Comparison of nonparametric methods in nonlinear mixed effects models

Julie Antic, Céline M. Laffont, Djalil Chafai, Didier Concordet

Comparison of nonparametric methods in nonlinear mixed effects models

J. Antic [a,b,*], C.M. Laffont [a], D. Chafaï [b], D. Concordet [b].

[a] *Institut de Recherches Internationales Servier, Courbevoie, France.*

[b] *UMR181 Physiopathologie et Toxicologie Expérimentales, INRA, ENVT, Toulouse, France.*

Abstract

During the drug development, nonlinear mixed effects models are routinely used to study the drug's pharmacokinetics and pharmacodynamics. The distribution of random effects is of special interest because it allows to describe the heterogeneity of the drug's kinetics or dynamics in the population of individuals studied. Parametric models are widely used, but they rely on a normality assumption which may be too restrictive. In practice, this assumption is often checked using the empirical distribution of random effects' empirical Bayes estimates. Unfortunately, when data are sparse (like in patients phase III clinical trials), this method is unreliable. In this context, nonparametric estimators of the random effects distribution are attractive. Several nonparametric methods (estimators and their associated computation algorithms) have been proposed but their use is limited. Indeed, their practical and theoretical properties are unclear and they have a reputation for being computationally expensive. The article evaluates four nonparametric methods in comparison with the usual parametric method. Statistical and computational features are reviewed and practical performances are compared in simulation studies mimicking real pharmacokinetic analyses. The nonparametric methods seemed very useful when data are sparse. On a simple pharmacokinetic model, all the nonparametric methods performed roughly equivalently. On a more challenging pharmacokinetic model, differences between the methods were clearer.

Key words: Nonlinear mixed effects models, Nonparametric, NPML, NPEM, NONMEM, SNP, EBEs.

* Corresponding author.
Email address: `j.antic@envt.fr` (J. Antic).

1 Introduction

Nonlinear mixed effects (NLME) models are increasingly used in several biomedical applications. Especially, they have become a routine tool for the analysis of pharmacokinetic (PK) and/or pharmacodynamic (PD) data obtained in clinical trials during drug development. Furthermore, the US Food and Drug Administration strongly recommends their use for regulatory review (FDA, 1999). For each individual, PK data include the dosing history (amount of drug administered, route and time of administration), measured drug concentrations in blood or in plasma, measurement times and values of clinically relevant covariates (e.g. sex, age, weight, renal function...). The drug's systemic concentration-time profile depends on fixed effects and on individual-specific random effects quantifying the drug's absorption, distribution and elimination. NLME models allow to estimate the distribution $\mathcal{P}^*$ of random effects, even when within-individual information is sparse. They are particularly interesting in phase II or phase III clinical trials that involve a large group of individuals representative of the target population, with often few information per individual. The objective of these analyses is to give a rationale for dosing recommendations and identify subpopulations of patients that require a dosing adjustment.

The most widely used model is parametric: it is usually assumed that $\mathcal{P}^*$ is a (log-) normal distribution. The model is said parametric because the distribution of random effects $\mathcal{P}^*$ depends only on a finite number of parameters: its mean and variance. In this framework, the maximum likelihood estimator (MLE) is generally favored because it is consistent and efficient. The nonlinear relationship between the observations and the random effects makes the likelihood not explicit. Several algorithms have been proposed to compute approximations of the MLE. The most popular are First Order Conditional Estimates with Interaction (FOCEI), First Order Conditional Estimates (FOCE) and First Order (FO) algorithms. They are implemented in the gold standard software NONMEM (Boeckmann, Sheiner & Beal, 2006).

However, the parametric assumption may be too restrictive to describe very heterogeneous populations. In practice, this can be checked by looking at the empirical distribution of individual predictions of random effects, known as empirical Bayes estimates (EBEs) (Pinheiro & Bates (2000), EMEA (2007)). The problem is that, when the number of measurements per individual is low (the dataset is qualified as sparse), the EBEs are strongly influenced by the parametric assumption and their empirical distribution is unreliable (Savic, Wilkins & al., 2006).

In that context, nonparametric (NP) estimators, which allow $\mathcal{P}^*$ to live in an infinite dimensional space, are attractive. Several NP methods have been

proposed (by method, we mean an estimator plus its computation algorithm). Laird (1978), Pfanzagl (1988) and Lindsay (1983) studied the MLE when no assumption is made on $\mathcal{P}^*$. Two algorithms have been introduced to compute it. Mallet (1986) exploited a duality with the D-optimal design theory to develop the Nonparametric Maximum Likelihood (**NPML**) algorithm. Schumitzky (1991) proposed a generalization of the EM algorithm (Dempster, Laird & Rubin, 1977): the Nonparametric EM (**NPEM**) algorithm. An advanced version of the NPEM algorithm, named NPAG, has been developed by the USC Laboratory of Applied Pharmacokinetics (USC Pack, 2002).
Gallant & Nychka (1987) and Davidian & Gallant (1993) investigated a Smooth (or Semi) Nonparametric (**SNP**) estimator assuming that $\mathcal{P}^*$ admits a smooth probability density function π^* with respect to the Lebesgue measure.
Kuhn (2003) proposed a consistent logspline estimator of π^*.
Finally, since 2006, NONMEM has introduced a discrete estimator whose support points are the EBEs (Boeckmann, Sheiner & Beal, 2006). In the present article, we will refer to it as **NPNM** which stands for NP in NONMEM.

Several articles evaluated one of these NP methods on simulated data: Roe, Vonesh & al. (1997) studied **NPML**; Bustad, Terziivanov & al. (2006) evaluated **NPEM**; Savic, Kjellsson & al. (2006) focused on **NPNM**. However, there is a lack of a comparison between the different NP methods. Besides these studies do not include comparison with the empirical distribution of EBEs which remains widely used, despite its weaknesses.

This article offers to carry out a detailed review and comparison of four NP methods among the most widely used or documented ones: **NPML**, **NPEM**, **SNP** and **NPNM**, with focus on PK applications. The empirical distribution of EBEs is also included in the study. The logspline estimator (Kuhn, 2003) is not further investigated here, because it is only defined for unidimensional random effects whereas the simplest PK model deals with bidimensional random effects.

The article is organized as follows. The model and the notations are presented in section 2. Section 3 describes the different methods studied. Section 4 is dedicated to the review of their asymptotic properties. Section 5 presents two PK simulation studies that empirically compare the performance of the methods. The article ends with a conclusion in section 6.

2 Notations and framework

We consider the following general NLME model:

$$Y_i = f_i(d_i, \theta^*, X_i) + \Gamma_i(d_i, \theta^*, X_i)\varepsilon_i \tag{1}$$

where

- $Y_i = (Y_{ij}, j = 1..n_i)$ is the n_i-vector of observations on individual i.
- d_i is the known experimental design for individual i.
- θ^* is an unknown p-vector of fixed parameters (or fixed effects).
- $X_i = (X_{1i}, ..., X_{qi})$ is the q-vector of real random effects associated with individual i. The X_is are assumed independent and identically distributed (iid) from the probability measure $\mathcal{P}^*$:

$$X_i \overset{iid}{\sim} \mathcal{P}^* \quad i = 1..N.$$

  We denote Π^* the cumulative distribution function (cdf) of X_i, and, if it exists, π^* its probability density function (pdf) with respect to the Lebesgue measure.
- f_i is a known real function depending on i, d_i, θ^* and X_i. An important feature of NLME models is the nonlinearity of f_i with respect to the random effects X_i.
- $\varepsilon_i = (\varepsilon_{ij}, j = 1..n_i)$ is a white noise:

$$\varepsilon_i \overset{iid}{\sim} \mathcal{N}(0, I_{n_i}) \quad i = 1..N.$$

  ε_i is independent of X_i.
- $\Gamma_i(d_i, \theta^*, X_i)$ is a known positive low triangular $n_i \times n_i$-matrix depending on d_i, θ^* and X_i.

The challenge of the statistical inference is the estimation of θ^* and $\mathcal{P}^*$ from the observation of $Y_i, i = 1..N$ while the X_is are unobserved.
In phase II or phase III clinical trials, the number of individual measurements (n_i) is usually not large enough to be under asymptotic conditions and can even be very low. Thus, the only asymptotic examined in the following will be the number N of individuals in the sample, whereas n_i will remain finite and possibly small (with possible differences between individuals).

It is important to note that, in our NP context, the first moment of $\mathcal{P}^*$ can be infinite. Consequently, it cannot be considered as a fixed effect as usually done in parametric models. One goal of the PK and/or PD analyses is to explain the mean variations of X_i with covariates (age, body weight, renal function...). Obviously in this case, it is necessary to assume that the first moment of $\mathcal{P}^*$ is finite. The ability of the NP methods to apply in this context will be discussed in the following.

