HAL
open science

Seeking for shortest paths between negative binomial distributions. Application to the statistical analysis of counts data.

Claude Manté

To cite this version:

Claude Manté. Seeking for shortest paths between negative binomial distributions. Application to the statistical analysis of counts data.. 2021. hal-03214474

HAL Id: hal-03214474

https://hal.science/hal-03214474

Preprint submitted on 1 May 2021

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.

Seeking for shortest paths between negative binomial distributions. Application to the statistical analysis of counts data.

Claude Manté

Aix-Marseille Université, Université du Sud Toulon-Var, CNRS/INSU, IRD, MIO,UM 110, Campus de Luminy, Case 901, F13288 Marseille Cedex 09, France

email: claude.mante@mio.osupytheas.fr, claude.mante@gmail.com

Abstract

The statistical analysis of counts of living organisms brings information about the collective behavior of species. This task is often implemented in a non-parametric setting, but parametric distributions such as the negative binomial (NB) distributions studied here, are also very useful for modeling populations abundance. Considering the Riemannian manifold $NB(D_{\mathcal{R}})$ of NB distributions equipped with the Rao metrics $D_{\mathcal{R}}$, one can compute geodesic distances between species, which can be considered as absolute. But computing such a distance requires solving a second-order nonlinear differential equation, whose solution cannot be always found in an acceptable length of time with enough precision. Manté and Kidé (2016) proposed numerical remedies to this problem, which are completed here by Poisson Approximation combined with Differential Geometry techniques. The performances of the proposed method are investigated, and it is illustrated by displaying distributions of counts of marine species through multidimensional scaling (MDS) of the table of computed Rao's distances between species.

Keywords: Riemannian manifold, geodesics, cut locus, fibers, Poisson Approximation, Multidimensional Scaling

Notations

Consider a Riemannian manifold $\mathfrak{M}$, and a parametric curve

$\alpha : [a,b] \to \mathfrak{M}$; its first derivative will be denoted $\dot{\alpha}$. A geodesic curve γ connecting two points p and q of $\mathfrak{M}$ will be denoted $p \curvearrowright q$, and $p \curvearrowright s \oplus s \curvearrowright q$ will denote the broken geodesic [1] connecting p to q with a "stopover" at s. 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 $T_\theta\mathfrak{M}$:

$$\forall\, V \in T_\theta\mathfrak{M},\ \|V\|_{\mathfrak{g}}(\theta) := \sqrt{V^t.\mathfrak{g}(\theta).V}. \tag{1}$$

The length of a curve α traced on $\mathfrak{M}$ will be denoted $\Lambda(\alpha)$. In addition, $\mathbb{R}^{+*} :=]0,+\infty[$, and $\|M\|_F$ will denote the Frobenius norm of the matrix M; logical propositions will be combined by using the classical connectors $\vee$ (or) and $\wedge$ (and).

A parametric probability distribution $\mathfrak{L}^i$ will be identified with its coordinates with respect to some chosen parametrization; for instance, we will write $\mathfrak{L}^i \equiv \left(\phi^i, \mu^i\right)$ for some negative binomial distribution.

1. Introduction

The statistical analysis of counts of living organisms brings information about the collective behavior of species (schooling, habitat preference, *etc*), possibly associated with their socio-biological characteristics (aggregation, growth rate, reproductive power, survival rate, *etc*). In the spirit of Manté et al. [2], we propose an original exploratory method, consisting in measuring the dissimilarity between species through the probability distribution of some characteristic, and analyzing the obtained dissimilarity table through MDS. In [2], this characteristic was the dispersion of each species while here it will be its abundance.

There is a wide range of statistical methods to deal with distributional data, fundamentally depending on the chosen metrics on the probabilities set. Recently, multivariate methods with a geometric dominance appeared in the literature, based on Riemannian structures equipping spaces of probability densities:

non-parametric Fisher-Rao metrics [3] or Wasserstein metrics [4]; see also [5, 6]. But all these methods were designed in a non-parametric setting, for absolutely continuous distributions, while our data are discrete. Furthermore, even if the parametric approach is quite sound from the ecological point of view (see [7] and the references therein), it is ill-suited for Exploratory Data Analysis (EDA): the visual distance between parameters of several distributions is misleading, because on the one hand it depends on the chosen parametrization and, on the other hand, these parameters are not commensurable in general (different ecological meaning, different ranges, ...).

In a seminal paper, Rao [8] 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 θ^1 and θ^2 is

$$D_{\mathcal{R}}\left(\theta^1, \theta^2\right) := \int_0^1 \sqrt{\dot{\gamma}^t(t) . \mathfrak{g}(\gamma(t)) . \dot{\gamma}(t)} dt \tag{2}$$

where γ is the **segment** (minimal length curve) connecting $\theta^1 = \gamma(0)$ to $\theta^2 = \gamma(1)$. Naturally, Rao [8, 9] proposed to use (2) as a distance between populations or for Goodness-Of-Fit (GOF) testing, followed by a number of authors [10, 11, 12, 13, 14, 15, 16, 17, 18].

The Rao's distance between members of a common family of distributions has been calculated in a number of classical cases [19] but it cannot be obtained in a closed form, generally. In such cases, like the NB distributions (when both parameters are undefined), $D_{\mathcal{R}}$ must be obtained by numerically solving a second-order nonlinear differential equation, frequently hard to integrate. Manté and Kidé [15] proposed numerical remedies to this issue, which are completed here by Poisson Approximation combined with Differential Geometry techniques.

2. Few elements of Riemannian geometry

According to the fundamental theorem of Riemannian geometry [1], there is a unique symmetric connection ∇ compatible with a given metrics $\mathfrak{g}$, giving

in our case the Rao's distance. It is noteworthy that other statistically sound (but not Riemannian) connections can be fruitfully considered (Amari et al. [20]). Geodesics with respect to ∇ are solutions of the Euler-Lagrange equation [21, 1, 19]:

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

where each Christoffel symbol $\Gamma_{i,j}^{k}$ only depends on $\mathfrak{g}$ and is defined in coordinates by:

$$\Gamma_{i,j}^{k} := \sum_{m=1}^{p} \frac{\mathfrak{g}^{\mathrm{km}}}{2} \left(\frac{\partial \mathfrak{g}_{jm}}{\partial \theta_i} + \frac{\partial \mathfrak{g}_{im}}{\partial \theta_j} - \frac{\partial \mathfrak{g}_{ij}}{\partial \theta_m} \right) \tag{4}$$

and $\mathfrak{g}^{\mathrm{im}}$ (resp. $\mathfrak{g}_{mk}$) is some entry of $\mathfrak{g}^{-1}$ (resp. $\mathfrak{g}$). The segment connecting $\mathfrak{L}^1$ to $\mathfrak{L}^2$ (if it exists) is necessarily a geodesic, but building it is not straightforward: a geodesic is not necessarily a segment, due to the possible existence of cut points.

Theorem 1. *[1, 22] Let $p = \alpha(0)$ be the initial point of a geodesic. Then there is some $0 < t_0 \leq +\infty$ such that α is a segment from p to $\alpha(t)$ for every $t \leq t_0$ and for $t > t_0$ thereafter never again a segment from p to any $\alpha(t)$ for $t > t_0$. This number t_0 is called the cut value of α and $\alpha(t_0)$ is called the* ***cut point*** *of α. There are only two possible reasons (which can occur simultaneously) for $\alpha(t_0)$ to be to be the cut point of α:*

- *there is a segment from p to $\alpha(t_0)$ different from α*
- *$\alpha(t_0)$ is the first conjugate point on α to p (i.e. $t_0\,\dot{\alpha}(0)$ is a critical point of the exponential map.*

In addition, the distance function $D_{\mathcal{R}}(p, \bullet)$ is not differentiable at $\alpha(t_0)$ [23, 1].

Remark 1. No matter the cause of the phenomenon, the main point for us is that if t_0 is a cut value of the unit-speed geodesic α, $\forall\, t \leq t_0$, $D_{\mathcal{R}}(p, \alpha(t)) = t$ while $\forall\, t > t_0$, $D_{\mathcal{R}}(\alpha(0), \alpha(t)) < t$. This is the basis of the method proposed by Manté and Kidé [15] for detecting cut points (see the Supplementary Material).

Remark 2. If $\alpha := p \curvearrowright q$ is a segment and $V_0 := \dot{\alpha}(0)$, because of uniqueness of geodesics, $\exp_p(V_0) := \alpha_{\mathcal{B}(V_0)}(1) = q$; reciprocally, if $V_1 := -\dot{\alpha}(1)$, we have also that $\exp_q(V_1) := \alpha_{\mathcal{B}(V_1)}(1) = p$.

