Most Likely Maximum Entropy for Population Analysis: a case study in decompression sickness prevention

Youssef Bennani, Luc Pronzato and Maria João Rendas

Univ. Nice Sophia Antipolis, CNRS, I3S, UMR 7271, 06900 Sophia Antipolis, France

Abstract. We estimate the density of a set of biophysical parameters from region censored observations. We propose a new Maximum Entropy (*maxent*) estimator formulated as finding the most likely constrained *maxent* density. By using the Rényi entropy of order two instead of the Shannon entropy, we are lead to a quadratic optimization problem with linear inequality constraints that has an efficient numerical solution. We compare the proposed estimator to the NPMLE and to the best fitting *maxent* solutions in real data from hyperbaric diving, showing that the resulting distribution has better generalization performance than NPMLE or *maxent* alone.

Keywords: Censored observations, non-parametric maximum likelihood, constrained maxent, regularisation.
PACS: <>

INTRODUCTION

The density estimation problem addressed in this paper is motivated by a problem of population analysis: we are interested on the distribution π_θ of the biophysical parameters θ of a mathematical model [1] for the instantaneous volume of micro-bubbles flowing through the right ventricle of a diver's heart when executing a decompression profile P:

$$(\theta, \{P(u)\}_{u\le t}) \to B(t, \theta, \{P(u)\}_{u\le t})).$$

The problem appears in the context of prevention of decompression sickness (DCS) in deep sea diving: since DCS is known to be highly correlated with the presence of gas bubbles in the blood, ability to correctly predict the probability that this volume becomes exceedingly high can be used to establish safety rules that avoid profiles that are dangerous for a non-negligible part of the population.

The instantaneous gas volume B is observed through periodic measurements of bubble grades G. Since it is usually accepted that DCS is related to the maximum observed grade, only strongly quantified versions of the peak volume $b(\theta,P) = \max_t B(t,\theta,\{P(u)\}_{u\le t})$ have been recorded:

$$G(\theta,P) = l \Leftrightarrow b(\theta,P) \in [\tau_l, \tau_{l+1}[, \;\; l \in \{0,\dots,L\}. \tag{1}$$

In our case $L = 4$, and thresholds $\tau = \{\tau_l\}_{l=1}^L$, where $\tau_0 = 0 < \tau_1 < \ldots < \tau_L < \tau_{L+1} = \infty$, are assumed known. Fig. 1 illustrates these definitions. A simplified model with $\theta \in \Theta \subset \mathbf{R}^2$ has been used, all other parameters of model [1] being held fixed. All θ in region $R_i \equiv \{\theta \in \Theta \,:\, b(\theta,P_i) \in [\tau_{G_i}, \tau_{G_i+1}[\}$, yield the same observed grade G_i for

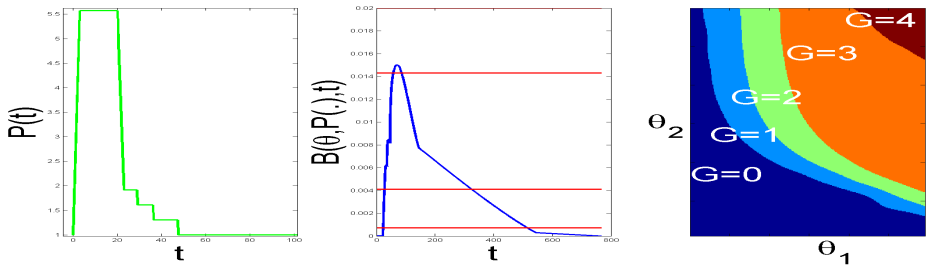

FIGURE 1. Left: profile $P(t)$. Centre: instantaneous volume B for one parameter value (blue) and thresholds (red). Right: regions corresponding to the 5 possible grades.

profile P_i. In Fig. 1 (right) we draw the regions corresponding to the profile P shown on the left. The plot in the centre shows $B(t,\theta,\{P(u)\}_{u\leq t})$ (thresholds τ are indicated by the horizontal red lines for a particular value of θ). All θ inside the orange region will yield a grade $G=3$ for this profile.

