On nonparametric maximum likelihood for a class of stochastic inverse problems

Djalil Chafai, Jean-Michel Loubes

On nonparametric maximum likelihood for a class of stochastic inverse problems

Djalil Chafaï & Jean-Michel Loubes

Preprint – November 2004

Abstract

We establish the consistency of a nonparametric maximum likelihood estimator for a class of stochastic inverse problems. We proceed by embedding the framework into the general settings of early results of Pfanzagl related to mixtures [23, 24].

Keywords: Inverse Problems; Nonlinear Models; Maximum Likelihood; EM Algorithm; Mixtures of Probability Measures; Repeated Measurements Data; Longitudinal Data.
Subject Classification MSC-2000: 62G05; 34K29.

Introduction

Let $(S_i, T_i)_{i\in\mathbb{N}^*}$ be a sequence of i.i.d. random variables with values in $\mathbb{R}^p \times \mathbb{R}^n_+$ and with common law $\mu_S \otimes \mu_T$. Let $(\varepsilon_i)_{i\in\mathbb{N}^*}$ be a sequence of i.i.d. standard normal random variables on $\mathbb{R}^n$, independent of the preceding sequence. We consider in the sequel the inverse problem which consists in estimating the law μ_S given the finite sequence $(Y_i, T_i)_{1\leqslant i\leqslant N}$ where

$$Y_i := f(S_i, T_i) + \sigma\,\varepsilon_i, \tag{1}$$

and where $f : \mathbb{R}^p \times \mathbb{R}^n \to \mathbb{R}^n$ is a known smooth function, which can be in particular nonlinear in the first variable. The asymptotic is taken in N, and n remains fixed. It is assumed that σ is some known non-negative variance parameter. We emphasise the fact that in the triplet (Y_i, T_i, S_i), we observe only the couple (Y_i, T_i), and we are interested in the estimation of the joint law of the unobserved random variables S_i.

In the sequel, $\mathcal{L}(Z)$ denotes the law of the random variable Z. For example, one has $\mathcal{L}(S_i, T_i) = \mu_S \otimes \mu_T$. In the same spirit, $\mathcal{L}(Z_1 \,|\, Z_2)$ denotes the conditional law of Z_1 given Z_2. Finally, we denote by $\mathcal{P}(\mathbb{R}^d)$ the convex set of probability measures

on $\mathbb{R}^d$ equipped with its Borel σ-field and with the $\mathcal{C}_b(\mathbb{R}^d, \mathbb{R})$ dual topology. We will sometimes denote S, T, Y for any random variable with law $\mu_S = \mathcal{L}(S_1)$, $\mu_T = \mathcal{L}(T_1)$, and $\mu_Y = \mathcal{L}(Y_1)$ respectively. Finally, we will denote by $y_i = (y_{i,1}, \dots, y_{i,n})$, $t_i = (t_{i,1}, \dots, t_{i,n})$ and $s_i = (s_{i,1}, \dots, s_{i,p})$ any realisation of the random variables Y_i, T_i and S_i respectively.

Before starting the mathematical analysis of the problem, let us give briefly some explanations regarding the notations and the motivations. The random variables $Y_i = (Y_{i,1}, \dots, Y_{i,n})$ represents the values measured for individual number i at times $T_i = (T_{i,1}, \dots, T_{i,n})$. The random variable S_i stands for the individual parameter and the random variable $\sigma\,\varepsilon_i$ models the (homoscedastic) random noise which is added to the possibly nonlinear true value $f(S_i, T_i)$. This kind of data is known as *repeated measurements*, or called *longitudinal* since each individual (from $i = 1$ to $i = N$) is observed $n_i := n$ times and provides a whole vector of consecutive observations $Y_i = (Y_{i,1}, \dots, Y_{i,n})$ performed at the corresponding individual times $(T_{i,1}, \dots, T_{i,n}) = T_i$. Since T_i is a sequence of measuring times, one can assume for simplicity that the law μ_T is a tensor product of uniform laws on disjoint consecutive compact intervals of the real half line $\mathbb{R}_+$. One can think about μ_T and n as the design of the experiment, whereas f and $\mathcal{L}(S_i \,|\, T_i)$ and $\mathcal{L}(\varepsilon_i)$ correspond to the model chosen for the inverse problem, relating the individual observation Y_i to the individual parameter S_i and to the individual measuring times T_i. Usually in applications, f is of the form

$$f(s, T_i) = (q_s(T_{i,1}), \dots, q_s(T_{i,n})), \tag{2}$$

where for any s in $\mathbb{R}^p$, $q_s : \mathbb{R} \to \mathbb{R}$ is a smooth function depending smoothly on the parameter s, for example a linear combination of time dependent exponentials with coefficients related to s. Function q_s represents in such a scheme the true evolution in time of the phenomenon of interest for an individual of parameter s.

Practical applications of models like (1) are numerous in signal transmission, in tomography, in econometrics, in geophysics, etc, cf. [22]. Let us give briefly a concrete example in Biology. We consider the decay of the concentration of a medicine in human blood. One has $p = 2$ and $q_{(A,\alpha)}(t) = A \exp(-\alpha t)$ in (2), where A stands for the quantity of medicine in the blood at time 0, and where α stands for the rate at which the medicine is eliminated. At the beginning of the experiment, the medicine is given to N independent patients. For patient number i, n measurements $(Y_{i,j})_{1 \leqslant j \leqslant n}$ of the concentration of the medicine in blood are made, at times $(t_{i,j})_{1 \leqslant j \leqslant n}$. One of the simplest model used in this context is

$$Y_{i,j} = q_{(A_i,\alpha_i)}(t_{i,j}) + \sigma \varepsilon_{i,j}, \text{ with } i = 1, \dots, N \text{ and } j = 1, \dots, n.$$

If we state $S_i := (A_i, \alpha_i)$, the random variables $S_1, \dots, S_N$ are i.i.d. and correspond to the biological specificity of each patient. We are interested in the estimation

of the distribution μ_S of the common law of these random variables (population pharmacokinetics). Deconvolution methods are useless since the required condition $n \to +\infty$ is unrealistic. The number of observations n for each individual remains small, a few units in practice. Our framework where the asymptotic is taken on the number of individuals N is the only mean to perform the estimation of the "population law" μ_S.

A stochastic inverse problem is an inverse problem for which the subject of the inversion is a probability measure, like in (1). The related theoretical and applied literature is huge, with many connected components. It contains in particular deconvolution problems, mixtures models, (non)linear mixed effects models, (non)linear filtering problems, etc. Even a common keyword or phrase like our "stochastic inverse problems" is most of the time missing and/or ignored. Therefore, it is quite hard to give a descent state of the art, but a bit less difficult is to show various natures of a particular subclass of problems.

