Maximum de vraisemblance non-paramétrique avec régions de censure. Application à un modèle biophysique de décompression[1]

Youssef Bennani & Luc Pronzato & Maria João Rendas

Laboratoire I3S, UNS-CNRS, Bât. Euclide, Les Algorithmes, 2000 route des Lucioles, 06903 Sophia Antipolis cedex
{*bennani, pronzato, rendas*}*@i3s.unice.fr*

Résumé. Nous considérons un problème d'estimation Non Paramétrique de densité au sens du Maximum de Vraisemblance (NPMV) avec données censurées, pour un modèle biophysique de formation de bulles en plongée sous-marine hyperbare. L'objectif ultime du projet porte sur la prédiction du volume de bulles dégagé pour un profil de plongée donné, afin de réduire les risques d'accident de décompression. Les observations (grades KM) correspondent au comptage quantifié du nombre de bulles circulant dans le sang (ventricule cardiaque droit), pour un ensemble de plongeurs ayant exploré des profils de plongées différents, chaque plongeur possédant ses propres paramètres biophysiques. Nous supposons connu le lien entre grades observés et volume de gaz dégagé, et estimons par maximum de vraisemblance la distribution des paramètres dans la population de plongeurs considérée. La quantification des données de comptage induit une censure des données, d'où des régions d'ambiguïté dans l'espace paramétrique, ici de formes complexes. Nous montrons que l'estimateur NPMV de la densité concentre sa masse dans quelques régions seulement, ce qui rend la méthode inadaptée à la prédiction de grades KM pour des profils de plongée absents du jeu de données initial. Différentes approches, reposant sur une régularisation par maximisation d'entropie, sont considérées afin d'obtenir une distribution plus dispersée.

Mots-clés. Maximum de vraisemblance, données censurées, maximum d'entropie.

Abstract. We consider a non-parametric Maximum Likelihood (ML) density estimation problem with censored observations for a biophysical model describing the production of nitrogen bubbles during deep-sea diving. The ultimate objective is to predict the bubble production associated with a diving profile, in order to prevent decompression sickness accidents. Observations correspond to quantized counts of bubbles circulating in the blood (right ventricule), called grades, for a series of divers that used various diving profiles, each diver having an individual parameter value for the biophysical model. Assuming that the relation between observed grades and volume of gaz produced is known, we estimate the distribution of the model parameters in the population of divers considered by maximum

[1]Travail partiellement financé par le projet DGA–RAPID "SAFE DIVE"

likelihood. Quantized counts induce data censoring and yield ambiguity regions in the parameter space of rather complicated forms. We show that the ML estimator of the density concentrates its mass on a few regions only, which makes the method inadequate for the prediction of grades associated with diving profiles not in the initial data set. Different approaches, based on entropy regularization, are considered to obtain a more dispersed distribution.

Keywords. Maximum likelihood estimation, censored data, maximum entropy.

1 Introduction

Notre étude repose sur l'observation de grades sur une série de plongées hyperbares utilisant des profils de compression/décompression connus. Un grade, compris entre 0 et 4, donne une mesure de la sévérité de la production de bulles circulant dans le sang du plongeur en phase de décompression, et nous admettrons connu le lien entre grades et volume maximum de gaz libéré. Nous utilisons un modèle paramétrique biophysique liant la production instantanée de gaz (micro bulles circulantes) au profil de plongée $P(t)$ (profondeur fonction du temps), soit

$$(\theta, P(\cdot)) \to B(\theta, P(\cdot), \cdot)$$

pour un jeu de paramètres θ ; voir (Hugon, 2010). Nous utiliserons ici $\theta = (\theta_1, \theta_2) \in \Theta \subset \mathbb{R}^2$, un vecteur de deux paramètres seulement, les autres variables du modèle étant fixées à des valeurs nominales. On notera par $b(\theta, P)$ le maximum du signal $B(\theta, P(\cdot), \cdot)$ au cours du temps, c'est-à-dire