The remark above shows that estimation of π_θ from observations $\{(G_i,P_i)\}_{i=1}^n$ is equivalent to the problem of estimating π_θ from observation of the set of regions $\{R_i\}_{i=1}^n$. We speak of "region-censoring". In the absence of knowledge about the expected dispersion of the biological parameters, we estimate π_θ non-parametrically, imposing no constrains on its shape.

Several facts are known about the Non-Parametric Maximum Likelihood Estimator (NPMLE) for interval-censored observations: *(i)* its support $\mathscr{S}_{\text{NPMLE}}=\{\theta,:\hat{\pi}_\theta(\theta)>0\}$ is confined to a finite number K of disjoint intervals (the so called "elementary regions"): $\mathscr{S}_{\text{NPMLE}}=\cup_{\ell=1}^K E_\ell, E_\ell\cap E_q=\emptyset, \ell\neq q$; *(ii)* all distributions that put the same probability mass $w_\ell\equiv\{\pi_\theta(E_\ell)\},\ell=1,\ldots,K$ in these intervals have the same likelihood; *(iii)* there is in general no unique assignment of probabilities $\{\hat{w}_\ell\}_{\ell=1}^K$ that maximizes the likelihood. Fact *(i)* shows that the NPMLE problem can be studied in the K-dimensional probability simplex $\mathbf{S}^K$, since $\hat{\pi}_\theta(\cdot)$ is determined only up to the probability vector $\hat{\mathbf{w}}=\{\hat{w}_1,\ldots,\hat{w}_K\}$. The two types of "non-uniqueness" of the NPMLE, *(ii)* and *(iii)*, have been first pointed out by Turnbull [2]. More recently, they were studied in detail for the multi-variate case in [3], where the authors coined the terms *representational* – *(ii)* – and *mixture* – *(iii)* – non-uniqueness, further showing that the set of NPMLE's $\hat{\pi}_\theta$ is a polytope.

THE NPMLE

We first introduce some notations. Let m denote the number of distinct profiles $\{P^{(j)}\}_{j=1}^m$ in the available dataset, and $\{R_i^{(j)}\}_{i=0,j=1}^{L,m}$ the corresponding regions $R_i^{(j)}=\left\{\theta\in\Theta\,:\,b(\theta,P^{(j)})\in[\tau_i,\tau_{i+1}[\right\}$. $\mathscr{Q}^{(j)}=\{R_i^{(j)}\}_{i=0}^L$ is a partition of Θ, see Fig. 1 (right). Let $\mathscr{Q}$ denote the smallest partition of Θ whose generated σ-algebra, $\sigma(\mathscr{Q})$, contains all partitions $\{\mathscr{Q}^{(j)}\}_{j=1}^m$. Denote by M its size. Let $\mathbf{E}_i^{(j)}$ be the set of elements of $\mathscr{Q}$ that intersect $R_i^{(j)}$, and L_A be the list of regions $R_i^{(j)}$ that contain $A\in\mathscr{Q}$.

Since the "elementary regions" $\{E_\ell\}_{\ell=1}^K$ are elements of $\mathscr{Q}$, notations $\mathbf{E}_i^{(j)}$ and L_{E_ℓ} are well defined. Let n be the total number of dives observed. In general, the same grade $G=i$ has been observed for the same profile $P=P^{(j)}$ more than once. Let n_j be the number of times $P^{(j)}$ has been executed, and $n_i^{(j)}$ the number of times grade i has been observed for $P^{(j)}$, such that $\sum_{i=0}^L n_i^{(j)} = n_j$ and $\sum_{j=1}^m n_j = n$. Denote by $\mathbf{f}^{(j)} = \{f_i^{(j)}\}_{i=0}^L$ the empirical distribution of the grades in $P^{(j)}$, $f_i^{(j)} = \frac{n_i^{(j)}}{n_j}$. Assuming that divers have been independently drawn in a population with probability distribution π_θ, the log-likelihood is

$$l_n(\pi_\theta; \{n_i^{(j)}, R_i^{(j)}\}) = \sum_{j=1}^m \sum_{i=0}^L n_i^{(j)} \log \pi_\theta(R_i^{(j)}) \quad . \tag{2}$$