We emphasise the fact that (1) is not a standard regression problem since f is *known* whereas the S_i and their law are *unknown*. Moreover, our problem (1) in not of Ibragimov and Hasminskii type since the S_i are not observed. Notice that when n is very large deconvolution techniques can give an estimation of each S_i. The approach developed recently in [13] in useless for our problem since we consider an asymptotic in N and not in n.

One of the common difficulties of stochastic inverse problems like (1) lies in the fact that they are ill-posed. The inverse of the underlying operator is not continuous in general, so that a small perturbation of the data may induce a large change for the common law of the unobserved random variable. If the unknown was a function in a Hilbert space instead of a probability density function, one could try a singular value decomposition (SVD), following for example Cavalier, Golubev, Picard and Tsybakov in [5].

Several authors have investigated nonparametric maximum likelihood estimation for stochastic inverse problems, and related Expectation Maximisation (EM, cf. [8]) algorithms. In the context of mixtures, Lindsay showed in [14, 15] by using elementary convex analysis that the fully nonparametric maximum likelihood is achieved by a discrete probability measure with finite number of atoms related to the sample size, connecting by this way this kind of problems with convex analysis algorithms (Simplex algorithm, Fedorov methods, etc). One can find some developments in [16, 17, 3, 2]. The consistency of such estimators was established at least by Pfanzagl in [23]. In [25], Schumitzky gave an EM like algorithm for Lindsay's estimator. In another direction, Eggermont and Lariccia have developed smoothing techniques for problems involving Fredholm integral operators, cf. [9] and references therein.

To sum up, our aim in this paper is to estimate μ_S, the common law of the unob-

served i.i.d. random variables S_i in (1), when μ_S belongs to some class $\mathcal{F}_S \subset \mathcal{P}(\mathbb{R}^p)$. The rest of the paper is divided as follows. Section 1 introduces a nonparametric Likelihood Estimator (NPML) for μ_S, and is devoted to establish its consistency up to identifiability. Section 2 presents finite dimensional and algorithmic approaches to approximate the NPML. Finally, in Section 3, various related questions are discussed.

1 An NPML and its consistency

Conditionally on the S_i, the Y_i are independent but not identically distributed, due to the dependence over T_i. However, since the individual observed datum consists in $X_i := (Y_i, T_i)$, it is quite natural to see S_i as the unique unobserved random variable in the triplet (Y_i, S_i, T_i). The law $\mathcal{L}(X_i) = \mathcal{L}(Y_i, T_i)$ is nothing else but

$$\int_{s\in\mathbb{R}^p} \gamma_{\sigma,n}(y - f(s,t))\, d\mu_T(t)\, d\mu_S(s)\, dy,$$

where "$(y,t) = x$" and where $\gamma_{\sigma,n}$ is the Gaussian probability density function on $\mathbb{R}^n$ given by $\gamma_{\sigma,n}(u) := (2\pi\sigma^2)^{-n/2} \exp(-\|x\|_2^2/2\sigma^2)$. Similarly, the law $\mathcal{L}(Y_i)$ of Y_i is the following mixture

$$\left[\int_{s\in\mathbb{R}^p}\int_{t\in\mathbb{R}^n_+} \gamma_{\sigma,n}(y - f(s,t))\, d\mu_T(t)\, d\mu_S(s)\right] dy,$$

where the mixing law is $\mu_S \otimes \mu_T$ and where the mixed family is the following f-deformed Gaussian location family

$$\big\{\gamma_{\sigma,n}(\bullet - f(s,t)) \text{ where } (s,t) \in \mathbb{R}^p \times \mathbb{R}^n\big\} = \gamma_{\sigma,n} * \big\{\delta_{f(s,t)} \text{ where } (s,t) \in \mathbb{R}^p \times \mathbb{R}^n\big\}.$$

Assume now that the law μ_T has a density ψ with respect to the Lebesgue measure on $\mathbb{R}^n_+$. Then, one has that the law $\mathcal{L}(X_i) = \mathcal{L}(Y_i, T_i)$ is absolutely continuous with respect to the Lebesgue measure on $\mathbb{R}^n \times \mathbb{R}^n_+$ with probability density function $\mathbf{K}(\mu_S)$ given by

$$\mathbf{K}(\mu_S)(y,t) := \psi(t) \int_{s\in\mathbb{R}^p} \gamma_{\sigma,n}(y - f(s,t))\, d\mu_S(s). \tag{3}$$

When μ_S has density φ with respect to Lebesgue's measure on $\mathbb{R}^p$, we will denote $\mathbf{K}(\varphi)$ instead of $\mathbf{K}(\mu_S)$, viewing by this way $\mathbf{K}$ as a linear operator over probability density functions.

$$\mathbf{K}(\varphi)(y,t) = \psi(t) \int_{s\in\mathbb{R}^p} \gamma_{\sigma,n}(y - f(s,t))\varphi(s)\, ds.$$

Here again, the law $\mathcal{L}(X_i) = \mathcal{L}(Y_i, T_i)$ is a mixture, with mixing law μ_S and mixed family

$$\{\psi(t)\, \gamma_{\sigma,n}(\bullet - f(s,t)) \text{ with } (s,t) \in \mathbb{R}^p \times \mathbb{R}^n\}.$$

Notice that $\mathbf{K}(\mu_S)(y,t)$ is always positive, and thus, $\log \mathbf{K}(\mu_S)$ always makes sense. The log-likelihood can be expressed by mean of the unknown law μ_S as follows

$$\mathbf{L}_N(\mu_S) := \mathbb{P}_N \log \mathbf{K}(\mu_S), \tag{4}$$

where $\mathbb{P}_N$ is the empirical measure of the sample $(X_i)_{1 \leqslant i \leqslant N}$ defined by

$$\mathbb{P}_N := \frac{1}{N} \sum_{i=1}^{N} \delta_{(Y_i,T_i)}. \tag{5}$$

Notice that we have used above the standard notation $\mathbb{P}_N F$ to denote the expectation of function F with respect to probability law $\mathbb{P}_N$. When f is of the form (2), the log-likelihood $\mathbf{L}_N$ defined above in (4) reads

$$\begin{aligned}
\mathbf{L}_N(\mu_S) &= \frac{1}{N} \sum_{i=1}^{N} \log \int_{s \in \mathbb{R}^p} \exp\left(-\frac{1}{2\sigma^2} \sum_{j=1}^{n} (Y_{i,j} - q_s(T_{i,j}))^2 \right) d\mu_S(s) \quad + C \\
&= \frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{n} \log \int_{s \in \mathbb{R}^p} \exp\left(-\frac{1}{2\sigma^2} (Y_{i,j} - q_s(T_{i,j}))^2 \right) d\mu_S(s) \quad + C,
\end{aligned}$$

where

$$C := -\frac{n}{2} \log(2\pi\sigma^2) + \frac{1}{N} \sum_{i=1}^{N} \log \psi(T_{i,1}, \ldots, T_{i,n}).$$