$$b(\theta, P) = \max_t B(\theta, P(\cdot), t).$$

La sévérité de la production du bulles est mesurée par des grades, correspondant à des seuils $\tau = \{\tau_l\}_{l=1}^L$ supposés connus pour la variable $b(\theta, P)$, avec $\tau_0 = 0 < \tau_1 < \cdots < \tau_L < \tau_{L+1} = \infty$ (et $L = 4$). Le grade G associé à une valeur $b(\theta, P)$ satisfait

$$G(\theta, P(.)) = l \Leftrightarrow \tau_l \leq b(\theta, P) < \tau_{l+1},$$

voir la figure 1. Connaissant les grades observés sur n plongées effectuées par des plongeurs supposés tirés au hasard (tirages indépendants) dans une population donnée, l'objectif est d'estimer la distribution des paramètres θ au sein de cette population.

Soient $\mathbf{P}_n = \{P_i\}_{i=1}^n$ les n profils (non nécessairement tous distincts) utilisées au cours des n plongées et $\mathbf{G}_n = \{G_i\}_{i=1}^n$ les grades observés correspondants. On suppose que la plongée i est réalisée par un plongeur de paramètres θ_i tirés au hasard, avec $\pi(\cdot)$ la distribution de θ au sein de la population de plongeurs considérée. L'observation du grade G_i pour la plongée i indique seulement que $\theta_i \in \mathcal{R}(G_i, P_i) = \{\theta \in \Theta \ : \ \tau_{G_i} \leq$

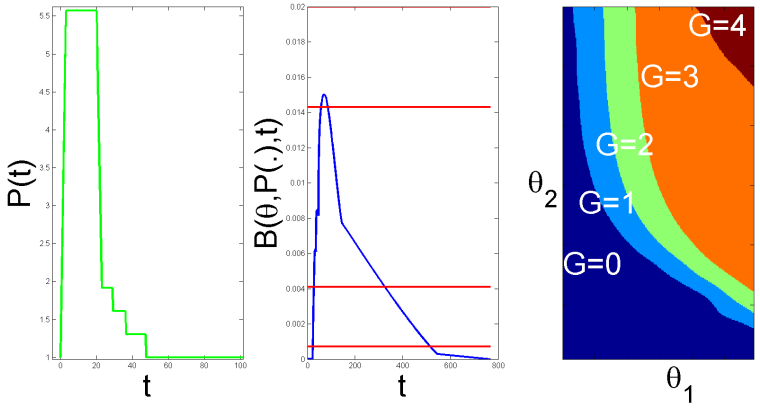

Figure 1: Gauche: Profil de plongée $P(t)$. Centre: réponse du modèle (en bleu) et seuils τ_l (lignes rouges horizontales), avec ici $G = 3$. Droite: régions de l'espace paramétrique correspondant aux 5 grades possibles.

$b(\theta, P_i)\} < \tau_{\{G_i+1\}}\}$. Nous sommes donc confrontés à un problème d'estimation de distribution de probabilité avec observations censurées, les région de censure pouvant être de forme arbitraire. Nous nous intéressons à l'estimateur non-paramétrique du Maximum de Vraisemblance (NPMV) $\hat{\pi}_{ML}$ de π, qui maximise la vraisemblance de $\mathbf{G}_n$, donnée par $\mathcal{L}(\pi; \mathbf{G}_n, \mathbf{P}_n) = \prod_{i=1}^n \mathbb{P}\left[G_i | P_i, \pi\right] = \prod_{i=1}^n \pi(\mathcal{R}(G_i, P_i))$.

La loi $\hat{\pi}_{ML}$ n'est pas unique, et parmi toutes les densités maximisant $\mathcal{L}(\pi; \mathbf{G}_n, \mathbf{P}_n)$ nous choisissons celle d'entropie maximale, que nous appellerons $\hat{\pi}_{MLME}$. Nous montrons que $\hat{\pi}_{MLME}$ tend à concentrer sa masse sur un petit domaine de l'espace paramétrique (§ 2 et 3), ce qui la rend inadaptée à une utilisation à des fins de prédiction (en particulier, de risque d'accident). D'autres régularisations, toujours à base de maximisation d'entropie sont alors proposées (§ 4).