From property *(i)* we have $\pi_\theta(R_i^{(j)}) = \sum_{E \in \mathbf{E}_i^{(j)}} \pi_\theta(E) \quad = \mathbf{B}_{i.}^{(j)} \mathbf{w}$, where $\mathbf{B}_{i.}^{(j)}$ is the i-th row of $\mathbf{B}^{(j)}$, the $(L+1) \times K$ binary matrix, with $\mathbf{B}^{(j)}(i,\ell) = 1 \Leftrightarrow E_\ell \in \mathbf{E}_i^{(j)}$, and $\mathbf{w} \in \mathbf{S}^K$ the vector of probabilities of the E_ℓ's: $\{w_\ell = \pi_\theta(E_\ell)\}_{\ell=1}^K$. We deduce that all π_θ leading to the same $\mathbf{w}$ have the same likelihood (property *(ii)*). In general, see [4], there is no single $\mathbf{w}$ maximizing l_n: let $\hat{\mathbf{w}}$ be an NPMLE, then all elements of $\mathscr{P} = \left\{ \mathbf{w}, \text{ s.t. } \forall j,\ \mathbf{B}^{(j)}\mathbf{w} = \mathbf{B}^{(j)}\hat{\mathbf{w}} \right\}$ are NPMLE's. We call $\mathscr{P}$ the NPMLE polytope.

Optimization

Several algorithms have been proposed to maximize the log-likelihood function l_n in 2, see e.g. [4]. It can be shown that the problem is equivalent to an optimal design problem, where $\mathbf{w}$ plays the role of the design measure, enabling application of a vast collection of efficient algorithms originating from optimal design. As shown in [5, 6, 7]:
(1) l_n is concave in $\mathbf{S}^K$.
(2) For any two probability measures $\mathbf{w}$ and $\mathbf{w}'$, the directional derivative defined as

$$\begin{aligned} F(\mathbf{w}, \mathbf{w}') &= \lim_{\alpha \to 0^+} \left(\frac{l_n((1-\alpha)\mathbf{w} + \alpha \mathbf{w}') - l_n(\mathbf{w})}{\alpha} \right) \\ &= \frac{1}{n} \sum_{j=1}^m \sum_{i=0}^L n_i^{(j)} \frac{\mathbf{B}_{i.}^{(j)} \mathbf{w}'}{\mathbf{B}_{i.}^{(j)} \mathbf{w}} - 1, \end{aligned}$$

exists.
(3) $\forall \ell = 1, \ldots, K, F(\hat{\mathbf{w}}, \mathbf{e}_\ell) \leq 0$ and $F(\hat{\mathbf{w}}, \mathbf{e}_\ell) = 0$ for every support point ℓ of $\hat{\mathbf{w}}$, with $\mathbf{e}_\ell$ the ℓ-th element of the canonical base of $\mathbf{R}^K$.
Based on properties *(1)-(3)*, several algorithms can be shown to maximize the log-likelihood.

The Vertex Direction Method:
The Vertex Direction Method (VDM) exploits the fact that if for $\ell = 1, \ldots, K$, we have $F(\mathbf{w}, \mathbf{e}_\ell) > 0$, then likelihood can be increased by putting more probability mass

at ℓ. Updating the mass vector is : $\mathbf{w}^{(r+1)} = (1-\alpha_{\max})\mathbf{w}^{(r)} + \alpha_{\max}\mathbf{e}_{\ell^\star}$, with $\ell^\star = \text{argmax}_{\ell=1,\ldots,K} F(\mathbf{w}^{(r)}, \mathbf{e}_\ell)$ and $\alpha_{\max}$ is the value of α maximizing the second-order approximation of $l_n((1-\alpha)\mathbf{w}^{(r)} + \alpha\mathbf{e}_{\ell^\star}) - l_n(\mathbf{w}^{(r)})$. The algorithm is stopped when $\max_{\ell=1,\ldots,K} F(\mathbf{w}^{(r)}, \mathbf{e}_\ell) < \delta \ll 1$, for a fixed $\delta > 0$ controlling distribution optimality.

