Comparison of nonparametric methods in nonlinear mixed effects models

Julie Antic, Céline Laffont, Djalil Chafai, Didier Concordet

To cite this version:

Julie Antic, Céline Laffont, Djalil Chafai, Didier Concordet. Comparison of nonparametric methods in nonlinear mixed effects models. Computational Statistics and Data Analysis, Elsevier, 2009, 53, pp.642-656. <10.1016/j.csda.2008.08.021>. <hal-00286487v2>

HAL Id: hal-00286487
https://hal.archives-ouvertes.fr/hal-00286487v2
Submitted on 3 Mar 2009

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
Comparison of nonparametric methods in nonlinear mixed effects models

J. Antic\textsuperscript{a,b,*}, C.M. Laffont\textsuperscript{a}, D. Chafaï\textsuperscript{b}, D. Concordet\textsuperscript{b}.

\textsuperscript{a}Institut de Recherches Internationales Servier, Courbevoie, France.
\textsuperscript{b}UMR181 Physiopathologie et Toxiciologie 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 computational 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).

Preprint submitted to Elsevier August 9, 2008 Draft version.
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 $P^*$ 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 $P^*$ is a (log-) normal distribution. The model is said parametric because the distribution of random effects $P^*$ 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 $P^*$ 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), Pflanzagl (1988) and Lindsay (1983) studied the MLE when no assumption is made on $P^*$. 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 $P^*$ admits a smooth probability density function $\pi^*$ with respect to the Lebesgue measure. Kuhn (2003) proposed a consistent logspline estimator of $\pi^*$. 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:

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

(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 \( i \).
- \( \theta^* \) is an unknown \( p \)-vector of fixed parameters (or fixed effects).
- \( X_i = (X_{i1}, ..., X_{iq}) \) is the \( q \)-vector of real random effects associated with individual \( i \). The \( X_i \)s are assumed independent and identically distributed (iid) from the probability measure \( P^* \):

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

We denote \( \Pi^* \) the cumulative distribution function (cdf) of \( X_i \), and, if it exists, \( \pi^* \) its probability density function (pdf) with respect to the Lebesgue measure.

- \( f_i \) is a known real function depending on \( i, d_i, \theta^* \) and \( X_i \). An important feature of NLME models is the nonlinearity of \( f_i \) with respect to the random effects \( X_i \).
- \( \epsilon_i = (\epsilon_{ij}, j = 1..n_i) \) is a white noise:

\[
\epsilon_i \overset{iid}{\sim} N(0, I_{n_i}) \quad i = 1..N.
\]

\( \epsilon_i \) 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 \( d_i, \theta^* \) and \( X_i \).

The challenge of the statistical inference is the estimation of \( \theta^* \) and \( P^* \) from the observation of \( Y_i, i = 1..N \) while the \( X_i \)s 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 \( P^* \) 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 \( P^* \) 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, \( P^* \) only depends on its mean and variance. These parameters and the fixed effects \( \theta^* \) are generally estimated by MLE:

\[
(\hat{\theta}^{\text{parametric}}_N, \hat{P}^{\text{parametric}}_N) = \operatorname{argsup}_{\theta, P \in \mathcal{N}_q} L_N(\theta, P),
\]

where \( \mathcal{N}_q \) denotes the set of the \( q \)-dimensional (log-)normal distributions and \( L_N(\theta, P) \) the likelihood of \( \theta \) and \( P \).

After this parametric estimation step, some predictions of the random effects \( X_i \), known as empirical Bayes estimates (EBEs), can be computed:

\[
EBE_i = \operatorname{argsup}_x \mathbb{P}[X_i = x | Y_i, \hat{P}^{\text{parametric}}_N, \hat{\theta}^{\text{parametric}}_N],
\]

where \( \mathbb{P}[X_i | Y_i, \hat{P}^{\text{parametric}}_N, \hat{\theta}^{\text{parametric}}_N] \) is the pdf of \( X_i \) conditionally to \( Y_i, \hat{\theta}^{\text{parametric}}_N \) and \( \hat{P}^{\text{parametric}}_N \). 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 \( P^* \): it is discrete with EBEs as support points and equal frequencies for all individuals (equal to \( 1/N \)). Its cdf \( \hat{\Pi}^{\text{EBEsED}}_N \) is:

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

where \( 1_{\{EBE_i \leq x\}} \) equals 1 when, for all \( j = 1..q, EBE_i(j) \leq x(j) \), 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 \( P^* \), the maximum likelihood estimator (MLE) \( \hat{P}^{\text{NP}}_N \) is defined as:

\[
\hat{P}^{\text{NP}}_N = \operatorname{argsup}_{P \in \mathcal{P}_q} L_N(\theta, P) = \operatorname{argsup}_{P \in \mathcal{P}_q} \prod_{i=1}^{N} \int \mathbb{P}(Y_i | X_i = x, \theta)P(dx)
\]
where $P_q$ 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 $\theta$ for fixed effects. The following theorem gives an handy way to compute $\hat{P}_N^{NP}$.