3 Studied methods

*3.1 The EBEs' empirical distribution (**EBEsED**)*

In population PK and/or PD analyses, it is common to assume the (log-) normality of X_i. In this parametric model, $\mathcal{P}^*$ only depends on its mean and variance. These parameters and the fixed effects θ^* are generally estimated by MLE:

$$(\hat{\theta}_N^{parametric}, \hat{\mathcal{P}}_N^{parametric}) = \underset{\theta, \mathcal{P} \in \mathcal{N}_q}{\operatorname{argsup}} \; L_N(\theta, \mathcal{P}),$$

where $\mathcal{N}_q$ denotes the set of the q-dimensional (log-)normal distributions and $L_N(\theta, \mathcal{P})$ the likelihood of θ and $\mathcal{P}$.
After this parametric estimation step, some predictions of the random effects X_i, known as empirical Bayes estimates (EBEs), can be computed:

$$EBE_i = \underset{x}{\operatorname{argsup}} \; \mathbb{P}[X_i = x | Y_i, \hat{\mathcal{P}}_N^{parametric}, \hat{\theta}_N^{parametric}],$$

where $\mathbb{P}[X_i|Y_i, \hat{\mathcal{P}}_N^{parametric}, \hat{\theta}_N^{parametric}]$ is the pdf of X_i conditionally to Y_i, $\hat{\theta}_N^{parametric}$ and $\hat{\mathcal{P}}_N^{parametric}$. In practice, the (log-) normality of X_i is often checked by looking at the EBEs' empirical distribution (**EBEsED**). In that context, **EBEsED** can be considered as an NP estimator of the random effects distribution $\mathcal{P}^*$: it is discrete with EBEs as support points and equal frequencies for all individuals (equal to $1/N$). Its cdf $\hat{\Pi}_N^{EBEsED}$ is:

$$\hat{\Pi}_N^{EBEsED}(x) = \frac{1}{N} \sum_{i=1}^{N} 1_{[EBE_i \leqslant x]},$$

where $1_{[EBE_i \leqslant x]}$ equals 1 when, for all $j = 1..q$, $EBE_i(j) \leqslant x(j)$, and 0 otherwise.
The drawback of **EBEsED** is that it gives the same weight ($1/N$) to all EBEs even if the accuracy of each EBE_i can be very different depending on the number of measurements per individual (n_i) and on the quality of the experimental design (d_i) (i.e. the choice of the measurement times).

*3.2 Nonparametric Maximum Likelihood (**NPML**)*

When no assumption is made on $\mathcal{P}^*$, the maximum likelihood estimator (MLE) $\widehat{\mathcal{P}}_N^{NP}$ is defined as:

$$\hat{\mathcal{P}}_N^{NP} = \underset{\mathcal{P} \in \mathbf{P}_q}{\operatorname{argsup}} \; L_N(\theta, \mathcal{P}) = \underset{\mathcal{P} \in \mathbf{P}_q}{\operatorname{argsup}} \; \prod_{i=1}^{N} \int \mathbb{P}(Y_i | X_i = x, \theta) \mathcal{P}(dx)$$

where $\mathbf{P}_q$ denotes the space of all probability measures on $\mathbb{R}^q$ and $\mathbb{P}(Y_i|X_i = x, \theta)$ the density of Y_i conditionally to $X_i = x$ with θ for fixed effects. The following theorem gives an handy way to compute $\hat{\mathcal{P}}_N^{NP}$.

Theorem 1 (Discreteness of $\hat{\mathcal{P}}_N^{NP}$ (Lindsay, 1983)) *Assuming that $\{(\mathbb{P}[Y_i|X_i = x, \theta], i = 1..N), x \in \mathbb{R}^q\}$ is a compact set, $\hat{\mathcal{P}}_N^{NP}$ can be expressed as a discrete measure with at most N support points.*

The proof is elementary. It deals with convexity and the theorem of Caratheodory. From a computational point of view, this discreteness has two useful consequences. First, for a given N, the likelihood maximization can be restricted to the finite dimensional set of all discrete measures with at most N support points. Secondly, the integral with respect to $\hat{\mathcal{P}}_N^{NP}$ is a finite sum, and thus the likelihood $L_N(\theta, \hat{\mathcal{P}}_N^{NP})$ has an explicit form:

$$\begin{aligned} L_N(\theta, \hat{\mathcal{P}}_N^{NP}) &= \prod_{i=1}^{N} \int \mathbb{P}[Y_i|X_i = x, \theta] d\hat{\mathcal{P}}_N^{NP}(x) \\ &= \prod_{i=1}^{N} \sum_{l=1}^{N} \mathbb{P}[Y_i|X_i = \mathcal{X}_l, \theta] \times p_l, \end{aligned}$$

where $\mathcal{X}_l, l = 1..N$ denote the support points of $\hat{\mathcal{P}}_N^{NP}$ and $p_l, l = 1..N$ their probabilities.

Despite the explicit writing of the likelihood, its maximum is still not explicit. Therefore, iterative algorithms have been developed to compute $\hat{\mathcal{P}}_N^{NP}$.

Mallet (1986) used an optimal design analogy to build the NPML algorithm. NPML is derived from the Fedorov algorithm (Fedorov , 1972). Assuming that, at iteration k, $\hat{\mathcal{P}}_N^k$ has L support points, **NPML** builds $\hat{\mathcal{P}}_N^{(k+1)}$ in 3 steps:

1. Support optimization step. The support of $\hat{\mathcal{P}}_N^{(k+1)}$ is built by adding a new point $\mathcal{X}_{L+1}^{(k+1)}$ to the support of $\hat{\mathcal{P}}_N^{(k)}$. $\mathcal{X}_{L+1}^{(k+1)}$ is defined as:

$$\mathcal{X}_{L+1}^{(k+1)} = \underset{\mathcal{X} \in \mathbb{R}^q}{\operatorname{argsup}} \frac{\partial L_N(\theta, \hat{\mathcal{P}}_N^{(k)})}{\partial \mathcal{X}},$$

   where $\frac{\partial L_N(\theta, \hat{\mathcal{P}}_N^{(k)})}{\partial \mathcal{X}}$ is the directional derivative of the likelihood in the direction $\mathcal{X}$.
   If $\underset{\mathcal{X} \in \mathbb{R}^q}{\operatorname{argmax}} \frac{\partial L_N(\theta, \hat{\mathcal{P}}_N^{(k)})}{\partial \mathcal{X}} = 0$, the algorithm can be stopped because $\hat{\mathcal{P}}_N^{(k)}$ maximizes the likelihood.
2. Frequencies optimization step. The $(L + 1)$ frequencies of $\hat{\mathcal{P}}_N^{(k+1)}$ are com-

puted by likelihood maximization:

$$(p_1^{(k+1)}, .., p_{L+1}^{(k+1)}) = \text{argsup } L_N(\theta, \hat{\mathcal{P}}_N^{(k+1)}),$$

where the *supremum* is over all $p_l, l = 1..L+1$ between 0 and 1 such that $\sum_{l=1}^{L+1} p_l = 1$.

3. Index limitation step. If $\hat{\mathcal{P}}_N^{(k+1)}$ has $N+1$ support points, a point can be removed from the support without a decreasing of the likelihood. Such a point $\mathcal{X}$ can be determined by the resolution of the linear system:

$$P^{Y|X} \times \mathcal{X} = b$$

where $P^{Y|X}$ is an $(N+1) \times (N+1)$-matrix and b an $(N+1)$-vector:

$$\begin{cases} P_{ij}^{Y|X} = \mathbb{P}[Y_i | X_i = \mathcal{X}_j^{(k+1)}] & i = 1..N,\ j = 1..N+1, \\ P_{ij}^{Y|X} = 1 & i = N+1,\ j = 1..N+1. \end{cases}$$

If $P^{Y|X}$ is singular then $b_i = 0\ \forall i = 1..N+1$, else $b_i = \frac{1}{\mathbb{P}[Y_i]}\ i = 1..N$ and $b_{N+1} = 0$.

Theorem 2 (convergence of NPML, Mallet (1986)) *Each iteration of NPML increases the sample likelihood. Besides, if the maximum likelihood estimator $\hat{\mathcal{P}}_N^{NP}$ is unique:*

$$\hat{\mathcal{P}}_N^{(k)} \underset{k \to +\infty}{\longrightarrow} \hat{\mathcal{P}}_N^{NP}$$

*where $\hat{\mathcal{P}}_N^{(k)}$ is the **NPML** estimator after k iterations.*

*3.3 Nonparametric Expectation Maximization (**NPEM**)*

As Mallet (1986), Schumitzky (1991) proposed an algorithm to compute the MLE $\hat{\mathcal{P}}_N^{NP}$ defined in the preceding section.

The Expectation Maximisation (EM) algorithm (Dempster, Laird & Rubin, 1977) is acknowledged to be very efficient for likelihood maximization with unobserved data. However, despite a wide area of applications, this algorithm does not permit to compute $\hat{\mathcal{P}}_N^{NP}$. Actually, the EM algorithm is restricted to absolutely continuous estimators of $\mathcal{P}^*$. Schumitzky (1991) extended the EM algorithm to the specific computation of $\hat{\mathcal{P}}_N^{NP}$. The Nonparametric EM (**NPEM**) algorithm consists of 2 steps. At iteration $(k+1)$, given $\hat{\mathcal{P}}_N^{(k)}$ with L support points ($L \leqslant N$), $\hat{\mathcal{P}}_N^{(k+1)}$ is defined in two steps:

1. Support optimization step. For $l = 1..L$, the support points $\mathcal{X}_l^{(k+1)}$ of $\hat{\mathcal{P}}_N^{(k+1)}$

are computed as:

$$\mathcal{X}_l^{(k+1)} = \underset{\mathcal{X}\in\mathbb{R}^q}{\text{argsup}} \sum_{i=1}^{N} \mathbb{P}[\mathcal{X}_l^{(k)}|Y_i, \hat{\mathcal{P}}_N^{(k)}, \theta] \times \ln \mathbb{P}[Y_i|X_i = \mathcal{X}, \theta],$$

where $\mathcal{X}_l^{(k)}, l = 1..L$ are the support points of $\hat{\mathcal{P}}_N^{(k)}$.