3. The special case of $NB(D_{\mathcal{R}})$

There is a large number of parametrizations for the NB distribution, and the most classical one is probably

$$P\left(X=j;(\phi,p)\right)=\left(\begin{array}{c}\phi+j-1\\ \phi-1\end{array}\right)p^{j}\left(1-p\right)^{\phi}\ j\geq 0 \tag{5}$$

with $(\phi,p)\in\mathbb{R}^{+}\times]0,1[$. Nevertheless, because of its orthogonality, we chose instead the parametrization used by Chua and Ong [24]:

$$P\left(X=j;(\phi,\mu)\right)=\left(\begin{array}{c}\phi+j-1\\ j\end{array}\right)\left(\frac{\mu}{\mu+\phi}\right)^{j}\left(1-\frac{\mu}{\mu+\phi}\right)^{\phi},\ j\geq 0 \tag{6}$$

$(\phi,\mu)\in\mathbb{R}^{+}\times\mathbb{R}^{+}$; here, μ is the mean of the distribution and ϕ is the so-called "index parameter". In these coordinates, the information matrix is:

$$\mathfrak{g}(\phi,\mu)=\left(\begin{array}{cc}G_{\phi\phi} & 0\\ 0 & G_{\mu\mu}\end{array}\right)$$

where $G_{\mu\mu}=\frac{\phi}{\mu(\mu+\phi)}$, while the expression of $G_{\phi\phi}$ is more complicated:

$$G_{\phi\phi}=-\frac{\mu+\phi\left(\mu+\phi\right)\left(\left(\phi/_{\mu+\phi}\right)^{\phi}-1\right)\psi^{1}(\phi)}{\phi\left(\mu+\phi\right)} \tag{7}$$

where ψ^{1} is the Trigamma function [25]. The reader will find in Burbea and Rao [19] the closed-form expression of the Rao's distance for a number of probability families; the Rao's distance between the Poisson distributions $\mathcal{P}(\lambda_1)$ and $\mathcal{P}(\lambda_2)$ is

$$D_{\mathcal{P}}(\lambda_1,\lambda_2):=2\left|\sqrt{\lambda_1}-\sqrt{\lambda_2}\right|. \tag{8}$$

We will denote $\mathcal{P}(D_{\mathcal{P}})$ the Riemannian manifold of Poisson distributions equipped with this distance. These authors also reported that **if the index parameter ϕ of two NB distributions is the same**,the Rao's distance is given by

$$D_{NB(p)}\left(\left(\phi,p^{1}\right),\left(\phi,p^{2}\right)\right):=2\sqrt{\phi}\ \cosh^{-1}\left(\frac{1-\sqrt{p^{1}\,p^{2}}}{\sqrt{\left(1-p^{1}\right)\left(1-p^{2}\right)}}\right) \tag{9}$$

in the parametrization (5). Of course, if $\mathfrak{L}^1 = NB\left(\phi, p^1\right)$ and $\mathfrak{L}^2 = NB\left(\phi, p^2\right)$, we have necessarily:

$$D_{\mathcal{R}}\left(\mathfrak{L}^1, \mathfrak{L}^2\right) \leq D_{NB(p)}\left(\mathfrak{L}^1, \mathfrak{L}^2\right). \tag{10}$$

Due to the complexity of (7), $D_{\mathcal{R}}\left(\mathfrak{L}^1, \mathfrak{L}^2\right)$ cannot be obtained in a closed-form. It must be computed by finding the numerical solution of (3) completed in the parametrization (6) by the conditions (boundary value problem)

$$\left\{\gamma(0) = \left(\phi^1, \mu^1\right), \gamma(1) = \left(\phi^2, \mu^2\right)\right\}. \tag{11}$$

3.1. Numerical approximation of $D_{\mathcal{R}}\left(\mathfrak{L}^1, \mathfrak{L}^2\right)$ [15]

From now, $\mathfrak{L}^i \equiv \left(\phi^i, \mu^i\right)$ will denote some NB distribution parametrized in the system (6), but notice that our purpose could be extended to **any** parametric family of probabilities.

Firstly, all the Christoffel symbols (4) were calculated from the expression (7) of $G_{\phi\phi}$, with the help of *Mathematica* [26] . Then, the differential equation (3) was numerically solved under the the boundary conditions (11), for a number of distributions of counts of marine species whose parameters had been estimated in [7]. In most cases a solution could be found in an acceptable time (four CPU minutes), with a good numerical precision (20 digits), but was each one of the geodesics found a segment? And what about failures met in computation?

We indeed had to face various problems detailed in [15], where numerical remedies were proposed. The main one consisted in inserting a well-placed "stopover" S between each pair of problematic distributions A and B, in such a way that $D_{\mathcal{R}}(A, S)$ and $D_{\mathcal{R}}(S, B)$ could be computed in a reasonable time, while $D_{\mathcal{R}}(A, B)$ could not. Furthermore, S was placed in order that $D_{\mathcal{R}}(A, S)+$ $D_{\mathcal{R}}(S, B)$ should by a good approximation of $D_{\mathcal{R}}(A, B)$. For sake of brevity, we moved to the Supplementary Material useful information and illustrations about this previous work. **All references to this supplement will be preceded by an S**.

3.2. Making computations easier through Differential Geometry techniques and Poisson Approximation

From the numerical side, it is noteworthy that the index parameter ϕ often takes large values, causing difficulties in the evaluation of quantities associated with $\Gamma(\phi)$, like formulas (6) and (7) or Christoffel's symbols (4).

From the statistical side, the convergence in distribution of some $\mathfrak{L} \equiv (\phi, \mu)$ towards a Poisson distribution $\mathcal{P}$ when $\phi \to \infty$ is well-known. Majsnerowska [27] proved the following result:

$$d_{TV}(\mathfrak{L}, \mathcal{P}(\lambda)) \leq \Delta(\phi, \mu) := \left(1 - e^{-\mu}\right) \frac{\mu}{\phi} \tag{12}$$

where $\lambda = \omega(\phi, \mu) := \frac{\phi\,\mu}{\phi+\mu}$ and d_{TV} denotes the total variation distance. Consequently, we can claim that $(\phi \gg \mu) \vee (\mu\, small) \Rightarrow \Delta(\phi, \mu)$ *small* and conclude that in such cases it may be quite impossible to find a difference between $\mathfrak{L}$ and $\mathcal{P}(\lambda)$, even when the index parameter is small or moderate! This fact suggests to replace the NB model by the Poisson one when both distributions are very close to each other. This is also biologically sound, since the former is well-suited for aggregative species, while the latter is associated to species with a random behavior (see [7, 2] and the references therein). But since $\mathcal{P}(D_{\mathcal{P}})$ is not a sub-manifold of $NB(D_{\mathcal{R}})$, there is no clear relationship between the associated Rao's distances: we cannot mix both types of distributions. To avoid this conceptual difficulty, instead of superseding the original distribution $\mathcal{L} \equiv (\phi, \mu)$ by $\mathcal{P}(\omega(\phi, \mu))$ when both distributions are very close, we propose to supersede $\mathcal{L}$ by another "equivalent" NB distribution, more easily manageable.

Let's focus now on the application

$$\begin{cases} \omega : \mathbb{R}^{+*} \times \mathbb{R}^{+*} \to \mathbb{R}^{+*} \\ (\phi, \mu) \mapsto \lambda \end{cases}$$

associating to any NB distribution (ϕ, μ) the corresponding $\mathcal{P}(\lambda)$.

Lemma 2. *The tangent application* $T_{(\phi,\mu)}\omega = \frac{1}{(\phi+\mu)^2} \begin{pmatrix} \mu^2 \\ \phi^2 \end{pmatrix}$ *is surjective.*

Proof. Let us fix some $\rho \in \mathbb{R}^{+*}$; one can easily show that the set of solutions of the equation $T_{(\phi,\mu)}\omega(x,y) = \rho$ is the line of equation $y = -\left(\frac{\mu}{\phi}\right)^2 x + \rho\left(1+\frac{\mu}{\phi}\right)^2$ □

As a consequence, ω is a surjective submersion and the fiber $F_\lambda := \omega^{-1}(\lambda)$ associated with any $\lambda \in \mathbb{R}^{+*}$ is a sub-manifold of $NB(D_{\mathcal{R}})$.

Proposition 3. *F_λ is defined by either equation:*

$$\begin{cases} \mu(\lambda;\phi) = \frac{\lambda}{1-\lambda/\phi} & : \phi > \lambda \\ \phi(\lambda;\mu) = \frac{\lambda}{1-\lambda/\mu} & : \mu > \lambda \end{cases}. \tag{13}$$

Proof. $F_\lambda := \left\{(\phi,\mu) : \frac{\phi\,\mu}{\phi+\mu} = \lambda\right\}$; thus, the strictly positive parameters λ, ϕ and μ are linked by the relationship $\phi\mu = \phi\lambda + \lambda\mu$, which proves that $\phi = \lambda + \lambda\frac{\phi}{\mu}$ and $\mu = \lambda + \lambda\frac{\mu}{\phi}$. Consequently, $\lambda < \min(\phi,\mu)$ and $\lim\limits_{\phi\to+\infty} \mu(\lambda;\phi) = \lim\limits_{\mu\to+\infty} \phi(\lambda;\mu) = \lambda$ □

Lemma 4. *One can easily verify that :*
$\forall\, \lambda \in \mathbb{R}^{+*},\ F_\lambda \neq \emptyset$
$\forall\ (\phi,\mu) \in \mathbb{R}^{+*} \times \mathbb{R}^{+*},\ (\phi,\mu) \in F_{\omega(\phi,\mu)}$
$\forall\ (\lambda_1,\lambda_2) \in \mathbb{R}^{+*} \times \mathbb{R}^{+*},\ \lambda_1 \neq \lambda_2 \Rightarrow F_{\lambda_1} \cap F_{\lambda_2} = \emptyset.$

Theorem 5. *Suppose $\mathfrak{L} \equiv (\phi,\mu) \in F_\lambda$ and $\Delta(\phi,\mu) \leq \delta$, where δ is some **fixed** threshold chosen for deciding whether $\mathfrak{L}$ can be identified with $\mathcal{P}(\lambda)$. Then, if $\mathfrak{L}' \equiv (\phi',\mu') \in F_\lambda$ is another distribution, such that $\phi' > \phi$, $\Delta(\phi',\mu') < \delta$ and $\mathfrak{L}'$ cannot be practically distinguished from $\mathcal{P}(\lambda)$ too.*

