Lasso-type estimators for Semiparametric Nonlinear Mixed-Effects Models Estimation

Ana Arribas-Gil, Karine Bertin, Cristian Meza, Vincent Rivoirard

Lasso-type estimators for Semiparametric Nonlinear Mixed-Effects Models Estimation

Ana Arribas-Gil
Departamento de Estadística
Universidad Carlos III de Madrid, Getafe, Spain.
E-mail: aarribas@est-econ.uc3m.es

Karine Bertin
CIMFAV-Facultad de Ingeniería
Universidad de Valparaíso, Valparaíso, Chile.
E-mail: karine.bertib@uv.cl

Cristian Meza
CIMFAV-Facultad de Ingeniería
Universidad de Valparaíso, Valparaíso, Chile.
E-mail: cristian.meza@uv.cl

Vincent Rivoirard
CEREMADE, CNRS-UMR 7534, Université Paris Dauphine, Paris, France.
E-mail:Vincent.Rivoirard@dauphine.fr

Abstract

Parametric nonlinear mixed effects models (NLMEs) are now widely used in biometrical studies, especially in pharmacokinetics research and HIV dynamics models, due to, among other aspects, the computational advances achieved during the last years. However, this kind of models may not be flexible enough for complex longitudinal data analysis. Semiparametric NLMEs (SNMMs) have been proposed by Ke and Wang (2001). These models are a good compromise and retain nice features of both parametric and nonparametric models resulting in more flexible models than standard parametric NLMEs. However, SNMMs are complex models for which estimation still remains a challenge. The estimation procedure proposed by Ke and Wang (2001) is based on a combination of log-likelihood approximation methods for parametric estimation and smoothing splines techniques for nonparametric estimation. In this work, we propose new estimation strategies in SNMMs. On the one hand, we use the Stochastic Approximation version of EM algorithm (Delyon et al., 1999) to obtain exact ML and REML estimates of the fixed effects and variance components. On the other hand, we propose a LASSO-type method to estimate the unknown nonlinear function. We derive oracle inequalities for this nonparametric estimator. We combine the two approaches in a general estimation procedure that we illustrate with simulated and real data.

1 Introduction

We consider the semiparametric nonlinear mixed effects model (SNMM) as defined by Ke and Wang (2001) in which we have n individuals and we observe:

$$y_{ij} = g(\boldsymbol{x}_{ij}, \boldsymbol{\phi}_i, f) + \varepsilon_{ij}, \quad \varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2) \text{ i.i.d.}, \quad i = 1 \ldots, N, \ j = 1, \ldots, n_i \tag{1}$$

where $y_{ij} \in \mathbb{R}$ is the jth observation in the ith individual, $\boldsymbol{x}_{ij} \in \mathbb{R}^d$ is a known regression variable, g is a common known function governing within-individual behaviour and f is an unknown nonlinear function to estimate. The random effects $\boldsymbol{\phi}_i \in \mathbb{R}^p$ satisfy

$$\boldsymbol{\phi}_i \quad = \quad \boldsymbol{A}_i\boldsymbol{\beta} + \boldsymbol{\eta}_i, \quad \boldsymbol{\eta}_i \sim \mathcal{N}(0, \boldsymbol{\Gamma}) \text{ i.i.d.}$$

where $\boldsymbol{A}_i \in \mathcal{M}_{p,q}$ are known design matrices, $\boldsymbol{\beta} \in \mathbb{R}^q$ is the unknown vector of fixed effects and we suppose that ε_{ij} and $\boldsymbol{\eta}_i$ are mutually independent. We use bold letters for vector and matrices.
The parameter of the model is $(\boldsymbol{\theta}, f)$, where $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\Gamma}, \sigma^2)$ belongs to a finite dimensional space whereas f belongs to an infinite dimensional space of functions denoted $\mathcal{H}$.

Ke and Wang (2001) consider the most common type of SNMM in practice, in which g is linear in f conditionally to $\boldsymbol{\phi}_i$,

$$g(\boldsymbol{x}_{ij}, \boldsymbol{\phi}_i, f) = a(\boldsymbol{\phi}_i; \boldsymbol{x}_{ij}) + b(\boldsymbol{\phi}_i; \boldsymbol{x}_{ij}) f(c(\boldsymbol{\phi}_i; \boldsymbol{x}_{ij})), \tag{2}$$

where a, b and c are known functions which may depend on i.

Different formulations of SNMM's have been recently used to model circadian rhythms (Wang and Brown (1996), Wang et al. (2003)), HIV dynamics (Wu and Zhang (2002), Liu and Wu (2007), Liu and Wu (2008)) or gene expression data (Luan and Li (2004)) among other applications.

Example 1 *The following model was proposed by Wang and Brown (1996) to fit human circadian rhythms:*

$$y_{ij} = \mu + \eta_{1i} + \exp(\eta_{2i})\, f\left(x_{ij} - \frac{\exp(\eta_{3i})}{1+\exp(\eta_{3i})}\right) + \varepsilon_{ij}, \qquad \varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2) \text{ i.i.d.}$$
$$\boldsymbol{\eta}_i \sim \mathcal{N}(0, \boldsymbol{\Gamma}) \text{ i.i.d.}$$

for $i = 1 \ldots, N$, $j = 1, \ldots, n_i$, where y_{ij} is the physiological response of individual ith at the jth time point x_{ij}. This model can be written in the general form (1) as:

$$\begin{aligned} y_{ij} &= g(x_{ij}, \boldsymbol{\phi}_i, f) + \varepsilon_{ij}, \quad \varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2) \text{ i.i.d.}, \quad i = 1 \ldots, N, \quad j = 1, \ldots, n_i \\ g(x_{ij}, \boldsymbol{\phi}_i, f) &= \phi_{1i} + \exp(\phi_{2i})\, f\left(x_{ij} - \frac{\exp(\phi_{3i})}{1+\exp(\phi_{3i})}\right) \\ \boldsymbol{\phi}_i &= (1,0,0)^T \mu + \boldsymbol{\eta}_i, \quad \boldsymbol{\eta}_i \sim \mathcal{N}(0, \boldsymbol{\Gamma}) \text{ i.i.d.} \end{aligned}$$

where $\boldsymbol{\phi}_i = (\phi_{1i}, \phi_{2i}, \phi_{3i})^T$ and $\boldsymbol{\eta}_i = (\eta_{1i}, \eta_{2i}, \eta_{3i})^T$. In this example f represents the common shape of the observed curves, and ϕ_{1i}, $\exp(\phi_{2i})$, and $\exp(\phi_{3i})/(1+\exp(\phi_{3i}))$ stand for the individual vertical shift, individual amplitude and individual horizontal shift respectively. Here $d = 1$, $p = 3$, $q = 1$ and the parameter of the model is $(\mu, \boldsymbol{\Gamma}, \sigma^2, f)$. This model was also used by Ke and Wang (2001) for modeling Canadian temperatures at different weather stations.

Let us introduce the following vectorial notations: $\boldsymbol{y}_i = (y_{i1}, \ldots, y_{in_i})'$, $\boldsymbol{y} = (\boldsymbol{y}_1', \ldots, \boldsymbol{y}_N')'$, $\boldsymbol{\phi} = (\phi_1', \ldots, \phi_N')'$, $\boldsymbol{\eta} = (\boldsymbol{\eta}_1', \ldots, \boldsymbol{\eta}_N')'$, $\boldsymbol{g}_i(\boldsymbol{\phi}_i, f) = (g(\boldsymbol{x}_{i1}, \boldsymbol{\phi}_i, f), \ldots, g(\boldsymbol{x}_{in_i}, \boldsymbol{\phi}_i, f))'$, $\boldsymbol{g}(\boldsymbol{\phi}, f) =$

$(\boldsymbol{g}_1(\boldsymbol{\phi}_1, f)', \ldots, \boldsymbol{g}_N(\boldsymbol{\phi}_n, f)')'$, $\boldsymbol{A} = (\boldsymbol{A}_1', \ldots, \boldsymbol{A}_N')'$, $\widetilde{\boldsymbol{\Gamma}} = \text{diag}(\boldsymbol{\Gamma}, \ldots, \boldsymbol{\Gamma})$ and $n = \sum_{i=1}^{N} n_i$. Then, model (1) can be written as:

$$\begin{aligned} \boldsymbol{y}|\boldsymbol{\phi} &\sim \mathcal{N}(\boldsymbol{g}(\boldsymbol{\phi}, f), \sigma^2 \mathbf{I}_n) \\ \boldsymbol{\phi} &\sim \mathcal{N}(\boldsymbol{A}\boldsymbol{\beta}, \widetilde{\boldsymbol{\Gamma}}) \end{aligned} \tag{3}$$

where $\mathbf{I}_n$ represents the identity matrix of dimension n.

The likelihood of observations $\boldsymbol{y}$ is:

$$\begin{aligned} & p(\boldsymbol{y}; (\boldsymbol{\theta}, f)) = \int p(\boldsymbol{y}|\boldsymbol{\phi}; (\boldsymbol{\theta}, f)) p(\boldsymbol{\phi}; (\boldsymbol{\theta}, f)) d\boldsymbol{\phi} \\ = & \int \frac{1}{(2\pi\sigma^2)^{\frac{n}{2}}} \exp\left\{\frac{-1}{2\sigma^2}\|\boldsymbol{y} - \boldsymbol{g}(\boldsymbol{\phi}, f)\|^2\right\} \frac{1}{(2\pi)^{\frac{Np}{2}}|\boldsymbol{\Gamma}|^{\frac{N}{2}}} \exp\left\{\frac{-1}{2}\|\widetilde{\boldsymbol{\Gamma}}^{-1/2}(\boldsymbol{\phi} - \boldsymbol{A}\boldsymbol{\beta})\|^2\right\} d\boldsymbol{\phi} \\ = & \frac{1}{(2\pi)^{\frac{n+Np}{2}}(\sigma^2)^{\frac{n}{2}}|\boldsymbol{\Gamma}|^{\frac{N}{2}}} \int \exp\left\{\frac{-1}{2}\left(\frac{1}{\sigma^2}\|\boldsymbol{y} - \boldsymbol{g}(\boldsymbol{\phi}, f)\|^2 + \|\widetilde{\boldsymbol{\Gamma}}^{-1/2}(\boldsymbol{\phi} - \boldsymbol{A}\boldsymbol{\beta})\|^2\right)\right\} d\boldsymbol{\phi}, \end{aligned} \tag{4}$$

where $\|\cdot\|$ is the L_2 norm. In their seminal paper, Ke and Wang consider a penalized maximum likelihood approach for the estimation of $(\boldsymbol{\theta}, f)$. That is, they propose to solve

$$\max_{\boldsymbol{\theta}, f} \{\ell(\boldsymbol{y}; (\boldsymbol{\theta}, f)) - n\lambda J(f)\} \tag{5}$$

where $\ell(\boldsymbol{y}; (\boldsymbol{\theta}, f))$ is the marginal log-likelihood, $J(f)$ is some roughness penalty and λ is a smoothing parameter. Moreover, they assume that f belongs to some reproducing kernel Hilbert space (RKHS) $\mathcal{H} = \mathcal{H}_1 \oplus \mathcal{H}_2$, where $\mathcal{H}_1$ is a finite dimensional space of functions, $\mathcal{H}_1 = span\{\psi_1, \ldots, \psi_M\}$, and $\mathcal{H}_2$ is a RKHS itself (see Section 2 of Ke and Wang (2001)). Since the nonlinear function f interacts in a complicated way with the random effects and the integral in (4) is intractable, they replace $\ell(\boldsymbol{y}; (\boldsymbol{\theta}, f))$ by a linear Laplace approximation $\tilde{\ell}(\boldsymbol{y}; (\boldsymbol{\theta}, f, \tilde{\boldsymbol{\phi}}))$, where $\tilde{\boldsymbol{\phi}}$ is some convenient value for $\boldsymbol{\phi}$ (see (10) in Section 3.1 of Ke and Wang (2001)). Then, they propose to estimate $(\boldsymbol{\theta}, f)$ with the following iterative procedure:

i) given an estimate of f, get estimates of $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$ by fitting the resultant nonlinear mixed model by linearizing the log-likelihood (replacing ℓ by $\tilde{\ell}$). Indeed, in practice they use the S-PLUS function `nlme`, Pinheiro and Bates (2000), to solve this step.

ii) given an estimate of $\boldsymbol{\theta}$, $\hat{\boldsymbol{\theta}}$, estimate f as the solution to

$$\max_{f \in \mathcal{H}} \left\{\ell(\boldsymbol{y}; (\hat{\boldsymbol{\theta}}, f)) - n\lambda J(f)\right\} \approx \max_{f \in \mathcal{H}} \left\{\tilde{\ell}(\boldsymbol{y}; (\hat{\boldsymbol{\theta}}, f, \tilde{\boldsymbol{\phi}})) - n\lambda J(f)\right\} = \max_{f \in \tilde{\mathcal{W}}_1} \left\{\tilde{\ell}(\boldsymbol{y}; (\hat{\boldsymbol{\theta}}, f, \tilde{\boldsymbol{\phi}})) - n\lambda J(f)\right\},$$

where $\tilde{\mathcal{W}}_1$ is some finite dimensional space whose particular definition depends on the set of points $\{c(\tilde{\boldsymbol{\phi}}_i; \boldsymbol{x}_{ij}), i = 1, \ldots, N,\ j = 1, \ldots, n_i\}$ in which the function f is evaluated. Indeed, since the approximated log-likelihood involves a bounded linear functional, the maximizer in $\mathcal{H}$ of $\tilde{\ell}(\boldsymbol{y}; (\hat{\boldsymbol{\theta}}, f, \tilde{\boldsymbol{\phi}})) - N\lambda J(f)$ belongs to $\tilde{\mathcal{W}}_1$ (see Section 4.1 of Ke and Wang (2001) and Wang (1998)). However, as it is pointed out by Lin and Zhang in their comment to Ke and Wang (2001), the solution to the original problem, namely (5), in such a space $\mathcal{H}$ might not exist, and if it exists, it may lie in an infinite dimensional space and might not be unique. This is the main difference with standard regression models in which the maximizer in $\mathcal{H}$ of the penalized log-likelihood belongs to a finite dimensional space (see Wahba (1990) for instance). This result

also holds for particular nonlinear nonparametric regression models (see Ke and Wang (2004)), but cannot be generally extended to SNMMs because of the interaction between the random effects and the nonlinear function f.
So in fact, the approach of Ke and Wang consists in choosing $\tilde{\mathcal{W}}_1$ as a finite-dimensional approximation of $\mathcal{H}$ to solve (5).
Also, it is important to point out some drawbacks of the approximated methods based on linearization of the log-likelihood, such as the Laplace's approximation used by Ke and Wang. It has been shown that they can produce inconsistent estimates of the fixed effects, in particular when the number of measurements per subject is not large enough (Ramos and Pantula (1995); Vonesh (1996)). In addition, simulation studies have shown unexpected increases in the type I error of the likelihood ratio and Wald tests based on these linearization methods (Ding and Wu (2001)).