The Vertex Exchange Method:
The Vertex Exchange Method (VEM) is based on the idea that the likelihood will grow faster if probability mass is moved from $\ell^\circ = \text{argmin}_{\ell=1,\ldots,K, w_\ell \neq 0} F(\mathbf{w}^{(r)}, \mathbf{e}_\ell)$ to $\ell^\star$, resulting in updates as : $\mathbf{w}^{(r+1)} = \mathbf{w}^{(r)} + \alpha_{\max} w^{(r)}_{\ell^\circ}(\mathbf{e}_{\ell^\star} - \mathbf{e}_{\ell^\circ})$, with $\alpha_{\max}$ maximizing a second-order approximation of $l_n(w^{(r)} + \alpha w^{(r)}_{\ell^\circ}(\mathbf{e}_{\ell^\star} - \mathbf{e}_{\ell^\circ})) - l_n(\mathbf{w}^{(r)})$. The same stopping condition as for VDM is used.

The EM method:
The application of the Expectation-Maximization (EM) method leads to the following multiplicative algorithm ([8, 9]): $w_\ell^{(t+1)} = \frac{1}{n}\left(\sum_{j=1}^m \sum_{i=0}^L n_i^{(j)} \frac{\mathbf{B}^{(j)}_{i\ell}}{\mathbf{B}^{(j)}_{i.}\mathbf{w}^{(t)}}\right) w_\ell^{(t)}$, initialized at some strictly positive $\mathbf{w}^{(0)}$. The same stopping condition is considered.

The speed of these 3 algorithms is improved by relying on the ability to detect the entries of $\mathbf{w}$ that will converge to zero, as shown in [10]. We denote by EM+, VDM+ and VEM+ these modified algorithms. For the dataset of our study, and with $\delta = 5.10^{-3}$, we observed a very fast convergence of EM+ algorithm, 115 iterations against 386 iterations for VDM+ algorithm and 583 for VEM+ algorithm with the same δ. When we choose $\delta = 10^{-4}$, the first algorithm to converge is the VEM+ with 600 iterations. The EM+ takes about 3000 iterations in contrast with VDM+ which takes more than 25000 iterations. So we deduce that for relatively medium levels of accuracy, EM+ will be the most efficient, while for high accuracy VEM+ is to be preferred.

REGULARIZED MAXENT

We now present the results obtained for a real dataset. The density estimator for a total of $m = 19$ profiles, repeated a number of times ranging from 12 to 41 (the most dangerous profiles have been executed a reduced number of times) is shown on the top left of Fig. 2, clearly displaying the singularities known affect NPMLE's. The white regions have zero probability mass, and the estimated density has large peaks in a few small dispersed regions.

Motivated by the context of risk assessment, we rely on the notion of entropy to select the element of the NPMLE polytope that is the least informative, and that will thus better reflect the possible diversity of the population analyzed. All $\mathbf{w} \in \mathscr{P}$ define the same measures $\hat{\mathbf{f}}^{(j)}$ over the partitions $\mathscr{Q}^{(j)}$ associated to the profiles $P^{(j)}$. We maximize the Rényi entropy of order 2 : $h_2(\pi_\theta) = -\log \int_\Theta \pi_\theta(\theta)^2 \, d\theta$, also called *extension entropy* [11]. Let $\hat{\mathbf{w}}$ be a solution obtained at convergence of the VEM+ algorithm and $\mathbf{B}$ the matrix that stacks the $\mathbf{B}^{(j)}, j = 1, \ldots, m$. The Rényi-*maxent* NPMLE probability vector

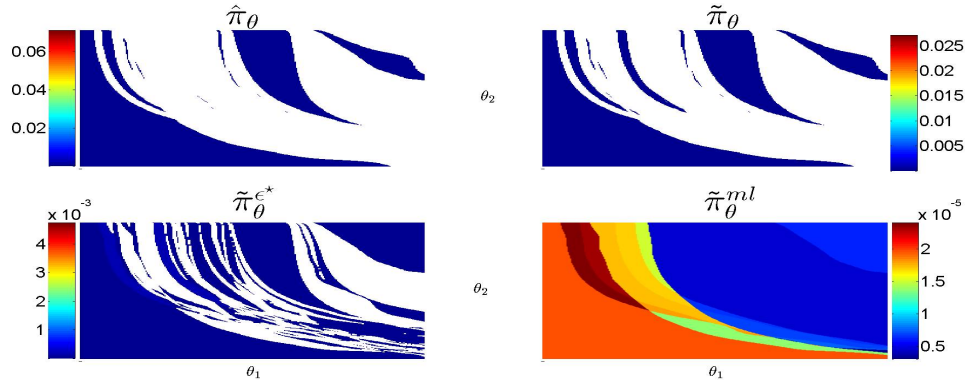

FIGURE 2. Estimates of π_θ. Top left: one NPMLE solution found by VEM+. Top right: Rényi-*maxent* NPMLE. Bottom left: Rényi-*maxent* $\tilde{\pi}_\theta^{\varepsilon^\star}$. Bottom right: ML-Rényi-*maxent* $\tilde{\pi}_\theta^{ml}$. White regions have zero probability mass.