The quantity C does not have any effect on the arg-maximum of the log-likelihood functional $\mathbf{L}_N$. In particular, the density ψ of μ_T does not play a direct role in the NPML (6) below since one can rewrite $\mathbf{L}_N$ as follows

$$\mathbf{L}_N(\mu_S) = \mathbb{P}_N \log \psi + \mathbb{P}_N \log \mathbf{K}^{\#}(\mu_S),$$

where

$$(\mathbf{K}^{\#}(\mu_S))(y,t) := \int_{s \in \mathbb{R}^p} \gamma_{\sigma,n}(y - f(s,t))\, d\mu_S(s).$$

On any set $\mathcal{F}$, the arg-maximum of $\mathbf{L}_N$ is equal to the arg-maximum of $\mathbf{L}_N^{\#}$ defined by

$$\mathbf{L}_N^{\#}(\mu_S) := \mathbb{P}_N \log \mathbf{K}^{\#}(\mu_S).$$

The functional $\mathbf{L}_N^{\#}$ does not depend on μ_T directly, but only implicitly via the sample $T_1, \ldots, T_N$ throughout $\mathbb{P}_N$. However, the law μ_T plays a role in identifiability, and the good choice of this law is always a crucial issue.

Definition 1.1 (Identifiability). We say that the mixture model (1) is *identifiable* if and only if $\mathbf{K}$ is injective, as a map from $\mathcal{F}_S$ to $\mathcal{P}(\mathbb{R}^n)$. Namely, for any couple $(\mu, \nu) \in \mathcal{F}_S \times \mathcal{F}_S$ with $\nu \neq \mu$, one has $\mathbf{K}(\mu) \neq \mathbf{K}(\nu)$ in $\mathcal{P}(\mathbb{R}^n)$. Similarly, we say that $\mu \in \mathcal{F}_S$ is *identifiable* in (1) if and only if $\mathbf{K}(\nu) \neq \mathbf{K}(\mu)$ in $\mathcal{P}(\mathbb{R}^n)$ for any $\nu \in \mathcal{F}_S$ with $\nu \neq \mu$.

Clearly, the model is identifiable if and only if every element of $\mathcal{F}_S$ is identifiable. Identifiability is essential for any estimation issue of the true mixing law μ_S. This condition is quite difficult to check in great generality. However, one can find some clues for example in [6] and references therein. In practice, and when possible, identifiability must be checked for the particular model considered, and is deeply related to the properties of function f and to the distribution μ_T of the observation times. We are now able to state the following Theorem.

Theorem 1.2 (Consistency of NPML). *Assume that $\mathcal{F}_S \subset \mathcal{P}(\mathbb{R}^p)$ is a compact convex subset of a linear space, that the model is identifiable, that $\mathcal{L}(T) = \psi(t)\,dt$, that $\mu_S \in \mathcal{F}_S$, and that for almost all $(y,t) \in \mathbb{R}^n \times \mathbb{R}_+^n$, the map $\mathbf{K}(\bullet)(y,t) : \mathcal{F}_S \to \mathbb{R}$ is continuous. Then, the NPML estimator $\widehat{\mu_{S,N}}$ given by*

$$\widehat{\mu_{S,N}} := \arg\max_{\mu \in \mathcal{F}_S} \mathbf{L}_N(\mu) \tag{6}$$

is well defined, unique, and converges almost surely toward μ_S when N goes to $+\infty$.

Proof. The random map $\mathbf{L}_N$ is a.s. continuous from $\mathcal{F}_S$ to $\mathbb{R}$ since the map $\mathbf{K}(\bullet)(y,t) : \mathcal{F}_S \to \mathbb{R}$ is continuous for any $(y,t) \in \mathbb{R}^n \times \mathbb{R}_+^n$. By linearity and identifiability of $\mathbf{K}$ and strict concavity of the logarithm, the map $\mathbf{L}_N$ is a.s. strictly concave. Thus, it achieves a.s. a unique sup over the compact convex set $\mathcal{F}_S$. The existence and unicity of the estimator $\widehat{\mu_{S,N}}$ is therefore proved. Finally, thanks to our choice of settings, the desired consistency result follows from [23, Theorem 3.4] and [23, Section 5], since the required hypotheses are fulfilled:

- **Condition 1**. $\mathcal{F}_S$ is a compact Hausdorff space, and a subset of a linear space.
- **Condition 2**. For almost all $(y_i, t_i)_{1 \leqslant i \leqslant N}$, the map $\prod_{i=1}^N \mathbf{K}(\bullet)(y_i, t_i)$ is continuous on $\mathcal{F}_S$ for the topology of $\mathcal{F}_S$.
- **Condition 3**. For almost all $(y,t) \in \mathbb{R}^n \times \mathbb{R}_+^n$, the map $\mathbf{K}(\bullet)(y,t)$ is concave on $\mathcal{F}_S$.

□

Remark 1.3. Let us give various remarks about Theorem 1.2 and its extensions.

1. **Identifiability.** Following again [23], one can relax the identifiability of the model to the identifiability of μ_S, but it is not really useful in practice since μ_S is unknown! For any $x := (y,t) \in \mathbb{R}^n \times \mathbb{R}^n$, let us denote by $k_x : \mathbb{R}^p \to \mathbb{R}_+$ the function $k_x(s) := \gamma_{n,\sigma}(y - f(s,t))$. Let $\mathcal{T}$ be the biggest open subset of $\mathbb{R}^n$ such that $\psi > 0$ over $\mathcal{T}$. Then, identifiability of the model corresponds to a condition on the set of functions $\mathcal{C} := \{k_x : \mathbb{R}^p \to \mathbb{R}_+ \text{ with } x \in \mathbb{R}^n \times \mathcal{T}\}$ appearing in the mixture (3). Namely, it must separate the elements of $\mathcal{F}_S$. In other words, when f is smooth, the $\mathcal{C}$ class must be large enough to fully characterise any element of $\mathcal{F}_S$ by duality as a set of test functions for a distribution of order zero in the sense of L. Schwartz distributions Theory. Such a necessary and sufficient separation condition relies on both f and $\mathcal{T}$ and can, depending on the particular choice of $\mathcal{F}$, be weaker than the full injectivity of f in the first variable when the second runs over $\mathcal{T}$. Notice that the smoothness of f together with its injectivity in the first variable induces in general a "degree of freedom" requirement on (n,p). If $\mathcal{F}_S \subset \mathcal{D}'(K)$ for some compact subset K of $\mathbb{R}^p$, then $\mathcal{C}$ separates the elements of μ_S as soon as the vector space spanned by $\mathcal{C}$ is dense in $\mathcal{C}^\infty(K)$ for the uniform topology.