In this paper we propose an alternative estimation procedure in SNMMs. On the one hand, for the parametric step we will focus on the maximization of the exact likelihood. We propose to use a stochastic version of the EM algorithm, the so-called SAEM algorithm introduced by Delyon et al. (1999) and extended by Kuhn and Lavielle (2005) for nonlinear mixed models, to estimate $\boldsymbol{\theta}$ without any approximation or linearization. This stochastic EM algorithm replaces the usual E step of EM algorithm (Dempster et al., 1977) by a simulation step and a stochastic procedure, and converges to a local maximum of the likelihood. The SAEM has been proved to be computationally much more efficient than other stochastic algorithms as for example the classical Monte Carlo EM (MCEM) algorithm (Wei and Tanner, 1990) thanks to a recycling of the simulated variables from one iteration to the next (see Kuhn and Lavielle (2005)). Indeed, previous attempts to perform exact ML estimation in SNMMs have been discarded because of the computational problems related to the use of an MCEM algorithm (see Liu and Wu (2007, 2008, 2009)). Moreover we use a Restricted Maximum Likelihood (REML) version of the SAEM algorithm to correct bias estimation problems of the variance parameters following the same strategy as Meza et al. (2007).
On the other hand, for the nonparametric step we will propose a LASSO-type method for the estimation of f. The popular LASSO estimator (least absolute shrinkage and selection operator, Tibshirani (1996)) based on ℓ_1 penalized least squares, has been extended in the last years to nonparametric regression (see for instance Bickel et al. (2009)). It has been also used by Schelldorfer et al. (2011) in high-dimensional linear mixed-effects models. In the nonparametric context, the idea is to reconstruct a sparse approximation of f with linear combinations of elements of a given set of functions $\{f_1, \ldots, f_M\}$, called dictionary. That is, we are implicitly assuming that f can be well approximated with a small number of those functions. In practice, for the nonparametric regression problem, the dictionary can be a collection of basis functions from different bases (splines with fixed knots, wavelets, Fourier, etc.). The advantage of this approach with respect to the penalized maximum likelihood estimation in an approximate space of functions, as proposed by Ke and Wang (2001), is that now the selection of the finite-dimensional space among a large collection of possible spaces spanned by very different functions is automatic and based on data. This approach allows to construct a good approximation of the nonparametric function which is sparse thanks to the large dictionary. The sparsity of the approximation gives a model more interpretable and since few coefficients have to be estimated, this minimizes the estimation error. The LASSO algorithm allows to use the dictionary approach to select a sparse approximation, unlike to wavelet thresholding or ℓ_0- penalization. Moreover the LASSO algorithm has a low computational cost since it is based on a convex penalty.

We can summarize our iterative estimation procedure as:

i) given $\hat{f}$, an estimate of f, get estimates of $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$ by fitting the resulting nonlinear mixed model with the SAEM algorithm (using ML or REML method).

ii) given estimates of $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$, solve the resulting nonparametric regression problem using a LASSO-type method.

In fact, since the SAEM algorithm is an iterative procedure itself, instead of running the whole SAEM algorithm until convergence for each given f at step i), we will rather perform only one iteration of the algorithm in order to update the $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$ estimates from the current value of $\hat{f}$. Then, the nonparametric estimation of step ii) will be performed at each iteration of the SAEM algorithm, as we will see in Section 4.

The rest of the article is organized as follows. In Section 2.1 we describe the SAEM algorithm and its REML version in the framework of SNMMs. In Section 3 we propose a LASSO-type method for the estimation of f in the resulting nonparametric regression problem after estimation of $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$. We derive oracle inequalities and subset selection properties for the proposed estimator. In Section 4, we describe the algorithm that combines both procedures to perform joint estimation of $(\boldsymbol{\theta}, f)$ in the SNMM. Finally, in Section 5, we illustrate our method through simulated and real data. We conclude the article in Section 6. The proofs of the results of Section 3 are in the Appendix.

2 Estimation of the finite-dimensional parameters

2.1 SAEM estimation of $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$

Let us focus on the first point of our procedure, which is performed by the Stochastic Approximation EM algorithm, SAEM (Delyon et al. (1999)). In this subsection we consider that we have an estimate of f, $\hat{f}$, obtained in the previous estimation step that does not change during the estimation of $\boldsymbol{\theta}$. Thus, we can proceed as if f was a known nonlinear function and we fall into the SAEM estimation of nonlinear mixed-effects model framework (see Kuhn and Lavielle (2005)). In fact, note that since the estimation of f is performed by solving a nonparametric regression problem with regression variables $c(\hat{\boldsymbol{\phi}}_i; \boldsymbol{x}_{ij}), i = 1, \ldots, N,\ j = 1, \ldots, n_i$ (see Section 3), it will depend on the estimated value of $\boldsymbol{\phi}$ at the precedent iteration. Then, we will note $\hat{f}_-$ the current estimated function.
The complete likelihood for model (1) is:

$$\begin{aligned} p(\boldsymbol{y}, \boldsymbol{\phi}; \boldsymbol{\theta}) &= p(\boldsymbol{y}|\boldsymbol{\phi}; \boldsymbol{\theta}) p(\boldsymbol{\phi}; \boldsymbol{\theta}) \\ &= \frac{1}{(2\pi\sigma^2)^{\frac{n}{2}}} \exp\left\{\frac{-1}{2\sigma^2}\|\boldsymbol{y} - \boldsymbol{g}(\boldsymbol{\phi}, \hat{f}_-)\|^2\right\} \frac{1}{(2\pi)^{\frac{Np}{2}}|\boldsymbol{\Gamma}|^{\frac{N}{2}}} \exp\left\{\frac{-1}{2}\|\widetilde{\boldsymbol{\Gamma}}^{-1/2}(\boldsymbol{\phi} - \boldsymbol{A}\boldsymbol{\beta})\|^2\right\} \\ &= \frac{1}{(2\pi)^{\frac{n+Np}{2}}(\sigma^2)^{\frac{n}{2}}|\boldsymbol{\Gamma}|^{\frac{N}{2}}} \exp\left\{\frac{-1}{2}\left(\frac{1}{\sigma^2}\|\boldsymbol{y} - \boldsymbol{g}(\boldsymbol{\phi}, \hat{f}_-)\|^2 + \|\widetilde{\boldsymbol{\Gamma}}^{-1/2}(\boldsymbol{\phi} - \boldsymbol{A}\boldsymbol{\beta})\|^2\right)\right\} \end{aligned}$$

where $n = \sum_{i=1}^{N} n_i$. The complete log-likelihood is:

$$\log p(\boldsymbol{y}, \boldsymbol{\phi}; \boldsymbol{\theta}) = \frac{-1}{2}\left\{C + n\log\sigma^2 + N\log|\boldsymbol{\Gamma}| + \frac{1}{\sigma^2}\|\boldsymbol{y} - \boldsymbol{g}(\boldsymbol{\phi}, \hat{f}_-)\|^2 + \|\widetilde{\boldsymbol{\Gamma}}^{-1/2}(\boldsymbol{\phi} - \boldsymbol{A}\boldsymbol{\beta})\|^2\right\} \quad (6)$$

where C is a constant that does not depend on $\boldsymbol{\theta}$.

The principle of the EM algorithm, Dempster et al. (1977), is to maximize at iteration k the conditional expectation of $\log p(\boldsymbol{y}, \boldsymbol{\phi}; \boldsymbol{\theta})$ given the observed data and the precedent value of $\boldsymbol{\theta}$, $\boldsymbol{\theta}^{(k)}$, that is

$$Q_{k+1}(\boldsymbol{\theta}) = \mathbb{E}\left(\log p(\boldsymbol{y}, \boldsymbol{\phi}; \boldsymbol{\theta}) | \boldsymbol{y}; \boldsymbol{\theta}^{(k)}\right).$$

This can be simplified if we assume that the distribution of the complete-data model belongs to the exponential family, that is, if

$$\log p(\boldsymbol{y}, \boldsymbol{\phi}; \boldsymbol{\theta}) = -\Psi(\boldsymbol{\theta}) + \langle S(\boldsymbol{y}, \boldsymbol{\phi}), \Phi(\boldsymbol{\theta}) \rangle$$

where $\langle \cdot, \cdot \rangle$ stands for the scalar product and $S(\boldsymbol{y}, \boldsymbol{\phi})$ is the sufficient statistics of the complete-data model. In that case, the EM algorithm consists in iterating the two following steps:

- E step: evaluate the quantity $s_{k+1} = \mathbb{E}[S(\boldsymbol{y}, \boldsymbol{\phi}) | \boldsymbol{y}; \boldsymbol{\theta}^{(k)}]$.
- M step: update the value of $\boldsymbol{\theta}$: $\boldsymbol{\theta}^{(k+1)} = \arg\max_{\boldsymbol{\theta}} \{-\Psi(\boldsymbol{\theta}) + \langle s_{k+1}, \Phi(\boldsymbol{\theta}) \rangle\}$.

One of the main drawbacks of the EM algorithm is that the computation in the E step is intractable in many cases. The SAEM algorithm replaces, at each iteration, the step E by a simulation step (S) of the missing data ($\boldsymbol{\phi}$) and an approximation step (A) of $Q_{k+1}(\boldsymbol{\theta})$. Then, at iteration k, the SAEM algorithm can be written as:

- S step: simulate m values of the random effects, $\boldsymbol{\phi}^{(k+1,1)}, \ldots, \boldsymbol{\phi}^{(k+1,m)}$, from the conditional law $p(\cdot | \boldsymbol{y}; \boldsymbol{\theta}^{(k)})$.
- A step: update s_{k+1} according to: $s_{k+1} = s_k + \gamma_k \left[\frac{1}{m} \sum_{l=1}^{m} S(\boldsymbol{y}, \boldsymbol{\phi}^{(k+1,l)}) - s_k \right]$.
- M step: update the value of $\boldsymbol{\theta}$: $\boldsymbol{\theta}^{(k+1)} = \arg\max_{\boldsymbol{\theta}} \{-\Psi(\boldsymbol{\theta}) + \langle s_{k+1}, \Phi(\boldsymbol{\theta}) \rangle\}$. (7)

The sequence $\{s_k\}$ is initialized at s_0 and γ_k is a decreasing sequence of positive numbers, as presented by Kuhn and Lavielle (2004), which accelerates the convergence.

For the approximation and the maximization steps, we need to define the quantities s_k. From (6), we have that

$$\log p(\boldsymbol{y}, \boldsymbol{\phi}; \boldsymbol{\theta}) = -\frac{1}{2} \left\{ C + n \log \sigma^2 + N \log |\boldsymbol{\Gamma}| + \frac{1}{\sigma^2} \|\boldsymbol{y} - \boldsymbol{g}(\boldsymbol{\phi}, \hat{f}_-)\|^2 + \sum_{i=1}^{N} (\boldsymbol{\phi}_i - \boldsymbol{A}_i \boldsymbol{\beta})' \boldsymbol{\Gamma}^{-1} (\boldsymbol{\phi}_i - \boldsymbol{A}_i \boldsymbol{\beta}) \right\}.$$

Then, the aproximation step reduces to updating the sufficient statistics for the complete model

$$\begin{aligned}
s_{1,i,k+1} &= s_{1,i,k} + \gamma_k \left[\frac{1}{m} \sum_{l=1}^{m} \boldsymbol{\phi}_i^{(k+1,l)} - s_{1,i,k} \right], \quad i = 1, \ldots, N \\
s_{2,k+1} &= s_{2,k} + \gamma_k \left[\frac{1}{m} \sum_{l=1}^{m} \sum_{i=1}^{N} \boldsymbol{\phi}_i^{(k+1,l)} \boldsymbol{\phi}_i^{(k+1,l)'} - s_{2,k} \right] \\
s_{3,k+1} &= s_{3,k} + \gamma_k \left[\frac{1}{m} \sum_{l=1}^{m} \|\boldsymbol{y} - \boldsymbol{g}(\boldsymbol{\phi}^{(k+1,l)}, \hat{f}_-)\|^2 - s_{3,k} \right].
\end{aligned}$$

Now, $\boldsymbol{\theta}^{(k+1)}$ is obtained in the maximization step as follows:

$$\begin{aligned}
\boldsymbol{\beta}^{(k+1)} &= \left(\sum_{i=1}^{N} \boldsymbol{A}_i' {\boldsymbol{\Gamma}^{(k)}}^{-1} \boldsymbol{A}_i\right)^{-1} \sum_{i=1}^{N} \boldsymbol{A}_i' {\boldsymbol{\Gamma}^{(k)}}^{-1} s_{1,i,k+1} \\
\boldsymbol{\Gamma}^{(k+1)} &= \frac{1}{N}\left(s_{2,k+1} - \sum_{i=1}^{N} \boldsymbol{A}_i \boldsymbol{\beta}^{(k+1)} s_{1,i,k+1}' - \sum_{i=1}^{N} s_{1,i,k+1}\left(\boldsymbol{A}_i \boldsymbol{\beta}^{(k+1)}\right)' + \sum_{i=1}^{N} \boldsymbol{A}_i \boldsymbol{\beta}^{(k+1)}\left(\boldsymbol{A}_i \boldsymbol{\beta}^{(k+1)}\right)'\right) \\
{\sigma^2}^{(k+1)} &= \frac{s_{3,k+1}}{n}.
\end{aligned}$$

When the simulation step cannot be directly performed, Kuhn and Lavielle (2004) propose to combine this algorithm with a Markov Chain Monte Carlo (MCMC) procedure. Then, the simulation step becomes:

- S step: using $\boldsymbol{\phi}^{(k,l)}$, draw $\boldsymbol{\phi}^{(k+1,l)}$ with transition probability $\Pi_{\boldsymbol{\theta}^{(k)}}(\cdot|\boldsymbol{\phi}^{(k,l)})$, $l = 1, \ldots, m$,

that is, $(\boldsymbol{\phi}^{(k+1,1)}), \ldots, (\boldsymbol{\phi}^{(k+1,m)})$ are m Markov chains with transition kernels $\left(\Pi_{\boldsymbol{\theta}^{(k)}}\right)$. In practice, these Markov chains are generated using a Hastings-Metropolis algorithm (see Kuhn and Lavielle (2005) for details).
With respect to the number of chains, the convergence of the whole algorithm to a local maximum of the likelihood is granted even for $m = 1$. Greater values of m can accelerate the convergence, but in practice m is always lower than 10. This is the main difference with the MCEM algorithm, in which very large samples of the random effects have to be generated in order for the algorithm to converge.

2.2 REML estimation of variance components

It is well known that the maximum likelihood estimator of variance components in mixed effects models can be biased downwards because it does not adjust for the loss of degrees of freedom caused by the estimation of the fixed effects. This is also true in the context of SNMMs as Ke and Wang (2001) point out in their paper.
Restricted maximum likelihood (REML), as originally formulated by Patterson and Thompson (1971) in the context of linear models, is a method that corrects this problem by maximizing the likelihood of a set of linear functions of the observed data that contain none of the fixed effects of the model. But this formulation does not directly extend beyond linear models, where in general it is not possible to construct linear functions of the observed data that do not contain any of the fixed effects. However, in the case of nonlinear models, other alternative formulations of REML have been proposed. Here, we will consider the approach of Harville (1974), that consists in the maximization of the likelihood after integrating out the fixed effects. The combination of this REML approach with the SAEM algorithm in the context of nonlinear mixed effects models has been studied recently by Meza et al. (2007). The authors showed the efficiency of the method against purely ML estimation performed by SAEM and against REML estimation based on likelihood approximation methods.