**Theorem 1 (Discreteness of $\hat{P}_N^{NP}$ (Lindsay, 1983))** Assuming that $\{\mathbb{P}[Y_i|X_i = x, \theta], i = 1..N, x \in \mathbb{R}^q\}$ is a compact set, $\hat{P}_N^{NP}$ 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 $\hat{P}_N^{NP}$ is a finite sum, and thus the likelihood $L_N(\theta, \hat{P}_N^{NP})$ has an explicit form:

$$L_N(\theta, \hat{P}_N^{NP}) = \prod_{i=1}^{N} \int \mathbb{P}[Y_i|X_i = x, \theta] d\hat{P}_N^{NP}(x)$$

$$= \prod_{i=1}^{N} \sum_{l=1}^{N} \mathbb{P}[Y_i|X_i = \mathcal{X}_l, \theta] \times p_l,$$

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

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 $k$, $P_N^k$ has $L$ support points, NPML builds $\hat{P}_N^{(k+1)}$ in 3 steps:

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

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

where $\frac{\partial L_N(\theta, \hat{P}_N^{(k)})}{\partial \mathcal{X}}$ is the directional derivative of the likelihood in the direction $\mathcal{X}$.

If $\text{argmax}_{\mathcal{X} \in \mathbb{R}^q} \frac{\partial L_N(\theta, \hat{P}_N^{(k)})}{\partial \mathcal{X}} = 0$, the algorithm can be stopped because $\hat{P}_N^{(k)}$ maximizes the likelihood.

2. Frequencies optimization step. The $(L + 1)$ frequencies of $\hat{P}_N^{(k+1)}$ are com-
puted by likelihood maximization:

\[
(p_i^{(k+1)}, \ldots, p_{L+1}^{(k+1)}) = \text{argsup}_L L_N(\theta, \hat{\mathcal{P}}_N^{(k+1)}),
\]

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:

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

If \( P^Y|X \) is singular then \( b_i = 0 \forall i = 1..N + 1 \), else \( b_i = \frac{1}{P[Y_i]} \) \( i = 1..N \) 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:

\[
\hat{\mathcal{P}}_N^{(k)} \longrightarrow_{k \to +\infty} \hat{\mathcal{P}}_N^{NP}
\]

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 \( (L \leq N) \), \( \hat{\mathcal{P}}_N^{(k+1)} \) 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:

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

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

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

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

**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 $\pi^*$ of $X_i$. More precisely, they considered the set $SNP_K$ of all density functions with finite mean $m$ of form:

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

where

- $K$ is a fixed truncation parameter,
- $N_q(x; M, V)$ 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,
- $\epsilon$ is small positive real number,
- $\pi_0$ is a strictly positive probability density function with zero mean, and such that $\|\pi_0\|_S$ is bounded, where $\|\cdot\|_S$ 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 $\hat{\theta}^{SNP}_N$ and $\hat{\pi}^{SNP}_N$ of $\theta^*$ and $\pi^*$ are defined as the MLE over the set $SNP_K$:

$$(\hat{\theta}^{SNP}_N, \hat{\pi}^{SNP}_N) = \arg \sup_{\theta, \pi \in SNP_K} L_N(\theta, \pi(x) dx)$$

The likelihood is not explicit, but it can be approximated by Gauss-Hermite quadrature. The maximization of the approximated likelihood with respect to $\theta$, the coefficients of the polynomial $P_K$ and the terms of $U$ under the constraints $\int \hat{\pi}^{SNP}_N(x) dx = 1$ and $\int x \hat{\pi}^{SNP}_N(x) dx = m$ 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 $p_1, \ldots, p_N$ associated with support points $EBE_1, \ldots, EBE_N$ are not necessarily the same for all individuals. They maximize the log-likelihood:

$$(p_1, \ldots, p_N) = \arg \sup \ln L_N(\hat{\theta}_{parametric}^N, \hat{P}_{NPNM}^N)$$

$$= \arg \sup N \sum_{i=1}^{N} \ln \left( \sum_{l=1}^{N} p_l \mathbb{P}[Y_i | X_i = EBE_l, \hat{\theta}_{parametric}^N] \right),$$

under the constraint $0 \leq p_l \leq 1$ 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 $\Pi^*$. 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 $\theta^*$ is known and all $n_i$ are large, the accuracy of this prediction is little affected by a misspecification of $P^*$ (see van der Vaart (1998) for instance): conditionally on $X_i = x_i$,

$$\| EBE_i - x_i \| \xrightarrow{a.s.} 0 \quad i = 1..N \quad \text{even if } P^* \text{ is misspecified.}$$