Proof. See Appendix 5 □

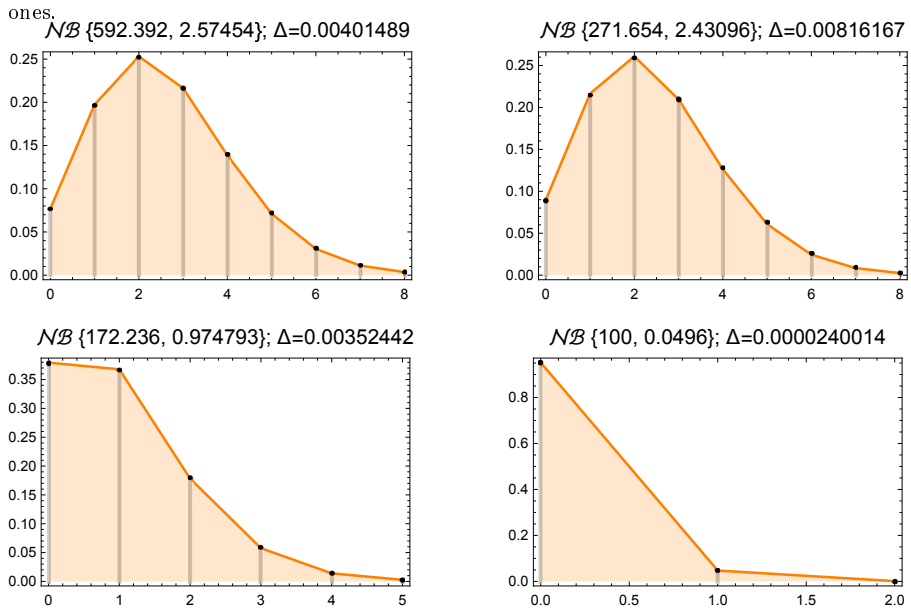
Corollary 6. *Consider $\mathfrak{L} \equiv (\phi,\mu)$, such that $\Delta(\phi,\mu) \leq \delta$. Then $\mathfrak{L} \in F_\lambda$, with $\lambda = \omega(\phi,\mu)$, and we can determine the "initial" distribution $\mathfrak{L}_*(\delta) := (\phi_*,\mu_*)(\delta)$ of F_λ defined by:*

$$\begin{cases} \phi_* = \underset{\phi:(\phi,\mu)\in F_\lambda}{\arg} (\Delta(\phi,\mu) = \delta) = \underset{\phi}{\arg}(\Delta(\phi,\mu(\lambda;\phi)) = \delta) \\ \mu_* = \frac{\lambda}{1-\lambda/\phi_*} \end{cases}. \tag{14}$$

Let us now fix $\mathfrak{L}^0 \equiv \left(\phi^0,\mu^0\right)$, such that $\Delta\left(\phi^0,\mu^0\right) \leq \delta$. $\mathfrak{L}^0$ can be identified with $\mathcal{P}\left(\lambda^0\right)$, as well as any distribution of the fiber whose index parameter is greater than ϕ^0_*, due to the propositions above. Consider now the following part of F_{λ^0}:

$$\mathring{\mathcal{P}}\left(\lambda^0,\delta\right) := \left\{(\phi,\mu) \in F_{\lambda^0} :\ \phi \geq \phi^0_*\right\}.$$

Figure 1: Four instances of Poisson-like distributions; Δ is given by Formula (12) and vertical bars are associated with NB probabilities while continuous curves are associated with Poisson ones.



Obviously, $\mathfrak{L}^0 \in \mathring{\mathcal{P}}\left(\lambda^0, \delta\right)$ but we have that, when $(\phi, \mu) \in \mathring{\mathcal{P}}\left(\lambda^0, \delta\right)$, $d_{TV}\left(\mathfrak{L}, \mathcal{P}\left(\lambda^0\right)\right) \leq \delta$ and $\left|\omega\left(\phi, \mu\right) - \lambda^0\right| \approx 0$, simultaneously. Thus, in such cases, $NB\left(\phi, \mu\right)$ and $\mathcal{P}\left(\lambda^0\right)$ are **practically indiscernible**.

Definition 7. We will say that $\mathfrak{L} \equiv (\phi, \mu)$ is **Poisson-like** if $(\phi, \mu) \in \mathring{\mathcal{P}}\left(\omega\left(\phi, \mu\right), \delta\right)$.

We displayed on Figure 1 four examples of such NB distributions (setting $\delta = 0.01$, say). Let us now denote $\overset{\delta}{\equiv}$ the following relation (δ has been **fixed**) between Poisson-like distributions:

$$\mathfrak{L}^1 \overset{\delta}{\equiv} \mathfrak{L}^2 \Leftrightarrow \exists \lambda : \ \mathfrak{L}^i \in \mathring{\mathcal{P}}\left(\lambda, \delta\right), \ i = 1, 2.$$

Corollary 8. *The relation $\overset{\delta}{\equiv}$ is an equivalence relation* ***between Poisson-like distributions****.*

Proof. Reflexive and symmetric properties are straightforward. Suppose now $\mathfrak{L}^1 \overset{\delta}{\equiv} \mathfrak{L}^2$ and $\mathfrak{L}^3 \overset{\delta}{\equiv} \mathfrak{L}^2$; there exists $\lambda_{1,2} : \mathfrak{L}^i \in \mathring{\mathcal{P}}\left(\lambda_{1,2}, \delta\right), \ i = 1, 2$ and $\lambda_{2,3} : \mathfrak{L}^i \in \mathring{\mathcal{P}}\left(\lambda_{2,3}, \delta\right), \ i = 2, 3$. Consequently, $\mathfrak{L}^2 \in F_{\lambda_{1,2}} \cap F_{\lambda_{2,3}}$ which is empty if $\lambda_{1,2} \neq \lambda_{2,3}$ (see Lemma 4), and these three Poisson-like distributions belong to the same fiber. Thus, $\mathfrak{L}^1 \overset{\delta}{\equiv} \mathfrak{L}^3$ and $\overset{\delta}{\equiv}$ is transitive $\square$

Suppose $\mathfrak{L}^1 \overset{\delta}{\equiv} \mathfrak{L}^2$ belong to a common fiber, $F_{\lambda_{1,2}}$. Being indiscernible, these distributions should be necessarily close to each other, and it would be sound to merely supersede $D_{\mathcal{R}}\left(\mathfrak{L}^1, \mathfrak{L}^2\right)$ by δ. The following corollaries about distributions belonging to different fibers also stem from Theorem 5.

Corollary 9. *Consider two Poisson-like distributions $\mathfrak{L}^1$ and $\mathfrak{L}^2$ belonging to different fibers, F_{λ_1} and F_{λ_2}. One can always determine a pair of Poisson-like distributions $\tilde{\mathcal{L}}^1 := \left(\tilde{\phi}, \tilde{\mu}_1\right) \in F_{\lambda_1}$ and $\tilde{\mathcal{L}}^2 := \left(\tilde{\phi}, \tilde{\mu}_2\right) \in F_{\lambda_2}$ such that $D_{\mathcal{R}}\left(\mathfrak{L}^1, \mathfrak{L}^2\right)$ can be superseded by $D_{\mathcal{R}}\left(\tilde{\mathcal{L}}^1, \tilde{\mathcal{L}}^2\right)$. In addition, thanks to formula (9), we can straightforwardly compute $D_{NB(p)}\left(\tilde{\mathcal{L}}^1, \tilde{\mathcal{L}}^2\right)$, which is an upper bound for $D_{\mathcal{R}}\left(\tilde{\mathcal{L}}^1, \tilde{\mathcal{L}}^2\right)$.*

Proof. We just have to compute, thanks to formulas (14) and (13)
$\tilde{\phi} := \max\left(\phi_*\left(\lambda^1, \delta\right), \phi_*\left(\lambda^2, \delta\right)\right)$, $\tilde{\mu}_1 := \mu\left(\lambda_1; \tilde{\phi}\right)$ and $\tilde{\mu}_2 := \mu\left(\lambda_2; \tilde{\phi}\right)$ □

Under the same conditions as in the preceding corollary, the following alternative strategy is always possible too.

Corollary 10. *Suppose $\phi_1 \leq \phi_2$; thanks to formula (13) we can determine $\check{\mu}_1 := \mu\left(\lambda_1; \phi_2\right)$, such that $\check{\mathfrak{L}}^1 := \left(\phi_2, \check{\mu}_1\right) \in F_{\lambda_1}$ is Poisson-like too (because of (5)) and belongs to the same class as $\mathfrak{L}^1$; $D_{\mathcal{R}}\left(\mathfrak{L}^1, \mathfrak{L}^2\right)$ can be superseded by $D_{\mathcal{R}}\left(\check{\mathfrak{L}}^1, \mathfrak{L}^2\right)$. In addition, we can easily compute $D_{NB(p)}\left(\check{\mathfrak{L}}^1, \mathfrak{L}^2\right)$, which is an upper bound for $D_{\mathcal{R}}\left(\check{\mathfrak{L}}^1, \mathfrak{L}^2\right)$.*

Since $\mathcal{P}(D_{\mathcal{P}})$ is not a sub-manifold of $NB(D_{\mathcal{R}})$, there is no clear relationship between the associated Rao's distances, in general (for instance, if $\mathfrak{L}^1 \neq \mathfrak{L}^2$ belong to the same fiber F_λ, $D_{\mathcal{R}}\left(\mathfrak{L}^1, \mathfrak{L}^2\right) > D_{\mathcal{P}}\left(\omega\left(\mathfrak{L}^1\right), \omega\left(\mathfrak{L}^2\right)\right) = 0$). Nevertheless, one can easily prove, with the same notation as in Proposition 10, that

$$\begin{cases} D_{NB(p)}\left(\tilde{\mathcal{L}}^1, \tilde{\mathcal{L}}^2\right) = \mathcal{C}\left(\tilde{\mu}_1, \tilde{\mu}_2\right)\ D_{\mathcal{P}}\left(\mathcal{P}(\lambda^1), \mathcal{P}(\lambda^2)\right) \\ D_{NB(p)}\left(\check{\mathfrak{L}}^1, \mathfrak{L}^2\right) = \mathcal{C}\left(\check{\mu}_1, \mu_2\right)\ D_{\mathcal{P}}\left(\mathcal{P}(\lambda^1), \mathcal{P}(\lambda^2)\right) \end{cases} \tag{15}$$

where $\mathcal{C}\left(\check{\mu}_1, \mu_2\right)$ and $\mathcal{C}\left(\tilde{\mu}_1, \tilde{\mu}_2\right) \geq 1$ are given by the function defined hereunder (up to a simple change of parametrization).