Then, following the ideas of Meza et al. (2007), we will note $\boldsymbol{z} = (\boldsymbol{\phi}, \boldsymbol{\beta})$ the random effects and $\tilde{\boldsymbol{\theta}} = (\boldsymbol{\Gamma}, \sigma^2)$ the new parameter of the model. As in the general case, the simulation step is performed through an MCMC procedure. Here, since we have to draw values from the joint distribution of $(\boldsymbol{\phi}, \boldsymbol{\beta})|\boldsymbol{y}; \tilde{\boldsymbol{\theta}}^{(k)}$, we use a Gibbs scheme, i.e., we iteratively draw values from the conditional distributions of $\boldsymbol{\phi}|\boldsymbol{y}, \boldsymbol{\beta}^{(k)}; \tilde{\boldsymbol{\theta}}^{(k)}$ and $\boldsymbol{\beta}|\boldsymbol{y}, \boldsymbol{\phi}^{(k)}; \tilde{\boldsymbol{\theta}}^{(k)}$. Then, we use again a Hastings-Metropolis algorithm to obtain approximations of these conditional distributions.

Finally, iteration k of the SAEM-REML algorithm for model (3) writes:

- S step: using $\boldsymbol{z}^{(k,l)} = (\boldsymbol{\phi}^{(k,l)}, \boldsymbol{\beta}^{(k,l)})$, simulate $\boldsymbol{z}^{(k+1,l)} = (\boldsymbol{\phi}^{(k+1,l)}, \boldsymbol{\beta}^{(k+1,l)})$, $l = 1, \ldots, m$ with a Metropolis-within-Gibbs scheme.

- A step: update $\tilde{s}_{k+1}$ according to $\tilde{s}_{k+1} = \tilde{s}_k + \gamma_k \left[\frac{1}{m} \sum_{l=1}^{m} \tilde{S}(\boldsymbol{y}, \boldsymbol{z}^{(k+1,j)}) - \tilde{s}_k \right]$, namely:

$$\begin{aligned} \tilde{s}_{1,k+1} &= \tilde{s}_{1,k} + \gamma_k \left[\frac{1}{m} \sum_{l=1}^{m} \sum_{i=1}^{N} \boldsymbol{\eta}_i^{(k+1,l)} \boldsymbol{\eta}_i^{(k+1,l)'} - \tilde{s}_{1,k} \right] \\ \tilde{s}_{2,k+1} &= \tilde{s}_{2,k} + \gamma_k \left[\frac{1}{m} \sum_{l=1}^{m} \| \boldsymbol{y} - \boldsymbol{g}(\boldsymbol{z}^{(k+1,l)}, \hat{f}_-) \|^2 - \tilde{s}_{2,k} \right] \end{aligned} \tag{8}$$

  where $\boldsymbol{\eta}_i^{(k+1,l)} = \boldsymbol{\phi}_i^{(k+1,l)} - A_i \boldsymbol{\beta}^{(k+1,l)}$.

- M step: update the value of $\tilde{\boldsymbol{\theta}}$ by $\tilde{\boldsymbol{\theta}}^{(k+1)} = \arg\max_{\tilde{\boldsymbol{\theta}}} \{ -\Psi(\tilde{\boldsymbol{\theta}}) + \langle \tilde{s}_{k+1}, \Phi(\tilde{\boldsymbol{\theta}}) \rangle \}$, namely:

$$\boldsymbol{\Gamma}^{(k+1)} = \frac{\tilde{s}_{1,k+1}}{N} \qquad \text{and} \qquad \sigma^{2^{(k+1)}} = \frac{\tilde{s}_{2,k+1}}{n}.$$

3 Estimation of the function f using a LASSO-type method

3.1 Estimation procedure

In this part, our objective is to estimate f in the model (1) using the observations $y_{i,j}$ and assuming that for $i = 1, \ldots, N$ we have $\boldsymbol{\phi}_i = \hat{\boldsymbol{\phi}}_i$ and $\sigma^2 = \hat{\sigma}^2$ where the estimates $\hat{\boldsymbol{\phi}}_i$ and $\hat{\sigma}^2$ have been obtained in the precedent SAEM step. Since g satisfies (2), model (1) can be rewritten as

$$\tilde{y}_{ij} = b(\boldsymbol{\phi}_i; \boldsymbol{x}_{ij}) f(\tilde{\boldsymbol{x}}_{ij}) + \varepsilon_{ij}, \quad i = 1 \ldots, N, \ j = 1, \ldots, n_i$$

with $\tilde{y}_{ij} = y_{ij} - a(\boldsymbol{\phi}_i; \boldsymbol{x}_{ij})$ and $\tilde{\boldsymbol{x}}_{ij} = c(\boldsymbol{\phi}_i; \boldsymbol{x}_{ij})$. Of course, since the $\hat{\boldsymbol{\phi}}_i$'s and $\hat{\sigma}^2$ depend on the observations, the distribution of $\hat{\sigma}^{-1} \tilde{y}_{ij}$ is no longer Gaussian. But in the sequel, to be able to derive theoretical results, we still assume that

$$\varepsilon_{ij} \overset{iid}{\sim} \mathcal{N}(0, \sigma^2), \tag{9}$$

where the value of σ^2 is given by $\hat{\sigma}^2$. Simulation studies of Section 5 show that this assumption is reasonable. However, note that (9) is true at the price of splitting the data set into two parts: the first part for estimating $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$, the second part for estimating f. Now, reordering the observations, it is equivalent to observing $(y_1, \ldots, y_n)$ with $n = \sum_{i=1}^{N} n_i$, such that

$$y_i = b_i f(x_i) + \varepsilon_i, \quad \varepsilon_i \sim \mathcal{N}(0, \sigma^2) \text{ i.i.d.} \tag{10}$$

where the b_i's and the design $(x_i)_{i=1,\ldots,n}$ are known and depend on the estimators of the precedent SAEM step and the ε_i's are random variables with variance σ^2 estimated by $\hat{\sigma}^2$. Note that the notation y_i, $i = 1, \ldots, n$, does not correspond to the original observations in the SNMM or to

any of the values introduced in the previous sections, and it is used in this section for the sake of simplicity. Without loss of generality, we suppose that $b_i \neq 0$ for all $i = 1, \ldots, n$.

In the sequel, our objective is then to estimate f nonparametrically in model (10). A classical method would consist in decomposing f on an orthornormal basis (Fourier basis, wavelets,...) and then to use a standard nonparametric procedure to estimate the coefficients of f associated with this basis (ℓ_0-penalization, wavelet thresholding,...). In the same spirit as Bertin et al. (2011) who investigated the problem of density estimation, we wish to combine a more general dictionary approach with an estimation procedure leading to fast algorithms. The dictionary approach consists in proposing estimates that are linear combinations of various types of functions. Typically, the dictionary is built by gathering together atoms of various classical orthonormal bases. This approach offers two advantages. First, with a more wealthy dictionary than a classical orthonormal basis, we aim at obtaining sparse estimates leading to few estimation errors of the coefficients. Secondly, if the estimator is sparse enough, interesting interpretations of the results are possible by using the set of the non-zero coefficients, which corresponds to the set of functions of the dictionary "selected" by the procedure. For instance, we can point out the frequency of periodic components of the signal if trigonometric functions are selected or local peaks if some wavelets are chosen by the algorithm. Both aspects are illustrated in the next sections. ℓ_0-penalization or thresholding cannot be combined with a dictionary approach if we wish to obtain fast and good algorithms. But LASSO-type estimators based on ℓ_1-penalization, leading to minimization of convex criteria, constitute a natural tool for the dictionary approach. Furthermore, unlike ridge penalization or more generally ℓ_p-penalization with $p > 1$, ℓ_1-penalization leads to sparse solutions for the minimization problem, in the sense that if the tuning parameter is large enough some coefficients are exactly equal to 0 (see Tibshirani (1996)).

There is now a very huge literature on LASSO-type procedures. From the theoretical point of view and in the specific context of the regression model close to (10), we mention that LASSO procedures have already been studied by Bunea et al. (2006), Bunea et al. (2007a), Bunea et al. (2007b), Bunea (2008), Bickel et al. (2009), van de Geer (2010), and Bühlmann and van de Geer (2011) among others.

In our setting, the proposed procedure is the following. For $M \in \mathbb{N}^*$, we consider a set of functions $\{\varphi_1, \ldots, \varphi_M\}$, called the *dictionary*. We denote for $\lambda \in \mathbb{R}^M$,

$$f_\lambda = \sum_{j=1}^{M} \lambda_j \varphi_j.$$

Our objective is to find good candidates for estimating f which are linear combinations of functions of the dictionary, i.e. of the form f_λ. We consider, for $\lambda \in \mathbb{R}^M$

$$\text{crit}(\lambda) = \frac{1}{n} \sum_{i=1}^{n} (y_i - b_i f_\lambda(x_i))^2 + 2 \sum_{j=1}^{M} r_{n,j} |\lambda_j|,$$

where $r_{n,j} = \sigma \|\varphi_j\|_n \sqrt{\frac{\gamma \log M}{n}}$ with $\gamma > 0$ and for a function h

$$\|h\|_n^2 = \frac{1}{n} \sum_{i=1}^{n} b_i^2 h^2(x_i).$$

We call the LASSO estimator $\hat{\lambda}$ the minimizer of $\lambda \longmapsto \text{crit}(\lambda)$ for $\lambda \in \mathbb{R}^M$ and we denote $\hat{f} = f_{\hat{\lambda}}$.

The function $\lambda \longmapsto \text{crit}(\lambda)$ is the sum of two terms: the first one is a goodness-of-fit criterion based on the ℓ_2-loss and the second one is a penalty term that can be viewed as the weighted ℓ_1-norm of λ.

Before going further, let us discuss the important issue of tuning. In our context, the tuning parameter is the constant γ. From a theoretical point of view (see Theorem 1), the benchmark value for γ is 2. In the sequel, γ will be chosen satisfying two criteria: to be as close as possible to this benchmark value and allowing the stability of the SAEM algorithm. In Section 5, we will see that sometimes we choose values of γ smaller than 2 but relatively close of it, in particular to obtain the convergence of the variance components estimates, which is always challenging in NLME models.

Once we have chosen a value for γ satisfying these two criteria, the numerical scheme of the nonparametric step is the following:

- Using the estimates of the ϕ_i's and of σ^2 obtained in the previous iteration of SAEM, compute for $i = 1, \ldots, n$, the observations y_i, the constants b_i and the design x_i.
- Evaluate the dictionary $\{\varphi_1, \ldots, \varphi_M\}$ at the design and calculate $r_{n,j}$.
- Obtain the LASSO estimates $\hat{\lambda}$ and $f_{\hat{\lambda}}$.

In practice, there exist many efficient algorithms to tackle this third point, namely, the minimization on λ of $crit(\lambda)$. For the implementation of our estimation procedure we have considered the approach used by Bertin et al. (2011) which consists in using the LARS algorithm.

3.2 Theoretical results

Numerical results of our procedure are presented in next sections but we now validate our approach from a theoretical point of view. More precisely, we consider the oracle approach.

3.2.1 Assumptions

As usual, assumptions on the dictionary are necessary to obtain oracle results for LASSO-type procedures. We refer the reader to van de Geer and Bühlmann (2009) for a good review of different assumptions considered in the literature for LASSO-type estimators and connections between them. The dictionary approach aims at extending results for orthonormal bases. Actually, our assumptions express the relaxation of the orthonormality property. To describe them, we introduce the following notation. For $l \in \mathbb{N}$, we denote

$$\nu_{\min}(l) = \min_{|J| \leq l} \min_{\substack{\lambda \in \mathbb{R}^M \\ \lambda_J \neq 0}} \frac{\|f_{\lambda_J}\|_n^2}{\|\lambda_J\|_{\ell_2}^2} \qquad \text{and} \qquad \nu_{\max}(l) = \max_{|J| \leq l} \max_{\substack{\lambda \in \mathbb{R}^M \\ \lambda_J \neq 0}} \frac{\|f_{\lambda_J}\|_n^2}{\|\lambda_J\|_{\ell_2}^2},$$

where $\|\cdot\|_{\ell_2}$ is the l_2 norm in $\mathbb{R}^M$. The notation λ_J means that for any $k \in \{1, \ldots, M\}$, $(\lambda_J)_k = \lambda_k$ if $k \in J$ and $(\lambda_J)_k = 0$ otherwise. Previous quantities correspond to the "restricted" eigenvalues of the Gram matrix $G = (G_{j,j'})$ with coefficients

$$G_{j,j'} = \frac{1}{n} \sum_{i=1}^{n} b_i^2 \varphi_j(x_i) \varphi_{j'}(x_i).$$

Assuming that $\nu_{\min}(l)$ and $\nu_{\max}(l)$ are close to 1 means that every set of columns of G with cardinality less than l behaves like an orthonormal system. We also consider the restricted correlations

$$\delta_{l,l'} = \max_{\substack{|J|\leq l \\ |J'|\leq l' \\ J\cap J'=\emptyset}} \max_{\lambda,\lambda'\in\mathbb{R}^M \atop \lambda_J\neq 0,\lambda'_{J'}\neq 0} \frac{\langle f_{\lambda_J}, f_{\lambda'_{J'}}\rangle}{\|\lambda_J\|_{\ell_2}\|\lambda'_{J'}\|_{\ell_2}},$$

where $\langle f,g\rangle = \frac{1}{n}\sum_{i=1}^n b_i^2 f(x_i)g(x_i)$. Small values of $\delta_{l,l'}$ means that two disjoint sets of columns of G with cardinality less than l and l' span nearly orthogonal spaces. We will use the following assumption considered in Bickel et al. (2009).

Assumption 1 *For some integer $1 \leq s \leq M/2$, we have*

$$\nu_{\min}(2s) > \delta_{s,2s}. \tag{A1(s)}$$

Oracle inequalities of the Dantzig selector were established under this assumption in the parametric linear model by Candès and Tao (2007) and for density estimation by Bertin et al. (2011). It was also considered by Bickel et al. (2009) for nonparametric regression and for the LASSO estimate.

Let us denote

$$\kappa_s = \sqrt{\nu_{\min}(2s)}\left(1 - \frac{\delta_{s,2s}}{\nu_{\min}(2s)}\right) > 0, \quad \mu_s = \frac{\delta_{s,2s}}{\sqrt{\nu_{\min}(2s)}}.$$

We will say that $\lambda \in \mathbb{R}^M$ satisfies the Dantzig constraints if for all $j = 1,\dots,M$

$$\left|(G\lambda)_j - \hat{\beta}_j\right| \leq r_{n,j}, \tag{11}$$

where

$$\hat{\beta}_j = \frac{1}{n}\sum_{i=1}^n b_i\varphi_j(x_i)Y_i.$$

We denote $\mathcal{D}$ the set of λ that satisfies (11). The classical use of Karush-Kuhn-Tucker conditions shows that the LASSO estimator $\hat{\lambda} \in \mathcal{D}$, so it satisfies the Dantzig constraint.

3.2.2 Oracle inequalities

We obtain the following oracle inequalities.