2. **Heteroscedasticity.** At least when the elements of $\mathcal{F}_S$ are compactly supported, Theorem 1.2 remains true for a class of heteroscedastic models of the form
$$Y_i = f(S_i, T_i) + \sigma\, \varepsilon_i + g(S_i, T_i) \cdot \varepsilon_i, \tag{7}$$
where $\sigma > 0$ is known, where $g : \mathbb{R}^p \times \mathbb{R}^n \to \mathbb{R}_+^n$ is a smooth function and where the dot mark $''\cdot''$ denotes the component-wise vectors multiplication. One can also incorporate a matrix between g and ε_i. Notice that condition $g \geqslant 0$ ensures that the variance of the conditional law is bounded below by σ^2 and thus, the mixture makes sense. The mixed family is a location-scale (f,g)-deformed Gaussian family:
$$\left\{\gamma_{(\sigma^2+g(s,t)^2)^{1/2},\,n}(\bullet - f(s,t)) \text{ where } (s,t) \in \mathbb{R}^p \times \mathbb{R}^n\right\}.$$
In concrete applications, it is quite usual to state that g and f are co-linear in the heteroscedastic model above, say $g = \sigma' f$, making the noise roughly proportional to the measured value.

3. **Non Gaussian noise.** Theorem 1.2 remains true when the Gaussian law of the noise ε_i in (1) is replaced by an absolutely continuous law with respect to the Lebesgue measure on $\mathbb{R}^n$. The related location mixed family is not Gaussian in that case, but this does not block the derivation of the consistency of the NPML.

4. **Non homogeneity via censure.** Let $(\mathbf{n}_i)_{i\in\mathbb{N}^*}$ be a sequence of i.i.d. random variables independent of $(S_i, T_i)_{i\in\mathbb{N}^*}$, with values in the set $\mathcal{N}_n$ of subsets of $\{1,\dots,n\}$, and with common law $p_\kappa := \mathbb{P}(\mathbf{n_i} = \kappa) > 0$ for any $\kappa \in \mathcal{N}_n$. Assume that for each i, one has access only to $Z_i := (Y_{i,j}, j \in \mathbf{n}_i)$ instead of the whole vector of measurements $Y_i := (Y_{i,1},\dots,Y_{i,n})$ itself. Then, the new inverse problem corresponds to the new sample

$$((Z_1, T_1, \mathbf{n}_1),\dots,(Z_N, T_N, \mathbf{n}_N))$$

which is the censored version of the original sample with unobserved S_i values

$$((Y_1, S_i, T_1),\dots,(Y_N, S_i, T_N)).$$

The problem is that the Z_i are not in the same space, but are still independent. Our goal then is to rewrite the problem in a i.i.d framework. One method consists in extending the data space to the larger direct sum space $E := \oplus_{\kappa\in\mathcal{N}_n} E_\kappa$, where E_κ is a copy of $\mathbb{R}^{|\kappa|}$ corresponding to the components present in κ, where $|\kappa| := \#\kappa$. It is then easy to write down the law of $(Z_i, T_i, \mathbf{n_i})$. Such a model is quite heavy to write down but gives rise to a simple extended log-likelihood:

$$\mathbf{L}_N(\mu_S) := \mathbb{P}_N \log p_\kappa + \mathbb{P}_N \log\psi + \mathbb{P}_N \log \mathbf{K}_\kappa(\mu_S),$$

where for any $\mu \in \mathcal{F}_S$

$$\mathbf{K}_\kappa(\mu)(z,t,\kappa) := \int_{s\in\mathbb{R}^p} \gamma_{\sigma,|\kappa|}(z - \pi_\kappa(f(s,t)))\, d\mu(s),$$

where π_κ is the projection of E on E_κ and where the empirical measure $\mathbb{P}_N$ is now

$$\mathbb{P}_N := \frac{1}{N}\sum_{i=1}^{N} \delta_{(Z_i,T_i,\mathbf{n}_i)}.$$

The $\mathbb{P}_N \log p_\kappa + \mathbb{P}_N \log\psi$ part of the log-likelihood does not depend on μ_S, and thus, it does not influence the arg-maximum of the log-likelihood and can be safely removed. For each i, the $T_{i,j}$ involved in the log-likelihood are those with $j \in \mathbf{n_i}$. Finally, one can notice that such type of independent censoring does not correspond to all realistic censure, since in practice, the $\mathbf{n}_i$ can depend on the Y_i it self via for example

$$\left(\mathrm{I}_{\{Y_{i,1}>\tau\}},\dots,\mathrm{I}_{\{Y_{i,n}>\tau\}}\right)$$

where τ is a detection threshold.

5. **Continuity of the operator.** The continuity assumption on $\mathbf{K}$ relies in general on function f, on the nature of $\mathcal{F}_S$, and on the law of the noise ε_i, which is Gaussian and homoscedastic here. Some concrete examples of $\mathcal{F}$ are given below.

6. **Full extension.** Mixing all the previous extensions is delicate.

Example 1.4. Consider for instance the set $\mathcal{F}_S \subset \mathcal{P}(\mathbb{R}^p)$ defined by

$$\mathcal{F}_S := \mathcal{F}_S^{M,A} := \big\{\varphi(s)\,ds;\ \text{where } \varphi \in \mathcal{C}^1_K([0,M]) \text{ and } \|\varphi\|_{\mathrm{L}^1} = 1,\ \||\nabla\varphi|\|_\infty \leqslant A\big\}, \tag{8}$$

where K is a fixed compact subset of $\mathbb{R}^p$ and where M, A are fixed non negative real numbers. Equipped with the L^∞ topology, this set is a compact convex subset of a linear space, as required by Theorem 1.2. Since the underlying mixture model is a "Gaussian position" one, we get for any couple $(\varphi_1, \varphi_2) \in \mathcal{F}_S \times \mathcal{F}_S$ and any $(y,t) \in \mathbb{R}^n \times \mathbb{R}^n$

$$|\mathbf{K}(\varphi_1)(y,t) - \mathbf{K}(\varphi_2)(y,t)| \leqslant \|\varphi_1 - \varphi_2\|_\infty \|\psi\|_\infty (2\pi\sigma^2)^{-n/2},$$

which gives the L^∞ continuity of $\mathbf{K}(\bullet)(y,t)$ for any couple $(y,t) \in \mathbb{R}^n \times \mathbb{R}^n$. Since we deal with a "Gaussian position model" (homoscedasticity), the operator norm does not depend on (y,t) and function f plays not role. The L^∞ a.s. consistency up to identifiability of the NPML follows then from Theorem 1.2.

