Kriging for non-parametric ML estimation from region-censored data

Y. Bennani
Université de Nice Sophia Antipolis

Affiliation: *Laboratoire I3S*

Email: `"bennani@i3s.unice.frBennani"<bennani@i3s.unice.fr>` – **URL**: `http://www.i3s.unice.fr/I3S/labos/labo4.html`

Master: M2 Modélisation Aléatoire, Université Paris 7

Ph.D. (2012-2015): Université de Nice Sophia Antipolis

Supervisor(s): Dr. Luc Pronzato (UNS-CNRS) and Dr. Maria-João Rendas (UNS-CNRS)

Abstract:

We consider a non-parametric density estimation problem with region-censored observations. The study is motivated by prevention of decompression sickness accidents through prediction of the amount of nitrogen bubbles produced during deep-sea diving. It relies on measurements of bubble grades – which reflect the peak gas volume in the diver's body – on a set of dives made by individuals in the population under analysis.

It is assumed that the instantaneous volume of gas $B(\theta, P(\cdot), t)$ flowing through the right-ventricule of a diver characterised by a set of biophysical parameters θ when executing dive profile $P(\cdot)$ (a function of time) is well described by a known mathematical model [1]. Bubble grades, $G \in \{0, \dots, 4\}$, are a strongly quantified version of peak gas volume (see Fig. 1):

$$G(\theta, P(.)) = i \Leftrightarrow \tau_i \leq \max_t (B(\theta, P(\cdot), t)\} < \tau_{i+1},$$

where $\tau_0 = 0 < \tau_1 < \cdots < \tau_4 < \tau_5 = \infty$ is a set of thresholds assumed known.

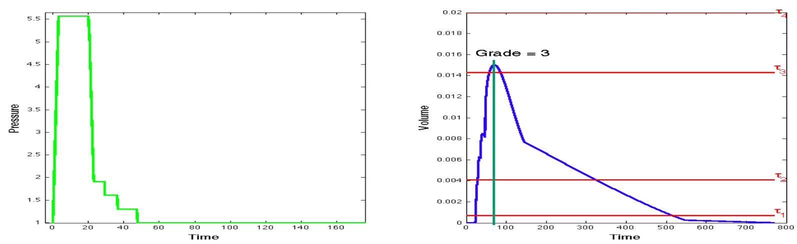

Figure 1: Left: dive profile $P(\cdot)$. Right: model response (blue) and grade computation (threshold are shown in red). Observed grade is equal to 3 in this case.

We address determination of $\hat{\pi}_\theta, \theta \in \Theta$, the non-parametric Maximum Likelihood estimate (NPMLE) of the distribution of θ in the population under study. Observation of grade $G = i$ when executing $P(\cdot)$ only indicates that $\theta \in \mathcal{R}_i(P(\cdot)) = \{\theta \in \Theta \ : \ \tau_i \leq \max_t (B(\theta, P(\cdot), t)\} < \tau_{i+1}\}$ and thus we face a problem of density estimation from (region-)censored observations. For interval-censored observations it is known [3, 4] that $\hat{\pi}_\theta$ is affected by several forms of indeterminacy, the major being that only the probability mass over the (finitely many) elements of a partition $\mathcal{P}$ of the parameter space, determined by the set $\mathcal{R}$ of observed regions $\mathcal{R}_i(P(\cdot))$, can be estimated. Moreover, $\hat{\pi}_\theta$ is concentrated on a subset of the elements of $\mathcal{P}$, that can be found from the intersection graph of $\mathcal{R}$, see Fig. 2. These features carry over to censoring by regions of arbitrary geometry, as in our case, requiring only a slightly more complex determination of the support of $\hat{\pi}_\theta$.