2. Frequencies optimization step. For $l = 1..L$, the probabilities $p_l^{(k+1)}$ of $\mathcal{X}_l^{(k+1)}$ have an explicit form:

$$p_l^{(k+1)} = \frac{1}{N} \sum_{i=1}^{N} \mathbb{P}[\mathcal{X}_l^{(k)}|Y_i, \hat{\mathcal{P}}_N^{(k)}, \theta]$$

Theorem 3 (Convergence of NPEM, Schumitzky (1991)) *Each iteration of **NPEM** increases the likelihood.*

NPEM often leads to measures with a low number of support points. Actually, **NPEM** does not allow to increase the number of support points, but some points can be removed if their probability becomes zero.

*3.4 Smooth Nonparametric (**SNP**)*

In most applications, the random effects X_i have absolutely continuous distributions. Davidian & Gallant (1993) proposed a smooth estimation of the density π^* of X_i. More precisely, they considered the set $\mathcal{SNP}_K$ of all density functions with finite mean m of form:

$$[P_K(U^{-1}(x-m))]^2 \times \mathbf{N}_q(x-m; 0, UU^T) + \epsilon\pi_0(x-m),$$

where

- K is a fixed truncation parameter,
- $\mathbf{N}_q(x; M, V)$ is the density of a q-dimensional normal random variable, with mean M and variance matrix V,
- $P_K(x)$ is a polynomial in $x = (x_1, ..., x_q)$ of degree K,
- U is a non singular upper triangular matrix, U^T is the transposed matrix of U,
- ϵ is small positive real number,
- π_0 is a strictly positive probability density function with zero mean, and such that $\|\pi_0\|_S$ is bounded, where $\|.\|_S$ is a weighted Sobolev norm (see Gallant & Nychka (1987) for details).

The term $\epsilon\pi_0(x-m)$ ensures that all the densities considered are strictly positive. In practice, this lower bound can generally be disregarded ($\epsilon = 0$) without affecting the behavior of the log-likelihood.

The **SNP** estimators $\hat{\theta}_N^{SNP}$ and $\hat{\pi}_N^{SNP}$ of θ^* and π^* are defined as the MLE over the set $\mathcal{SNP}_K$:

$$(\hat{\theta}_N^{SNP}, \hat{\pi}_N^{SNP}) = \underset{\theta,\pi\in\mathcal{SNP}_K}{\text{argsup}}\ L_N(\theta, \pi(x)dx)$$

The likelihood is not explicit, but it can be approximated by Gauss-Hermite quadrature. The maximization of the approximated likelihood with respect to θ, the coefficients of the polynomial P_K and the terms of U under the constraints $\int \hat{\pi}_N^{SNP}(x)dx = 1$ and $\int x\hat{\pi}_N^{SNP}(x)dx = m$ can be performed thanks to standard algorithms. Davidian & Gallant (1993) suggested for instance the Sequential Quadratic Programming algorithm.

*3.5 Nonparametric in NONMEM (**NPNM**)*

The software NONMEM (version VI) has recently introduced a nonparametric subroutine. As EBEsED, this estimator, referred to as **NPNM**, is a discrete measure with the EBEs used as support points. However, in contrast to **EBEsED**, the frequencies $p_1, .., p_N$ associated with support points $EBE_1, .., EBE_N$ are not necessarily the same for all individuals. They maximize the log-likelihood:

$$\begin{aligned}(p_1, .., p_N) &= \text{argsup}\ \ln L_N(\hat{\theta}_N^{parametric}, \hat{\mathcal{P}}_N^{NPNM}) \\ &= \text{argsup}\ \sum_{i=1}^{N} \ln(\sum_{l=1}^{N} p_l \mathbb{P}[Y_i|X_i = EBE_l, \hat{\theta}_N^{parametric}]),\end{aligned}$$

under the constraint $0 \leqslant p_l \leqslant 1$ for all $l = 1..N$ and $\sum_{l=1}^{N} p_l = 1$. It shall be noticed that, thanks to the discreteness of **NPNM**, the log-likelihood is explicit. Besides, this optimization is a convex problem.

4 Asymptotic behaviors

*4.1 The EBEs' empirical distribution (**EBEsED**)*

In this section, the consistency of **EBEsED**, considered as an NP estimator of $\mathcal{P}^*$, is investigated.
If all X_i $(i = 1..N)$ are observed, their empirical cdf is an unbiased consistent estimator of the cdf Π^*. In general mixed effects models, X_i are not observed but EBE_i, defined in section 3, can give an accurate prediction of X_i. Besides,

if θ^* is known and all n_i are large, the accuracy of this prediction is little affected by a misspecification of $\mathcal{P}^*$ (see van der Vaart (1998) for instance): conditionally on $X_i = x_i$,

$$\|EBE_i - x_i\| \xrightarrow[n_i \to +\infty]{\text{a.s.}} 0 \quad i = 1..N \quad \text{even if } \mathcal{P}^* \text{ is misspecified.}$$

Consequently, when θ^* is known and all n_i are large, **EBEsED** is consistent in the sense that, if for all i, $n_i \sim +\infty$,

$$\hat{\Pi}_N^{EBEsED}(x) = \frac{1}{N}\sum_{i=1}^{N} 1_{EBE_i \leqslant x} \xrightarrow[N \to +\infty]{\text{a.s.}} \Pi^*(x), \quad \forall x \in \mathbb{R}^q.$$

However, when data are sparse, the properties of **EBEsED** can be questioned. In a general NLME model, when n_i remains fixed, the rigorous study of the asymptotic (in N) behavior of **EBEsED** appears to be a hard matter. However, for a specific linear homoscedastic mixed effects model, the following theorem shows that **EBEsED** is not consistent.

Theorem 4 (non consistency of EBEsED) *Consider the model:*

$$Y_i = X_i + \theta^* \varepsilon_i, \ i = 1..N,$$

with the same notations as for model 1. Besides, we suppose that n_i is the same for all individuals i. Under the assumptions:

1. $|\mathbb{E}[X_i]| < \infty$ *and* $var[X_i] < \infty$,
2. $\exists M < \infty, \lim_{N \to +\infty} \sup_{i=1..N} |EBE_i| \leqslant M$.

If $\longrightarrow$ means weak convergence, we have:

$$\hat{\mathcal{P}}_N^{EBEsED} \underset{N \to +\infty}{\not\longrightarrow} \mathcal{P}^*.$$

Proof In this linear mixed effects model, the EBE_i, $\hat{\mathcal{P}}_N^{EBEsED}$, its mean and its variance are explicit. They depend on the parametric estimators of θ^*, $\mathbb{E}[X_i]$ and $\text{var}[X_i]$, which, in this linear mixed effects model, are consistent even if $\mathcal{P}^*$ is misspecified. Then, by the law of large numbers, assumption 1, the continuous mapping theorem, one can easily show that:

$$\mathbb{E}[\hat{\mathcal{P}}_N^{EBEsED}] \underset{N \to +\infty}{\longrightarrow} \mathbb{E}[\mathcal{P}^*]$$

and

$$var[\hat{\mathcal{P}}_N^{EBEsED}] \underset{N \to +\infty}{\longrightarrow} \frac{1}{1 + \frac{1}{n_i}\frac{\theta^{*2}}{var[\mathcal{P}^*]}} var[\mathcal{P}^*] \neq var[\mathcal{P}^*]$$

By assumption 2, it implies the non weak consistency of $\hat{\mathcal{P}}_N^{EBEsED}$ (van der Vaart, 1998, example 2.21). □

The shrinkage phenomenon, described by Savic, Wilkins & al. (2006), is an illustration of this result: when the data are sparse (that is the number of measurement per individual is small) or 'uninformative' (the variance of the residual error is high in comparison with the variance of X_i), the variance of **EBEsED** shrinks, i.e. it is smaller than the variance of $\mathcal{P}^*$.
As expected, when individual information is important (n_i is large), the bias of var$[\hat{\mathcal{P}}_N^{EBEsED}]$ is insignificant.
To our knowledge, the (non) consistency of **EBEsED** has not been established in sparse general NLME models.

*4.2 **NPML** and **NPEM***

NPML and **NPEM** aim at computing the same estimator, $\hat{\mathcal{P}}_N^{NP}$, defined as:

$$\hat{\mathcal{P}}_N^{NP} = \underset{\mathcal{P}\in\mathbf{P}_q}{\text{argsup}}\ L_N(\theta, \mathcal{P}).$$

Here $\mathcal{P}$ is unrestricted since $\mathbf{P}_q$ denotes the set of all probability measures on $\mathbb{R}^q$.
For a specific NLME model, Pfanzagl (1988) gave the asymptotic properties of $\hat{\mathcal{P}}_N^{NP}$. His result has been extended to a non-equally distributed (n_i and d_i are not the same for all individuals) and heteroscedastic NLME model by Chafaï & Loubes (2006).

Theorem 5 (Consistency of $\hat{\mathcal{P}}_N^{NP}$ (Chafaï & Loubes, 2006)) *Under the assumptions*

1. $\forall(\mathcal{P}_1, \mathcal{P}_2) \in \boldsymbol{P}_q \times \boldsymbol{P}_q$, $\mathcal{P}_1 \neq \mathcal{P}_2 \Rightarrow$

$$L_N(\theta, \mathcal{P}_1)dy_1..dy_N \neq L_N(\theta, \mathcal{P}_2)dy_1..dy_N \text{ in } \boldsymbol{P}_N \times \boldsymbol{P}_N,$$

2. the fixed effects θ^* are known.