$\tilde{\mathbf{w}}$ is the solution of the following quadratic program with linear equality constraints, for which efficient solutions exist $\tilde{\mathbf{w}} = \text{argmin}_{\mathbf{w}\in\mathbf{S}^K} \left(\sum_{\ell=1}^K \frac{1}{\nu(E_\ell)} w_\ell^2 \right)$, s.t. $B\mathbf{w} = B\hat{\mathbf{w}}$. The Rényi-*maxent* NPMLE computed using the routine *quadprog* of Matlab is displayed in the top right of Fig. 2. We can see that the support of $\tilde{\pi}_\theta$ is larger than the support of $\hat{\pi}_\theta$, but that restriction of the solution to the NPMLE polytope still forces the density to be concentrated in a set of small disconnected regions, with large areas of zero measure. This is inherent to the likelihood criterion, that favours the most concentrated densities that are able to explain the observed data.

Rényi-*maxent*

To avoid the singular behaviour of densities in the NPMLE polytope we must estimate π_θ with a criterion other than Maximum Likelihood. We propose to estimate π_θ as the Rényi-*maxent* distribution that best matches the *observed frequencies* for each profile, $\mathbf{f}^{(j)}$, which, as we saw above, can be written as empirical averages of the indicator functions of the elements of partitions $\mathscr{Q}^{(j)}$. Note that while in the previous section the constraints were determined from the NPMLE polytope, here they are directly obtained from the data.

If there exists a π that can satisfy all constraints, the corresponding $\mathbf{w}$ belongs to the NPMLE polytope $\mathscr{P}$. However, the m constraints will in general be inconsistent and, as in [12], we consider entropy maximization under relaxed constraints. Let $\varepsilon^\star \geq 0$ be the smallest value of ε for which there exists a solution to the problem

$$\tilde{\pi}_\theta^\varepsilon = \text{argmax}_\pi \left(h_2(\pi) \right) \text{s.t.} \quad \left\| \Sigma^{(j)^{-1/2}} \left(\mathrm{E}_\pi[\breve{\mathbf{f}}^{(j)}] - \breve{\mathbf{f}}^{(j)} \right) \right\|_\infty \leq \varepsilon, \ \ \forall j \ ,$$

where $\Sigma^{(j)}$ is the $L \times L$ covariance matrix of the empirical distribution of the vector $\breve{\mathbf{f}}^{(j)}$ obtained from $\mathbf{f}^{(j)}$ by removing the last entry. Note that the equivalence between relaxed *maxent* and penalized likelihood used in [12] does not hold in our case.

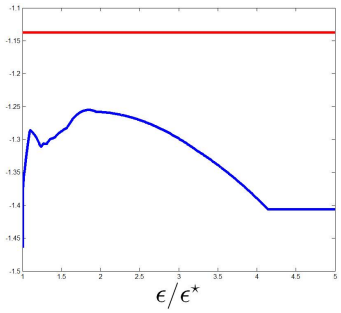

FIGURE 3. Variation of $l_n(\tilde{\pi}_\theta^\varepsilon)$ with $\varepsilon/\varepsilon^\star$. Red line: $l_n(\hat{\pi}_\theta)$.

For our hyperbaric dataset, the Rényi-*maxent* solution corresponding to $\varepsilon^\star$ is shown in the bottom left of Fig. 2. As we can see, the support of this best fitting *maxent* density is still composed of a number of disjoint regions, and does not seem a plausible model of a biological population.