Lemma 11. *The function $\mathcal{C}$ is defined in the parametrization (5) by*

$$\mathcal{C}\left(p^1,p^2\right)=\frac{\cosh^{-1}\left(\frac{1-\sqrt{p^1p^2}}{\sqrt{(p^1-1)(p^2-1)}}\right)}{\sqrt{-2\sqrt{p^1p^2}+p^1+p^2}}\geq 1$$

if $p^1 \neq p^2$, and $\mathcal{C}\left(p,p\right):=1$.

Proof. In the parametrization (5), the mean $\mu=\frac{K\,p}{1-p}$ and thus $p=\frac{\mu}{\phi+\mu}$, while $\phi=K$. Consequently, $\lambda^i=K\,p^i$ and $D_{\mathcal{P}}\left(\mathcal{P}(\lambda^1),\mathcal{P}(\lambda^2)\right)=2\sqrt{K}\left|\sqrt{p^1}-\sqrt{p^2}\right|$. Then, because of formula (9), we have:

$$\frac{D_{NB(p)}\left(\check{\mathfrak{L}}^1,\check{\mathfrak{L}}^2\right)}{D_{\mathcal{P}}\left(\mathcal{P}(\lambda^1),\mathcal{P}(\lambda^2)\right)}=\mathcal{C}\left(p^1,p^2\right)$$

□

A detailed example

Lets us fix $\delta:=0.01$, and consider

$\mathfrak{L}^1=(172.236,0.974793)$ and $\mathfrak{L}^2=(6,0.05)$. Both these distributions are Poisson-like, with $\mathfrak{L}^1\in F_{0.0495868}$ and $\mathfrak{L}^2\in F_{0.969307}$. We plotted on Figure 2 interesting portions of these fibers. On each one of the panels, the big gray point (of coordinates (λ,λ)) corresponds to the lower bound of ϕ and μ, while $\mathfrak{L}_*:=(\phi_*,\mu_*)\,(\lambda,\delta)$ is the "initial" distribution given by (14). All the distributions situated on the right of $\mathfrak{L}_*$ are Poisson-like. It is the case of $\mathfrak{L}^1$ and $\mathfrak{L}^2$, represented on Figure 2 by small gray points.

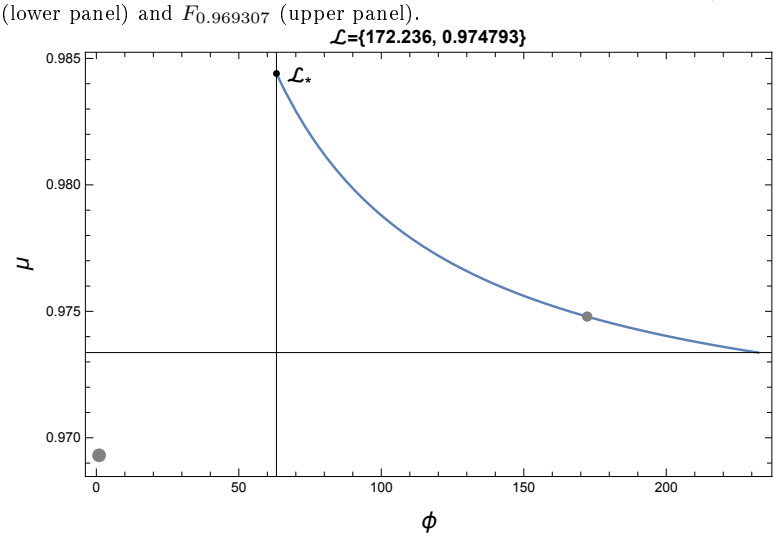
We found that $\phi_*^1=63.2096<\phi^1$ and $\phi_*^2=0.412537$; thus, $\tilde{\phi}=\phi_*^1$ and $\tilde{\mathcal{L}}^1=\mathfrak{L}_*^1$, while $\tilde{\mathcal{L}}^2\neq\mathfrak{L}_*^2$. Next, in accordance with Corollary 9, we computed $D_{NB(p)}\left(\tilde{\mathcal{L}}^1,\tilde{\mathcal{L}}^2\right)\approx 1.53375$, which is rather close to $D_{\mathcal{P}}\left(\mathcal{P}(\lambda_1),\mathcal{P}(\lambda_2)\right)\approx 1.52371$.

Since $\phi^2<\phi^1$, it is also possible to determine the distribution $\check{\mathfrak{L}}^2:=(\phi_1,\check{\mu}_2)\overset{\delta}{\equiv}\mathfrak{L}^2$ (black point on the lower panel) and, thanks to Corollary 10, to compute $D_{NB(p)}\left(\check{\mathfrak{L}}^2,\mathfrak{L}^1\right)\approx 1.52737$ (very close to $D_{\mathcal{P}}\left(\mathcal{P}(\lambda_1),\mathcal{P}(\lambda_2)\right)$ too).

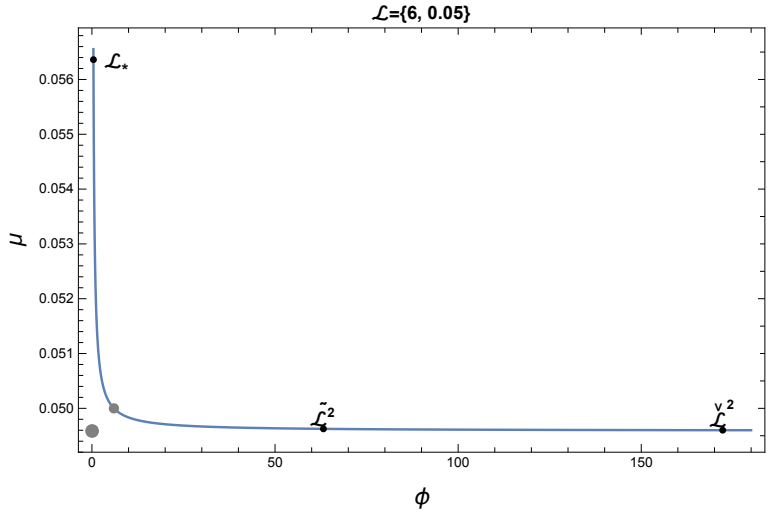
3.2.1. Application to EDA

Let $\mathfrak{L}^1$ and $\mathfrak{L}^2\in NB(D_{\mathcal{R}})$; three cases may be met: both of them are Poisson-like, only one of them is Poisson-like, or none of them is so.

Figure 2: Portions of fibers associated with two Poisson-like distributions ($\delta = 0.01$): $F_{0.0495868}$ (lower panel) and $F_{0.969307}$ (upper panel).



Out[ ]=



Suppose first $\mathfrak{L}^1$ and $\mathfrak{L}^2$ belong to distinct fibers F_{λ^1} and F_{λ^2} and each $\mathfrak{L}^i \in \mathring{\mathcal{P}}\left(\lambda^i, \delta\right)$. Then $\mu^i \approx \lambda^i$ and we can use each one of the Corollaries 6, 9 or 10 to determine a pair of equivalent distributions, whose distance would be easier to compute. So, one will successively (try to) compute $D_{\mathcal{R}}\left(\mathfrak{L}^1, \mathfrak{L}^2\right)$, $D_{\mathcal{R}}\left(\mathfrak{L}^1_*, \mathfrak{L}^2_*\right)$, either $D_{NB(p)}\left(\check{\mathfrak{L}}^1, \mathfrak{L}^2\right)$ or $D_{NB(p)}\left(\check{\mathfrak{L}}^2, \mathfrak{L}^1\right)$, or finally $D_{NB(p)}\left(\tilde{\mathcal{L}}^1, \tilde{\mathcal{L}}^2\right)$, stopping as soon as possible in order to alter the data as little as possible.

Suppose now $\mathfrak{L}^1$ is Poisson-like while $\mathfrak{L}^2$ is not; if $\phi^1 \leq \phi^2$ (or even if $\phi^1_* \leq \phi^2$), we can again consider $\check{\mathfrak{L}}^1 := \left(\phi^2, \check{\mu}^1\right)$. One will successively (try to) compute three distances: $D_{\mathcal{R}}\left(\mathfrak{L}^1, \mathfrak{L}^2\right)$, $D_{\mathcal{R}}\left(\mathfrak{L}^1_*, \mathfrak{L}^2\right)$ and $D_{NB(p)}\left(\check{\mathfrak{L}}^1, \mathfrak{L}^2\right)$, stopping as soon as possible.

Three instances

Look at Figures S1, S3 and S4 of the Supplementary Material. In all these plots, only one of the two distributions is Poisson-like. In the first case, $\mathfrak{L}^1 = (0.00487399, 0.262591)$, and $\mathfrak{L}^2 = (592.392, 2.57454) \overset{\delta}{\equiv} (3.5634, 9.13442) = \mathfrak{L}^2_*$. Since $\phi^2_* = 3.5634 \nleq 0.00487399$ we cannot consider $\check{\mathfrak{L}}^2 := \left(\phi^1, \check{\mu}^2\right)$ and compute $D_{NB(p)}\left(\check{\mathfrak{L}}^2, \mathfrak{L}^1\right)$, but $D_{\mathcal{R}}\left(\mathfrak{L}^2_*, \mathfrak{L}^1\right) = 3.53253$ could be computed (simple configuration, no cut point), while the original approximation of $D_{\mathcal{R}}\left(\mathfrak{L}^1, \mathfrak{L}^2\right)$ (intricate configuration with a cut point) was 45.1321 (definitions of **simple** and **intricate configurations** are reminded in the supplementary material - for further details, see [15]).

In the next case (Figure S3) $\mathfrak{L}^1 = (0.00996246, 0.121282)$, while the second distribution is the same as in the previous case. The original distance corresponded to an intricate configuration with a cut point, and to the upper bound 43.1519. We found instead $D_{\mathcal{R}}\left(\mathfrak{L}^2_*, \mathfrak{L}^1\right) = 3.48809$.