The maximum likelihood estimator $\hat{\mathcal{P}}_N^{NP}$ is weakly consistent:

$$\hat{\mathcal{P}}_N^{NP} \underset{N\to+\infty}{\longrightarrow} \mathcal{P}^*.$$

Assumption 1 ensures identifiability of the model.
Assumption 2 is very restrictive because θ^* is generally unknown. Pfanzagl (1990) showed that assumption 2 can be relaxed if θ^* is consistently estimated whatever $\mathcal{P}^*$. But, to the best of our knowledge, such an estimation of θ^* is not currently available for a general NLME model. Lai & Shih (2003) proposed to jointly estimate θ^* and $\mathcal{P}^*$ by likelihood maximization. This procedure is

consistent in an Ibragimov-Has'minski framework:
$\exists I, 1 \leqslant I \leqslant N, I \underset{N\to+\infty}{\longrightarrow} +\infty$ such that $\forall i = 1..I, n_i \sim +\infty$.

As noticed in section 2, it is usual in PK and/or PD analyses to constrain the mean of X_i ($\mathbb{E}[X_i] = m < \infty$) in order to explain between subjects variations with the available covariates. Unfortunately, to the best of our knowledge, generalization of theorem 5 when the mean of X_i is constrained is not trivial. Mallet, Mentré & al. (1988) circumvent this difficulty by considering all covariates as random effects. In this case, the distribution of interest is the distribution of X_i conditionally to covariates. The joint distribution of X_i and the covariates is first estimated by unconstrained likelihood maximization (with **NPML** or **NPEM**). The distribution of X_i conditionally to covariates is then easily obtained by the Bayes formula. The method provides a consistent estimator of the distribution of X_i conditionally to covariates. Besides, the specification of a regression model between X_i and the covariates, and estimation of the regression parameters, are no longer necessary. On the other hand, this method increases the dimension of the distribution to be estimated. This highly increases the computation time and requires very large samples because the estimation of a joint distribution is much more demanding than the estimation of the distribution of X_i.

*4.3 Smooth Nonparametric (**SNP**)*

The estimators $\hat{\theta}_N^{SNP}$ and $\hat{\pi}_N^{SNP}$ have strong asymptotic properties.

Theorem 6 (consistency of SNP, Gallant & Nychka (1987)) *Let $(\hat{\theta}_N^{SNP}, \hat{\pi}_N^{SNP})$ be the MLE defined as:*

$$(\hat{\theta}_N^{SNP}, \hat{\pi}_N^{SNP}) = \underset{(\theta,\pi)\in\Theta\times\mathcal{SNP}_K}{argmax} L_N(\theta, \pi(x)dx),$$

where $\Theta \subset \mathbb{R}^p$ is compact and $\mathcal{SNP}_K$ is the set of smooth density functions defined in section 3.4. Under the assumptions:

1. *$\mathbb{E}[X_i] = m$ is finite. The actual density π^* of X_i can be written as:*

$$\pi^*(x) = h^2(x - m) + \epsilon h_0(x - m),$$

   where ϵ is a positive real number, h and h_0 are probability density functions with zero mean. $\|h\|_S$ and $\|h_0\|_S$ are bounded, where $\|.\|_S$ denotes a weighted Sobolev norm, see Gallant & Nychka (1987). h_0 is strictly positive. The actual fixed effect θ^* lives in Θ.
2. *$K \underset{N\to+\infty}{\longrightarrow} +\infty$*

3. *there is a function* $L(\theta, \pi, \theta^*, \pi^*)$ *that is continuous in* (θ, π) *with respect to* $\|(\theta, \pi)\| = (\|\theta\|^2 + \|\pi\|^2_{S\infty})^{1/2}$ *such that, with probability 1,*

$$\lim_{N\to\infty} \sup_{\theta\times\pi} |L(Y_i, i = 1..N|\pi(x)dx, \theta) - L(\theta, \pi, \theta^*, \pi^*)| = 0$$

where $\|.\|_{S\infty}$ *is an infinite weighted Sobolev norm (see Gallant & Nychka (1987) for details).*

4. *the model is identifiable.*

The following convergences hold:

$$\|\hat{\pi}_N^{SNP} - \pi^*\|_{S\infty} \xrightarrow[N\to+\infty]{a.s.} 0 \text{ and } \|\hat{\theta}_N^{SNP} - \theta^*\| \xrightarrow[N\to+\infty]{a.s.} 0$$

Consistency with respect to $\|.\|_{S\infty}$ implies L1-consistency of $\hat{\pi}_N^{SNP}$ and of its derivatives up to a certain order, and consistency of the first moments of X_i.

Assumption 1 imposes regularity conditions on π: the bound on $\|h\|_S$ restricts violent oscillatory behavior and tail's thickness, conditions on h_0 bound the tails from below. This assumption seems realistic for most of the applications because numerous type of densities, like skewed, leptokurtic, platykurtic, or multi-modal, are allowed.
Assumption 2 ensures that the potential complexity of $\hat{\pi}_N^{SNP}$ increases with the sample size. The truncation parameter K thus plays a key role in the **SNP** estimation. In practice, its determination for a given sample size N is not obvious. Fenton & Gallant (1996) investigated deterministic rules. When X_i is observed and univariate, they found that the optimal convergence rate (with respect to the L1-norm) was achieved for $K = N^{\frac{1}{5}}$. To our knowledge, there has been no extension of this result to NLME models. In these models, Davidian & Gallant (1993) recommend to 'inspect plots [...] for all models between those selected by Schwarz and Akaike criteria inclusively and make a visual selection'. An automatic rule can be necessary when an *a priori* knowledge or the subjectivity of the analyst could distort the figures interpretation. In that case, Davidian & Gallant (1993) advise to use the Hannan-Quinn criterion which gives good results in others Hermite series expansion studies.

It is noteworthy that theorem 6 proves the consistency of the fixed effects' estimation jointly with the distribution of random effects' estimation. Such a result is not established for the others NP methods.

It shall be noticed that theorem 6 ensures that $\hat{\pi}_N^{SNP}$ and $\hat{\theta}_N^{SNP}$ remain consistent when the mean of X_i is constrained: this allows to use covariates for explaining the mean variations of X_i.

Given that the (log-) normal distribution is a special case of **SNP** estimation (for $K = 0$), the (log)-normality assumption can be tested with, for instance,

the likelihood ratio test.

*4.4 Nonparametric in NONMEM (**NPNM**)*

To the best of our knowledge, asymptotic properties of **NPNM** have not been documented.

5 Comparison on simulation studies

In order to compare the various NP methods in practical situations, we performed two simulation studies. This section is organized in 3 parts: part 1 describes the design of the simulations, part 2 presents the various *criterion* used to compare the methods' performance and part 3 is devoted the results.

5.1 Design of simulations

General strategy

The simulations were conducted in a PK framework. In order to simulate realistic datasets, the model's settings were chosen according to real population PK studies of phenobarbital in infants. The phenobarbital is a barbiturate generally used to treat neonatal seizures. Two simulations were performed: simulation 1 illustrated a simple PK analysis, simulation 2 illustrates a more challenging PK analysis.

Simulation 1 was inspired by the phenobarbital PK analysis of Grasela & Donn (1985). The PK of phenobarbital after intravenous administration is described with a one compartment model with first order elimination, as follows: for $j = 1..n_i$,

$$[f_i(d_i, \theta^*, X_i)]_j = \frac{D}{X_{i2}} \times \exp^{-X_{i1} \times t_{ij}/X_{i2}} .$$

The known experimental design d_i consists of the administered dose D at time 0 and the measurement times t_{ij}. The 2 random effects are: the clearance (X_{i1}) and the volume of distribution (X_{i2}). The measurement error is assumed to be proportional to the expected concentration:

$$\Gamma_i(d_i, \theta^*, X_i) = \theta^* \times \text{diag}(f_i(d_i, \theta^*, X_i)),$$

where θ^* is the only fixed effect. The clearance was simulated with a bimodal distribution. Clearance and volume of distribution were considered independent. In order to investigate the influence of data sparseness, a rich and a

sparse experimental design were simulated (in PK, a dataset is usually considered rich if, for each individual i, $n_i \geqslant q$, q being the number of random effects).

Simulation 2 was inspired by the phenobarbital PK analysis of Yukawa, Suematsu & al. (2005). The PK of phenobarbital after oral administration is described with a one compartment model with first-order elimination and first-order absorption, as follows: for $j = 1..n_i$

$$[f_i(d_i, \theta^*, X_i)]_j = \frac{DX_{i3}}{X_{i2}(X_{i3} - X_{i1}/X_{i2})}(\exp^{-X_{i1} \times t_{ij}/X_{i2}} - \exp^{-X_{i3} \times t_{ij}}).$$

The known experimental design d_i consists of the administered dose D at time 0 and the measurement times t_{ij}. The 3 random effects are: the apparent clearance (X_{i1}), the apparent volume of distribution (X_{i2}) and the absorption rate constant (X_{i3}). The measurement error is assumed to be proportional to the expected concentration. The apparent clearance was simulated with a bimodal distribution. Apparent clearance and volume of distribution were considered correlated.