Theorem 1 *Let $\gamma > 2$. With probability at least $1 - M^{1-\gamma/2}$, for any integer $s < n/2$ such that (A1(s)) holds, we have for any $\alpha > 0$,*

$$\|\hat{f} - f\|_n^2 \leq \inf_{\lambda\in\mathbb{R}^M} \inf_{\substack{J_0\subset\{1,\dots,M\} \\ |J_0|=s}} \left\{\|f_\lambda - f\|_n^2 + \alpha\left(1 + \frac{2\mu_s}{\kappa_s}\right)^2 \frac{\Lambda(\lambda, J_0^c)^2}{s} + 16s\left(\frac{1}{\alpha} + \frac{1}{\kappa_s^2}\right) r_n^2\right\} \tag{12}$$

where

$$r_n = \sup_{j=1,\dots,M} r_{n,j},$$

$$\Lambda(\lambda, J_0^c) = \|\lambda_{J_0^C}\|_{\ell_1} + \frac{\left(\|\hat{\lambda}\|_{\ell_1} - \|\lambda\|_{\ell_1}\right)_+}{2},$$

for any $x \in \mathbb{R}$ $x_+ := \max(x, 0)$ and $\|\cdot\|_{\ell_1}$ is the l_1 norm in $\mathbb{R}^M$.

Theorem 2 *Let $\gamma > 2$. With probability at least $1 - M^{1-\gamma/2}$, for any integer $s < n/2$ such that (A1(s)) holds, we have for any $\alpha > 0$,*

$$\|\hat{f} - f\|_n^2 \leq \inf_{\lambda \in \mathcal{D}} \inf_{\substack{J_0 \subset \{1,\dots,M\} \\ |J_0| = s}} \left\{ \|f_\lambda - f\|_n^2 + \alpha \left(1 + \frac{2\mu_s}{\kappa_s}\right)^2 \frac{\|\lambda_{J_0^C}\|_{\ell_1} + \|\hat{\lambda}_{J_0^C}\|_{\ell_1}}{s} + 32 s \left(\frac{1}{\alpha} + \frac{1}{\kappa_s^2}\right) r_n^2 \right\}. \tag{13}$$

Similar oracle inequalities were established by Bunea et al. (2006), Bunea et al. (2007a), Bunea et al. (2007b), or van de Geer (2010). But in these works, the functions of the dictionary are assumed to be bounded by a constant independent of M and n. Let us comment the right-hand side of inequalities (12) and (13) of Theorems 1 and 2. The first term is an approximation term which measures the closeness between f and f_λ and that can vanish if f is a linear combination of the functions of the dictionary. The second term can be considered as a bias term. In both theorems, the term $\|\lambda_{J_0^C}\|_{\ell_1}$ corresponds to the cost of having λ with a support different of J_0. For a given λ, this term can be minimized by choosing J_0 as the set of largest coordinates of λ. Note that if the function f has a sparse expansion on the dictionary, that is $f = f_\lambda$ where λ is a vector with s non-zero coordinates, then by choosing J_0 as the set of the s non-zero coordinates, the approximation term and the term $\|\lambda_{J_0^C}\|_{\ell_1}$ vanish. In Theorem 1, the term $\left(\|\hat{\lambda}\|_{\ell_1} - \|\lambda\|_{\ell_1}\right)_+$ will be smaller as the ℓ_1-norm of the LASSO estimator is small and this term is equal to 0 if $\|\hat{\lambda}\|_{\ell_1} \leq \|\lambda\|_{\ell_1}$, which is frequently the case. In Theorem 2, given a vector λ such that f_λ approximates well f, the term $\|\hat{\lambda}_{J_0^C}\|_{\ell_1}$ will be small if the LASSO estimator selects the largest coordinates of λ. The last term can be viewed as a variance term corresponding to the estimation of f as linear combination of s functions of the dictionary (see Bertin et al. (2011) for more details). Finally, the parameter α calibrates the weights given for the bias and variance terms.

The following section deals with estimation of sparse functions.

3.2.3 The support property of the LASSO estimate

Let $\gamma > 2$. In this section, we apply the LASSO procedure with $\tilde{r}_{n,j}$ instead of $r_{n,j}$, with

$$\tilde{r}_{n,j} = \sigma \|\varphi_j\|_n \sqrt{\frac{\tilde{\gamma} \log M}{n}}, \quad \tilde{\gamma} > \gamma.$$

We assume that the regression function f can be decomposed on the dictionary: there exists $\lambda^* \in \mathbb{R}^M$ such that

$$f = \sum_{j=1}^{M} \lambda_j^* \varphi_j.$$

We denote S^* the support of λ^*:

$$S^* = \left\{ j \in \{1, \dots, M\} : \quad \lambda_j^* \neq 0 \right\},$$

and by s^* the cardinal of S^*. We still consider the LASSO estimate $\hat{\lambda}$ and, similarly, we denote $\hat{S}$ the support of $\hat{\lambda}$:

$$\hat{S} = \left\{ j \in \{1, \dots, M\} : \quad \hat{\lambda}_j \neq 0 \right\}.$$

One goal of this section is to show that with high probability, we have:

$$\hat{S} \subset S^*.$$

We have the following result.

Theorem 3 *We define*

$$\rho(S^*) = \max_{k \in S^*} \max_{j \neq k} \frac{| < \varphi_j, \varphi_k > |}{\|\varphi_j\|_n \|\varphi_k\|_n}$$

and we assume that there exists $c \in (0, 1/3)$ *such that*

$$s^* \rho(S^*) \leq c.$$

If we have

$$\frac{\sqrt{\tilde{\gamma}} + \sqrt{\gamma}}{\sqrt{\tilde{\gamma}} - \sqrt{\gamma}} \leq \frac{1-c}{2c},$$

then

$$\mathbb{P}\left\{\hat{S} \subset S^*\right\} \geq 1 - 2M^{1-\gamma/2}.$$

A similar result was established by Bunea (2008) in a slightly less general model. However, her result is based on strong assumptions on the dictionary, namely each function is bounded by a constant L (see Assumption (A2)(a) in Bunea (2008)). This assumption is mild when considering dictionaries only based on Fourier bases. It is no longer the case when wavelets are considered and Bunea's assumption is satisfied only in the case where L depends on M and n on the one hand and is very large on the other hand. Since L plays a main role in the definition of the tuning parameters of the method, with too rough values for L, the procedure cannot achieve satisfying numerical results for moderate values of n even if asymptotic theoretical results of the procedure are good. In the setting of this paper, where we aim at providing calibrated statistical procedures, we avoid such assumptions.

Finally, we have the following corollary.

Corollary 1 *We suppose that* $A1(s^*)$ *is satisfied and that there exists* $c \in (0, 1/3)$ *such that*

$$s^* \rho(S^*) \leq c.$$

If we have

$$\frac{\sqrt{\tilde{\gamma}} + \sqrt{\gamma}}{\sqrt{\tilde{\gamma}} - \sqrt{\gamma}} \leq \frac{1-c}{2c},$$

then, with probability at least $1 - 4M^{1-\gamma/2}$,

$$\|\hat{f} - f\|_n^2 \leq \frac{32 s^* \tilde{r}_n^2}{\kappa_{s^*}},$$

where

$$\tilde{r}_n = \sup_{j=1,\ldots,M} \tilde{r}_{n,j}.$$

This corollary is a simple consequence of Theorem 2 with $\lambda = \lambda^*$ and $J_0 = S^*$. Taking $\lambda = \lambda^*$ implies that the approximation term vanishes. Taking $J_0 = S^*$ implies that the bias term vanishes since the support of the LASSO estimator is included in the the support of λ^*. In this case, assuming that $\sup_j \|\varphi_j\|_n < \infty$, the rate of convergence is the classical rate $\frac{s^* \log M}{n}$.

4 Estimation algorithm and inferences

We propose the following estimation procedure for semiparametric estimation of $(\boldsymbol{\theta}, f)$ in model (3), combining the algorithms described in sections 2.1 and 3.1:

Estimation Algorithm - ML version: at iteration k,

- Given the current estimate of $\boldsymbol{\theta}$, $\boldsymbol{\theta}^{(k)} = (\boldsymbol{\beta}^{(k)}, \boldsymbol{\Gamma}^{(k)}, \sigma^{2(k)})$, and m sampled values of the random effects $\boldsymbol{\phi}^{(k,l)}$, $l = 1, \ldots, m$, update the estimates of f, $f^{(k,l)}$, $l = 1, \ldots, m$, with the algorithm described in Section 3.1.

- Given the current estimates of f, $f^{(k,l)}$, $l = 1, \ldots, m$, sample m values of the random effects $\boldsymbol{\phi}^{(k,l)}$, $l = 1, \ldots, m$, and update the value of $\boldsymbol{\theta}$, $\boldsymbol{\theta}^{(k+1)} = (\boldsymbol{\beta}^{(k+1)}, \boldsymbol{\Gamma}^{(k+1)}, \sigma^{2(k+1)})$ with algorithm (7).

(14)

Estimation Algorithm - REML version: at iteration k,

- Given the current estimate of $\tilde{\boldsymbol{\theta}}$, $\tilde{\boldsymbol{\theta}}^{(k)} = (\boldsymbol{\Gamma}^{(k)}, \sigma^{2(k)})$, and m sampled values of the missing data $\boldsymbol{z}^{(k,l)} = (\boldsymbol{\phi}^{(k,l)}, \boldsymbol{\beta}^{(k,l)})$, $l = 1, \ldots, m$, update the estimates of f, $f^{(k,l)}$, $l = 1, \ldots, m$, with the algorithm described in Section 3.1.

- Given the current estimates of f, $f^{(k,l)}$, $l = 1, \ldots, m$, sample m values of the missing data $\boldsymbol{z}^{(k+1,l)} = (\boldsymbol{\phi}^{(k+1,l)}, \boldsymbol{\beta}^{(k+1,l)})$, $l = 1, \ldots, m$, and update the value of $\tilde{\boldsymbol{\theta}}$, $\tilde{\boldsymbol{\theta}}^{(k+1)} = (\boldsymbol{\Gamma}^{(k+1)}, \sigma^{2(k+1)})$ with algorithm (8).

(15)

As it is explained in Section 2.1, for parametric estimation (SAEM or SAEM-REML algorithms alone) the number of chains, m, can be set to 1, which still guarantees the convergence towards a local maximum of the log-likelihood. Higher values of m, may accelerate the convergence of the algorithms (but in practice, m is always lower than 10).
For the global semiparametric estimation procedure, we extend this idea of "parallel chains" of values to the estimation of f. Indeed, at iteration k, the estimation of f depends on the value of the missing data, and thus, from m sampled values $\boldsymbol{z}^{(k,1)}, \ldots, \boldsymbol{z}^{(k,m)}$ we obtain m estimates of f, $f^{(k,1)}, \ldots, f^{(k,m)}$ (see Section 3). Then, in the second step, we use each one of these different estimates of f in parallel to perform parametric estimation (using $f^{(k,l)}$ to sample $\boldsymbol{z}^{(k+1,l)}$ and replacing $\hat{f}_-$ by $f^{(k,l)}$ in (8) for the estimation of $\tilde{\boldsymbol{\theta}}$). This is in the case of the REML version of the algorithm, but the same idea underlies the ML version.

Inferences on model and individual parameters, $\boldsymbol{\beta}, \boldsymbol{\Gamma}, \sigma^2$ and $\boldsymbol{\phi}$, are performed as in NLMEs (see Kuhn and Lavielle (2005) and Meza et al. (2007)). For inferences on the nonlinear function f, we propose an empirical approach based on the fact that our algorithm automatically provides large samples of estimates of f.
Indeed, at each iteration of algorithms (14) and (15) we obtain m estimates of f. The last iterations of the algorithms typically correspond to small values of γ_k in algorithms (7) and (8), see Section 5 for the details. This can be seen as a phase in which the estimates of parameters are stabilized since we assume that convergence has been reached. Let us note by K and $L < K$ the total number of iterations and the number of iterations in the "stabilization phase" of the algorithm. Then, by considering the last $L_0 < L$ iterations of the algorithm, we get a large

sample of estimates of f: $f^{(k,l)}$, $l = 1, \ldots, m$, $k = L_0 + 1, \ldots, K$. These $m \times L_0$ estimates of f are obtained conditionally to values of $\boldsymbol{\theta}$ which are supposed to be close to the corresponding ML or REML estimates. Then, we obtain a point estimate for f as:

$$\hat{f} = \frac{1}{m \times L_0} \sum_{k=K-L_0+1}^{K} \sum_{l=1}^{m} f^{(k,l)} \tag{16}$$

and an empirical pointwise $(1-\alpha)100\%$ confidence interval for $f(x)$ as:

$$\left(\hat{f}(x) - z_{\frac{\alpha}{2}} \sqrt{\frac{S^2_{f(x)}}{m \times L_0}} \ , \ \hat{f}(x) + z_{\frac{\alpha}{2}} \sqrt{\frac{S^2_{f(x)}}{m \times L_0}} \right),$$

where $S^2_{f(x)} = \dfrac{1}{m \times L_0 - 1} \sum_{k=K-L_0+1}^{K} \sum_{l=1}^{m} (f^{(k,l)}(x) - \hat{f}(x))^2$ and $z_{\frac{\alpha}{2}}$ is the $1 - \frac{\alpha}{2}$ percentile of a standard normal distribution. This interval is of course not a true $(1-\alpha)100\%$ confidence interval for $f(x)$ but constitutes a good approximation of it. In the same way, we can also construct approximated confidence intervals for the expected response at a given point for a given individual. This approach is an alternative to the Bayesian confidence intervals proposed by Ke and Wang (2001). The idea is similar to bootstrap confidence intervals, with the advantage that here, the samples of estimates are automatically generated by the estimation algorithm.

5 Application to synthetic and real data

5.1 First simulation study: parametric estimation

As a first step, we want to validate through simulation our parametric estimation strategy alone, based on the SAEM algorithm, and to compare it, in the framework of SNMMs, to the approximate method *nlme* of Ke and Wang (2001). In order to be able to asses only the differences induced by the use of different parametric estimation algorithms, we will use the same nonparametric estimation algorithm for the estimation of f, namely the procedure proposed by Ke and Wang (2001). In the next section we will compare the whole versions, including nonparametric estimation, of both approaches.
To this end, we realized the following simulation study. As in Example 1, data were generated from the model:

$$y_{ij} = \phi_{1i} + \exp(\phi_{2i}) 2 f\left(\frac{j}{N} - \frac{\exp(\phi_{3i})}{1+\exp(\phi_{3i})}\right) + \varepsilon_{ij}, \quad i = 1, \cdots, N, \ j = 1, \cdots, J,$$

where $\varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$ and $\phi_i = (\phi_{1i}, \phi_{2i}, \phi_{3i})^T \sim \mathcal{N}(\mu, \Gamma)$ with $\mu = (\mu_1, \mu_2, \mu_3)^T$. Here, the nonlinear function was set to $f(t) = \sin(2\pi t)$. The following parameter values were used for simulation:

$$N = J = 10,\ \mu = (1, 0, 0)^T,\ \sigma^2 = 1 \text{ and } \Gamma \text{ is diagonal with } diag(\Gamma) = (1, 0.25, 0.16).$$

These data were analyzed using two semiparametric procedures: our SAEM based method combined with the nonparametric algorithm of Ke and Wang's (called *semi-SAEM*) and Ke and Wang's procedure for semiparametric models (called *snm*). For the SAEM algorithm, we used 80 iterations and the following sequence (γ_k): $\gamma_k = 1$ for $1 \leq k \leq 50$ and $\gamma_k = 1/(k-50)$

	Method	μ_1	μ_2	μ_3
True Value		1	0	0
Mean	semi-SAEM	1.06	0.31	0.27
	snm	1.05	0.26	-0.01
MSE	semi-SAEM	0.12	0.16	0.10
	snm	0.12	0.11	0.01
95 % C.I.	semi-SAEM	[0.99;1.12]	[0.27;0.36]	[0.23;0.30]
	snm	[0.99;1.12]	[0.22;0.30]	[-0.02;0.01]

Table 1: ML procedure: Mean, MSE and 95% confidence interval of mean components.