In the last case (Figure S4), $\mathfrak{L}^1 = (0.938781, 9.86571)$, and

$\mathfrak{L}^2 = (172.236, 0.974793) \overset{\delta}{\equiv} (63.2096, 0.984403) = \mathfrak{L}^2_*$. Since $\phi^2_* = 63.2096 \nleq 0.938781$, we cannot consider $D_{NB(p)}\left(\check{\mathfrak{L}}^2, \mathfrak{L}^1\right)$, but we found that $D_{\mathcal{R}}\left(\mathfrak{L}^2_*, \mathfrak{L}^1\right) = 12.5294$ (intricate configuration with an acceptable rough solution and a stopover), while the original $D_{\mathcal{R}}\left(\mathfrak{L}^1, \mathfrak{L}^2\right)$ (intricate configuration with linear interpolation) gave rise to the upper distance 21.351.

3.3. EDA of field data: representation of counts distributions of marine species

The Mauritanian coast, situated on the Atlantic side of the northwestern African continent, embeds a wide long continental shelf of about 750 km and $36000 km^2$, with an Exclusive Economic Zone (the MEEZ) of $230000 km^2$. Manté et al. [7] considered the abundance of species of fish and invertebrates collected in the MEEZ during annual scientific trawl surveys since 1997 to now. Because the spatial distribution of groundfish species is strongly influenced by the physical environment, we split this set into an optimal number (four) of subsets (typical habitats) associated with homogeneous physical conditions determined by available environmental variables. The counts associated with each species found in each one of the four habitats were then gathered, and fitted by a NB distribution (for further information, see [7]).

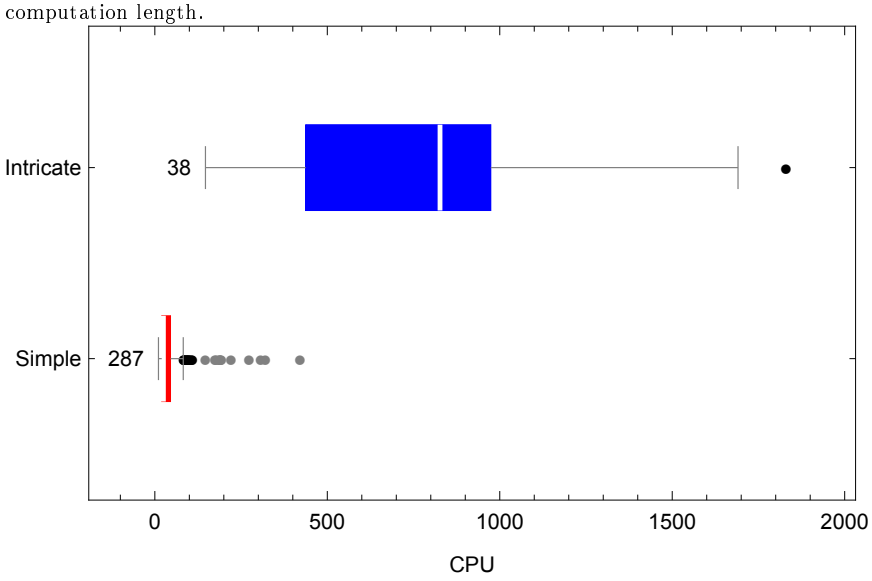
Table 1: Global results obtained in the four habitats of the MEEZ ($\delta = 0.01$).

Habitat	Number of species (well-fitted)	Simple configurations	Intricate (Rough, Linear)	Cut points	Poisson-like distributions
$C1$	30	400	(35,0)	1	3
$C2$	19	147	(24,0)	0	3
$C3$	26	309	(16,0)	0	5
$C4$	26	304	(21,0)	0	1

3.3.1. Benefits of the proposed method: speeding up computations

Processing these data, Manté and Kidé [15] found an overwhelming proportion of simple configurations (more than 70%), while numerical cut points were quite rare. In the intricate cases, the rough solution was generally accepted (more than 90% of occurrences). We display on Figure 3 statistics about the computational cost of the 26×26 distance matrix corresponding to $C4$. Among the 325 distances computed, 287 were simple cases, with a median computation cost of 25"; the remaining 38 cases were intricate, with a median cost of 826". Superseding each Poisson-like distribution $\mathfrak{L}$ by either $\mathfrak{L}_*$ (which always exists), or $\tilde{\mathcal{L}}$, or $\check{\mathfrak{L}}$ (depending on the situation - see Section 3.2) , we found that the proportion of simple configurations was greater than 90%, excepted for the

Figure 3: Statistics of the computational burden for processing (without Poisson approximation) species collected in zone 4 of the MEEZ: number of distances of each type, box-plots of computation length.



second type of habitat, $C2$ (85%) (see Table 1). Furthermore, in the intricate cases, the rough solution was always accepted. In our previous study, numerical cut points were rare (less than a pair per class), but now we detect a single numerical cut point (see Table 1)! Poisson-like distributions were quite rare, but they were generally so "pathological" that their replacement by equivalent better-suited NB distributions changed a lot the results. This is shown by the statistics displayed on Figure 4. Among the 325 distances computed, 305 were simple cases, with a median computation cost of 20"; the remaining 20 cases were intricate, with a median cost of 404".

3.3.2. Parametric representation of species from the habitat C4

We display on Figure 5 the estimated parameters of the counts distribution of species sampled in the zone 4 of the MEEZ. This region is of paramount importance: it is a high plankton productivity area, supporting a large variety of fish communities, with many commercial species that sustain fishing activities. We distinguished two categories of species on Figure 5, according to the index parameter: $\phi > 1$ or $\phi \leq 1$. Species belonging to the second category being very

Figure 4: Statistics of the computational cost for processing the same species as in Figure 3, with Poisson approximation.

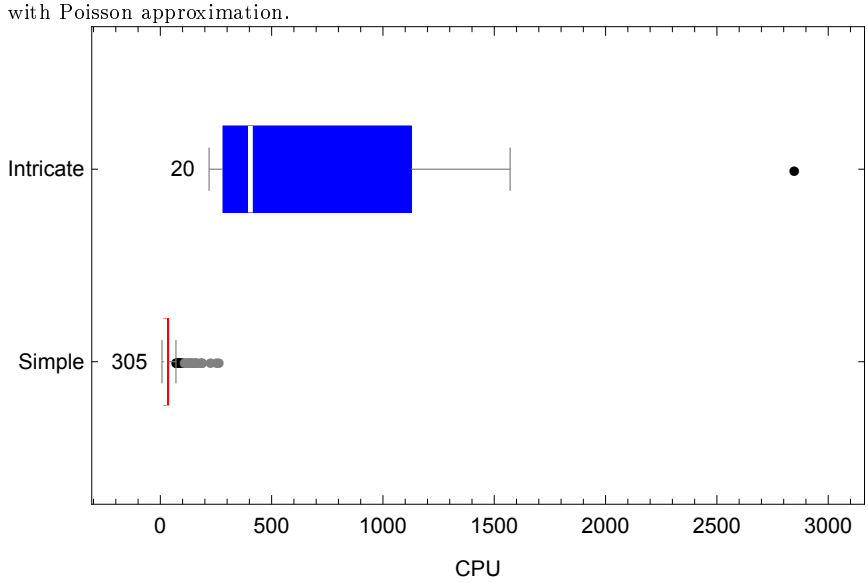


Figure 5: Species collected in zone 4 of the MEEZ; the green dotted (resp. orange dashed) closed curve corresponds to the confidence region of 0.99 level associated with the spatial median of the first (resp. second) category species .

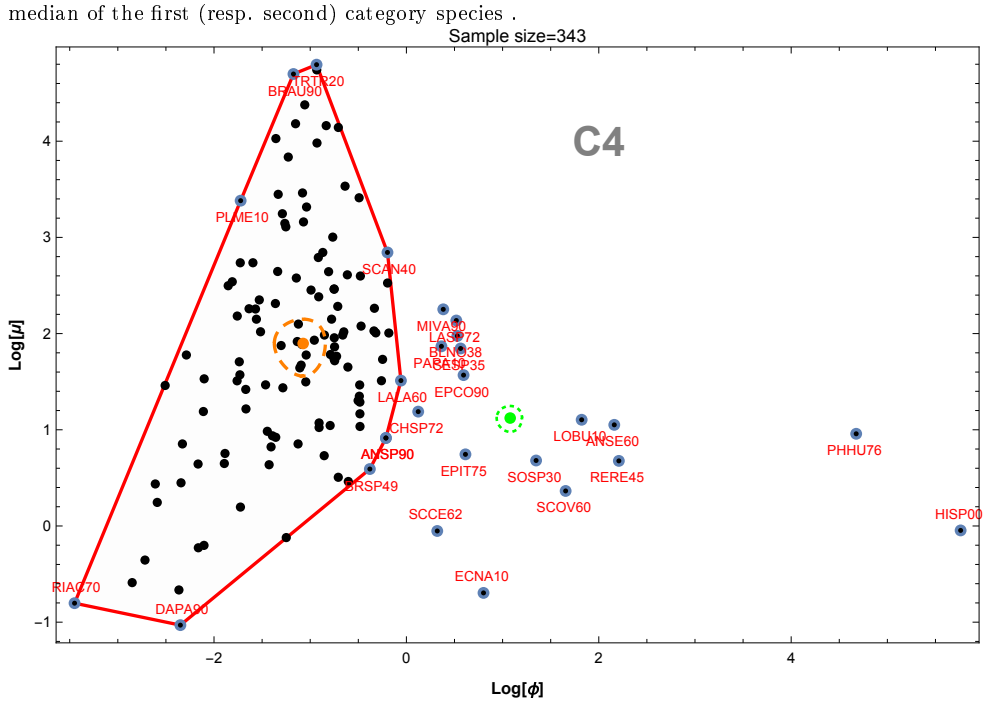
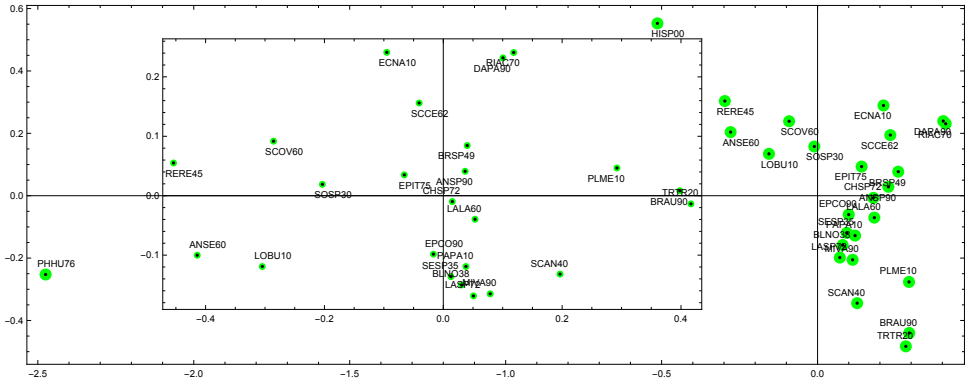


Figure 6: MDS of the Rao's distance between the 26 selected species (big points); the inset graph corresponds to the same analysis, performed after removing both the species "HISP00" and "PHHU76". .



numerous, we only kept for MDS those which are situated on the convex envelope on the associated cloud. Twenty-six species were selected this way (while 138 species are represented) for computation of the Rao's distances and subsequent MDS of the table of distances. We also plotted on Figure 5 theoretical confidence ellipses centered on the spatial medians [28] of both species categories.