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

$$\hat{\Pi}_{EBE_{i}ED}^N(x) = \frac{1}{N} \sum_{i=1}^{N} 1_{EBE_i \leq x} \xrightarrow{a.s.} \Pi^*(x), \quad \forall x \in \mathbb{R}^q.$$

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

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

$$\hat{P}_{N}^{EBE_{i}ED} \xrightarrow{N \to +\infty} P^*.$$

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

$$E[\hat{P}_{N}^{EBE_{i}ED}] \xrightarrow{N \to +\infty} E[P^*]$$

and

$$\text{var}[\hat{P}_{N}^{EBE_{i}ED}] \xrightarrow{N \to +\infty} \frac{1}{1 + \frac{\theta^2}{n_i \text{var}[P^*]}} \text{var}[P^*] \neq \text{var}[P^*]$$

By assumption 2, it implies the non weak consistency of $\hat{P}_{N}^{EBE_{i}ED}$ (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 $P^*$.

As expected, when individual information is important ($n_i$ is large), the bias of $\text{var}[\hat{\mathcal{P}}^{EBESE}_{i}]$ 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, $\hat{\mathcal{P}}^{NP}_N$, defined as:

$$\hat{\mathcal{P}}^{NP}_N = \operatorname{arg}
\sup_{\mathcal{P} \in \mathcal{P}_q} L_N(\theta, \mathcal{P}).$$

Here $\mathcal{P}$ is unrestricted since $\mathcal{P}_q$ denotes the set of all probability measures on $\mathbb{R}^q$.

For a specific NLME model, Pfanzagl (1988) gave the asymptotic properties of $\hat{\mathcal{P}}^{NP}_N$. 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 $\hat{\mathcal{P}}^{NP}_N$ (Chafaï & Loubes, 2006)) Under the assumptions

1. $\forall (P_1, P_2) \in \mathcal{P}_q \times \mathcal{P}_q, P_1 \neq P_2 \Rightarrow L_N(\theta, P_1)dy_1..dy_N \neq L_N(\theta, P_2)dy_1..dy_N$ in $\mathcal{P}_N \times \mathcal{P}_N$,

2. the fixed effects $\theta^*$ are known.

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

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

Assumption 1 ensures identifiability of the model.

Assumption 2 is very restrictive because $\theta^*$ is generally unknown. Pfanzagl (1990) showed that assumption 2 can be relaxed if $\theta^*$ is consistently estimated whatever $\mathcal{P}^*$. But, to the best of our knowledge, such an estimation of $\theta^*$ is not currently available for a general NLME model. Lai & Shih (2003) proposed to jointly estimate $\theta^*$ and $\mathcal{P}^*$ by likelihood maximization. This procedure is
consistent in an Ibragimov-Has’minski framework:
\[ \exists I, 1 \leq I \leq N, I \xrightarrow{N \to +\infty} +\infty \text{ such that } \forall i = 1..I, n_i \sim +\infty. \]

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 \( \hat{\theta}_\text{SNP}^N \) and \( \hat{\pi}_\text{SNP}^N \) have strong asymptotic properties.

**Theorem 6 (consistency of SNP, Gallant & Nychka (1987))** Let \((\hat{\theta}_\text{SNP}^N, \hat{\pi}_\text{SNP}^N)\) be the MLE defined as:
\[
(\hat{\theta}_\text{SNP}^N, \hat{\pi}_\text{SNP}^N) = \arg\max_{(\theta, \pi) \in \Theta \times \mathcal{SNP}_K} L_N(\theta, \pi(x)dx),
\]
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 \( \pi^* \) of \( X_i \) can be written as:
\[
\pi^*(x) = h^2(x - m) + \epsilon h_0(x - m),
\]
where \( \epsilon \) 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 \( \|\cdot\|_S \) denotes a weighted Sobolev norm, see Gallant & Nychka (1987). \( h_0 \) is strictly positive. The actual fixed effect \( \theta^* \) lives in \( \Theta \).
2. \( K \xrightarrow{N \to +\infty} +\infty \)
3. there is a function $L(\theta, \pi, \theta^*, \pi^*)$ that is continuous in $(\theta, \pi)$ 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:

$$\| \hat{\pi}^{SNP}_N - \pi^* \|_{S^\infty} \overset{a.s.}{\longrightarrow} 0 \quad \text{and} \quad \| \hat{\theta}^{SNP}_N - \theta^* \| \overset{a.s.}{\longrightarrow} 0$$

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

Assumption 1 imposes regularity conditions on $\pi$: 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 types of densities, like skewed, leptokurtic, platykurtic, or multi-modal, are allowed.