	Method	γ_1	γ_2	γ_3	σ^2
True Value		1	0.25	0.16	1
Mean	semi-SAEM	0.86	0.24	0.16	0.95
	snm	0.89	0.19	0.14	0.99
MSE	semi-SAEM	0.22	0.02	0.01	0.03
	snm	0.22	0.02	0.01	0.03
95 % C.I.	semi-SAEM	[0.77;0.95]	[0.21;0.27]	[0.14;0.17]	[0.92;0.98]
	snm	[0.80;0.98]	[0.17;0.21]	[0.13;0.16]	[0.96;1.02]

Table 2: ML procedure: Mean, MSE and 95% confidence interval of variance components obtained with *semi-SAEM* and *snm*.

for $51 \leq k \leq 80$. We also considered $m = 5$ chains in each iteration. For the nonparametric estimation algorithm common to both procedures, following Ke and Wang (2001) we considered that f is periodic with period equal to 1 and $\int_0^1 f = 0$, i.e. $f \in W_2^0(per) = W_2(per) \ominus span\{1\}$ where $W_2(per)$ is the periodic Sobolev space of order 2 in L^2 and $span\{1\}$ represents the set of constant functions.
The same initial values were used for both methods:

$$\mu_0 = (1, 0, 0),\ \sigma_0^2 = 2 \text{ and } \text{diag}(\Gamma_0) = (\gamma_1^0, \gamma_2^0, \gamma_3^0) = (1, 0.3, 0.1).$$

Tables 1 and 2 sumarize the performance of both methods over 100 simulated data sets. For each parameter we show the sample mean, the mean squared error ($MSE(\hat{\theta}) = \frac{1}{100}\sum_{i=1}^{n}(\theta - \hat{\theta}_i)^2$), and a 95% confidence interval computed over the total number of simulations.

We also compared the REML estimates obtained with our method and with *snm* (using the REML version of *nlme*) for the same simulated data sets. The results are summarized in Table 3. It can be seen that the mean values for the REML estimates obtained with both procedures were closer to the simulated values, especially for parameters γ_1. Moreover, the individual confidence intervals of REML estimates of this parameter, at a 95% level, include the true value for these parameters on the contrary to the ML estimates, showing that REML versions of the algorithms were able to correct the bias observed with ML. If we compare our method and *snm*, for both procedures ML and REML, we obtained results that are similar but it seems that our REML estimates are closer to the simulated values than those obtained with Ke and Wang's method.

An important issue to discuss is the convergence of estimates with this kind of iterative

	Method	γ_1	γ_2	γ_3	σ^2
True Value		1	0.25	0.16	1
Mean	semi-SAEM	0.99	0.25	0.16	0.95
	snm	0.92	0.19	0.15	1.02
MSE	semi-SAEM	0.21	0.03	0.01	0.03
	snm	0.23	0.02	0.01	0.03
95 % C.I.	semi-SAEM	[0.89;1.08]	[0.22;0.28]	[0.14;0.18]	[0.92;0.98]
	snm	[0.83;1.02]	[0.17;0.22]	[0.13;0.17]	[0.98;1.05]

Table 3: REML procedure: Mean, MSE and 95% confidence interval of variance components obtained with *semi-SAEM* and *snm*.

maximization algorithms. It is well known that approximate methods for maximum likelihood estimation often present numerical problems and even fail to converge in the framework of NLME estimation (see (Hartford and Davidian, 2000) for instance). An advantage of the exact likelihood method is exactly to avoid those convergence problems as it was established by Kuhn and Lavielle (2005). In this simulation study, we have to say that both *semi-SAEM* and *snm* achieved convergence for all the data sets. However, we also tried to fit a nonlinear mixed effects model to the simulated data, that is, assuming that f was known and estimating only the fixed and random effects with *SAEM* and *nlme*, and in that case the second algorithm failed to converge for several data sets. It seems that in this case the combination of *nlme* with a nonparametric algorithm to perform semiparametric estimation solves the numerical problems encountered by *nlme* on its own. However, this is not true in general as we will see in the next simulation study.

5.2 Second simulation study: semiparametric estimation

In order to test our LASSO-based estimator, we modified the model introduced in section 5.1 as follows:

$$y_{ij} = \phi_{1i} + \exp(\phi_{2i}) 2 f\left(\frac{j}{N} - \frac{\exp(\phi_{3i})}{1+\exp(\phi_{3i})}\right) + \varepsilon_{ij}, \quad i = 1, \cdots, N,\ j = 1, \cdots, J,$$

where $\varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$ and $\phi_i = (\phi_{1i}, \phi_{2i}, \phi_{3i})^T \sim \mathcal{N}(\mu, \Gamma)$ with $\mu = (\mu_1, \mu_2, \mu_3)^T$. Here, $f(\cdot)$ is a mixture of one trigonometric function and two Laplace densities (see Figure 1).

$$f(t) = 0.6 \times \sin(2\pi t) + 0.2 \times \left(\frac{e^{-40|t-0.75|}}{2 \times \int_0^1 e^{-40|t-0.75|}}\right) + 0.2 \times \left(\frac{e^{-40|t-0.8|}}{2 \times \int_0^1 e^{-40|t-0.80|}}\right).$$

Data were simulated using the following parameters:

$N = 10$, $J = 20$, $\mu = (1, 0, 0)^T$, $\sigma^2 = 0.4$ and Γ is diagonal with $diag(\Gamma) = (0.25, 0.16, 0.04)$.

Now data were analyzed using the two following semiparametric procedures: our SAEM and LASSO based method (called *LASSO-SAEM*) and Ke and Wang's procedure for semiparametric models, still denoted *snm*. For both methods we obtained the REML estimates of parameters. It is necessary to specify several values in order to run our algorithm, such as the choice of

the LASSO's tunning parameter γ and the inputs of the SAEM algorithm (initial values, step sizes γ_k, number of chains in the MCMC step, number of burn-in iterations, and total number of iterations). For the latter, we used again 80 iterations with $\gamma_k = 1$ for $1 \leq k \leq 50$ and $\gamma_k = 1/(k-50)$ for $51 \leq k \leq 80$, and we considered $m = 5$ chains in each iteration. The initial values, which were also used with *snm*, were:

$\mu_0 = (1, 0, 0)$, $\sigma_0^2 = 2$ and $\text{diag}(\Gamma_0) = (\gamma_1^0, \gamma_2^0, \gamma_3^0) = (1, 0.3, 0.1)$.
The nonparametric LASSO step has been performed with $\gamma = 1/3$. Larger values of γ did not allow, for some datasets, to stabilizing the convergence of some parameters, in particular the variance γ_2, and smaller values of γ provided similar results to the one presented here. The dictionary chosen combined very different orthonormal families, namely Fourier functions with Haar wavelets, which ensured a sufficiently incoherent design in the spirit of Section 3. More precisely, our dictionary was composed by the following Fourier functions $\{t \mapsto 1; t \mapsto \cos(\pi t); t \mapsto \sin(\pi t); t \mapsto \cos(2\pi j t), t \mapsto \sin(2\pi j t), j = 1, \cdots, 5\}$ and by the Haar wavelet basis with resolution between 2^4 and 2^7, with a total size of 245 functions. Note that the data $\tilde{\boldsymbol{x}}_{ij} = c(\boldsymbol{\phi}_i; \boldsymbol{x}_{ij})$ belongs approximately to $[-0.4, 1.6]$. For *snm*, it seemed reasonable to consider that $f \in W_2^0(per)$ since if we look at a simulated data set (see Figure 3 for example), we can see clearly the periodic structure in the data.
In Figures 2 and 3, we can see the estimates of f and the fitted data with the two methods for a specific simulated data set.

Results for REML estimates obtained with *LASSO-SAEM* and *snm* for 100 simulated data sets are summarized in Table 4. We can see that the means of the estimates obtained with our method are close to their real values except for the variance of the error, σ^2, since our method tends to overestimate that parameter. However, we get overall better results than using the *snm* methodology (except for γ_1).

	Method	γ_1	γ_2	γ_3	σ^2
True Value		0.25	0.16	0.04	0.4
Mean	LASSO-SAEM	0.18	0.14	0.03	0.69
	snm	0.21	0.11	0.03	0.90
MSE	LASSO-SAEM	0.01	0.01	4.0e-4	0.12
	snm	0.02	0.01	5.9e-4	0.27
95 % C.I.	LASSO-SAEM	[0.16;0.20]	[0.12;0.15]	[0.030;0.037]	[0.66;0.73]
	snm	[0.18;0.25]	[0.09;0.14]	[0.028;0.042]	[0.86;0.94]

Table 4: REML procedure: Mean, MSE and 95% confidence interval of variance components obtained with *LASSO-SAEM* and *snm*.

An important issue for this kind of problem is the estimation of the nonlinear function f. Then, to evaluate the accuracy of the estimation, we calculated the Integrated Square Error (ISE) of $\hat{f}$ for each simulated data set. Figure 4 provides a summary of estimates of f using *LASSO-SAEM* and *snm*. We computed the ISE for each estimate of f and plotted the estimates corresponding to (a) the minimum, (b) 1/4 quantile, (c) median, (d) 3/4 quantile and (e) maximum ISEs. We can see that our method outperforms *snm* in the estimation of f, in the sense that our estimates are able to detect the presence of the peaks in the original function.
As for the functions of the dictionary selected with our LASSO method, it is interesting to note

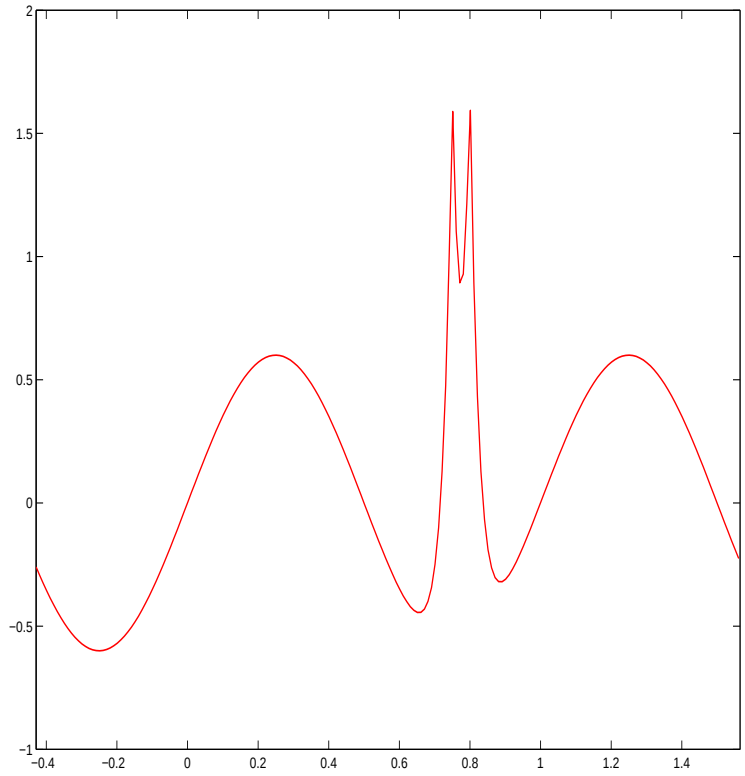

Figure 1: Real function f in the semiparametric simulation study.

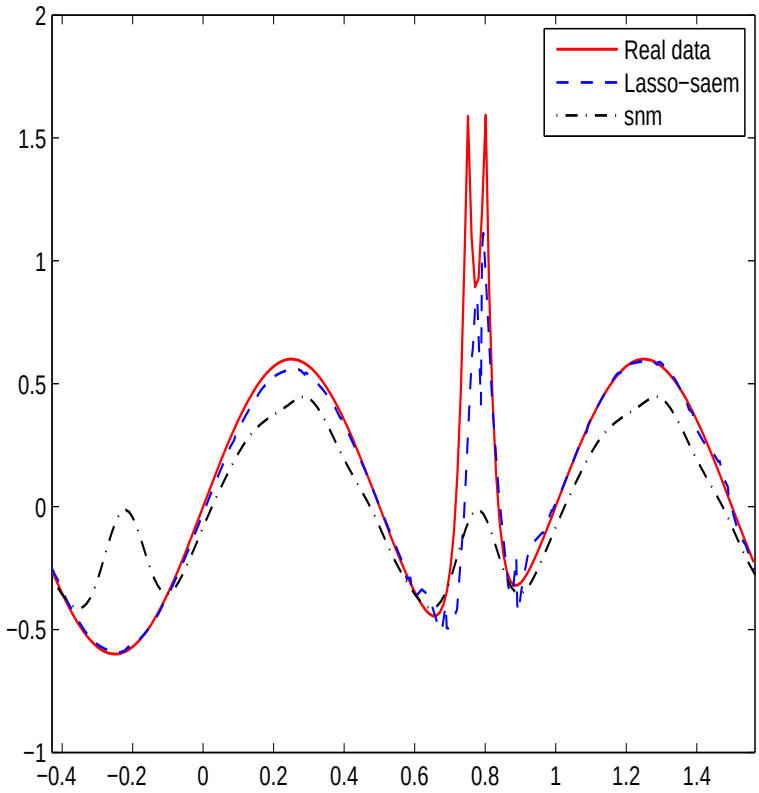

Figure 2: Real function f (solid line) and its estimates obtained with *LASSO-SAEM* (dashed line) and *snm* (dash-dotted line) for a particular data set in the semiparametric simulation study.