3.3.3. Representation of the Rao's distance table

Only one of the 26 species retained in C4 was Poisson-like: "HISP00" (*Hippocampus sp*). Consequently, the distance table was defined this way: $\Delta_{i,j} = D_{\mathcal{R}}\left(\mathfrak{L}^i, \mathfrak{L}^j\right)$, excepted when one of these species was "HISP00" (index h, say). In these cases, we used Poisson Approximation, as in Section 3.2.1; MDS of the 26 species is thus represented on Figure 6. Clearly, the counts of the species "HISP00" and "PHHU76" (*Physiculus huloti*, a type of cod) are distributed in a very special manner, which was not so obvious on Figure 5. In the inset region we represented the other 24 species separately analyzed by MDS of the restricted table Δ^{AC} (no point from the first MDS is hidden).

4. Results and discussion

Motivated by the analysis of a large data set of marine species counts collected in the MEEZ, we developed a parameter-free method to compare species

counts distributions in the setting of the Riemannian manifold $NB(D_{\mathcal{R}})$ of negative binomial distributions, equipped with $D_{\mathcal{R}}$.

We focused first [15] on numerical problems met in computing $D_{\mathcal{R}}\left(\mathfrak{L}^1, \mathfrak{L}^2\right)$: lengthy computations could result from the presence of a cut point on the geodesic $\mathfrak{L}^1 \curvearrowright \mathfrak{L}^2$, requiring to determine a stopover S somewhere between these distributions. $D_{\mathcal{R}}\left(\mathfrak{L}^1, \mathfrak{L}^2\right)$ was then bounded by $D_{\mathcal{R}}\left(\mathfrak{L}^1, S\right) + D_{\mathcal{R}}\left(S, \mathfrak{L}^2\right)$.

In this work, we have essentially shown that Poisson Approximation, combined with Differential Geometry techniques, can be used to evaluate more efficiently $D_{\mathcal{R}}\left(\mathfrak{L}^1, \mathfrak{L}^2\right)$ when one (at least) of the distributions involved is "Poisson-like". Superseding original NB distributions by equivalent ones (in the sense of Definition 7), we could obtain lower upper bounds of the distances than with the former strategy, with a lower computational cost. More precisely, this refinement enabled us to get around computation issues: the number of intricate configurations has been approximately divided by two, the computational cost of the corresponding distances has been approximately divided by two, and numerical cut points were nearly eliminated.

5. Appendix

Proof of Theorem 5

Notice first that on the fiber F_λ, because of (13), the expression of Δ defined in (12) is $\frac{\lambda\left(e^{\frac{\lambda\phi}{\lambda-\phi}}-1\right)}{\lambda-\phi}$, giving:

$\frac{\partial\Delta}{\partial\phi}(\phi) = \frac{\lambda\left(e^{\frac{\lambda\phi}{\lambda-\phi}}\left(\lambda^2+\lambda-\phi\right)-\lambda+\phi\right)}{(\lambda-\phi)^3}$. Since $\phi > \lambda$, the denominator of this expression is always negative while the numerator is clearly positive, excepted potentially if $\left(\lambda^2+\lambda-\phi\right) < 0$. Substituting $\lambda^2+\lambda+\zeta$ to ϕ (with $\zeta > 0$) in the equation, we get a simpler expression for $\frac{\partial\Delta}{\partial\phi}$:

$$-\frac{\lambda\left(-\zeta\, e^{\frac{\zeta}{\zeta+\lambda^2}-\lambda-1}+\zeta+\lambda^2\right)}{\left(\zeta+\lambda^2\right)^3}$$

whose sign depends on the sign of $\left(1-e^{\frac{\zeta}{\zeta+\lambda^2}-\lambda-1}\right)$. Since the only solutions of $\frac{\zeta}{\zeta+\lambda^2}-\lambda-1=0$ are $\lambda=0 \wedge \zeta \neq 0$ and $\zeta = -\left(\lambda+\lambda^2\right) \wedge \lambda \neq 0$, this expression

is negative, and $\left(1 - e^{\frac{\varsigma}{\varsigma+\lambda^2} - \lambda - 1}\right) \geq 0$. Consequently, $\frac{\partial \Delta}{\partial \phi}(\phi)$ is negative and $\Delta(\phi, \mu)$ is a decreasing function of ϕ on a fiber.

Acknowledgments

We thank the Mauritanian Institute of Oceanographic Research and Fisheries (especially S. O. Kidé) and the Department of Cooperation and Cultural Action of the Embassy of France in Mauritania for their support for this study. We also thank all scientists who contributed to field surveys and data collection.

References

[1] M. Berger, A Panoramic View of Riemannian Geometry, Springer, Berlin, Heidelberg, 2003.

[2] C. Manté, J. P. Durbec, J. C. Dauvin, A functional data-analytic approach to the classification of species according to their spatial dispersion. Application to a marine macrobenthic community from the Bay of Morlaix (Western English Channel), Journal of Applied Statistics 32 (2005) 831–840.

[3] A. Srivastava, I. Jermyn, S. Joshi, Riemannian Analysis of Probability Density Functions with Applications in Vision, in: 2007 IEEE Conference on Computer Vision and Pattern Recognition, 2007, pp. 1–8.

[4] V. Seguy, M. Cuturi, Principal Geodesic Analysis for Probability Measures under the Optimal Transport Metric, in: C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, R. Garnett (Eds.), Advances in Neural Information Processing Systems 28, Curran Associates, Inc., 2015, pp. 3312–3320.

[5] A. Petersen, H.-G. Muller, Functional data analysis for density functions by transformation to a Hilbert space, The Annals of Statistics 44 (2016) 183–218.

[6] E. Cazelles, V. Seguy, J. Bigot, M. Cuturi, N. Papadakis, Geodesic PCA versus Log-PCA of Histograms in the Wasserstein Space, SIAM Journal on Scientific Computing 40 (2018) B429–B456.

[7] C. Manté, S. O. Kidé, A.-F. Yao-Lafourcade, B. Merigot, 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 (2016) 359–385.

[8] C. R. Rao, Information and the Accuracy Attainable in the Estimation of Statistical Parameters, Resonance-Journal of Science Education 20 (2015) 78–90.

[9] C. R. Rao, Comment to Kass' paper, Statistical Science 4 (1989) 229–231.

[10] K. M. Carter, R. Raich, W. G. Finn, A. O. Hero, FINE: Fisher Information Nonparametric Embedding, IEEE Transactions on Pattern Analysis and Machine Intelligence 31 (2009) 2093–U195.

[11] G. Galanis, P. C. Chu, G. Kallos, Y.-H. Kuo, C. T. J. Dodson, Wave height characteristics in the north Atlantic ocean: a new approach based on statistical and geometrical techniques, Stochastic Environmental Research and Risk Assessment 26 (2012) 83–103.

[12] C. T. J. Dodson, Some illustrations of information geometry in biology and physics, in: J. Leng, W. W. Sharrock (Eds.), Handbook of research on computational science and engineering, Engineering Science Reference, 2012, pp. 287–315.

[13] M. Cubedo, A. Minarro, J. M. Oller, A dissimilarity based on relevant population features, Journal of Statistical Planning and Inference 143 (2013) 346–355.

[14] I. Ilea, L. Bombrun, C. Germain, I. Champion, R. Terebes, M. Borda, Statistical Hypothesis Test for Maritime Pine Forest Sar Images Classification

Based on the Geodesic Distance, in: 2015 IEEE International Geoscience and Remote Sensing Symposium (igarss), IEEE, New York, 2015, pp. 3215–3218.

[15] C. Manté, S. O. Kidé, Approximating the Rao's distance between negative binomial distributions. Application to counts of marine organisms, in: Proceedings of COMPSTAT 2016, A. Colubi, A. Blanco and C. Gatu., Oviedo (Spain), 2016, pp. 37–47. URL: `https://hal.archives-ouvertes.fr/hal-01357264`.

[16] R. E. Kass, The geometry of asymptotic inference, Statistical Science 4 (1989) 188–234.

[17] M. Menendez, D. Morales, L. Pardo, M. Salicru, Statistical Tests Based on Geodesic Distances, Applied Mathematics Letters 8 (1995) 65–69.

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

[19] J. Burbea, C. R. Rao, Informative geometry of probality spaces, Expo. Math. 4 (1986) 347–378.

[20] S.-i. Amari, H. Nagaoka, D. Harada, Methods of information geometry, number 191 in Translations of mathematical monographs, nachdr. ed., American Math. Soc. [u.a.], Providence, RI, 2007.

[21] A. Gray, Modern differential geometry of curves and surfaces with Mathematica, 2nd ed ed., CRC Press, Boca Raton, 1998.

[22] M. P. d. Carmo, Riemannian geometry, Mathematics. Theory & applications, Birkhauser, Boston, 1992.

[23] J. I. Itoh, T. Sakai, Cut loci and distance functions, Math. J. Yokohama Univ. 49 (2007) 65–92.