Most likely Rényi-*maxent*

In [13] duality arguments were used to show that, for the Shannon entropy, the solution to the problem above, when the constraints $f_i^{(j)}$ on the expected values are all obtained using the same empirical distribution – derived from an underlying data set $\{\theta_i\}_{i=1}^N$ – is the same as a L_1-penalized maximum likelihood estimate of π from data $\{\theta_i\}_{i=1}^N$ in the Gibbs family. For the censored data problem considered in this paper, the constraints are empirical averages, derived from independent datasets, and this equivalence is lost. Moreover, since we rely on the L_∞ metric to evaluate deviation of the modeled distributions (by $\tilde{\pi}_\theta$) with respect to the empirical $\mathbf{f}^{(j)}$, and L_∞ is not equivalent to the (Riemannian) metric induced by Maximum Likelihood estimation for the exponential family, we cannot guarantee that likelihood is monotonic on the degree of regularization, i.e. $l_n(\tilde{\pi}_\theta^\varepsilon) < l_n(\tilde{\pi}_\theta^{\varepsilon^\star})$, for $\varepsilon > \varepsilon^\star$. In fact, as shown in Fig. 3 that plots the likelihood of $\tilde{\pi}_\theta^\varepsilon$ as a function of $\varepsilon/\varepsilon^\star$, this is not true.

We propose to use likelihood to select the degree of regularization using density $\tilde{\pi}_\theta^{ml}$, the most likely Rényi-*maxent* solution:

$$\tilde{\pi}_\theta^{ml} = \text{argmax}_{\tilde{\pi}_\theta^\varepsilon,\ \varepsilon>\varepsilon^\star}\ \ l_n(\tilde{\pi}_\theta^\varepsilon; \{n^{(j)}, R^{(j)}\})\ \ .$$

This estimate is displayed in the bottom left of Fig. 2. We can see that the support of $\tilde{\pi}_\theta^{ml}$ is now the entire Θ, with a smoother distribution of the probability mass, being a more plausible characterization of the natural variation within a biological population.

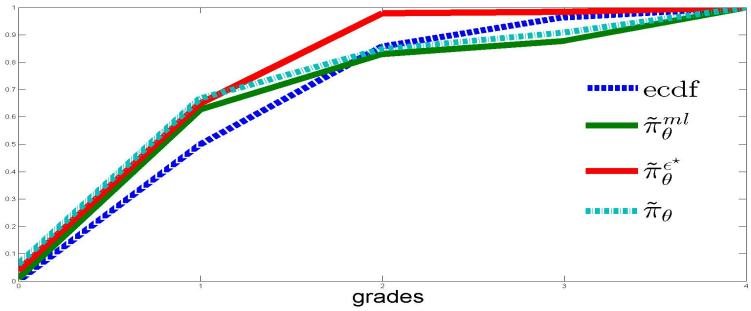

FIGURE 4. The empirical cumulative distribution function (ecdf) and the the 3 estimated cumulative distribution functions of grades for one of the profiles in the dataset.

TABLE 1. Statistics about d_{K} and d_{TV} for the 19 datasets of the *Leave-on-out-cross-validation* model.

	$d_{\mathbf{K}}$			$d_{\mathbf{TV}}$		
	$\tilde{\pi}_\theta$	$\tilde{\pi}_\theta^{\varepsilon^\star}$	$\tilde{\pi}_\theta^{ml}$	$\tilde{\pi}_\theta$	$\tilde{\pi}_\theta^{\varepsilon^\star}$	$\tilde{\pi}_\theta^{ml}$
min	0.1068	0.0540	0.0389	0.0887	0.0511	0.0396
mean	0.2196	0.2051	0.1833	0.2309	0.1937	0.1809
standard deviation	0.0895	0.1046	0.0829	0.1163	0.1008	0.0975

Cross-validation model comparing $\tilde{\pi}_\theta$, $\tilde{\pi}_\theta^{\varepsilon^\star}$ and $\tilde{\pi}_\theta^{ml}$

To assess the predictive power of the three estimators $\tilde{\pi}_\theta$, $\tilde{\pi}_\theta^{\varepsilon^\star}$ and $\tilde{\pi}_\theta^{ml}$, we performed *Leave-on-out-cross-validation*. It consists in removing at each time data relative to one profile, and computing the three estimators using data for the remaining 18 profiles. Estimated and observed frequencies of grades for the retained profile are then compared. This comparison is performed by computing the distance between the empirical cumulative distribution function (ecdf) of grades and the distribution determined by the three density estimators (see Fig. 4). Two common measures of the difference between two distributions are the *Kolmogorov* and the *Total Variation* distances. The Kolmogorov distance d_{K} is the maximum value of the absolute difference between the two cumulative distributions while Total Variation distance d_{TV} is the sum of all absolute differences [14]. We compute the Kolmogorov and the Total Variation distances for the three estimators. In 13 datasets among the 19 available, $\tilde{\pi}_\theta^{ml}$ predicted distribution that were the closest to the ecdf in the sense of Kolmogorov distance. For Total Variation distance, $\tilde{\pi}_\theta^{ml}$ was the best 8 times out of the 19.