For each simulation, sample of different sizes were simulated (N = 50, 100, 200, 300, 400) in order to assess the consistency of the methods.
For each simulation and sample size, 100 independent datasets were simulated and analyzed with the various methods studied (**EBEsED**, **NPML**, **NPEM**, **SNP**, **NPNM**).

More details on simulations 1, on simulation 2 and on computational settings are provided in the following sections.

Details on simulation 1

In order to choose the values of θ^* and of the first moments of X_i, we analyzed the real dataset (Grasela & Donn, 1985), without taking into account the covariates and assuming that X_i is log-normal. The estimation was performed using NONMEM software (subroutine ADVAN1 TRANS2, method FOCEI). The estimated values were:

- 12.8% for θ^*,
- 0.00613 liters per hour (L/h) for the mean clearance and 34.3% $\times$ 0.00613 for its standard deviation,
- 1.60 liters (L) for the mean volume of distribution and 47.3% $\times$ 1.60 for its standard deviation.

We simulated X_i with the following distribution with a minor (30%) and a major (70%) sub-population (see figure 1):

$$X_{i1} \overset{iid}{\sim} \begin{cases} \ln\mathcal{N}(0.00735, (14.84\% \times 0.00735)^2) & \text{with probability } 70\% \\ \ln\mathcal{N}(0.00330, (20.27\% \times 0.00330)^2) & \text{with probability } 30\% \end{cases}$$

$$\text{and} \quad X_{i2} \overset{iid}{\sim} \ln\mathcal{N}(1.60, (47.32\% \times 1.60)^2),$$

where $\ln\mathcal{N}(m, v)$ denotes the log-normal distribution with mean m and variance v, X_{i1} and X_{i2} were independently simulated. The values of modes and proportions were arbitrary chosen in such a way that the overall mean and variance were similar to those estimated from the real dataset used by Grasela & Donn (1985).

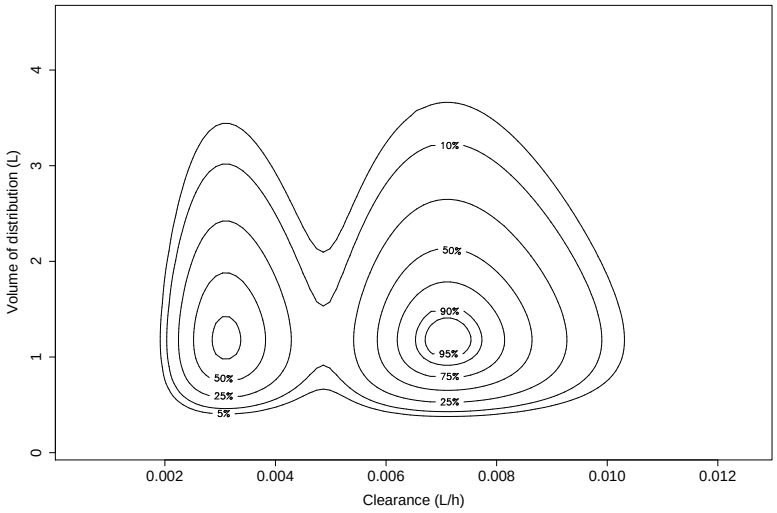

Figure 1. The distribution of random effects chosen in simulation 1: contour plots of the joint pdf of clearance and volume of distribution at quantiles 5%, 10%, 25%, 50%, 75%, 90% and 95%. The pdf of clearance is bimodal. Clearance and volume of distribution are independent.

In the simulations, we supposed that a single dose D was administered at time $t = 0$ to all individuals. The value of D was set to the mean of the initial doses of the real dataset used by Grasela & Donn (1985): $D = 25.7\mu g$. We simulated some rich and some sparse datasets. For rich datasets, we simulated:

- 2 observations per individual for 85% of the individuals: one early and one late,
- 3 observations per individual for 15% of the individuals: one early, one intermediate and one late.

For sparse datasets, we simulated:

- 1 single observation per individual for 85% of the individuals. This observation was early for half of the individuals, late for others individuals,
- 3 observations per individual for 15% of the individuals: one early, one intermediate and one late.

The observation times were: 12 minutes (min) for the early observation, 6 hours (h) for the intermediate one and 246 h for the late one. These times respectively correspond to the minimum, the median and the maximum of the observation times following intravenous administration in the real dataset.

Details on simulation 2

The value of θ^* used for the simulations equals the one found by Yukawa, Suematsu & al. (2005): 25.2% for oral administration. Yukawa, Suematsu & al. (2005) also found that, for an individual with a mean body weight of 2.9 kg and a mean postnatal age of 21 days: the apparent clearance has a mean of 0.0363 L/h and a standard deviation of 32%$\times$0.0363 L/h, the apparent volume of distribution has a mean of 7.75 L and a standard deviation of 54% $\times$ 7.75 L, the absorption rate constant equals $50h^{-1}$ (it was considered as a known fixed effect).
We simulated X_i with the following distribution (see figure 2):

$$X_{i1} \overset{iid}{\sim} \begin{cases} \ln\mathcal{N}(0.0465, (43.01\% \times 0.0465)^2) & \text{with probability } 70\% \\ \ln\mathcal{N}(0.0124, (47.72\% \times 0.0124)^2) & \text{with probability } 30\%, \end{cases}$$

$$X_{i2} \overset{iid}{\sim} \ln\mathcal{N}(7.75, 17.5)$$

$$\text{and} \quad X_{i3} \overset{iid}{\sim} \ln\mathcal{N}(5.00, 2.25).$$

Within each sub-population, the apparent clearance and volume of distribution were correlated:

$$\text{correlation}[\ln(X_{i1}), \ln(X_{i2})] = 0.6.$$

The absorption rate constant (X_{i3}) was independent of apparent clearance (X_{i1}) and volume of distribution (X_{i2}). The values of modes and proportions were arbitrary chosen in such a way that: the mean of X_{i1} and X_{i2} and the standard deviation of X_{i2} were similar to the value found by Yukawa, Suematsu & al. (2005). The X_{i1} coefficient of variation was increased from 32% to 64% to obtain a larger between individuals variability, very usual in PK applications. The absorption rate constant was considered as a random effect in order to have a more complex model. Its mean value was decreased from $50h^{-1}$ to $5h^{-1}$.

The value of the dose administered (D) was set to the mean of the real dataset's doses: $D = 12,400\mu g$. In contrast to simulation 1, we only simulated sparse datasets, with:

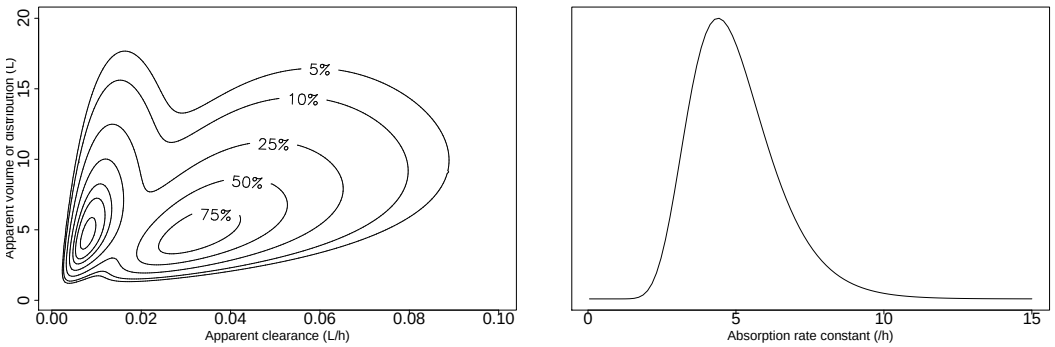

Figure 2. The distribution of random effects chosen in simulation 2. On the left, contour plot of the joint pdf of the apparent clearance and the apparent volume of distribution at quantiles 5%, 10%, 25%, 50%, 75%, 90% and 95%. The pdf of apparent clearance is bimodal. The apparent clearance and volume of distribution are correlated. On the right, the absorption rate constant's marginal pdf: it was independently simulated as a log-normal random variable.

- 2 observations per individual for 85% of the individuals. One of these observations was around the maximum of the concentration-time profile. The other observation was early for half of the individuals, and late for others individuals.
- 4 observations per individual for 15% of the individuals: one early, one intermediate, one late and one around the maximum of the concentration-time profile.

The early observation time was equal to 12 min, the intermediate one to 6 h and the late one to 246 h. The time around the maximum of the concentration-time profile was set to 1h 24 min (this time achieved the maximum of the population mean concentration-time profile).

5.1.1 Computational settings

In order to compute **EBEsED**,**NPNM** and an estimation $\hat{\theta}_N^{parametric}$ of θ^*, we assumed that $\mathcal{P}^*$ was a log-normal distribution and used the software NONMEM VI with the $ESTIMATION (FOCEI method for simulation 1, FO method for simulation 2), option POSTHOC and $NONPARAMETRIC routines (Boeckmann, Sheiner & Beal (2006)). FO has been used in simulation 2 because FOCEI failed too often.