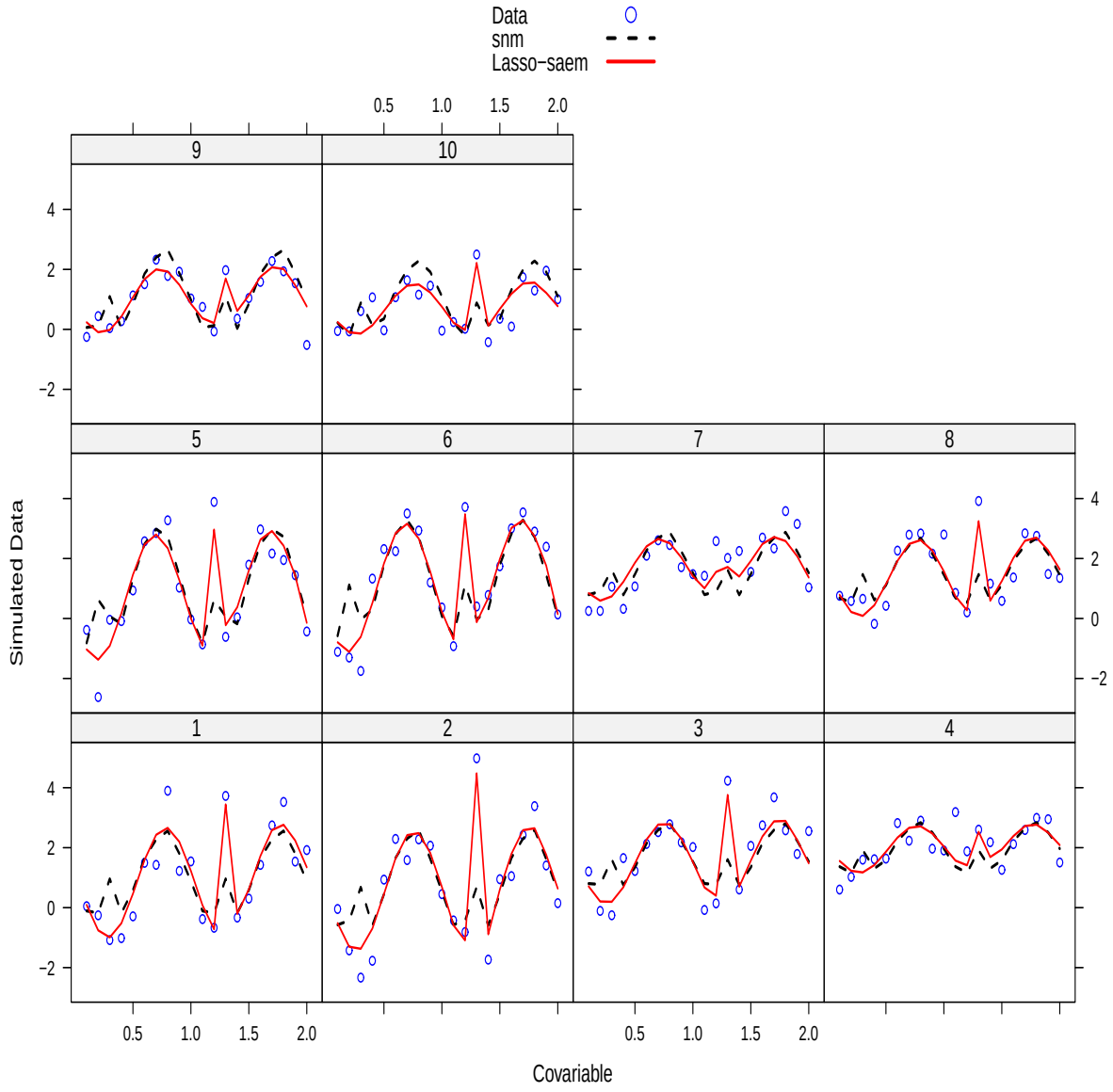

Figure 3: Simulated data and fitted curves obtained with *LASSO-SAEM* (solid line) and *snm* (dashed line) for a particular data set in the semiparametric simulation study.

that the 100 linear combinations of functions of the dictionary obtained for each one of the 100 data sets have a length which varies between 10 and 32 functions, with an average length equal to 20. Furthermore, in 98% of the cases, the method selects the function $\sin(2\pi t)$ with the highest coefficient. For the remaining two data sets, the functions $\sin(6\pi t)$ and $\sin(10\pi t)$ are selected. For all the replicates, in addition to these sine functions, the rest of the selected functions are related to the Haar wavelets with smaller coefficients.

It is very important to point out that the results obtained with *snm* are based only on 51 data sets since the function did not reach convergence in 46 data sets and in other 3 data sets we obtained incoherent estimation of the nonlinear function, when using the default setup of the *snm* algorithm (REML estimation and Generalized Cross Validation for the choice of the penalized parameter). By contrast, our method achieved convergence for all simulated data sets with the specific setup used here (choice of γ, initial values, number of chains, step sizes γ_k, number of iterations, etc ...).

5.3 Application to on-line auction data

Modelling of price paths in on-line auction data has received a lot of attention in the last years (Shmueli and Jank, 2005; Jank and Shmueli, 2006; Shmueli et al., 2007; Liu and Müller, 2008). One of the reasons is the availability of huge amounts of data made public by the on-line auction and shopping website eBay.com, which has become a global market place in which millions of people worldwide buy and sell products. The price evolution during an auction can be thought as a continuous process which is observed discretely and sparsely only at the instants in which bids are placed. In fact, bids tend to concentrate at the beginning and at the end of the auction, responding to two typically observed phenomena, "early bidding" and "bid sniping" (a situation in which "snipers" place their bids at the very last moment).

To our knowledge, Reithinger et al. (2008) provide the first attempt to model price paths taking into account the dependence among different auctions. This is an important consideration, since in practice bidders can participate in different auctions that take place simultaneously. They propose a semiparametric additive mixed model with a boosting estimation approach. In the same line, but considering a more complex interaction of the random effects and the unknown nonlinear function, we propose the following shape-invariant model for the price paths:

$$y_{ij} = \phi_{1i} + \exp(\phi_{2i}) f(t_{ij} - \phi_{3i}) + \varepsilon_{ij}, \quad i = 1, \cdots, N,\ j = 1, \cdots, n_i,$$

where $\varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$ and $\phi_i = (\phi_{1i}, \phi_{2i}, \phi_{3i})^T \sim \mathcal{N}(\mu, \Gamma)$ with $\mu = (\mu_1, \mu_2, \mu_3)^T$. We introduce an individual random horizontal shift, ϕ_{3i}, to model the possible delay of the price dynamics in some auctions with respect to the rest.

We analyzed a set of 183 eBay auctions for Palm M515 Personal Digital Assistants (PDA), of a fixed duration of seven days, that took place between March and May, 2003. This is the dataset used in Reithinger et al. (2008) and it is publicly available at
`http://www.rhsmith.umd.edu/digits/statistics/data.aspx`. We were interested in modelling the live bids, that is, the actual prices that are shown by eBay during the live auction. Note that these are different from the bids placed by bidders during the auction, which are the prices recorded in the bid history published by eBay after the auction closes. Then, a transformation on the bid records is required to recover the live bids (see Shmueli and Jank (2005) for details).

The live bids range from \$0.01 to \$300 and form a sequence of non decreasing prices for each auction. We typically observe between 10 and 30 bids per auction, although there are auctions

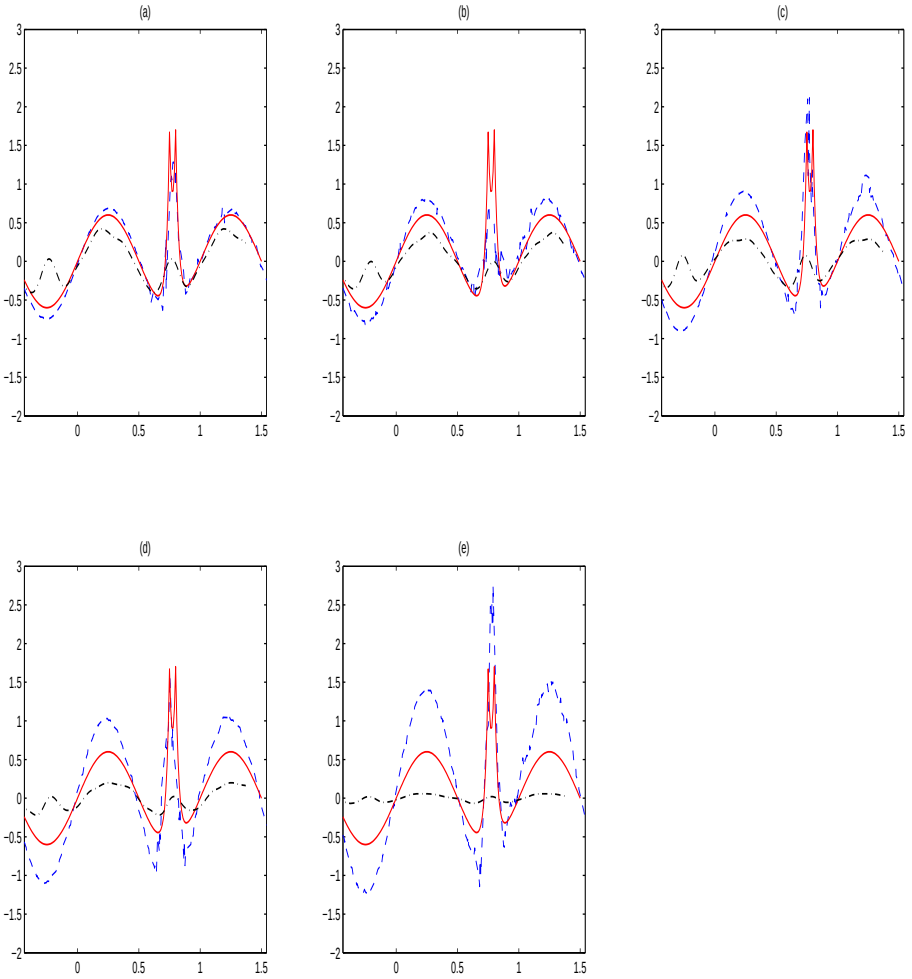

Figure 4: Estimated functions corresponding to the five quantiles of ISE ((a) minimum, (b) 1/4 quantile, (c) median, (d) 3/4 quantile and (e) maximum) obtained with *LASSO-SAEM* (dashed line) and *snm* (dash-dotted line) compared to the true function f (solid line) for the total of the 100 simulated data sets in the semiparametric simulation study.

	ϕ_1	ϕ_2	ϕ_3	
Mean	1.04	0.18	-0.06	
Correlation	1 (7.68)	-0.02	0.41	ϕ_1
Matrix	-0.02	1 (0.19)	0.37	ϕ_2
(variances)	0.41	0.37	1 (0.23)	ϕ_3
σ^2		1.93		

Table 5: Estimated mean vector and covariance matrix of the random effects and estimated error variance in the on-line auction dataset.

with only two bids. We have a total of 3280 bids for the 183 auctions. Following Reithinger et al. (2008), we considered the square root of live bids to reduce the price variability. We run the REML version of our *LASSO-SAEM* algorithm, of which we performed 100 iterations with the following sequence of decreasing steps (γ_k): $\gamma_k = 1$ for $1 \leq k \leq 60$ and $\gamma_k = 1/(k-60)$ for $61 \leq k \leq 100$. We also considered $m = 3$ chains in each iteration. The dictionary for nonparametric estimation was composed by a combination of B-splines of degrees three and four, with 17 knots unequally spaced so that most of the knots were in those places with more data observed (at the beginning, at the end and at the middle of the interval), 10 power functions, 10 exponential functions and 5 logit functions, with a total size of 64 functions. The estimate of f is monotone, as expected by the nature of the data, and presents two steepest parts at the beginning and at the end of the interval. At each iteration of the algorithm the estimated function at the nonparametric step is a sparse combination of the functions of the dictionary. In fact, the set of functions selected by the LASSO method at the last iterations of the algorithm is almost constant, containing mainly two functions, $\varphi(x) = x^{0.35}$ and $\varphi(x) = \exp(0.9\,x)$, and in some iterations a small component of a cubic B-spline around the middle of the interval. In Figure 5 we present the last 24 estimates $f^{(k,l)}$ from which we have obtained $\hat{f}$ as in (16), and $\hat{f}$, together with a 95% pointwise confidence band. These results have been obtained with $\gamma = 2$ as the value for the tunning parameter in the LASSO estimation step.
The estimates for μ and Γ are presented in Table 5. In Figure 6 we present the observed live bids and the model fit for 18 chosen auctions with different price profiles. We can appreciate how the fitted model provides in general an accurate fit of the final price, even in the cases when "bid sniping" is present.

6 Conclusions and discussion

Semiparametric nonlinear mixed effects models cover a wide range of situations and generalize a large class of models, such as nonlinear mixed effects models or self-modelling nonlinear regression models among others. We have proposed a new approach for estimation in SNMMs combining an exact likelihood estimation algorithm with a LASSO-type procedure. Our strategy relies on an iterative procedure to estimate $\boldsymbol{\theta}$ conditioned on f and vice versa, which allow us to tackle the parametric and the nonparametric problem independently. This makes possible the use of fast algorithms providing an accurate and computationally efficient estimation method.
Concerning parametric estimation, our simulation results illustrate our method and point out some important advantages of using an exact likelihood estimation algorithm instead of likelihood approximation methods, such as convergence of the estimates. The REML version of our algorithm, corrects the estimation of variance components accounting for the loss of degrees of

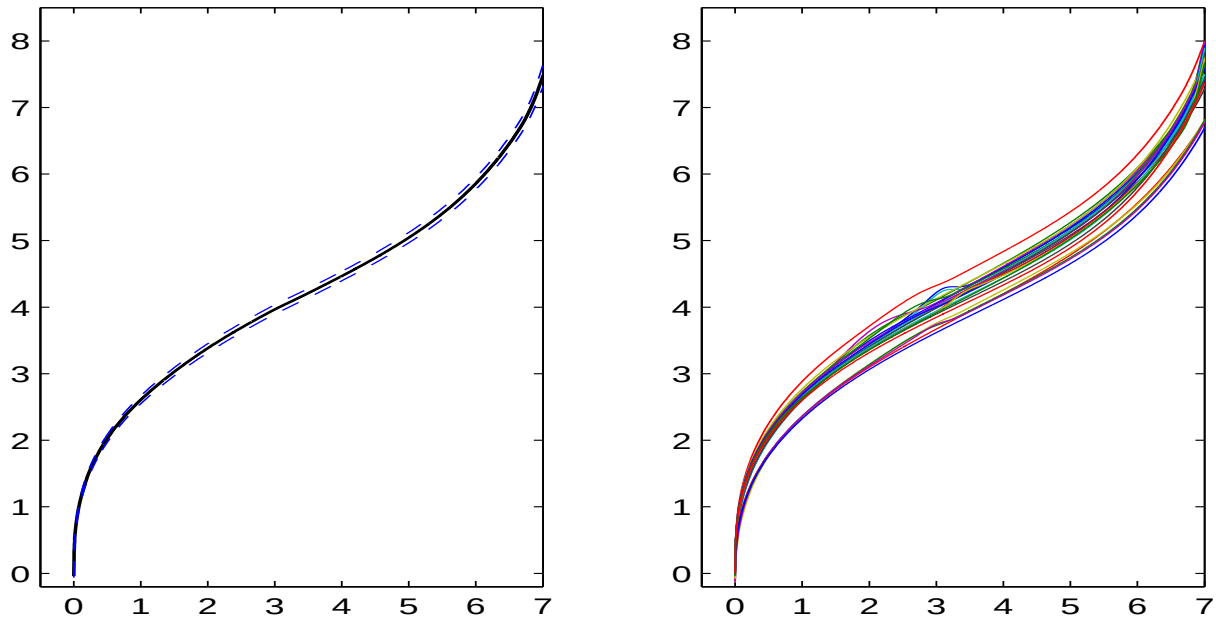

Figure 5: Left: Estimated nonlinear function $\hat{f}$ (solid line) and 95% confidence band (dashed lines) in the on-line auction dataset. Right: Last 24 LASSO estimates whose empirical mean provides $\hat{f}$.

freedom from estimating the fixed effects and provide satisfactory results. However, as it was already pointed out in the comments to Ke and Wang (2001), it will be important to define a REML estimator that can also take into account the loss of degrees of freedom from estimating the nonlinear function f.

As for nonparametric estimation, the dictionary approach allows us to obtain interesting interpretation with respect to the functions of the dictionary selected by the procedure. For instance, we can detect trends, frequencies of sinusoids or location and heights of peaks of the common shape represented by the estimated function f. We have observed that our LASSO estimate achieves good theoretical and numerical results if the dictionary is wealthy and incoherent enough. From the theoretical point of view, incoherence is expressed, in this paper, by Assumption A1(s) or by the quantity $\rho(S^*)$ defined in Section 3.2.3. These incoherence assumptions are hard to check in practice and we do not know if they can be relaxed in our setting. We mention that the method is quite sensitive to the choice of the dictionary. Indeed, in our application to on-line auction data we have detected that differences in the size of the dictionary, but not necessarily in the nature of the function families therein included, may lead to slightly different estimated functions, in the sense that we may obtain rougher and smoother versions of a similar function.