Example 1.5. Consider the set $\mathcal{G}_S \subset \mathcal{P}(\mathbb{R}^p)$ defined by

$$\mathcal{G}_S := \mathcal{G}_S^{A,\alpha} := \{\varphi(s)\,ds;\ \text{where } \varphi \in \mathrm{H}^\alpha(K) \text{ with } \|\varphi\|_{\mathrm{L}^1} = 1 \text{ and } \|\varphi\|_{H^\alpha} \leqslant A\},$$

where K is a fixed compact subset of $\mathbb{R}^p$, A is a fixed non negative real number and $\mathrm{H}^\alpha(K)$ is the Sobolev space over the compact K. Provided that $\alpha > \frac{1}{2} - \frac{1}{p}$, Rellich-Sobolev embedding Theorem yields that $\mathcal{F}_S$ is a compact convex subset of a linear space for the L^2 topology, cf. [1, 19], as required by Theorem 1.2. Since the underlying mixture model is a "Gaussian position" one, we get for any couple $(\varphi_1, \varphi_2) \in \mathcal{F}_S \times \mathcal{F}_S$ and any $(y,t) \in \mathbb{R}^n \times \mathbb{R}^n$

$$|\mathbf{K}(\varphi_1)(y,t) - \mathbf{K}(\varphi_2)(y,t)| \leqslant \|\varphi_1 - \varphi_2\|_2 \|\psi\|_\infty (4\pi\sigma^2)^{-n/4},$$

which gives the L^2 continuity of $\mathbf{K}(\bullet)(y,t)$ for any couple $(y,t) \in \mathbb{R}^n \times \mathbb{R}^n$. Since we deal with a "Gaussian position model" (homoscedasticity), the operator norm does not depend on (y,t) and function f plays not role. The L^2 a.s. consistency up to identifiability of the NPML follows then from Theorem 1.2.

2 Algorithms for the NPML

2.1 Finite dimensional approximation

The first step towards a practical implementation is to transform the maximum $\widehat{\mu_{S,N}}$ of the log-likelihood $\mathbf{L}_N$ over the whole infinite dimensional class $\mathcal{F}_S$ into a maximum $\widehat{\mu_{S,N,m}}$ over a finite dimensional convex subset $\mathcal{F}_{S,m}$, where $(\mathcal{F}_{S,m})_{m\in\mathbb{N}^*}$ is an exhaustive sequence of subsets of $\mathcal{F}_S$, i.e. $\mathbf{adh}(\cup_{m\in\mathbb{N}^*}\mathcal{F}_m) = \mathcal{F}$.

Theorem 2.1. *Assume that $\mathcal{F}_S$ is a metric space. Let $(\mathcal{F}_{S,m})_{m\in\mathbb{N}^*}$ be an exhaustive sequence of finite dimensional closed convex subsets of $\mathcal{F}_S$. Under the assumptions of Theorem 1.2, and for any fixed sample of size N, the approximated NPML estimator $\widehat{\mu_{S,N,m}}$ given by*

$$\widehat{\mu_{S,N,m}} := \arg\max_{\mu\in\mathcal{F}_{S,m}} \mathbf{L}_N(\mu). \tag{9}$$

is well defined, unique, and converges toward the NPML $\widehat{\mu_{S,N}}$ when m goes to $+\infty$.

Proof. We proceed at fixed N. Since $\mathcal{F}_{S,m}$ is a compact convex subset, the approximated NPML estimator $\widehat{\mu_{S,N,m}}$ exists, as it was the case for the NPML estimator $\widehat{\mu_{S,N}}$ in Theorem 1.2. Let us now establish the convergence. By the definition of $\widehat{\mu_{S,N,m}}$ and $\widehat{\mu_{S,N}}$ one has that

$$\mathbf{L}_N(\widehat{\mu_{S,N,m}}) \leqslant \mathbf{L}_N(\widehat{\mu_{S,N}}).$$

In the other hand, there exists a sequence $(\mu_m)_{m\in\mathbb{N}^*}$ converging towards $\widehat{\mu_{S,N}}$ in $\mathcal{F}_S$ and such that $\mu_m \in \mathcal{F}_{S,m}$ for any $m \in \mathbb{N}^*$. Hence, lower semi continuity of $\mathbf{L}_N$ induces that, for any $\varepsilon > 0$, there exists $m_\varepsilon \in \mathbb{N}^*$ such that for any $m \geqslant m_\varepsilon$,

$$\mathbf{L}_N(\widehat{\mu_{S,N}}) - \varepsilon \leqslant \mathbf{L}_N(\mu_m).$$

But by definition of $\widehat{\mu_{S,N,m}}$ we have

$$\mathbf{L}_N(\mu_m) \leqslant \mathbf{L}_N(\widehat{\mu_{S,N,m}}).$$

As a result, the following bound holds for any $\varepsilon > 0$ and any $m > m_\varepsilon$

$$\mathbf{L}_N(\widehat{\mu_{S,N}}) - \varepsilon \leqslant \mathbf{L}_N(\widehat{\mu_{S,N,m}}) \leqslant \mathbf{L}_N(\widehat{\mu_{S,N}}). \tag{10}$$

If $\mu^* \in \mathcal{F}_S$ is an adherence value of the sequence $(\widehat{\mu_{S,N,m}})_{m\in\mathbb{N}^*}$, corresponding to the limit point of a subsequence $(\widehat{\mu_{S,N,m_k}})_{k\in\mathbb{N}^*}$, then $\mu^* = \widehat{\mu_{S,N}}$. Namely, if it was not the case, then (10) will implies that $(\mathbf{L}(\widehat{\mu_{S,N,m_k}}))_{k\in\mathbb{N}^*}$ converges toward $\mathbf{L}_N(\widehat{\mu_{S,N}})$, and thus that $\mathbf{L}_N(\mu^*) = \mathbf{L}_N(\widehat{\mu_{S,N}})$, which contradicts the unicity of $\widehat{\mu_{S,N}}$ as a maximum of $\mathbf{L}_N$ over $\mathcal{F}_S$. Hence, $\widehat{\mu_{S,N}}$ is the unique adherence value of the sequence $(\widehat{\mu_{S,N,m}})_{m\in\mathbb{N}^*}$, and the compacity of $\mathcal{F}_S$ yields finally that $(\widehat{\mu_{S,N,m}})_{m\in\mathbb{N}^*}$ converges towards $\widehat{\mu_{S,N}}$, which is exactly the desired result. □

Remark 2.2. The rate of convergence of $(\widehat{\mu_{S,N,m}})_{m\in\mathbb{N}^*}$ towards $\widehat{\mu_{S,N}}$ when m goes to $+\infty$ depends on the regularity of $\mathcal{F}_{S,m}$ and $\mathbf{L}_N$.

