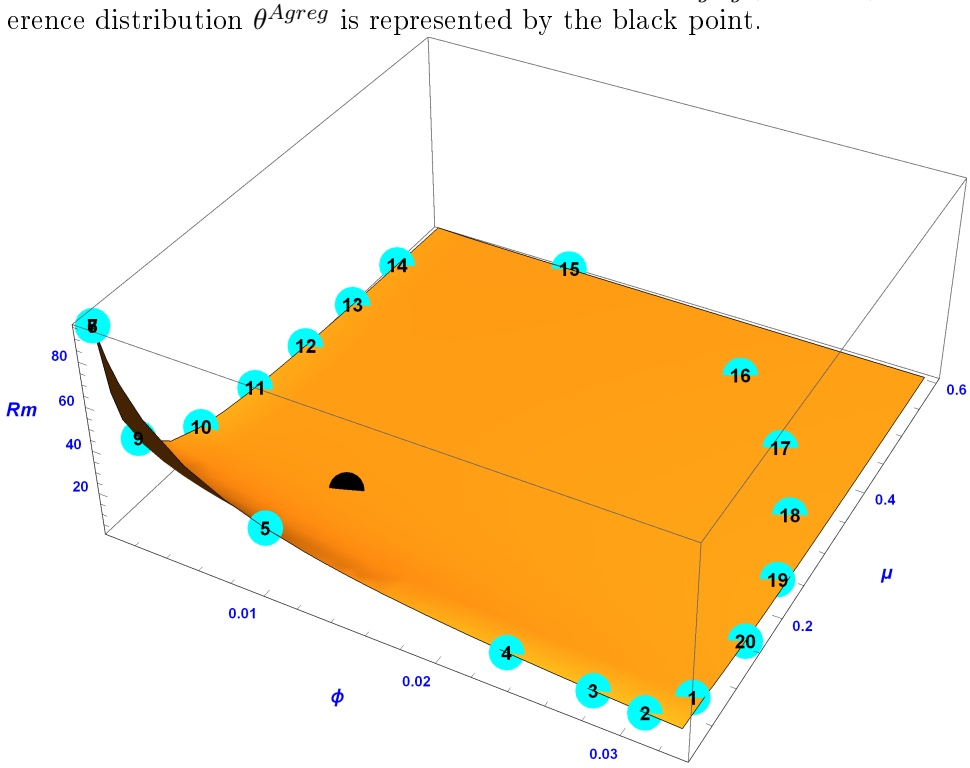
The first four distributions were already exemplified in Manté *et al.* (2016, Appendix 2).
Just as Critchley & Marriott (2016), we fixed the sample size to $N=100$; then, for each one of the test distributions above, we determined a set $\mathcal{SE}_{\theta^0}\left(\alpha,N\right)$ of 20 regularly-sampled points upon the ellipsoid $\mathcal{E}_{\theta^0}\left(\alpha,N\right)\subset\Theta$, for $\alpha\in\{0.1,\ 0.99\}$. By "regularly-sampled", we mean that the arc length between two neighbor points should be $Length\left(\mathcal{E}_{\theta^0}\left(\alpha,N\right)\right)/20$. Afterwards, for each pair on points $\left\{\theta^1,\theta^2\right\}$ of each $\mathcal{SE}_{\theta^0}\left(\alpha,N\right)$, $D_{\mathcal{R}}\left(\theta^1,\theta^2\right)$ was computed by solving (3) under (8); in addition each $D_{\mathcal{R}}\left(\theta^1,\theta^0\right)$ was computed too. This gave rise to a 21×21 table $\Delta\left(\theta^0,\alpha,N\right)$ submitted to MDS.
The reader can examine on Figures 2&3 several sampled metric "spheres" obtained through MDS. On these plots, the 21st point corresponds to the reference $\mathfrak{L}^{\theta^0}$, and should occupy the center of the circle; we superimposed the "true" circle obtained through MDS from the regularly sampled circle, with $N=100$.
For $\alpha=0.1$, the representation was always rather good (even in the case of Agreg: see Figure 3). With a lower confidence level ($\alpha=0.99$), the spheres are more or less deformed (see the right panel of Figure 2) , or

Figure 3: The case of Agreg. Left (resp. right) panel: representation of $\mathfrak{P}\left(\mathcal{SE}_{Agreg}\left(0.1,100\right)\right)$ (resp. $\mathfrak{P}\left(\mathcal{SE}_{Agreg}\left(0.99,100\right)\right)$).



squarely destroyed in the case of borderline distributions like "Agreg" or "Boundary" (see the right panel of Figure 3).