In Section 3, the particular structure of the observations (where we have n_i observations for each individual i) is not used for applying the standard LASSO-procedure. But a natural and possible extension of this work would be to take into account this structure and then to apply a more sophisticated LASSO-type procedure inspired, for instance, by the group-LASSO proposed by Yuan and Lin (2006) to achieve better results. This is a challenging research axis we wish to investigate from a theoretical and practical point of view.

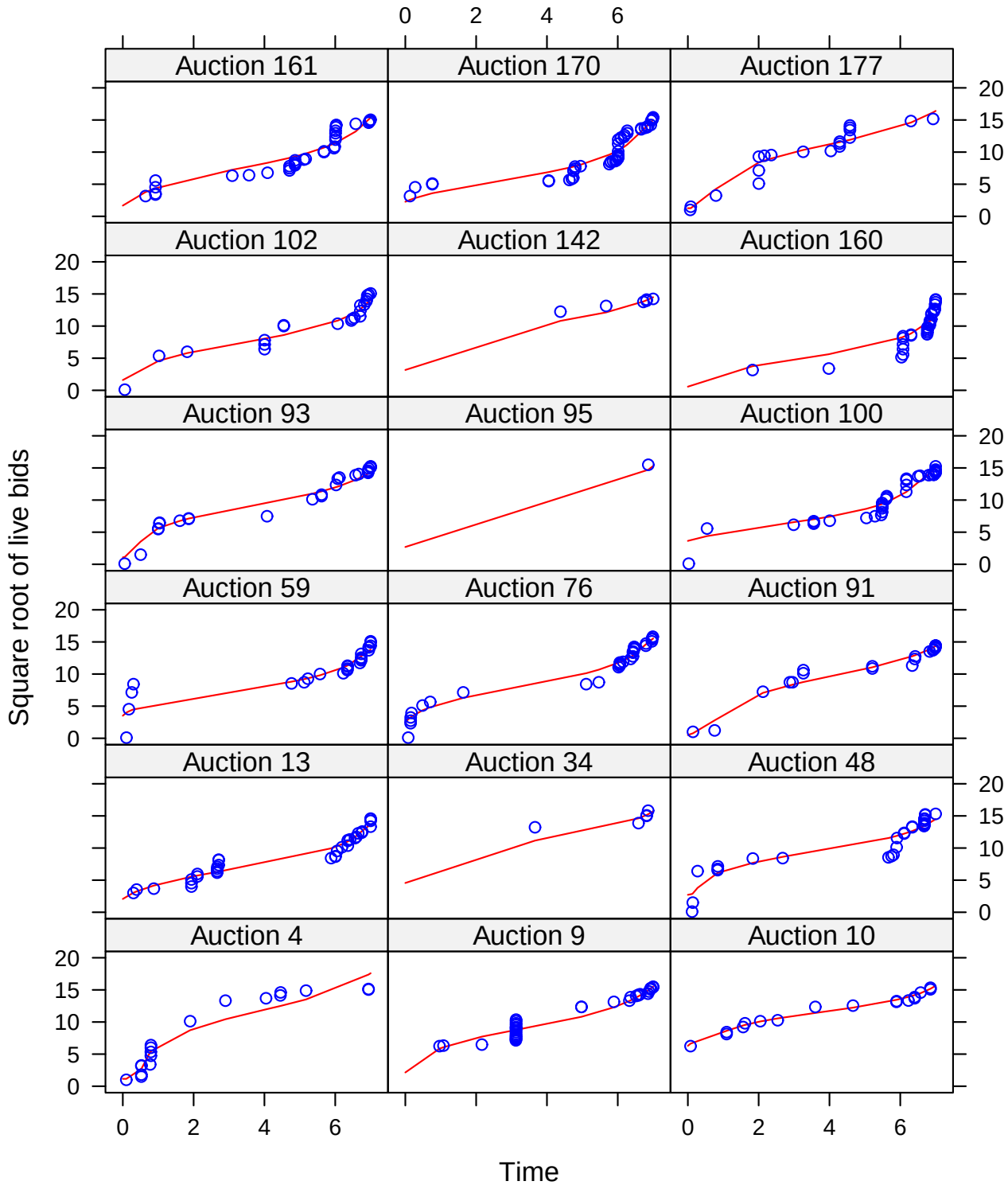

Figure 6: Observed live bids (circles) and fitted price curves (solid lines) for a subset of 18 auctions.

Among other possible extensions of this work, a very promising one would be the use of the nonparametric techniques herein described for density estimation (in the spirit of (Bertin et al., 2011)) of the ramdom errors, assuming that they do not need to be normal. Indeed, the recent work of Comte and Samson (2012) deals with this problem in the case of a linear mixed effects model. Its generalization to NLMEs or even SNMMs is a real challenge.

Appendix. The proofs

Preliminary lemma

Lemma 1 *For $1 \leq j \leq M$, we consider the event $\mathcal{A}_j = \{|V_j| < r_{n,j}\}$ where $V_j = \frac{1}{n}\sum_{i=1}^n b_i \varphi_j(x_i)\varepsilon_i$. Then,*

$$\mathbb{P}\left(\mathcal{A}_j\right) \geq 1 - M^{-\gamma/2}.$$

Proof of Lemma 1: We have

$$\begin{aligned}
\mathbb{P}\left(\mathcal{A}_j^c\right) &\leq \mathbb{P}\left(\sqrt{n}|V_j|/(\sigma\|\varphi_j\|_n) \geq \sqrt{n} r_{n,j}/(\sigma\|\varphi_j\|_n)\right) \\
&\leq \mathbb{P}\left(|Z| \geq \sqrt{\gamma \log M}\right) \\
&\leq M^{-\gamma/2}
\end{aligned}$$

where Z is a standard normal variable. □

Proof of Theorem 1

Let $\lambda \in \mathbb{R}^M$ and J_0 such that $|J_0| = s$. We have

$$\|f_\lambda - f\|_n^2 = \|\hat{f} - f\|_n^2 + \|f_\lambda - \hat{f}\|_n^2 + \frac{2}{n}\sum_{i=1}^n b_i^2 \left(\hat{f}(x_i) - f(x_i)\right)\left(f_\lambda(x_i) - \hat{f}(x_i)\right).$$

We have $\|f_\lambda - \hat{f}\|_n^2 = \|f_\Delta\|_n^2$ where $\Delta = \lambda - \hat{\lambda}$. Moreover

$$A = \frac{2}{n}\sum_{i=1}^n b_i^2 \left(\hat{f}(x_i) - f(x_i)\right)\left(f_\lambda(x_i) - \hat{f}(x_i)\right) = 2\sum_{j=1}^M (\lambda_j - \hat{\lambda}_j)\left[(G\hat{\lambda})_j - \beta_j\right],$$

where

$$\beta_j = \frac{1}{n}\sum_{i=1}^n b_i^2 \varphi_j(x_i) f(x_i).$$

Since $\hat{\lambda}$ satisfies the Dantzig constraint, we have with probability at least $1 - M^{1-\gamma/2}$, for any $j \in \{1, \ldots, M\}$,

$$|(G\hat{\lambda})_j - \beta_j| \leq |(G\hat{\lambda})_j - \hat{\beta}_j| + |\hat{\beta}_j - \beta_j| \leq 2r_{n,j}$$

and $|A| \leq 4r_n\|\Delta\|_1$. This implies that

$$\|\hat{f} - f\|_n^2 \leq \|f_\lambda - f\|_n^2 + 4r_n\|\Delta\|_1 - \|f_\Delta\|_n^2.$$

Moreover using Lemma 1 and Proposition 1 of Bertin et al. (2011) (where the norm $\|\cdot\|_2$ is replaced by $\|\cdot\|_n$), we obtain that

$$\left(\|\Delta_{J_0^C}\|_{\ell_1} - \|\Delta_{J_0}\|_{\ell_1}\right)_+ \leq 2\|\lambda_{J_0^C}\|_{\ell_1} + \left(\|\hat{\lambda}\|_{\ell_1} - \|\lambda\|_{\ell_1}\right)_+ \tag{17}$$

and

$$\begin{aligned}\|f_\Delta\|_n &\geq \kappa_s\|\Delta_{J_0}\|_{\ell_2} - \frac{\mu_s}{\sqrt{|J_0|}}\left(\|\Delta_{J_0^C}\|_{\ell_1} - \|\Delta_{J_0}\|_{\ell_1}\right)_+ \\ &\geq \kappa_s\|\Delta_{J_0}\|_{\ell_2} - 2\frac{\mu_s}{\sqrt{|J_0|}}\Lambda(\lambda, J_0^c).\end{aligned}$$

Note that Proposition 1 of Bertin et al. (2011) is obtained using Lemma 2 and Lemma 3 of Bertin et al. (2011). In our context, Lemma 2 and Lemma 3 can be proved in the same way by replacing the norm $\|\cdot\|_2$ by $\|\cdot\|_n$ and by considering $P_{J_{01}}$ as the projector on the linear space spanned by $(\varphi_j(x_1), \ldots, \varphi_j(x_n))_{j\in J_{01}}$.

Now following the same lines as Theorem 2 of Bertin et al. (2011), replacing κ_{J_0} by κ_s and μ_{J_0} by μ_s, we obtain the result of the theorem.

Proof of Theorem 2

We consider $\hat{\lambda}^D$ defined by

$$\hat{\lambda}^D = \text{argmin}_{\lambda\in\mathbb{R}^M}\|\lambda\|_{\ell_1} \quad \text{such that } \lambda \text{ satisfies the Dantzig constraint (11).}$$

Denote by $\hat{f}^D$ the estimator $f_{\hat{\lambda}^D}$. Following the same lines as in the proof of Theorem 1, it can be obtained that, with probability at least $1-M^{1-\gamma/2}$, for any integer $s < n/2$ such that (A1(s)) holds, we have for any $\alpha > 0$,

$$\|\hat{f}^D - f\|_n^2 \leq \inf_{\lambda\in\mathbb{R}^M}\inf_{\substack{J_0\subset\{1,\ldots,M\}\\|J_0|=s}}\left\{\|f_\lambda - f\|_n^2 + \alpha\left(1+\frac{2\mu_s}{\kappa_s}\right)^2\frac{\Lambda(\lambda,J_0^c)^2}{s} + 16s\left(\frac{1}{\alpha}+\frac{1}{\kappa_s^2}\right)r_n^2\right\},$$

where here

$$\Lambda(\lambda, J_0^c) = \|\lambda_{J_0^C}\|_{\ell_1} + \frac{\left(\|\hat{\lambda}^D\|_{\ell_1} - \|\lambda\|_{\ell_1}\right)_+}{2}.$$

If the infimum is only taken over the vectors λ that satisfy the Dantzig constraint, then, with the same probability we have

$$\|\hat{f}^D - f\|_n^2 \leq \inf_{\lambda\in\mathcal{D}}\inf_{\substack{J_0\subset\{1,\ldots,M\}\\|J_0|=s}}\left\{\|f_\lambda - f\|_n^2 + \alpha\left(1+\frac{2\mu_s}{\kappa_s}\right)^2\frac{\|\lambda_{J_0^C}\|_{l_1}^2}{s} + 16s\left(\frac{1}{\alpha}+\frac{1}{\kappa_s^2}\right)r_n^2\right\}. \quad (18)$$

Following the same lines as the proof of Theorem 1, replacing λ by $\hat{\lambda}^D$, we obtain, with probability at least $1 - M^{1-\gamma/2}$,

$$\|\hat{f} - f\|_n^2 \leq \|\hat{f}^D - f\|_n^2 + 4r_n\|\Delta\|_1 - \|f_\Delta\|_n^2,$$

with $\Delta = \hat{\lambda} - \hat{\lambda}^D$. Applying (17) where $\hat{\lambda}$ plays the role of λ and $\hat{\lambda}^D$ the role of $\hat{\lambda}$, the vector Δ satisfies

$$\left(\|\Delta_{J_0^C}\|_{\ell_1} - \|\Delta_{J_0}\|_{\ell_1}\right)_+ \leq 2\|\hat{\lambda}_{J_0^C}\|_{\ell_1}.$$

Following the same lines as in the proof of Theorem 1, we obtain that for each $J_0 \subset \{1,\ldots,M\}$ such that $|J_0| = s$

$$\|\hat{f} - f\|_n^2 \leq \left\{\|\hat{f}^D - f\|_n^2 + \alpha\left(1+\frac{2\mu_s}{\kappa_s}\right)^2\frac{\|\hat{\lambda}_{J_0^C}\|_{l_1}^2}{s} + 16s\left(\frac{1}{\alpha}+\frac{1}{\kappa_s^2}\right)r_n^2\right\}. \quad (19)$$

Finally, (18) and (19) imply the theorem.

Proof of Theorem 3

We first state the following lemma.

Lemma 2 *We have for any $u \in \mathbb{R}^M$,*

$$crit(\hat{\lambda} + u) - crit(\hat{\lambda}) \geq \left\| \sum_{k=1}^{M} u_k \varphi_k \right\|_n^2 .$$

Proof of Lemma 2: Since for any λ,

$$\mathrm{crit}(\lambda) = \frac{1}{n} \sum_{i=1}^{n} (y_i - b_i f_\lambda(x_i))^2 + 2 \sum_{j=1}^{M} \tilde{r}_{n,j} |\lambda_j|,$$

$$\begin{aligned}
\mathrm{crit}(\hat{\lambda} + u) - \mathrm{crit}(\hat{\lambda}) &= \frac{1}{n} \sum_{i=1}^{n} \left(y_i - b_i \sum_{k=1}^{M} \hat{\lambda}_k \varphi_k(x_i) - b_i \sum_{k=1}^{M} u_k \varphi_k(x_i) \right)^2 + 2 \sum_{j=1}^{M} \tilde{r}_{n,j} |\hat{\lambda}_j + u_j| \\
&\quad - \frac{1}{n} \sum_{i=1}^{n} \left(y_i - b_i \sum_{k=1}^{M} \hat{\lambda}_k \varphi_k(x_i) \right)^2 - 2 \sum_{j=1}^{M} \tilde{r}_{n,j} |\hat{\lambda}_j| \\
&= \frac{1}{n} \sum_{i=1}^{n} b_i^2 \left(\sum_{k=1}^{M} u_k \varphi_k(x_i) \right)^2 + 2 \sum_{j=1}^{M} \tilde{r}_{n,j} \left(|\hat{\lambda}_j + u_j| - |\hat{\lambda}_j| \right) \\
&\quad - \frac{2}{n} \sum_{i=1}^{n} \left(y_i - b_i \sum_{k=1}^{M} \hat{\lambda}_k \varphi_k(x_i) \right) b_i \sum_{k=1}^{M} u_k \varphi_k(x_i) \\
&= \frac{1}{n} \sum_{i=1}^{n} b_i^2 \left(\sum_{k=1}^{M} u_k \varphi_k(x_i) \right)^2 + 2 \sum_{j=1}^{M} \tilde{r}_{n,j} \left(|\hat{\lambda}_j + u_j| - |\hat{\lambda}_j| \right) \\
&\quad + \frac{2}{n} \sum_{i=1}^{n} b_i^2 \sum_{j=1}^{M} \hat{\lambda}_j \varphi_j(x_i) \sum_{k=1}^{M} u_k \varphi_k(x_i) - \frac{2}{n} \sum_{i=1}^{n} b_i y_i \sum_{k=1}^{M} u_k \varphi_k(x_i) \\
&= \frac{1}{n} \sum_{i=1}^{n} b_i^2 \left(\sum_{k=1}^{M} u_k \varphi_k(x_i) \right)^2 + 2 \sum_{j=1}^{M} \tilde{r}_{n,j} \left(|\hat{\lambda}_j + u_j| - |\hat{\lambda}_j| \right) \\
&\quad + \frac{2}{n} \sum_{i=1}^{n} \sum_{k=1}^{M} u_k \varphi_k(x_i) \left(b_i^2 \sum_{j=1}^{M} \hat{\lambda}_j \varphi_j(x_i) - b_i y_i \right).
\end{aligned}$$