2.2 A Gradient algorithm for log-likelihood maximisation

Since for any $m \in \mathcal{F}_{S,m}$ and any couple (μ, ν) in $\mathcal{F}_{S,m} \times \mathcal{F}_{S,m}$,

$$\mathbf{L}_N(\mu) - \mathbf{L}_N(\nu) = \mathbb{P}_N \log \frac{\mathbf{K}(\mu)}{\mathbf{K}(\nu)},$$

the sieves log-likelihood estimator $\widehat{\mu_{S,N,m}}$ defined in (9) can be viewed as the solution of the following optimisation issue:

$$\text{find } \widehat{\mu_{S,N,m}} \text{ such that } \forall \mu \in \mathcal{F}_{S,m},\ \mathbb{P}_N \log \frac{\mathbf{K}(\mu)}{\mathbf{K}(\widehat{\mu_{S,N,m}})} \leqslant 0. \tag{11}$$

By using the concavity of the objective function, Pfanzagl has proved in [24] that one may switch, in the definition of the estimator in (11), from the log function to any other function $L : \mathbb{R}_+^* \to \mathbb{R}$, provided that it is concave, strictly increasing, with $L(1) = 0$.

$$\text{find } \widehat{\mu_{S,N,m}} \text{ such that } \forall \mu \in \mathcal{F}_{S,m}, \quad \mathbb{P}_N L\left[\frac{\mathbf{K}(\mu)}{\mathbf{K}(\widehat{\mu_{S,N,m}})}\right] \leqslant 0. \tag{12}$$

As a result defining the estimator for a particular L is enough to get inequality (12) for all "contrast" function L satisfying the previous assumptions. In particular, the estimator $\widehat{\mu_{S,N,m}}$ can be obtained for the special choice $L(t) = t - 1$, which corresponds exactly to the definition of the EM algorithm iteration. Hence, maximising the estimator can be practically computed via the EM algorithm, while Theorem 2.1 still applies, proving consistency of the estimator. This invariance in L relies on the "concavity" of the model, as explained in [24].

3 Discussion

3.1 Heuristics for the NPML in Theorem 1.2

As usual for maximum log-likelihood, the strong law of large numbers yields that $(\mathbb{P}_N)_{N \in \mathbb{N}^*}$ converges a.s. toward $\mathbf{K}(\mu_S)$ in $\mathcal{P}(\mathbb{R}^n)$. In other words, $\mathcal{L}(Y) = \mathbf{K}(\mu_S)$. Consequently, for any $\mu \in \mathcal{F}_S$, $(\mathbf{L}_N(\mu))_{N \in \mathbb{N}^*}$ converges toward

$$\mathbf{L}_\infty(\mu) := -\mathbf{Ent}(\mathbf{K}(\mu_S) \,|\, \mathbf{K}(\mu)) + \mathbf{H}(\mathbf{K}(\mu_S)),$$

where $\mathbf{Ent}(\mathbf{K}(\mu_S) \,|\, \mathbf{K}(\mu)) = \int (\log \mathbf{K}(\mu_S) - \log \mathbf{K}(\mu)) \mathbf{K}(\mu_S)$ is the Kullback-Leibler relative entropy of $\mathbf{K}(\mu_S)$ with respect to $\mathbf{K}(\mu)$ and where $\mathbf{H}(\mathbf{K}(\mu_S)) = \mathbf{L}_\infty(\mu_S)$ is the Shannon entropy of $\mathbf{K}(\mu_S)$. In other words, the log-likelihood random functional $\mathbf{L}_N$ converges toward the deterministic functional $\mathbf{L}_\infty$ when N goes to $+\infty$. This deterministic limit $\mathbf{L}_\infty$ is the relative entropy functional $\mathbf{Ent}(\bullet \,|\, \mathbf{K}(\mu_S))$, up to

the additive constant $\mathbf{H}(\mathbf{K}(\mu_S))$ which does not play any role for the arg-maximum problem. Since $\mathbf{K}$ is injective (identifiability), $\mathbf{L}_\infty$ is strictly concave with unique maximum achieved at point μ_S. The NPML estimator replaces the asymptotic arg-maximum μ_S with the finite N arg-maximum $\widehat{\mu_{S,N}}$. The non-asymptotic log-likelihood $\mathbf{L}_N$ is not a relative entropy, but remains strictly concave. The EM algorithm $\mu_{N,k+1} = \mathbf{F}_N(\mu_{N,k})$ consists in approximating $\widehat{\mu_{S,N}}$ by finding an entropic lower bound functional for $\mathbf{L}_N$ which touches $\mathbf{L}_N$ at the current step $\mu_{N,k}$. The EM algorithm in this context can be seen also as a gradient like algorithm $\mu_{N,k+1} = \mu_{N,k} + \mathbf{G}_N(\mu_{N,k})$ for the concave functional $\mathbf{L}_N$, where $\mathbf{G}_N$ is the Gâteau directional derivative of $\mathbf{L}_N$. It turns out that this gradient like approach appears as a fixed point iteration $\mu_{N,k+1} = \mathbf{F}_N(\mu_{N,k})$ where $\mathbf{F}_N = \mathbf{G}_N + \mathrm{Id}$. The fixed point problem $\mathbf{F}_N(\mu) = \mu$ corresponds exactly to Bayes rule where the unknown μ_S is replaced by the current step μ and where $\mathcal{L}(Y) = \mathbf{K}(\mu_S)$ is replaced by the first marginal of $\mathbb{P}_N$. Here again, $(\mathbf{F}_N)_{N\in\mathbb{N}^*}$ converges point-wise toward $\mathbf{F}_\infty$ which admits μ_S as unique fixed point. One of the main feature of EM is the monotonicity of the objective function $\mathbf{L}_N$ along the algorithm. The drawback with such a basic EM approach for nonparametric NPML is the fact that the support is non increasing along the algorithm.

3.2 Destruction the log-likelihood concavity for mixtures models

The log-likelihood of mixtures models is a concave functional of the unknown mixing probability measure. However, this structure is very sensitive. Lindsay has showed in [14] by simply using Minkowski-Caratheodory Theorem that the fully nonparametric NPML for mixtures models like (1) is achieved by an atomic probability measure with at most $N+1$ atoms. By fully nonparametric, we mean that $\mathcal{F}_S = \mathcal{P}(\mathbb{R}^p)$. This observation is enough robust to remain valid for heteroscedastic models as in Remark 1.3. Unfortunately, the parametrisation of such discrete probability measures in terms of weights and support points destroys the concavity of the log-likelihood objective function $\mathbf{L}_N$. This lack of concavity cannot be fixed by the introduction of a stochastic ordering on the set of discrete probability measures with at most $N+1$ atoms.