Figure 4: Plot of the Riemannian measure inside $\mathcal{SE}_{Agreg}\left(0.99,100\right)$ (truncated ellipsoid); the reference distribution θ^{Agreg} is represented by the black point.



Let's examine a little further the case of "Agreg". On Figure 4 we plotted the Riemannian measure $dV_{\mathfrak{M}}\left(\theta\right) := \sqrt{\det\left(\mathfrak{g}_{i\,j}\left(\theta\right)\right)}\; d\theta_1 \cdots d\theta_p$ associated with the metrics $\mathfrak{g}$, for values inside $\mathcal{E}_{Agreg}\left(0.99,100\right)$ cut by the axes $\phi = 0$ and $\mu = 0$. The points far from the frontiers are labeled by $\{1, 15, \cdots, 20\}$; we can see on the right panel of Figure 3 that these points are exactly those which appear on an arc of ellipsoid while the other points

collapse, forming two clusters: $\{2,4,5\}$, associated with the $\mu = 0$ borderline, and $\{8,\cdots,14\}$, associated with the $\phi = 0$ borderline. In fact, as Critchley & Marriott (2016) noticed, problems are met near the frontier $\left(\Theta - \mathring{\Theta}\right)$. Incidentally, a similar phenomenon happened in large neighbohoods of the "MED" distribution (see Figure 1).

6. Conclusions

We have shown that:

- the critical locus of level $1-\alpha$ associated with the GOF test $\mathfrak{L}^{\hat{\theta}} \stackrel{?}{=} \mathfrak{L}^{\theta^0}$ is an ellipsoid $\mathcal{E}_{\theta^0}(\alpha) \subset \Theta$ which depends (centre, eccentricity) on θ^0
- the image of $\mathcal{E}_{\theta^0}(\alpha)$ under $\mathfrak{P} : \Theta \to NB(D_{\mathcal{R}})$ associating to θ the probability $\mathfrak{L}^{\theta}$ is theoretically a metric sphere
- $\mathfrak{P}(\mathcal{E}_{\theta^0}(\alpha))$ can be considered as spherical only when $\mathcal{E}_{\theta^0}(\alpha)$ is "far enough" from the frontier of Θ (see Figure 1, or Critchley & Marriott (2016)).

References

Critchley, F. & Marriott, P. (2016). Computing with Fisher geodesics and extended exponential families. Stat Comput, 26, 325-332.

Berger, M. (2003). A Panoramic View of Riemannian Geometry, Springer-Verlag, Berlin.

Burbea, J. (1986). Informative geometry of probality spaces. Expo. Math., 4, 347-378.

Chua, K. C. & Ong, S. H. (2013). Test of mispecification with application to Negative Binomial distribution. Computational Statistics, 993-1009.

Cubedo, M. & Oller, J.M. (2002). Hypothesis testing: a model selection approach. Journal of Statistical Planning and Inference, 108, 3-21.

Manté, C., Kidé, O. S., Yao, A. F. & Mérigot, B. (2016). Fitting the truncated negative binomial distribution to count data. A comparison of estimators, with an application to groundfishes from the Mauritanian Exclusive Economic Zone. Environmental and Ecological Statistics, 23, 359-385..

Manté, C., & Kidé, S. O. (2016). Approximating the Rao's distance between negative binomial distributions. Application to counts of marine organisms. In: A. Colubi, A. Blanco and C. Gatu. (Eds.), Proceedings of COMPSTAT 2016, pp. 37-47.

Menendez, M.L., Morales, D., Pardo, L. & Salicru, M. (1995). Statistical tests based on geodesic distances. Appl. Math. Lett., 8, 1, 65-69.

Rao, C. R. (1945). Information and accuracy attainable in the estimation of statistical parameters. Bull. Calcutta Math. Soc., 37, 81-91.

Rao, C. R. (1965). Linear statistical inference and its application. John Wiley & sons, New York.

Serfling, R. (2004). Nonparametric multivariate descriptive measures based on spatial quantiles. Journal of Statistical Planning and Inference, 123, 259-278.