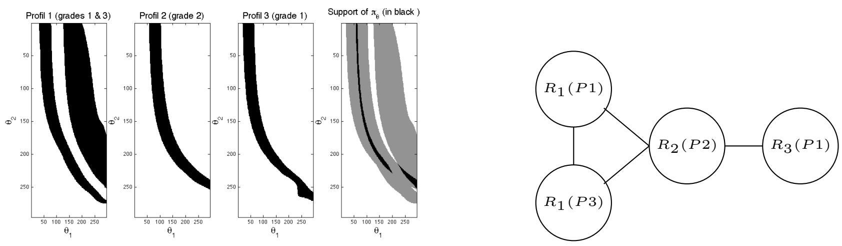

Figure 2: Definition of $\mathcal{P}$ for the observation of grades 1 and 3 for profile P_1, grade 2 for P_2 and grade 1 for P_3 ($\theta \in \Theta \subset R^2$). Left: observed regions and partition $\mathcal{P}$ (the support of the $\hat{\pi}_\theta$ is indicated in black). Right: intersection graph of $\mathcal{R}$ (support of $\hat{\pi}_\theta$ is determined by its cliques).

Central to the determination of the NPMLE is the identification of the regions $\mathcal{R}_i(P(\cdot))$, that must resort to numerical methods, requiring computation of the model response to $P(\cdot)$ over a dense grid covering Θ. In our case, we have 444 measures of grades along 48 different decompression profiles, rendering impractical direct use of the biophysical model. To overcome this problem, we rely on a set of kriged observation models, that predict the value of $\max_t B(\theta, P(\cdot), t)$, from the model response over a sparse (11×11) grid. The response surface was estimated by simple kriging for an isotropic Mattern kernel using the package STK [2].

We present the NPML estimate of the probability mass of π_θ over the elements of $\mathcal{P}$, which is based on a fast multiplicative algorithm [5]. We illustrate the pathological behaviour of this estimator, in particular its sensitivity to the detailed geometry of $\mathcal{P}$, and propose alternative (regularised) solutions that account for the entropy of the estimated distribution.

Our ultimate goal is to predict the distribution of grades for an arbitrary profile $P(\cdot)$: $\hat{p}_P(i) = \hat{\pi}_\theta(\mathcal{R}_i(P(\cdot))) = \sum_{A\in\mathcal{P}} \hat{\pi}_\theta(A \cap \mathcal{R}_i(P(\cdot)))$. These estimates are affected by two distinct uncertainties: *(a)* we do not know $\hat{\pi}_\theta(A \cap \mathcal{R}_i(P(\cdot)))$ since $\mathcal{R}_i(P(\cdot)) \notin \mathcal{P}$ for new profiles; *(b)* the identification of $\mathcal{R}_i(P(\cdot))$ relies on kriging and is thus uncertain. We present upper and lower bounds on each $\hat{p}_P(i)$ that take into account *(a)*. Assessment of uncertainty source *(b)* concerns the determination of level sets based on kriging, as approached e.g. in [6] using the notions of Vorob'ev expectation and deviation and will be considered in the near future.

References

[1] Julien Hugon. *Vers une modélisation biophysique de la décompression*. PhD Dissertation, University of Aix-Marseille 2, 2010.

[2] Small Tool for Kriging (STK), `http://sourceforge.net/projects/kriging/`

[3] Bruce W. Turnbull. *The empirical distribution function with arbitrarily grouped, censored and truncated data*. Journal of the Royal Statistical Society. Series B (Methodological), 1976, 290–295.

[4] Robert Gentleman and Alain C. Vandal. *Nonparametric estimation of the bivariate CDF for arbitrarily censored data*. Canadian Journal of Statistics, Vol.30, No. 4, 2002, 557–571.

[5] Radoslav Harman and Luc Pronzato. *Improvements on removing nonoptimal support points in D-optimum design algorithms*. Statistics & probability letters, Vol. 77, No. 1, 2007, 90–94.

[6] Clément Chevalier, David Ginsbourger, Julien Bect, Ilya Molchanov *Estimating and quantifying uncertainties on level sets using the Vorob'ev expectation and deviance with Gaussian process models*. In proceeding of: mODa 10, Advances in Model-Oriented Design and Analysis, Physica-Verlag HD, 35-43.

Short biography – After obtaining an engineer diploma (2008) and the master degree (2009), Y. Bennani worked in banking before starting a PhD thesis at Laboratory I3S (2012) on the estimation of the risk of decompression accidents among deep-sea divers. His thesis is conducted in the framework of contract DGA-DGCIS SAFE DIVE, a joint partnership between the company BF-Systèmes, Institut Langevin (ESPCI Paristech), and the laboratory I3S (UNS-CNRS).