In order to perform **NPML** and **NPEM**, an estimation of θ^* is needed: we chose $\hat{\theta}_N^{parametric}$. **NPML** and **NPEM** algorithms were implemented in C^{++}. **NPNM** was used as starting point for **NPML** and **NPEM**. A Fletcher-Reeves algorithm with 10 random different initializations was used to perform the support optimization step of **NPML**. The frequencies optimization step of **NPML** was performed with an EM algorithm as described by McLachlan &

Peel (2004). **NPEM** was stopped when:

$$\frac{\ln L_N(\hat{\theta}_N^{parametric}, \hat{\mathcal{P}}^{(k+1)}) - \ln L_N(\hat{\theta}_N^{parametric}, \hat{\mathcal{P}}^{(k)})}{|\ln L_N(\hat{\theta}_N^{parametric}, \hat{\mathcal{P}}^{(k)})|} < 10^{-4},$$

where $\ln L_N(\hat{\theta}_N^{parametric}, \hat{\mathcal{P}}^{(k)})$ is the **NPEM** likelihood at iteration k. This stopping criteria was not convenient for **NPML**, because the likelihood improvement between 2 NPML iterations is always very small (each iteration of **NPML** modifies only 1 support point whereas **NPEM** modifies N support points). We considered the likelihood improvement between N iterations: **NPML** was stopped when:

$$\frac{\ln L_N(\hat{\theta}_N^{parametric}, \hat{\mathcal{P}}^{(k+N)}) - \ln L_N(\hat{\theta}_N^{parametric}, \hat{\mathcal{P}}^{(k)})}{|\ln L_N(\hat{\theta}_N^{parametric}, \hat{\mathcal{P}}^{(k)})|} < 10^{-4},$$

where $\ln L_N(\hat{\theta}_N^{parametric}, \hat{\mathcal{P}}^{(k)})$ is the **NPML** likelihood at iteration k.

A fortran 77 implementation of the **SNP** method, called nlmix (Davidian & Gallant, 1991), is available in the public domain (statlib, Carnegie-Mellon University). The **SNP** estimations of θ^*, $\hat{\theta}_N^{SNP}$, and of $\mathcal{P}^*$ were performed with nlmix. The likelihood was computed by quadrature (with 15 quadrature points in simulation 1 and, to save computation time, only 10 in simulation 2). Following Davidian & Gallant (1993) and Davidian & Gallant (1991) recommendations, the matrix U was set diagonal for $K > 0$. Besides, for each dataset, several truncation parameters K (K=0,2,3 and 4) and, for each K, several initialization points (10 in simulation 1, 5 in simulation 2) were assessed. For a given K, we selected the initialization which performed the highest likelihood. The best K was then selected using the Hannan-Quinn criterion. Besides, when K was greater than 0, the matrix U was set diagonal.

5.2 Performance evaluation

Comparing the estimation of a distribution in inverse problems is not an straightforward task. Indeed, ways of comparing estimated and actual probability distribution of random effects are numerous, and in general not equivalent. Besides, in such inverse problems the objective in not always the estimation of the random effects' distribution. For instance in this study, the ability to detect the sub-population was of specific interest. Similarly, when the aim of the statistical study is more predictive than descriptive, the ability to predict the observed data Y_i or the unobserved random effects X_i could be a pertinent information. Therefore, it is necessary to use several *criteria* to compare the methods.

5.2.1 Estimation of $\mathcal{P}^*$

To compare the estimated and the true distribution, the T1 distance (also known as Wasserstein or Kantorovich or transport distance) is convenient here because it can be computed for multidimensional discrete and continuous distributions. The T1-error made by $\hat{\mathcal{P}}_N$ (with cdf $\hat{\Pi}_N$) in estimating $\mathcal{P}^*$ (with cdf Π^*) is defined as:

$$\text{T1-error}(\hat{\mathcal{P}}_N) = \int |\hat{\Pi}_N(x) - \Pi^*(x)| dx,$$

T1-error represents the absolute volume between the estimated and the true cdf.

5.2.2 Prediction of random effects

A prediction of unobserved random effects X_i can be computed after an estimation $\hat{\mathcal{P}}_N$ of $\mathcal{P}^*$ and $\hat{\theta}_N$ of θ^*:

$$EBE_i(\hat{\theta}_N, \hat{\mathcal{P}}_N) = \arg\sup_{X_i} \mathbb{P}[X_i | Y_i, \hat{\theta}_N, \hat{\mathcal{P}}_N].$$

It is noteworthy that $EBE_i(\hat{\theta}_N^{parametric}, \textbf{EBEsED})$ equals the parametric EBE_i defined in section 3.1.
For each dataset, the sum of absolute relative errors between $EBE_i(\hat{\theta}_N, \hat{\mathcal{P}}_N)$ and the actual realization x_i of X_i, gives the average percentage error (denoted X-error) for the prediction of X_i:

$$\text{X-error}_p(\hat{\theta}_N, \hat{\mathcal{P}}_N) = \frac{1}{N} \sum_{i=1}^{N} \left| \frac{EBE_i(\hat{\theta}_N, \hat{\mathcal{P}}_N)_p - x_{ip}}{x_{ip}} \right|.$$

5.2.3 Prediction of observed data

In order to evaluate the errors made in predicting the individual concentration time profile of interest, we computed, for each dataset:

$$\text{Y-error}(\hat{\theta}_N, \hat{\mathcal{P}}_N) = \frac{1}{N} \sum_{i=1}^{N} \int_{t=0}^{+\infty} |f(t, EBE_i(\hat{\theta}_N, \hat{\mathcal{P}}_N)) - f(t, x_i)| dt,$$

where $f(t, x) = f_i(d, \theta, x)$, where d represents the experimental design with dose D and measurement time $t_i = (t)$. The Y-error represents the average absolute area between the predicted and the real concentration-time profiles.

5.3 Results

Computation

The NP methods were only run when the NONMEM $ESTIMATION routine converged without warnings. Table 1 reports the number of these datasets. On these datasets, all NP methods converged (**SNP** was considered convergent if at least one initialization converged).

Table 1
Number of analyzed datasets over the number of simulated datasets. A dataset was analyzed only if the NONMEM $ESTIMATION routine converged without warnings.

N	Simulation 1 Rich datasets	Simulation 1 Sparse datasets	Simulation 2 Sparse datasets
50	93/100	98/100	38/100
100	95/100	100/100	51/100
200	93/100	95/100	53/100
300	93/100	100/100	58/100
400	96/100	97/100	69/100

Figure 3 displays the average computation times. The computation time of each method consists of the computation time needed to compute the fixed effects estimation ($\hat{\theta}_N^{parametric}$ for **EBEsED**, **NPNM**, **NPML** and **NPEM**, $\hat{\theta}_N^{SNP}$ for **SNP**) and of the time needed to compute the estimation of $\mathcal{P}^*$. For **SNP**, the computation time consists of the computation time needed to perform the MLE with $K = 0$, 2, 3 and 4, and for each K several initializations (10 for simulation 1, 5 for simulation 2).
As expected, the computation time of all methods increased with the sample size and with the number of random effects. **SNP** appeared particularly sensitive to an increase of the random effects dimension: this can be explained by the use of quadratures to approximate the likelihood.

Estimation of $\mathcal{P}^*$

Figure 4 and 5 display boxplots of the T1-errors.
For rich datasets (simulation 1 only), **EBEsED** performed as well as the NP methods. However, for sparse datasets (for simulations 1 and 2), it produced much more T1-error than the NP methods and its T1-error did not really decrease with the sample size: it seemed not consistent with respect to T1 metric.
On the contrary, in all situations, the T1-errors of all the NP methods de-

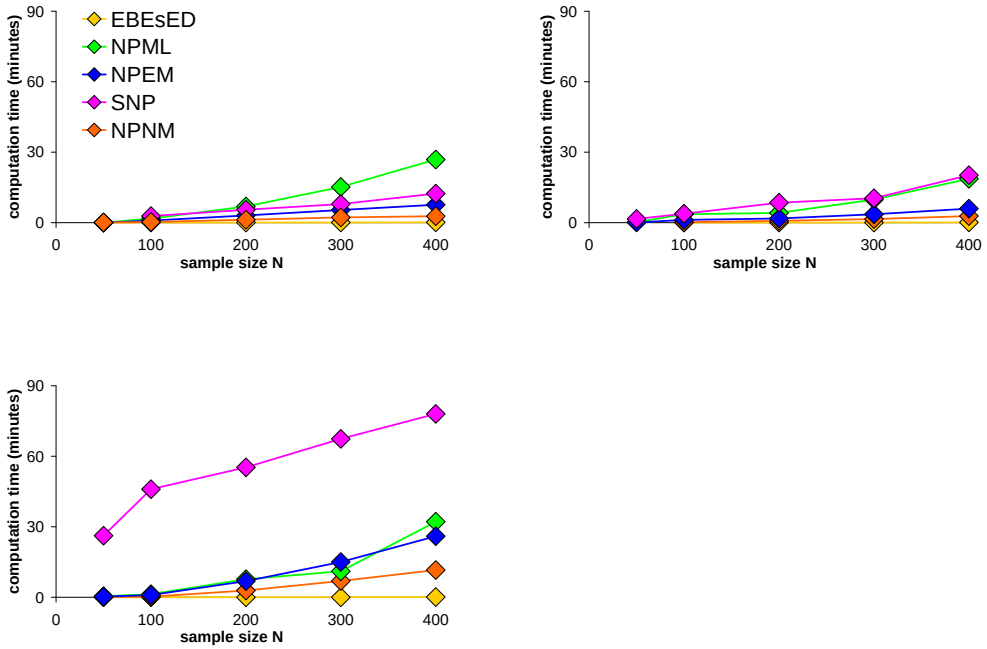

Figure 3. Average computation times with respect to the sample size N: on top, for simulation 1, on the left, for rich datasets, on the right for sparse datasets, at the bottom, for simulation 2 (for sparse datasets only).

creased closed to 0 with the sample size: all the NP methods seemed convergent with respect to T1 metric.