[24] K. C. Chua, S. H. Ong, Test of misspecification with application to negative binomial distribution, Computational Statistics 28 (2013) 993–1009.

[25] M. Abramowicz, I. A. Stegun, Handbook of mathematical functions with formulas, graphs and mathematical tables, Knovel, London, 2002.

[26] W. R. Inc., Mathematica, Version 11.3, 2017. Champaign, IL, 2018.

[27] M. Majsnerowska, A note on Poisson approximation by w-functions, Applicationes mathematicae 25 (1998) 387–392.

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

1. Numerical cut points and broken geodesics

Consider a Riemannian manifold $\mathfrak{M}$; we have the following plain proposition.

Proposition 1. *Let $\gamma : I \to \mathfrak{M}$ be a geodesic* ***with respect to the metric connection*** ∇. *Then γ has constant speed in the local norm M1*

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

and, for any $[a,b] \subseteq I$, we have:

$$\int_a^b \sqrt{\dot{\gamma}^t(t).\mathfrak{g}(\gamma(t)).\dot{\gamma}(t)}dt = (b-a)\ \|\dot{\gamma}\|_{\mathfrak{g}}.$$

Suppose that a satisfactory solution $\gamma = \mathfrak{L}^1 \equiv \left(\phi^1, \mu^1\right) \curvearrowright \left(\phi^2, \mu^2\right) \equiv \mathfrak{L}^2$ of problem (M3) under the boundary condition (M11) has been found; according to Proposition 1 above, $D_{\mathcal{R}}\left(\mathfrak{L}^1, \mathfrak{L}^2\right)$ is naturally approximated by $\|\dot{\gamma}\|_{\mathfrak{g}}$. But notice that $\|\dot{\gamma}\|_{\mathfrak{g}}$ is an upper bound which is attained only when there is no cut point in $\gamma([0,1])$ (Theorem M1). That is why we needed some test to detect a possible cut point on some geodesic curve. Suppose a cut point $\left(\phi^{c(1,2)}, \mu^{c(1,2)}\right)$ has been detected on γ. Then, it is natural [1] to supersede γ by the broken geodesic

$$\left(\phi^1, \mu^1\right) \curvearrowright \left(\phi^{c(1,2)}, \mu^{c(1,2)}\right) \oplus \left(\phi^{c(1,2)}, \mu^{c(1,2)}\right) \curvearrowright \left(\phi^2, \mu^2\right)$$

whose length is shorter than $\Lambda(\gamma)$, provided that $\left(\phi^{c(1,2)}, \mu^{c(1,2)}\right) \curvearrowright \left(\phi^2, \mu^2\right)$ is also a segment. Since in our case geodesics are obtained by numerically solving a differential equation, we are only able to locate **numerical** cut points. Such points were called (J,ε)-cut points, since their detection depends on two parameters: the chosen number J of sub-intervals of $[0,1]$ determining the accuracy of computations, and some fixed threshold $\varepsilon \in]0,1[$.

1.1. Locating some (J,ϵ)- cut point on a geodesic γ

For that purpose, the unit interval is first divided into J intervals: $[0,1] = \bigcup_{i=1}^{J} \delta_i$, with $\delta_i := [\frac{i-1}{J}, \frac{i}{J}[$. Suppose there exists a cut point $\gamma(t_c)$ on γ, such that $t_c \in \delta_{i_c}$. Consider the set

$$\mathfrak{C}_J(\gamma) := \left\{\gamma_1 := \gamma\left(\frac{1}{J}\right), \cdots, \gamma_k := \gamma\left(\frac{k}{J}\right), \cdots, \gamma_{J-1} := \gamma\left(\frac{J-1}{J}\right)\right\} \subset \mathfrak{M}$$

and, for each $1 \leq i \leq J$ the geodesic $\alpha_i := \gamma_{i-1} \curvearrowright \gamma_i$ obtained by solving (M3) **under the constraints**

$$\{\alpha_i(0) = \gamma_{i-1},\ \alpha_i(1) = \gamma_i\}.$$

Because of the uniqueness of segments (Proposition 1 and Remark M1), $\forall\, i < i_c,\ \frac{\|\dot{\gamma}\|_{\mathfrak{g}}}{J} = \Lambda(\alpha_i) = \|\dot{\alpha_i}\|_{\mathfrak{g}}$. On the contrary, when $i \geq i_c$, the distance between γ_{i-1} and γ_i **along** γ is $\frac{\|\dot{\gamma}\|_{\mathfrak{g}}}{J}$ yet, while $\|\dot{\alpha_i}\|_{\mathfrak{g}}$ should be smaller. More precisely, if the resolution $\frac{1}{J}$ is small enough (for instance, smaller than the injectivity radius of $\mathfrak{M}$ [1]), $\gamma_{i-1} \curvearrowright \gamma_i$ is a segment and we may write:

$$\begin{cases} \forall\, i < i_c,\ \frac{\|\dot{\gamma}\|_{\mathfrak{g}}}{J} - \|\dot{\alpha_i}\|_{\mathfrak{g}} = 0 \\ \qquad\qquad \frac{\|\dot{\gamma}\|_{\mathfrak{g}}}{J} - \|\dot{\alpha_{i_c}}\|_{\mathfrak{g}} > 0 \\ \forall\, i > i_c,\ \frac{\|\dot{\gamma}\|_{\mathfrak{g}}}{J} - \|\dot{\alpha_i}\|_{\mathfrak{g}} \geq 0. \end{cases}$$

Thus, after fixing some (small) ϵ, we can locate possible cut points, with a precision depending on (J, ϵ) and lay a definition for numerical cut points.

Definition 2. We will say that $\gamma_{i_c} \in \mathfrak{C}_J(\gamma)$ is a (J, ϵ)- cut point on γ if

$$i_c = \arg \min_{1 \leq i \leq J-1} \left(\left| \frac{\|\dot{\gamma}\|_{\mathfrak{g}}}{J} - \|\dot{\alpha_i}\|_{\mathfrak{g}} \right| > \epsilon \right).$$

2. Possible numerical issues

Various numerical problems were met in computing $(\phi^1, \mu^1) \curvearrowright (\phi^2, \mu^2)$:

(P1) no solution was found (due to time limit)

(P2) an unsuitable solution was found: for some $t \in [0, 1]$, $(\phi(t), \mu(t)) \notin \mathbb{R}^+ \times \mathbb{R}^+$

(P3) the boundary condition (M11) was not fulfilled with a satisfactory precision.

Thus, two kinds of configuration were distinguished: **simple**, when none of the above issues is met, or **intricate** in the alternative. For further information, see [2].

2.1. Simple configurations

When none of the above issues is met, we first check that there is no (J,ε)-cut point on $\gamma = \left(\phi^1,\mu^1\right) \curvearrowright \left(\phi^2,\mu^2\right)$. Then, the solution found is accepted, and we can write:

$$D_{\mathcal{R}}\left(\mathfrak{L}^1,\mathfrak{L}^2\right) = \Lambda\left(\gamma\right) = \left\|\dot{\gamma}\right\|_{\mathfrak{g}}. \tag{1}$$

If a (J,ε)-cut point $\left(\phi^{c(1,2)},\mu^{c(1,2)}\right)$ is detected on γ, and if $\left(\phi^{c(1,2)},\mu^{c(1,2)}\right) \curvearrowright \left(\phi^2,\mu^2\right)$ is free of (J,ε)-cut point, we adopt as an upper bound for $D_{\mathcal{R}}\left(\mathfrak{L}^1,\mathfrak{L}^2\right)$

$$\Lambda\left(\left(\phi^1,\mu^1\right) \curvearrowright \left(\phi^{c(1,2)},\mu^{c(1,2)}\right)\right) + \Lambda\left(\left(\phi^{c(1,2)},\mu^{c(1,2)}\right) \curvearrowright \left(\phi^2,\mu^2\right)\right).$$

2.2. Intricate configurations

When (P1) or (P2) is met, we consider that the best achievable solution consists in breaking $\gamma = \left(\phi^1,\mu^1\right) \curvearrowright \left(\phi^2,\mu^2\right)$ by inserting a well-placed "stopover". But since γ is undetermined, how should this stopover $\left(\phi^{S(1,2)},\mu^{S(1,2)}\right)$ be placed? We proposed two heuristics for that purpose:

1. compute a "rough solution" $\widetilde{\gamma_R}$ to the original problem M3, contenting ourselves with low-precision (5 digits), and substitute $\widetilde{\gamma_R}$ for γ to search for $\left(\phi^{S(1,2)},\mu^{S(1,2)}\right)$
2. when $\widetilde{\gamma_R}$ cannot be obtained, merely use instead $\widetilde{\gamma_L}\left(t\right) := t\,\left(\phi^1,\mu^1\right) + \left(1-t\right)\,\left(\phi^2,\mu^2\right)$.