3.3 Semi-parametric estimation

The convexity structure of the NPML problem is destroyed by the incorporation of fixed effects estimation. This is typically the case for mixed-effects models where a linear model structure is imposed to μ_S and where σ is unknown in (1). In such cases, the global log-likelihood, seen as a functional of both random and fixed effects, is not concave and has potentially many local maxima. The semi-parametric approach developed in [24] is useless since we do not have a consistent estimator of the fixed effects regardless of the random effect.

Recall that a typical *mixed effects model* corresponds to some particular structure (a linear model in general) on the S_i in (1). Namely, $S_i = \Theta V_i + \eta_i$, where V_i is an observed vector of per-individual co-variables (sex, weight, etc), where Θ is an unknown matrix parameter giving the trend (fixed effect), and where η_i is the random effect of unobserved data. In such a model, the $(V_i)_{i\in\mathbb{N}^*}$ and the $(\eta_i)_{i\in\mathbb{N}^*}$ are i.i.d., and the $\{T_i, V_i, \eta_i, \varepsilon_i, \text{ where } i \in \mathbb{N}^*\}$ are mutually independent random variables. The goal is then to estimate the Θ matrix and the common law μ_η of the $(\eta_i)_{i\in\mathbb{N}^*}$. Such models are used for example in Biology to let the measurements take into account the known specificity of each individual while conducting a survey. The pattern, which is determined by physiological rules is given by the function f, while the specificity of each individual is modelled by the random variables $(S_i)_{1\leqslant i\leqslant N}$. If we write $S_i = \Theta V_i + m + \eta_i'$ where m is a fixed parameter to be estimated and where η_i' is a centred random effect, one can first estimate the law of the centred random effect η' and then estimate the fixed effects Θ and m. However, this approach must be adapted when the coefficient σ in (1) is not known, since it appears in that case as a new fixed effect to be estimated. We believe that a semi-parametric extension of our method can be made, providing an estimation of (Θ, μ_η). The approach presented in [24] does not help since we do not have a consistent estimator for the fixed effects. Despite the fact that numerous nonparametric techniques were developed for mixtures models, the widely used approach in applications of nonlinear mixed effects models is quite rough and consists in a fully parametric estimation of the first two moments of the law μ_η of the random effect η, where it is arbitrarily assumed that this law is normal or log-normal, cf. [20, 21] and [7] for example. Even if they speed up the effective computations, such fully parametric approaches are not satisfactory since the consequences it terms of decision are highly sensitive to the arbitrarily chosen structure for the random effect law (not robust).

3.4 No rates

To obtain rates of convergence for the maximum likelihood estimator, we consider a neighbourhood of the true distribution μ_S, defined by the topology chosen according to fulfils the conditions of Theorem 1.2. Write $V(\mu_S)$ this neighbourhood, then using compacity there exist a finite sequence of neighbourhood $V(\mu_k)$, $k = 1, \ldots, r_N$ such that

$$\mathcal{F}_S - V(\mu_S) \subset \cup_{k=1}^{r_N} V(\mu_k).$$

Hence, finding the rate of convergence of nonparametric maximum likelihood estimator implies studying the deviation probability

$$\begin{aligned}\mathbf{P}(\widehat{\mu_{S,N}} \notin V(\mu_S)) &\leqslant \sum_{k=1}^{r_N} \mathbf{P}(\widehat{\mu_{S,N}} \in V(\mu_k)) \\ &\leqslant \sum_{k=1}^{r_N} \mathbf{P}\left(\sup_{\mu \in V(\mu_k)} \frac{1}{N} \sum_{i=1}^{N} \log \left[2 \left(1 + \frac{(\mathbf{K}^{\#}(\mu_S))(X_i)}{(\mathbf{K}^{\#}(\mu)(X_i)} \right)^{-1} \right] \geqslant \log \gamma \right)\end{aligned}$$

for $0 < \gamma < 1$ as it is quoted in [24]. Bounding this deviation inequality requires two main ingredients. First a bound for the entropy of the mixture class. Recent works by van der Vaart, see for instance [10] and [12], give upper bounds for the entropy of such classes and hence provide a control over r_N. Second, to conclude, there is a need for a deviation inequality over the previous empirical process. Unfortunately, to our concern, concentration bounds in this framework are very difficult to obtain, preventing further calculations to obtain rates of convergence. Work in this direction was conducted by van de Geer in [27] but can not be applied in this framework. Thus, it seems rather difficult to obtain rates of convergence for nonparametric maximum likelihood estimator using this settings.

3.5 No sieves

In order to construct a practical maximum likelihood estimator, one needs to construct a family of finite dimensional spaces undergoing the assumptions of Theorem (2.1). Two main choices are investigated in the statistical literature, but none fulfils all the needed requirements.

On the one hand, we could consider sieves constructed on log bases. Indeed, for a basis $(\psi_\lambda)_{\lambda \in \Lambda}$ of an Hilbert space, consider for a fixed integer m the set

$$\mathcal{F}_{S,m} := \left\{ \varphi \in \mathcal{F}_S, \text{ s.t. } \log \varphi = \sum_{\lambda \in \Lambda_m} \beta_\lambda \psi_\lambda \right\},$$

where $\Lambda_m \subset \Lambda$ with $|\Lambda_m| \leqslant m$. If we have taken spline basis for our initial choice of ψ_λ, we get the traditional log-spline model, well studied by Stone in [26]. Such sets are made of densities but are not compact for the chosen topology.

On the other hand consider a Multiresolution analysis, see for instance [18], constructed using a wavelet basis, $(\zeta_\lambda)_{\lambda \in \Lambda}$. Hence the finite dimensional sets corresponding to the approximation spaces are defined by $\mathcal{F}_{S,m} = \left\{ \varphi = \sum_{\lambda \in \Lambda_m} \beta_\lambda \zeta_\lambda \right\}$. Notice that $\mathcal{F}_{S,m}$ is a closed convex subset of an Hilbert space. However, it is not a subset of $\mathcal{F}_S$, set of the densities. This drawback appears frequently when estimating densities by wavelet estimators: the estimate is not a density. This defect, which in standard issues is not redhibitory, prevents here the use of Theorem (2.1).