For simulation 1, the NP methods performed roughly equivalently with respect to the T1-error. As expected, these errors were lower for the rich than for the sparse datasets. It is noteworthy that, in this simulation, **NPNM** performed nearly as well as the others methods, which is surprising as its support points suffer from shrinkage.

For simulation 2, the difference between the NP methods was more perceptible. In particular, **NPNM** appeared less successful than **NPML** and **NPEM**: we suppose that, because of shrinkage, the EBEs did not provide a convenient support to describe $\mathcal{P}^*$ in this 3-dimensional model. The performance of **SNP** was not as good as in the 2-dimensional model: we speculate that the 10 quadrature points were not enough to give an accurate approximation of the likelihood and/or that the 5 initializations assessed were not sufficient to reach the global maximum in this high dimensional optimization problem (e.g. for $K = 4$, the dimension equals 41). The difference between **NPML** and **NPEM** was very slight.

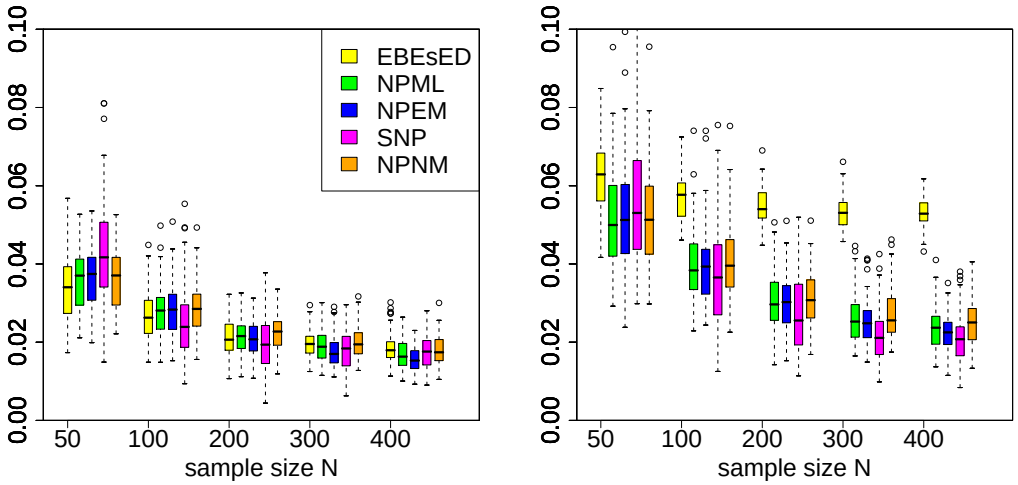

Figure 4. For simulation 1, boxplots of T1-errors for various methods with respect to the sample size N: on the left for the rich datasets, on the right for the sparse datasets.

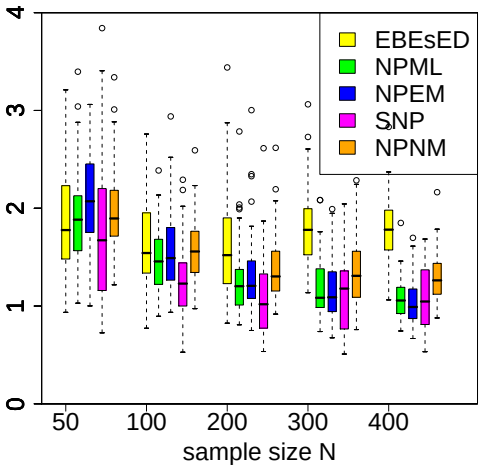

Figure 5. For simulation 2, boxplots of T1-errors with respect to the sample size N (the datasets are sparse).

Maximization of the likelihood

In order to evaluate the performances of the algorithms for likelihood maximization, we looked at the likelihood obtained with the various methods. The likelihood of **EBEsED** considered here is not the approximated likelihood given by NONMEM software after the parametric estimation, but the (explicit) likelihood of the NP discrete measure with EBEs as support points and equal frequencies. In that sense, it is comparable with the likelihood of **NPNM**, **NPML** and **NPEM**. The likelihood of **SNP** is not considered here since it is not comparable with the others because the estimation of θ^* is dif-

ferent. Table 2 reports the average log-likelihood of the various methods for samples of size $N = 400$. Results for the other sample sizes are not presented here but led to the same conclusions.

Table 2
Average log-likelihood of the various discrete NP estimators, for samples of size $N = 400$. * the likelihood of EBEsED computed here is not the parametric likelihood but the NP likelihood considering EBEsED as a discrete NP estimator with EBEs as support points and equals frequencies.

method	Simulation 1 Rich datasets	Simulation 1 Sparse datasets	Simulation 2 Sparse datasets
EBEsED*	-2403.3	-1518.6	-7260.7
NPNM	-2377.9	-1450.5	-7235.9
NPML	-2376.7	-1450.1	-7183.4
NPEM	-2376.3	-1449.1	-7166.9

In theory, for a given estimation of θ^*, the likelihood of **NPML** and **NPEM** should be greater than the likelihood of **EBEsED** and **NPNM**, since these algorithms maximize the likelihood over the set of all probability measures. Besides by definition, the likelihood of **NPNM** cannot be less than the **EBEsED**'s one. Our results are in agreement with the theory for all the datasets.
For simulation 1, the likelihood improvement due to the optimization of support points (difference between **NPML**/**NPEM** and **NPNM**) is slight in comparison with the likelihood increasing due to the frequencies optimization (the difference between **NPNM** and **EBEsED**), in both sparse and rich datasets. On the contrary, for simulation 2, the optimization of the support points provided a real improvement of likelihood. Overall, these results are coherent with the T1-error results presented above.
In all situations, in average, **NPEM** reached a higher likelihood than **NPML**, with a lower computational cost. Therefore, it can be concluded that the **NPEM** algorithm was, in average, more efficient than the **NPML** algorithm, under the studied conditions.

Prediction abilities

Figure 6 displays the boxplots of prediction errors for the clearance (X-error$_1$), for simulation 1, the sparse datasets and samples of size $N = 400$. In the overall population, **EBEsED** produced roughly as much X-error$_1$ as the NP methods, despite its poor estimation of $\mathcal{P}^*$. Actually, when looking at the 2 sub-populations separately, we can see that: in the major sub-population, **EBEsED** seemed better than the discrete NP methods, whereas in the minor

sub-population, the discrete NP methods seemed better than **EBEsED**. For all methods, the X-error$_1$ was more important in the minor than in the major sub-population. Surprisingly, **SNP** performed equivalently to **EBEsED** with respect to X-error$_1$, even in the minor sub-population.

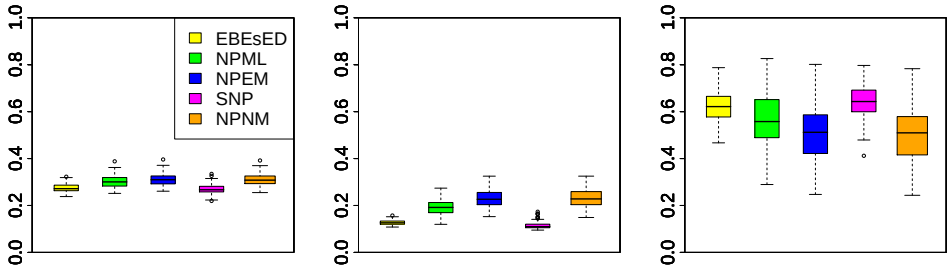

Figure 6. Boxplots of prediction errors for the clearance (X-error$_1$), for simulation 1, sparse datasets and samples of size $N = 400$: on the left for the overall population, in the middle within the major sub-population (around 70% of the individuals), on the right within the minor sub-population.

Figure 7 displays boxplots of prediction errors (Y-error), for simulation 1, the sparse datasets and samples of size $N = 400$. In the overall population, **EBEsED** made less Y-error than the discrete NP methods. Within the major sub-population, **EBEsED** was better than all the NP methods. However, within the minor sub-population, all the NP methods appeared better than **EBEsED**. It is noteworthy that **SNP** was the best in this sub-population, despite its poor prediction of clearance (X-error$_1$).

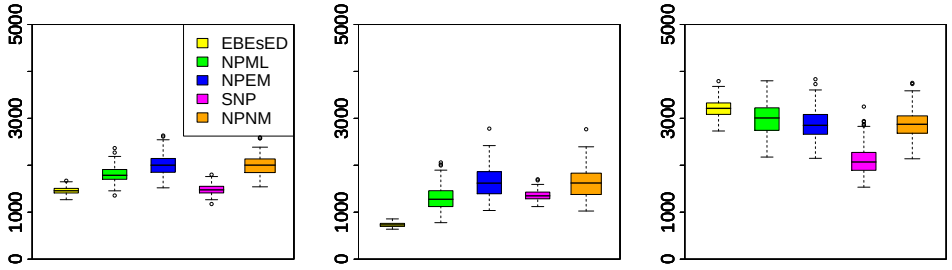

Figure 7. Boxplots of prediction errors of observed data (Y-error), for simulation 1, sparse datasets and samples of size $N = 400$: on the left for the overall population, in the middle within the major sub-population (around 70% of the individuals), on the right within the minor sub-population.