In the second case, the stopover S naturally corresponds to the shortest broken geodesic determined by:

$$\begin{cases} \left(\phi^{S(1,2)},\mu^{S(1,2)}\right) = \widetilde{\gamma_L}\left(\frac{k_L}{J}\right) \\ \textit{with } k_L := \arg\min\limits_{1\le k\le J-1}\left(\Lambda\left(\left(\phi^1,\mu^1\right) \curvearrowright \widetilde{\gamma_L}\left(\frac{k}{J}\right)\right) + \Lambda\left(\widetilde{\gamma_L}\left(\frac{k}{J}\right) \curvearrowright \left(\phi^2,\mu^2\right)\right)\right). \end{cases} \tag{2}$$

In the first case, two eventualities must be considered:

1. a (J,ε)-cut point $\left(\phi^{c(1,2)},\mu^{c(1,2)}\right)$ is detected on $\widetilde{\gamma_R}\left([0,1]\right)$: then $\left(\phi^{S(1,2)},\mu^{S(1,2)}\right)=\left(\phi^{c(1,2)},\mu^{c(1,2)}\right)$, and the length of the broken geodesic is computed in full precision
2. if no (J,ε)-cut point is detected, proceed like in (2)

$$\begin{cases}\left(\phi^{S(1,2)},\mu^{S(1,2)}\right)=\widetilde{\gamma_R}\left(\frac{k_R}{J}\right)\\ \textit{with } k_R:=\arg\min\limits_{1\leq k\leq J-1}\left(\Lambda\left(\left(\phi^1,\mu^1\right)\curvearrowright\widetilde{\gamma_R}\left(\frac{k}{J}\right)\right)+\Lambda\left(\widetilde{\gamma_R}\left(\frac{k}{J}\right)\curvearrowright\left(\phi^2,\mu^2\right)\right)\right).\end{cases} \tag{3}$$

2.3. Boundary issues

(P3) is easy to solve: we just have to complete (1) by the corrective boundary error term:

$$BE\left(\gamma\right):=\left\|\gamma(0)-\mathfrak{L}^1\right\|_{\mathfrak{g}}\left(\mathfrak{L}^1\right)+\left\|\gamma(1)-\mathfrak{L}^2\right\|_{\mathfrak{g}}\left(\mathfrak{L}^2\right). \tag{4}$$

Finally, whatever the selected geodesic (broken, or not) may be, we obtain the upper bound:

$$D_{\mathcal{R}}\left(\mathfrak{L}^1,\mathfrak{L}^2\right)\leq\Lambda\left(\gamma\right)+BE\left(\gamma\right). \tag{5}$$

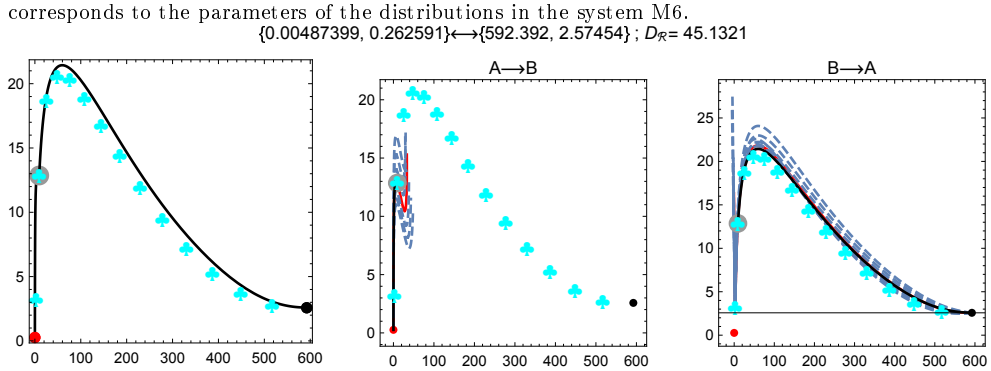
3. A bestiary of computational problems

Before processing data, it was necessary to tune the accuracy parameters; we fixed: $(J,\epsilon)=(10,0.05)$.

We will display in this section typical cases of intricate configurations encountered in computing $D_{\mathcal{R}}$ between marine species count distributions. Each illustration will be composed of three panels. On the left one we superimposed to the rough solution of (3), $\widetilde{\gamma_R}$ (suits), the final solution: a broken geodesic (black curve). On both the right panels, we featured the structure of each component of the broken geodesic in the neighborhood of the stopover S, with the help of the exponential map: remember that geodesics can be computed by solving (M3) under the alternative constraints (initial value problem)

$$\left\{\gamma\left(0\right)=\left(\phi^1,\mu^1\right),\,\dot{\gamma}\left(0\right)=V\in\mathbb{R}^2\right\}. \tag{6}$$

Figure S1: Numerical cut point ($\widetilde{\gamma_R}$ satisfactory). Left panel: first guess (suits) and final solution (black curve). Right panels: plot of the two bundles of geodesics issued from A or B. Red curve (color figure online): $\theta = 0$ in equation (7); dashed curves: $\theta \neq 0$. The header corresponds to the parameters of the distributions in the system M6.



Here V is the initial velocity of the geodesic, and this solution is associated with the exponential map at $\left(\phi^1, \mu^1\right)$. We first determined this way $\gamma_1 = A \curvearrowright S$ (resp. $\gamma_2 = B \curvearrowright S$) by solving equation (M3) under the constraints (M11). We afterward considered

$$\{V_i\left(\theta_k\right) := \rho\left(\theta_k\right) \centerdot \mathcal{B}\left(\dot{\gamma}_i(0)\right) : \; i = 1, 2\}, \tag{7}$$

where the angle of the rotation ρ acting on the initial direction $\mathcal{B}\left(\dot{\gamma}_i(0)\right)$ is $\theta_k \in \{0, \pm 0.1, \pm 0.2, \pm 0.3\}$ (in degrees). Equation (M3) was then solved under the constraints (6) with $V = V_i\left(\theta_k\right)$, giving rise to two bundles of seven geodesics. In all these plots, the red (color figure online) point will be "A" and the black one will be "B", while the stopover is represented by the big gray point; exponential maps corresponding to $\theta = 0$ (unrotated) are plotted in red (color figure online).

On Figures S1 and S2 are represented typical cases of intricate situations where $\widetilde{\gamma_R}$ was a satisfactory first guess for the geodesic. The final solution

$A \curvearrowright S \oplus S \curvearrowright B$ is similar to $\widetilde{\gamma_R}$. The structure of $A \curvearrowright S$ and $B \curvearrowright S$ is investigated on both right panels of each sub-figure. Notice they look like genuine cut points ("arriving at the cut locus means some kind of catastrophe" [1, p. 279]; cf. also the last statement of Theorem M1).

On Figures S3 and S4 are represented worse cases, where the first guess $\widetilde{\gamma_R}$

Figure S2: Numerical cut point ($\widetilde{\gamma_R}$ satisfactory). Same structure as in S1.

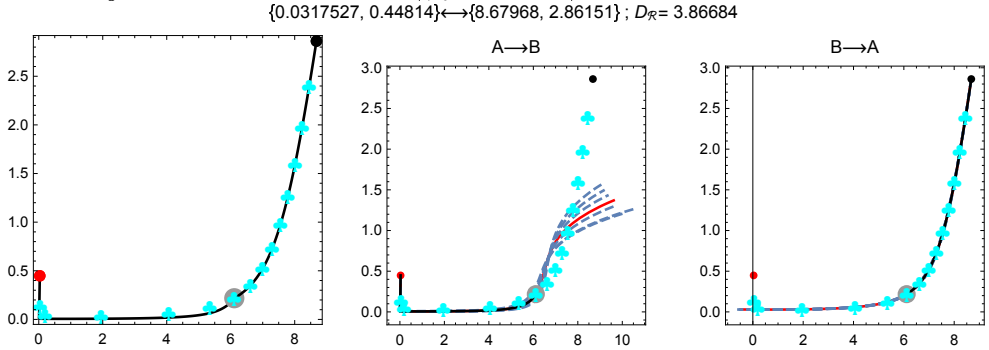
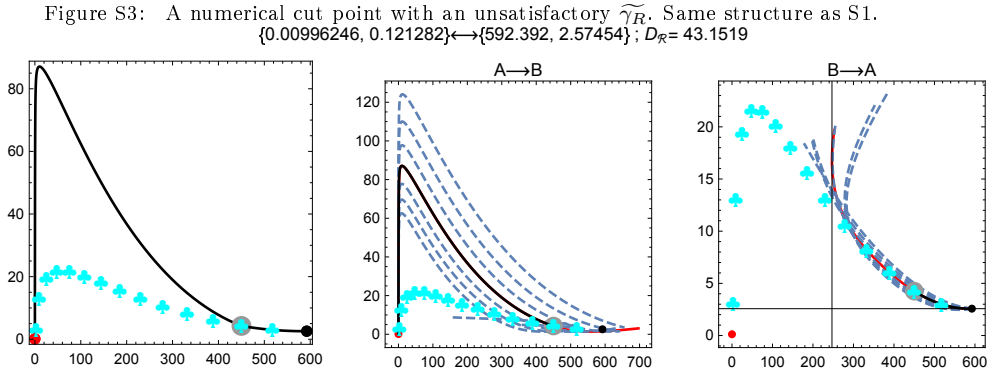
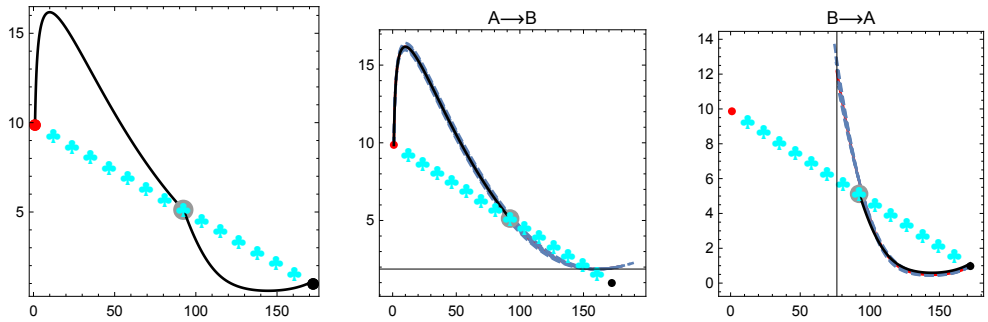


Figure S3: A numerical cut point with an unsatisfactory $\widetilde{\gamma_R}$. Same structure as S1.



was either unsatisfactory, or could not even be computed.

Figure S4: Linear first guess. Same structure as S1.
{0.938781, 9.86571}⟷{172.236, 0.974793} ; $D_{\mathcal{R}}$= 21.351



References

[1] M. Berger, A Panoramic View of Riemannian Geometry, Springer, Berlin, Heidelberg, 2003.

[2] C. Manté, S. O. Kidé, Approximating the Rao's distance between negative binomial distributions. Application to counts of marine organisms, in: Proceedings of COMPSTAT 2016, A. Colubi, A. Blanco and C. Gatu., Oviedo (Spain), 2016, pp. 37–47. URL: `https://hal.archives-ouvertes.fr/hal-01357264`.