Conclusion

We have shown that the nonparametric maximum likelihood estimator for (1) is consistent. However, the practical construction of usable sieves in the spirit of Section 2 is questionable. Improvements and rates of convergence are difficult to obtain in these setting. In the case where a large number of observations for each subject are available, i.e. $n \to +\infty$, the problem can be divided in two sub-issues: first estimate the random effect and then build a nonparametric estimator of its density. This point of view is tackled for example in [4] or [11]. However, when there is no hope for more data, in particular when dealing with medical data for which typically n is less than 5, we believe that other types of estimators should be considered.

References

[1] Robert A. Adams, *Sobolev spaces*, Academic Press [A subsidiary of Harcourt Brace Jovanovich, Publishers], New York-London, 1975, Pure and Applied Mathematics, Vol. 65. MR 56 #9247

[2] Dankmar Böhning, *A review of reliable maximum likelihood algorithms for semiparametric mixture models*, J. Statist. Plann. Inference **47** (1995), no. 1-2, 5–28, Statistical modelling (Leuven, 1993). MR 96h:62056

[3] ———, *Computer-assisted analysis of mixtures and applications*, Monographs on Statistics and Applied Probability, vol. 81, Chapman & Hall/CRC, Boca Raton, FL, 1999, Meta-analysis, disease mapping and others. MR 2001a:62001

[4] I. Castillo and J-M. Loubes, *Estimation of the distribution of random shifts deformation*, Prépublications de l'université d'Orsay (2004).

[5] L. Cavalier, G.K. Golubev, D. Picard, and A.B. Tsybakov, *Oracle inequalities for inverse problems.*, Ann. Stat. **30** (2002), no. 3, 843–874.

[6] D. Concordet and Nunez, *When is a nonlinear mixed-effects model identifiable?*, Preprint, `http://biostat.envt.fr/~dconcordet/`, 2002.

[7] Mary Davidian and David Giltinan, *Nonlinear Models for Repeated Measurement Data: An Overview and Update*, Journal of Agricultural, Biological, and Environmental Statstics **8** (2003), 387–419, `http://www4.stat.ncsu.edu/~davidian/`.

[8] A. P. Dempster, N. M. Laird, and D. B. Rubin, *Maximum likelihood from incomplete data via the EM algorithm*, J. Roy. Statist. Soc. Ser. B **39** (1977), no. 1, 1–38, With discussion. MR 58 #18858

[9] P. P. B. Eggermont and V. N. LaRiccia, *Maximum penalized likelihood estimation. Vol. I*, Springer Series in Statistics, Springer-Verlag, New York, 2001, Density estimation. MR 2002j:62050

[10] S. Ghosal and A. van der Vaart, *Posterior convergence rates of dirichlet mixtures of normal distributions for smooth densities*, preprint (2003).

[11] C. Giutys, *Adaptative density estimation in deconvolution*, JASA (1997).

[12] P. Groeneboom, G. Jongbloed, and J. A. Wellner, *The support reduction algorithm for computing nonparametric function estimates in mixture models*, preprint (2002).

[13] Tze Leung Lai and Mei-Chiung Shih, *Nonparametric estimation in nonlinear mixed effects models*, Biometrika **90** (2003), no. 1, 1–13. MR 2004b:62093

[14] Bruce G. Lindsay, *The geometry of mixture likelihoods: a general theory*, Ann. Statist. **11** (1983), no. 1, 86–94. MR 85m:62008a

[15] ———, *The geometry of mixture likelihoods. II. The exponential family*, Ann. Statist. **11** (1983), no. 3, 783–792. MR 85m:62008b

[16] ———, *Mixture Models: Theory, Geometry, and Applications*, Institute of Mathematical Statistics and the American Statistical Association, 1995.

[17] Bruce G. Lindsay and Mary L. Lesperance, *A review of semiparametric mixture models*, J. Statist. Plann. Inference **47** (1995), no. 1-2, 29–39, Statistical modelling (Leuven, 1993). MR 96h:62075

[18] S. Mallat, *A wavelet tour of signal processing*, Academic Press Inc., San Diego, CA, 1998. MR 99m:94012

[19] Vladimir G. Maz'ja, *Sobolev spaces*, Springer Series in Soviet Mathematics, Springer-Verlag, Berlin, 1985, Translated from the Russian by T. O. Shaposhnikova. MR 87g:46056

[20] France Mentré and Alain Mallet, *Handling covariates in population pharmacokinetics*, Int. J. Biomed. Comp. **36** (1994), 25–33.

[21] France Mentré, Alain Mallet, and Doha Baccar, *Optimal design in random-effects regression models*, Biometrika **84** (1997), no. 2, 429–442. MR 1 467 058

[22] Finbarr O'Sullivan, *A statistical perspective on ill-posed inverse problems (with discussion).*, Stat. Sci. **1** (1986), 502–527.

[23] J. Pfanzagl, *Consistency of maximum likelihood estimators for certain nonparametric families, in particular: mixtures*, J. Statist. Plann. Inference **19** (1988), no. 2, 137–158. MR 89g:62063

[24] ———, *Large deviation inequality for maximum likelihood estimators for certain nonparametric families, in particular: mixtures*, Ann. of Stats. **19** (1988), no. 2, 137–158. MR 89g:62063

[25] Alan Schumitzky, *Nonparametric EM algorithms for estimating prior distributions*, Appl. Math. Comput. **45** (1991), no. 2, part II, 143–157. MR 92g:62047

[26] C. Stone, *Large-sample inference for log-spline models*, Ann. Statist. **18** (1990), no. 2, 717–741. MR 91m:62073

[27] Sara van de Geer, *Rates of convergence for the maximum likelihood estimator in mixture models.*, J. Nonparametric Stat. **6** (1996), no. 4, 293–310 (English).

Djalil Chafaï.
Address: UMR 181 INRA/ENVT, École Nationale Vétérinaire de Toulouse, 23 Chemin des Capelles, B.P. 87614, F-31076, Toulouse, Cedex 3, France.
E-mail: mailto:d.chafai@envt.fr.nospam
Address: UMR 5583 CNRS/UPS, Institut de Mathématiques de Toulouse, Université Paul Sabatier, 118 route de Narbonne, F-31062, Toulouse, Cedex 4, France.
E-mail: mailto:chafai@math.ups-tlse.fr.nospam
Web-site: http://www.lsp.ups-tlse.fr/Chafai/

Jean-Michel Loubes.
Address: UMR 8628 CNRS/Paris-Sud, Bâtiment 425, Département de Mathématiques d'Orsay, Université d'Orsay Paris XI, F-91425, Orsay, Cedex, France.
E-mail: mailto:Jean-Michel.Loubes@math.u-psud.fr.nospam
Web-site: http://www.math.u-psud.fr/~loubes/