Assumption 2 ensures that the potential complexity of $\hat{\pi}^{SNP}_N$ 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 $K = N^{1/2}$. 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 $\hat{\pi}^{SNP}_N$ and $\hat{\theta}^{SNP}_N$ 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 $j = 1..n_i$,

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

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 $\theta^*$ 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 $i$, $n_i \geq q$, $q$ 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 $j = 1..n_i$

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

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

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

and

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

where $\text{ln} \mathcal{N}(m, v)$ 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 \( \theta^* \) 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% × 0.0363 L/h, the apparent volume of distribution has a mean of 7.75 L and a standard deviation of 54% × 7.75 L, the absorption rate constant equals 50 h\(^{-1}\) (it was considered as a known fixed effect).

We simulated \( X_i \) with the following distribution (see figure 2):

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

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

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 50 h\(^{-1}\) to 5 h\(^{-1}\).

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

In order to perform $\text{NPML}$ and $\text{NPEM}$, an estimation of $\theta^*$ is needed: we chose $\hat{\theta}_N^{\text{parametric}}$. NPML and NPEM algorithms were implemented in C++. $\text{NPNM}$ was used as starting point for $\text{NPML}$ and $\text{NPEM}$. A Fletcher-Reeves algorithm with 10 random different initializations was used to perform the support optimization step of $\text{NPML}$. The frequencies optimization step of $\text{NPML}$ was performed with an EM algorithm as described by McLachlan &
Peel (2004). NPEM was stopped when:

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

where $\ln L_N(\hat{\theta}_N^{\text{parametric}}, \hat{\mathcal{P}}^{(k)})$ 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:

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

where $\ln L_N(\hat{\theta}_N^{\text{parametric}}, \hat{\mathcal{P}}^{(k)})$ 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 $\theta^*$, $\hat{\theta}_N^{\text{SNP}}$, 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 $P^*$

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 $\hat{P}_N$ (with cdf $\hat{\Pi}_N$) in estimating $P^*$ (with cdf $\Pi^*$) is defined as:

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

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 $\hat{P}_N$ of $P^*$ and $\hat{\theta}_N$ of $\theta^*$:

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

It is noteworthy that $EBe_i(\hat{\theta}_{\text{parametric}}, EBe\text{ED})$ equals the parametric $EBe_i$ defined in section 3.1.

For each dataset, the sum of absolute relative errors between $EBe_i(\hat{\theta}_N, \hat{P}_N)$ and the actual realization $x_i$ of $X_i$, gives the average percentage error (denoted X-error) for the prediction of $X_i$:

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

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:

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

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.

<table>
<thead>
<tr>
<th>N</th>
<th>Simulation 1</th>
<th>Simulation 1</th>
<th>Simulation 2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Rich datasets</td>
<td>Sparse datasets</td>
<td>Sparse datasets</td>
</tr>
<tr>
<td>50</td>
<td>93/100</td>
<td>98/100</td>
<td>38/100</td>
</tr>
<tr>
<td>100</td>
<td>95/100</td>
<td>100/100</td>
<td>51/100</td>
</tr>
<tr>
<td>200</td>
<td>93/100</td>
<td>95/100</td>
<td>53/100</td>
</tr>
<tr>
<td>300</td>
<td>93/100</td>
<td>100/100</td>
<td>58/100</td>
</tr>
<tr>
<td>400</td>
<td>96/100</td>
<td>97/100</td>
<td>69/100</td>
</tr>
</tbody>
</table>

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 ($\hat{\theta}_N^{\text{parametric}}$ for EBEsED, NPNM, NPML and NPEM, $\hat{\theta}_N^{\text{SNP}}$ for SNP) and of the time needed to compute the estimation of $P^*$. 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 $P^*$

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 $P^*$ 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 \( \theta^* \) is dif-
ficient. 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.

<table>
<thead>
<tr>
<th>method</th>
<th>Simulation 1</th>
<th>Simulation 1</th>
<th>Simulation 2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Rich datasets</td>
<td>Sparse datasets</td>
<td>Sparse datasets</td>
</tr>
<tr>
<td>EBEsED*</td>
<td>-2403.3</td>
<td>-1518.6</td>
<td>-7260.7</td>
</tr>
<tr>
<td>NPNM</td>
<td>-2377.9</td>
<td>-1450.5</td>
<td>-7235.9</td>
</tr>
<tr>
<td>NPML</td>
<td>-2376.7</td>
<td>-1450.1</td>
<td>-7183.4</td>
</tr>
<tr>
<td>NPEM</td>
<td>-2376.3</td>
<td>-1449.1</td>
<td>-7166.9</td>
</tr>
</tbody>
</table>

In theory, for a given estimation of $\theta^*$, 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 $P^*$. 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 was more important in the minor than in the major sub-population. Surprisingly, SNP performed equivalently to EBEsED with respect to X-error, even in the minor sub-population.

Figure 6. Boxplots of prediction errors for the clearance (X-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.

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).

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


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

29