2 Support de l'estimateur NPMV

Le cas de régions de censure correspondant à des intervalles est étudié par Turnbull (1976). Le support de la loi estimée par NPMV se réduit alors à quelques intervalles disjoints : seuls ceux représentés par des cliques maximales dans le graphe d'intersection des observations reçoivent une masse non nulle. Gentleman et Vandal (2001) montrent que ce résultat reste valable pour des intervalles multivariés (c'est-à-dire des parallélépipèdes en dimension d) ; plusieurs algorithmes ont été proposés pour construire les cliques maximales dans ce cas (Gentleman et Vandal, 2001 ; Maathuis, 2003 ; Tomita *et al.*, 2004 ; Liu, 2005). Notons que l'estimation par NPMV conduit à deux types d'ambiguïté :

(*i*) la répartition des masses à l'intérieur des intervalles associés à des cliques maximales est indifférente ;

(*ii*) le vecteur des masses totales des intervalles correspondant à l'estimateur NPMV n'est pas toujours unique.

Quand les régions de censure sont des intervalles, l'intersection deux-à-deux d'un ensemble d'intervalles implique l'intersection complète de l'ensemble. Ceci n'est pas le cas en revanche pour des régions de forme arbitraire, et on peut montrer que le support de l'estimateur NPMV correspond alors à un ensemble de cliques non nécessairement maximales. Ainsi, le cas de régions de censure de forme arbitraire est plus délicat, mais les mêmes ambiguïtés demeurent. Nous proposons une extension de l'algorithme de construction présenté dans (Tomita *et al.*, 2004) pour la génération des cliques maximales.

Notons $\{\Theta_k\}_{k=1}^K$ les intersections de régions $\mathcal{R}(G_i, P_i)$ qui sont susceptibles de recevoir une masse non nulle $w_k = \pi(\Theta_k)$ par l'estimateur NPMV, la vraisemblance s'écrit alors

$$\mathcal{L}(\pi; \mathbf{G}_n, \mathbf{P}_n) \;=\; \prod_{i=1}^{n} \pi(\mathcal{R}(G_i, P_i)) = \prod_{i=1}^{n} \left(\sum_{\Theta_k \subset \mathcal{R}(G_i, P_i)} \pi(\Theta_k) \right).$$

Nous n'avons accès qu'aux masses $w_k = \pi(\Theta_k)$, voir le point (i) ci-dessus, et nous noterons $\mathbf{w}$ le vecteur des masses $(w_1, \ldots, w_K)^\mathsf{T}$, qui appartient au simplexe de probabilité K-dimensionnel $\mathbb{S}^K = \{\mathbf{w} \in \mathbb{R}^K : w_k \geq 0 \text{ et } \sum_k w_k = 1\}$, et $\mathbf{A}$ la matrice $n \times K$ définie par $\{\mathbf{A}\}_{ik} = 1_{\{\Theta_k \subset \mathcal{R}(G_i, P_i)\}}$. La log-vraisemblance (normalisée) de π s'écrit alors $\mathscr{L}(\mathbf{w}; \mathbf{G}_n, \mathbf{P}_n) = (1/n) \sum_{i=1}^n \log \mathbf{A}_{\mathbf{i}\cdot} \mathbf{w}$, avec $\mathbf{A}_{\mathbf{i}\cdot}$ la i-ème ligne de $\mathbf{A}$. Notons que $\mathbf{A}$ a ses K colonnes distinctes de par la définition des $\{\Theta_k\}_{k=1}^K$.