In Tab.1, we can see that $\tilde{\pi}_\theta^{ml}$ has the smallest minimum and the smallest mean of the Kolmogorov and the Total Variation distances over the 19 datasets used in the cross-validation. It has as well as the smallest standard deviation for the two distances. This study shows that $\tilde{\pi}_\theta^{ml}$ is the most efficient estimator in terms of prediction for the problem studied here. This is confirmed by Fig.5 that shows the cumulative curves of d_{K} and d_{TV} for the three estimators.

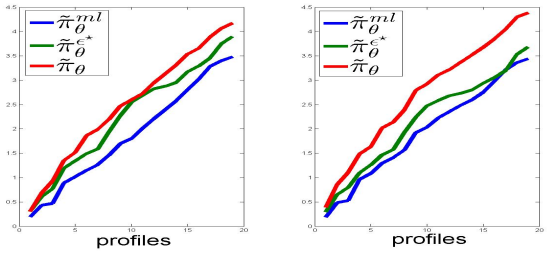

FIGURE 5. Left : The cumulative curves of the 19 distances in the sens of d_{K} corresponding to $\tilde{\pi}_\theta$, $\tilde{\pi}_\theta^{\varepsilon^\star}$ and $\tilde{\pi}_\theta^{ml}$. Right : The same curves in the sens of d_{TV}.

SUMMARY

The paper studied identification of a probability density from region-censored observations, with application to modeling of decompression sickness during hyperbaric diving. Expressing counts of the censored observations as empirical means of a set of binary features, we derive the *maxent* solution that best approximates the empirical distributions. The degree of fitting to the observed frequencies is chosen by selecting the *maxent* solution that has largest likelihood. The tests conducted show that the proposed most likely Rényi-*maxent* estimator has superior behaviour compared to the simpler relaxed-constraints *maxent*, being able to approximate the observed dataset while at the same time being compatible with description of a natural population.

REFERENCES

1. J. Hugon, *Vers une modélisation biophysique de la décompression*, Ph.D. thesis, Aix Marseille 2 (2010).
2. B. Turnbull, *Journal of the Royal Statistical Society. Series B (Methodological)* pp. 290–295 (1976).
3. R. Gentleman, and A. Vandal, *Journal of Computational and Graphical Statistics* **10**, 403–421 (2001).
4. X. Liu, *Nonparametric estimation with censored data: a discrete approach*, Ph.D. thesis, McGill University, Montreal (2005).
5. B. G. Lindsay, "Properties of the maximum likelihood estimator of a mixing distribution," in *Statistical Distributions in Scientific Work*, Springer, 1981, pp. 95–109.
6. D. Bohning, et al., *The Annals of Statistics* **10**, 1006–1008 (1982).
7. N. P. Jewell, *The Annals of Statistics* pp. 479–484 (1982).
8. S. Silvey, D. Titterington, and B. Torsney, *Communications in Statistics-Theory and Methods* **7**, 1379–1389 (1978).
9. B. Torsney, "A moment inequality and monotonicity of an algorithm," in *Semi-Infinite Programming and Applications*, Springer, 1983, pp. 249–260.
10. R. Harman, and L. Pronzato, *Statistics & probability letters* **77**, 90–94 (2007).
11. A. Renyi, *Fourth Berkeley Symposium on Mathematical Statistics and Probability* pp. 547–561 (1961).
12. M. Dudik, M. Phillips, and R. Schapire, *Proc. 7*th *Annual Conf. on Comp. Learning Theory* (2004).
13. M. Dudik, *Maximum entropy density estimation and modeling geographic distributions of species*, Ph.D. thesis, Princeton University, Department of Computer Science (2007).
14. V. Strassen, *The Annals of Mathematical Statistics* pp. 423–439 (1965).