6 Conclusion

The goal of this article was to compare several NP methods for estimation of the random effects' distribution $\mathcal{P}^*$ in NLME models, with applications to PK analyses. We studied four NP methods: **NPNM**, **NPML**, **NPEM** and **SNP**. **EBEsED** which is often used to check the parametric assumption, has also been included in the study.
As expected, **EBEsED** appeared to be an efficient estimator of $\mathcal{P}^*$ when the experimental design was rich but was unreliable when the experimental design was sparse. In that case, the need for NP methods was confirmed.
The **NPML** and **NPEM** algorithms aim at computing the MLE when no assumption is made on $\mathcal{P}^*$. This MLE is a discrete measure with at most N support points. This discreteness has the advantage of making the likelihood explicit. On the other hand, it can make the interpretation difficult since $\mathcal{P}^*$ is generally continuous. The consistency of this MLE is established, but under restrictive assumptions: all fixed effects are known and no covariates are used to explain mean variations of X_i. In our simulation studies, **NPML** and **NPEM** gave satisfactory results even if the fixed effect (relative to residual error) was previously estimated using a parametric model. In general, **NPEM** reached a higher likelihood with a lower computational cost. Both **NPML** and **NPEM** present the risk of being trapped in local *maxima* since they are deterministic algorithms.
In contrast to these discrete NP methods, the **SNP** method provides a continuous estimator of $\mathcal{P}^*$. Its statistical properties are established under mild assumptions. In particular, the fixed effects can be estimated and covariates can be used to explain the mean variations of X_i. Besides, the (log-)normality can be tested. **SNP** have two main weaknesses: the selection of the truncation parameter and the likelihood optimization. To select the truncation parameter, no *criterion* has well established properties in NLME models. As for the likelihood optimization, this is difficult for two reasons. First, the likelihood is not explicit: here, it was approximated with quadratures which is computationally very expensive when the number of random effects is important. Second, the likelihood certainly has several local *maxima*: the algorithm used here was deterministic and so several initializations were needed to avoid local *maxima*. Investigations on stochastic algorithms would certainly be fruitful.
Lastly, **NPNM** is an easy-to-compute discrete estimator. On a simple PK model, it performed nearly as good as the others NP estimators, but was less satisfactory in a more complex PK model. Since its support points can be affected by shrinkage, its statistical properties are questionable.
Finite mixtures of (log-) normal distributions could also be an alternative to the parametric assumption (Lemenuel-Diot, Mallet & al. (2005), Wang, Schumitzky & al. (2007)). If $\mathcal{P}^*$ is really a finite mixture, the MLE is consistent and some computational algorithms are available. If not, one could think that increasing the number of mixture components would allow to approximate any

kind of distribution, but, to our knowledge, such a result is not established. In conclusion, all the studied NP methods seemed promising but would probably require more investigations on statistical and/or computational issues in order to be more widely used in population PK and/or PD analyses.

References

Beal S.L. & Sheiner L.B., 1982. Estimating Population Pharmacokinetics. CRC critical reviews in Biomedical Engineering. 8, 195–222.

Boeckmann A. J., Sheiner L. B. & Beal S. L., 2006. NONMEM Users Guide.

Bustad A., Terziivanov D., Leary R., Port R., Schumitzky A. & Jelliffe R., 2006. Parametric and Non parametric Population Methods : Their Comparative Performance in Analysing Clinical Dataset and Two MonteCarlo Simulation Studies. Clinical Pharmacokinetics. 45: 4, 365–383.

Chafaï D. & Loubes J.M., 2006. On Nonparametric Maximum Likelihood for a class of stochastic inverse problems. Statistics & Probability Letters. 76, 1125–1237.

Davidian M. & Gallant A.R., 1991. Nlmix: A Program for Maximum Likelihood Estimation of Nonlinear Mixed Effects Model with a Smooth Random Effects Density. User's Guide. Revised May 1993.

Davidian M. & Gallant A.R., 1993. The nonlinear mixed effects model with a smooth random effects density. Biometrika. 80: 3, 475–488.

Delyon B., Lavielle M. & Moulines E., 1999. Convergence of a stochastic approximation version of the EM algorithm Annals of Statistics. 27: 1, 94-128.

Dempster A.P., Laird N.M. & Rubin D.B., 1977. Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society, Series B, Methodological. 39: 1, 1–38.

Eggermont P., 1999. Nonlinear smoothing and the EM algorithm for positive integral equation of the first kind. Applied Mathematics and Optimization. 39: 1, 75–91.

Eggermont P. LaRiccia V., 1995. Maximum smoothed likelihood density estimation for inverse problems. Annals of Statistics. 23: 1, 199–220.

European Medecines Agency, 2007. Guideline on reporting the results of population pharmacokinetic analyses.

Fattinger, K.E., Sheiner L.B. & Verotta D., 1995. A new method to explore the distribution of interindividual random effects in non-linear mixed effects models. Biometrics. 51: 4, 1236–1251.

Fedorov V.V., 1972. Theory of Optimal Experiments. New York: Academic Press.

Fenton V.M., Gallant A.R., 1996. Convergence rates of SNP density estimators Econometrica. 64: 3, 719–727.

Food and drug administration, 1999. Guidance for industry: population kinetics.

Gallant A. R. & Nychka D. W. 1987. Semi-nonparametric maximum likelihood estimation. Econometrica. 55: 2, 363–390.

Grasela, T.H. & Donn S.M., 1985. Neonatal population pharmacokinetics of phenobarbital derived from routine clinical data. Developmental Pharmacology and Therapeutics. 8: 6, 374–383.

Hartford A. & Davidian M., 2000. Consequences of misspecifying assumptions in nonlinear mixed effects models. Computational Statistics and Data Analysis. 34: 2, 139–164.

Ingelman-Sunberg M., 2006. Drug Metabolic Enzymes: Genetic Polymorphisms. Encyclopedia of Life Sciences. John Wiley & Sons, Ltd.

Kuhn E., 2003. Estimation par maximum de vraisemblance dans des problèmes inverses non linéaires. phD thesis.

Kuhn E. & Lavielle M., 2005. Maximum likelihood estimation in nonlinear mixed effects models. Computational Statistics and Data Analysis. 49: 4, 1020–1038.

Lai T.L. & Shih M.C., 2003. Nonparametric estimation in nonlinear mixed effects models. Biometrika. 90: 1, 1–13.

Laird N., 1978. Nonparametric Maximum Likelihood Estimation of a Mixing Distribution. Journal of the American Statistical Association. 73: 364, 805-811.

Lemenuel-Diot A., Mallet A., Laveille C. & Bruno R., 2005. Estimating Heterogeneity in Random Effects Models for Longitudinal Data. Biometrical Journal. 47: 3, 329–345.

Lindsay B.G., 1983. The geometry of mixture likelihoods : a general theory. The annals of statistics. 11: 1, 86–94.

McLachlan G., Peel D., 2004. Finite Mixture Models. John Wiley & Sons.

Mallet A., 1986. A maximum likelihood method for random coefficient regression models. Biometrika. 73: 3, 645–656.

Mallet A., Mentré F., Gilles J., Kelman A.W., Thomson A.H., Bryson S.M. & Whiting B., 1988. Handling Covariates in Population Pharmacokinetics, with an Application to Gentamicin. Journal of Biomedical Measurement, Informatics and Control. 2, 138–146.

Pfanzagl J., 1988. Consistency of maximum likelihood estimators for certain nonparametric families, in particular: mixtures. Journal of Statistical Planning and Inference. 19: 2, 137–158.

Pfanzagl J., 1990. Large deviation probabilities for certain nonparametric maximum likelihood estimators. The Annals of Statistics. 18: 4, 1868–1877.

Pinheiro J.C. & Bates D.M., 2000. Mixed-Effects Models in S and S-PLUS. Statistics and Computing. Springer.

Roe D.J., Vonesh E.F., Wolfinger R.D., Mesnil F. & Mallet A., 1997. Comparison of population pharmacokinetic modeling using simulated data: results from the population modeling workgroup. Statistics in Medicine. 16: 11, 1241–1262

Savic R., Kjellsson M. & Karlsson M., 2006. Evaluation of the nonparametric method in NONMEM $VI\beta$. Population Approach Group in Europe.

Savic, R.M., Wilkins J.J. & Karlsson M.O., 2006. (Un)informativeness of empirical Bayes estimate-based diagnostics. The American Association of Pharmaceutical Scientists. 8: S2.

Schumitzky A., 1991. Nonparametric EM algorithms for estimating prior distributions. Applied Mathematics and Computation. 45: 2, 143–157.

Laboratory of Applied Pharmacokinetics, USC School of Medicine, 2002. USC*PACK PC Pharmacokinetic Programs.

van der Vaart A.W., 1998. Asymptotic Statistics. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.

Wang X., Schumitzky A., D'Argenio D., 2007. Nonlinear random effects mixture models: Maximum likelihood estimation via the EM algorithm. Computational Statistics and Data Analysis. 51: 12, 6614–6623.

Yukawa E., Suematsu F., Yukawa M. and Minemoto M., 2005. Population pharmacokinetic investigation of phenobarbital by mixed effect modelling using routine clinical pharmacokinetic data in Japanese neonates and infants. Journal of Clinical Pharmacy and Therapeutics. 30: 159–163.