3 Maximisation de la vraisemblance

De nombreux algorithmes ont été proposés dans la littérature, nous utilisons ici l'équivalence entre l'estimation du maximum de vraisemblance pour un problème de mélange et la construction d'un plan d'expérience D-optimal, voir par exemple (Lindsay, 1983 ; Mallet, 1986 ; Böhning, 1989). Ceci nous permet d'exploiter la propriété démontrée dans (Harman et Pronzato, 2007) qui permet de supprimer au cours de l'optimisation des points θ_k dont on est sûr que leur masse doit être nulle pour l'estimateur NPMV.

Supposons que les n plongées conduisent à $m \leq n$ régions distinctes $\mathcal{R}(G_i, P_i)$ (plusieurs profils P_i pouvant être identiques), notées $\{R_j\}_{j=1}^m$. Soit n_j le nombre d'observations correspondant à la région R_j, avec $\sum_{j=1}^m n_j = n$; notons $\mathbf{f} = \{f_j\}_{j=1}^m$ avec $f_j = n_j/n$. On peut alors écrire $\mathscr{L}(\mathbf{w}; \mathbf{G}_n, \mathbf{P}_n) = \sum_{j=1}^m f_j \log \mathbf{B}_{j\cdot} \mathbf{w}$, avec $\mathbf{B}$ la matrice $(m \times K)$ donnée par $\mathbf{B}_{jk} = 1_{\{\theta_k \in R_j\}}$. Toutes les lignes et colonnes de $\mathbf{B}$ sont distinctes. Nous utilisons un algorithme multiplicatif, avec des itérations de la forme

$$w_k^{(t+1)} = \left(\sum_{j=1}^{m} f_j \frac{\mathbf{B}_{jk}}{\mathbf{B}_{j\cdot} \mathbf{w}^{(t)}} \right) w_k^{(t)},$$

initialisé en un $\mathbf{w}^{(0)}$ de $\mathbb{S}^K$ tel que $w_k^{(0)} > 0$ pour tout k, voir (Silvey *et al.*, 1978 ; Torsney, 1983), et forçons $w_k^{(t+1)}$ à zéro quand la condition de (Harman et Pronzato, 2007) est satisfaite. La concavité de la log-vraisemblance permet de donner une règle d'arrêt liée à l'écart entre $\mathscr{L}(\mathbf{w}^t; \mathbf{G}_n, \mathbf{P}_n)$ et $\mathscr{L}_n^* = \max_{\mathbf{w}\in\mathbb{S}^K} \mathscr{L}(\mathbf{w}; \mathbf{G}_n, \mathbf{P}_n)$.

4 Régularisation

Pour faire face à l'ambiguïté (i) du § 2, nous considérerons toujours la loi d'entropie maximale associée à un vecteur $\mathbf{w}$, qui distribue la masse w_i uniformément sur chaque domaine concerné. L'ambiguïté (ii) est également parfois présente, et nous construisons dans ce cas la loi $\hat{\pi}_{MLME}$ d'entropie maximale parmi toutes celles qui maximisent la vraisemblance. Soit $\mathbf{w}_{MLME}$ le vecteur de poids associé. Il satisfait $\mathscr{L}(\mathbf{w}_{MLME}; \mathbf{G}_n, \mathbf{P}_n) = \mathscr{L}_n^* = \mathscr{L}(\mathbf{w}_{ML}; \mathbf{G}_n, \mathbf{P}_n)$, avec $\mathbf{w}_{ML}$ une solution quelconque au sens du maximum de vraisemblance. Du fait de la stricte concavité de la fonction logarithme, ceci est équivalent à $\mathbf{B}_{j\cdot}\mathbf{w}_{MLME} = \mathbf{B}_{j\cdot}\mathbf{w}_{ML}$ pour tout $j = 1, \ldots, m$, et $\mathbf{w}_{MLME}$ (qui appartient au simplexe $\mathbb{S}^K$) est donc contraint par des égalités et inégalités linéaires. Le choix de l'entropie de Rényi d'ordre 2 $H_2(\pi)$ permet de déterminer $\mathbf{w}_{MLME}$ en résolvant un problème de programmation quadratique. En effet,

$$H_2(\pi) = -\log \int_\Theta \pi^2(\theta)\,\mathrm{d}\theta = -\log \sum_{k=1}^K \int_{\Theta_k} \pi^2(\theta)\,\mathrm{d}\theta = -\log \sum_{k=1}^K \frac{w_k^2}{\mathrm{vol}(\Theta_k)}$$

avec $\mathbf{w}$ le vecteur de poids associé à la densité π et $\mathrm{vol}(\Theta_k)$ le volume du domaine Θ_k, qui est connu (peut être calculé numériquement) quand les G_i et P_i, et donc les régions associées, sont connus. Par conséquent, $\mathbf{w}_{MLME}$ d'entropie de Rényi d'ordre 2 maximale maximise $\sum_{k=1}^K w_k^2/\mathrm{vol}(\Theta_k)$ sous des contraintes linéaires.

Bien que d'entropie maximale, $\hat{\pi}_{MLME}$ tend à concentrer sa masse sur quelques régions Θ_k seulement. Ceci rend les prédictions de grades pour un nouveau profil de plongée P_* très instables vis-à-vis de la position des régions $\mathcal{R}(G_i, P_*)$ par rapport aux Θ_k et nous avons considéré une méthode de régularisation différente, permettant de s'éloigner des singularités de l'estimateur NPMV. Nous posons une loi *a priori* de Dirichlet sur les f_j, qui revient à supposer que chaque grade $G = 0, \ldots, 4$ a été observé un certain nombre de fois pour chaque profil de plongée considéré. Notons Σ_n la covariance de la loi a posteriori des f_j. On peut alors construire la distribution π d'entropie $H_2(\pi)$ maximale telle que $\|\Sigma_n^{-1/2}(\mathbf{B}\mathbf{w} - \mathbf{f})\|_\infty \leq \epsilon$ pour un ϵ donné, suffisamment grand pour que ce nouveau problème de programmation quadratique ait une solution en $\mathbf{w}$.

La construction sera illustrée par une application à des données réelles recueillies sur une série de 444 plongées suivant 48 profils différents (données fournies par la société BF-Systèmes dans le cadre du projet DGA–RAPID "SAFE DIVE").

References

[1] D. Böhning. Likelihood inference for mixtures: geometrical and other constructions of monotone step-length algorithms. *Biometrika*, 76(2):375–383, 1989.

[2] R. Gentleman and A.C. Vandal. Computational algorithms for censored-data problems using intersection graphs. *Journal of Computational and Graphical Statistics*, 10(3):403–421, 2001.

[3] R. Harman and L. Pronzato. Improvements on removing nonoptimal support points in d-optimum design algorithms. *Statistics & Probability Letters*, 77(1):90–94, 2007.

[4] J. Hugon. *Vers une modélisation biophysique de la décompression*. PhD thesis, Aix Marseille 2, 2010.

[5] B.G. Lindsay. The geometry of mixture likelihoods: a general theory. *The Annals of Statistics*, 11(1):86–94, 1983.

[6] X. Liu. *Nonparametric estimation with censored data: a discrete approach*. PhD thesis, McGill University, Montreal, 2005.

[7] M.H. Maathuis. Nonparametric maximum likelihood estimation for bivariate censored data. *Master's thesis, Delft University of Technology, The Netherlands. Available at http://www. stat. washington. edu/marloes/papers*, 2003.

[8] A. Mallet. A maximum likelihood estimation method for random coefficient regression models. *Biometrika*, 73(3):645–656, 1986.

[9] S.D. Silvey, D.H. Titterington, and B. Torsney. An algorithm for optimal designs on a design space. *Communications in Statistics – Theory and Methods*, 7(14):1379–1389, 1978.

[10] E. Tomita, A. Tanaka, and H. Takahashi. The worst-case time complexity for generating all maximal cliques. In *Computing and Combinatorics*, pages 161–170. Springer, 2004.

[11] B. Torsney. A moment inequality and monotonicity of an algorithm. In *Semi-Infinite Programming and Applications*, pages 249–260. Springer, 1983.

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