Since $\hat{\lambda}$ minimizes $\lambda \longmapsto \mathrm{crit}(\lambda)$, we have for any k,

$$0 = \frac{2}{n} \sum_{i=1}^{n} \varphi_k(x_i) \left(b_i^2 \sum_{j=1}^{M} \hat{\lambda}_j \varphi_j(x_i) - b_i y_i \right) + 2 \tilde{r}_{n,k} s(\hat{\lambda}_k),$$

where $|s(\hat{\lambda}_k)| \leq 1$ and $s(\hat{\lambda}_k) = \mathrm{sign}(\hat{\lambda}_k)$ if $\hat{\lambda}_k \neq 0$. So,

$$\frac{2}{n} \sum_{i=1}^{n} \sum_{k=1}^{M} u_k \varphi_k(x_i) \left(b_i^2 \sum_{j=1}^{M} \hat{\lambda}_j \varphi_j(x_i) - b_i y_i \right) = -2 \sum_{k=1}^{M} u_k \tilde{r}_{n,k} s(\hat{\lambda}_k)$$

and

$$\begin{aligned}
\mathrm{crit}(\hat{\lambda}+u)-\mathrm{crit}(\hat{\lambda}) &= \frac{1}{n}\sum_{i=1}^{n} b_i^2\left(\sum_{k=1}^{M} u_k\varphi_k(x_i)\right)^2 + 2\sum_{j=1}^{M}\tilde{r}_{n,j}\left(|\hat{\lambda}_j+u_j|-|\hat{\lambda}_j|\right) \\
&\qquad -2\sum_{k=1}^{M} u_k\tilde{r}_{n,k}s(\hat{\lambda}_k) \\
&= \frac{1}{n}\sum_{i=1}^{n} b_i^2\left(\sum_{k=1}^{M} u_k\varphi_k(x_i)\right)^2 + 2\sum_{j=1}^{M}\tilde{r}_{n,j}\left(|\hat{\lambda}_j+u_j|-|\hat{\lambda}_j|-u_j s(\hat{\lambda}_j)\right) \\
&\geq \frac{1}{n}\sum_{i=1}^{n} b_i^2\left(\sum_{k=1}^{M} u_k\varphi_k(x_i)\right)^2,
\end{aligned}$$

which proves the result. □

Now, still with $s^* = \mathrm{card}(S^*)$, we consider for $\mu \in \mathbb{R}^{s^*}$

$$\mathrm{critS}^*(\mu) = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - b_i\sum_{j\in S^*}\mu_j\varphi_j(x_i)\right)^2 + 2\sum_{j\in S^*}\tilde{r}_{n,j}|\mu_j|,$$

and

$$\tilde{\mu} = \arg\min_{\mu\in\mathbb{R}^{s^*}} \mathrm{critS}^*(\mu).$$

Then we set

$$\mathcal{S} = \bigcap_{j\notin S^*}\left\{\left|\frac{1}{n}\sum_{i=1}^{n} y_i b_i\varphi_j(x_i) - \sum_{k\in S^*}\tilde{\mu}_k < \varphi_j,\varphi_k>\right| < \tilde{r}_{n,j}\right\}$$

and we state the following lemma.

Lemma 3 *On the set $\mathcal{S}$, the non-zero coordinates of $\hat{\lambda}$ are included into S^*.*

Proof of Lemma 3: Recall that $\hat{\lambda}$ is a minimizer of $\lambda \longmapsto \mathrm{crit}(\lambda)$. Using standard convex analysis arguments, this is equivalent to say that for any $1 \leq j \leq M$,

$$\begin{cases}
\frac{1}{n}\sum_{i=1}^{n} y_i b_i\varphi_j(x_i) - \sum_{k=1}^{M}\hat{\lambda}_k < \varphi_j,\varphi_k > \;= \tilde{r}_{n,j}\mathrm{sign}(\hat{\lambda}_j) & \text{if } \hat{\lambda}_j \neq 0, \\
\left|\frac{1}{n}\sum_{i=1}^{n} y_i b_i\varphi_j(x_i) - \sum_{k=1}^{M}\hat{\lambda}_k < \varphi_j,\varphi_k >\right| \leq \tilde{r}_{n,j} & \text{if } \hat{\lambda}_j = 0.
\end{cases}$$

Similarly, on $\mathcal{S}$, we have

$$\begin{cases}
\frac{1}{n}\sum_{i=1}^{n} y_i b_i\varphi_j(x_i) - \sum_{k\in S^*}\tilde{\mu}_k < \varphi_j,\varphi_k > \;= \tilde{r}_{n,j}\mathrm{sign}(\tilde{\mu}_j) & \text{if } j\in S^* \text{ and } \tilde{\mu}_j \neq 0, \\
\left|\frac{1}{n}\sum_{i=1}^{n} y_i b_i\varphi_j(x_i) - \sum_{k\in S^*}\tilde{\mu}_k < \varphi_j,\varphi_k >\right| \leq \tilde{r}_{n,j} & \text{if } j\in S^* \text{ and } \tilde{\mu}_j = 0, \\
\left|\frac{1}{n}\sum_{i=1}^{n} y_i b_i\varphi_j(x_i) - \sum_{k\in S^*}\tilde{\mu}_k < \varphi_j,\varphi_k >\right| < \tilde{r}_{n,j} & \text{if } j\notin S^*.
\end{cases}$$

So, on $\mathcal{S}$, the vector $\hat{\mu}$ such $\hat{\mu}_j = \tilde{\mu}_j$ if $j \in S^*$ and $\hat{\mu}_j = 0$ if $j \notin S^*$ is also a minimizer of $\lambda \longmapsto \text{crit}(\lambda)$. Using Lemma 2, we have for any $1 \leq i \leq n$:

$$\sum_{k=1}^{M} (\hat{\lambda}_k - \hat{\mu}_k)\varphi_k(x_i) = 0.$$

So, for $j \notin S^*$,

$$\left| \frac{1}{n} \sum_{i=1}^{n} y_i b_i \varphi_j(x_i) - \sum_{k=1}^{M} \hat{\lambda}_k < \varphi_j, \varphi_k > \right| < \tilde{r}_{n,j}.$$

Therefore, on $\mathcal{S}$, the non-zero coordinates of $\hat{\lambda}$ are included into S^*. □

Lemma 3 shows that we just need to prove that

$$\mathbb{P}\{\mathcal{S}\} \geq 1 - 2M^{1-\gamma/2}$$

$$\begin{aligned} \mathbb{P}\{\mathcal{S}^c\} &\leq \sum_{j \notin S^*} \mathbb{P}\left\{ \left| \frac{1}{n} \sum_{i=1}^{n} y_i b_i \varphi_j(x_i) - \sum_{k \in S^*} \tilde{\mu}_k < \varphi_j, \varphi_k > \right| \geq \tilde{r}_{n,j} \right\} \\ &\leq A + B, \end{aligned}$$

with

$$\begin{aligned} A &= \sum_{j \notin S^*} \mathbb{P}\left\{ \left| \frac{1}{n} \sum_{i=1}^{n} [y_i b_i \varphi_j(x_i) - \mathbb{E}(y_i b_i \varphi_j(x_i))] \right| \geq r_{n,j} \right\} \\ &= \sum_{j \notin S^*} \mathbb{P}\left\{ \left| \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i b_i \varphi_j(x_i) \right| \geq r_{n,j} \right\} \\ &= \sum_{j \notin S^*} \mathbb{P}\{|V_j| \geq r_{n,j}\} \end{aligned}$$

(see Lemma 1) and

$$\begin{aligned} B &= \mathbb{P}\left[\bigcup_{j \notin S^*} \left\{ \left| \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}(y_i b_i \varphi_j(x_i)) - \sum_{k \in S^*} \tilde{\mu}_k < \varphi_j, \varphi_k > \right| \geq \tilde{r}_{n,j} - r_{n,j} \right\} \right] \\ &= \mathbb{P}\left[\bigcup_{j \notin S^*} \left\{ \left| < \varphi_j, f_{\lambda^*} > - \sum_{k \in S^*} \tilde{\mu}_k < \varphi_j, \varphi_k > \right| \geq \tilde{r}_{n,j} - r_{n,j} \right\} \right] \\ &= \mathbb{P}\left[\bigcup_{j \notin S^*} \left\{ \left| \sum_{k \in S^*} (\lambda_k^* - \tilde{\mu}_k) < \varphi_j, \varphi_k > \right| \geq \tilde{r}_{n,j} - r_{n,j} \right\} \right] \\ &\leq \mathbb{P}\left[\bigcup_{j \notin S^*} \left\{ \rho(S^*) \|\varphi_j\|_n \sum_{k \in S^*} |\lambda_k^* - \tilde{\mu}_k| \, \|\varphi_k\|_n \geq \tilde{r}_{n,j} - r_{n,j} \right\} \right] \end{aligned}$$

since

$$\rho(S^*) = \max_{k \in S^*} \max_{j \neq k} \frac{| < \varphi_j, \varphi_k > |}{\|\varphi_j\|_n \|\varphi_k\|_n}.$$

Using notation of Lemma 3, we have:

$$\begin{aligned}\|f_{\lambda^*} - f_{\hat{\mu}}\|_n^2 &= \|\sum_{k\in S^*}(\lambda_k^* - \hat{\mu}_k)\varphi_k\|_n^2 \\ &= \sum_{k\in S^*}(\lambda_k^* - \hat{\mu}_k)^2\|\varphi_k\|_n^2 + \sum_{k\in S^*}\sum_{j\in S^*,\ j\neq k}(\lambda_k^* - \hat{\mu}_k)(\lambda_j^* - \hat{\mu}_j) < \varphi_j, \varphi_k >,\end{aligned}$$

and

$$\begin{aligned}\sum_{k\in S^*}(\lambda_k^* - \hat{\mu}_k)^2\|\varphi_k\|_n^2 &\leq \|f_{\lambda^*} - f_{\hat{\mu}}\|_n^2 + \rho(S^*)\sum_{k\in S^*}\sum_{j\in S^*,\ j\neq k}|\lambda_k^* - \hat{\mu}_k|\|\varphi_k\|_n \times |\lambda_j^* - \hat{\mu}_j|\|\varphi_j\|_n \\ &\leq \|f_{\lambda^*} - f_{\hat{\mu}}\|_n^2 + \rho(S^*)\left(\sum_{k\in S^*}|\lambda_k^* - \hat{\mu}_k|\|\varphi_k\|_n\right)^2.\end{aligned}$$

Finally,

$$\begin{aligned}\left(\sum_{k\in S^*}|\lambda_k^* - \hat{\mu}_k|\|\varphi_k\|_n\right)^2 &\leq s^*\sum_{k\in S^*}(\lambda_k^* - \hat{\mu}_k)^2\|\varphi_k\|_n^2 \\ &\leq s^*\left(\|f_{\lambda^*} - f_{\hat{\mu}}\|_n^2 + \rho(S^*)\left(\sum_{k\in S^*}|\lambda_k^* - \hat{\mu}_k|\|\varphi_k\|_n\right)^2\right),\end{aligned}$$

which shows that

$$\left(\sum_{k\in S^*}|\lambda_k^* - \hat{\mu}_k|\|\varphi_k\|_n\right)^2 \leq \frac{s^*}{1-\rho(S^*)s^*}\|f_{\lambda^*} - f_{\hat{\mu}}\|_n^2.$$

Now,

$$\begin{aligned}&\frac{1}{n}\sum_{i=1}^n\left(y_i - b_i\sum_{j\in S^*}\tilde{\mu}_j\varphi_j(x_i)\right)^2 + 2\sum_{j\in S^*}\tilde{r}_{n,j}|\tilde{\mu}_j| \leq \\ &\qquad \frac{1}{n}\sum_{i=1}^n\left(y_i - b_i\sum_{j\in S^*}\lambda_j^*\varphi_j(x_i)\right)^2 + 2\sum_{j\in S^*}\tilde{r}_{n,j}|\lambda_j^*|.\end{aligned}$$

So,

$$\begin{aligned}&\|\sum_{j\in S^*}\tilde{\mu}_j\varphi_j\|_n^2 - \frac{2}{n}\sum_{i=1}^n b_i y_i\sum_{j\in S^*}\tilde{\mu}_j\varphi_j(x_i) + 2\sum_{j\in S^*}\tilde{r}_{n,j}|\tilde{\mu}_j| \leq \\ &\qquad \|\sum_{j\in S^*}\lambda_j^*\varphi_j\|_n^2 - \frac{2}{n}\sum_{i=1}^n b_i y_i\sum_{j\in S^*}\lambda_j^*\varphi_j(x_i) + 2\sum_{j\in S^*}\tilde{r}_{n,j}|\lambda_j^*|,\end{aligned}$$

and using previous notation,

$$\begin{aligned}&\|f_{\hat{\mu}}\|_n^2 - \frac{2}{n}\sum_{i=1}^n b_i y_i\sum_{j\in S^*}\tilde{\mu}_j\varphi_j(x_i) + 2\sum_{j\in S^*}\tilde{r}_{n,j}|\tilde{\mu}_j| \leq \\ &\qquad \|f_{\lambda^*}\|_n^2 - \frac{2}{n}\sum_{i=1}^n b_i y_i\sum_{j\in S^*}\lambda_j^*\varphi_j(x_i) + 2\sum_{j\in S^*}\tilde{r}_{n,j}|\lambda_j^*|.\end{aligned}$$

Therefore,

$$\begin{aligned}
\|f_{\lambda^*} - f_{\hat{\mu}}\|_n^2 &= \|f_{\hat{\mu}}\|_n^2 + \|f_{\lambda^*}\|_n^2 - 2 < f_{\hat{\mu}}, f_{\lambda^*} > \\
&\leq 2\|f_{\lambda^*}\|_n^2 - 2 < f_{\hat{\mu}}, f_{\lambda^*} > + \frac{2}{n}\sum_{i=1}^n b_i y_i \sum_{j\in S^*}(\tilde{\mu}_j - \lambda_j^*)\varphi_j(x_i) + 2\sum_{j\in S^*}\tilde{r}_{n,j}(|\lambda_j^*| - |\tilde{\mu}_j|) \\
&= \frac{2}{n}\sum_{i=1}^n b_i y_i (f_{\hat{\mu}}(x_i) - f_{\lambda^*}(x_i)) - \frac{2}{n}\sum_{i=1}^n b_i^2 f_{\lambda^*}(x_i)(f_{\hat{\mu}}(x_i) - f_{\lambda^*}(x_i)) \\
&\qquad +2\sum_{j\in S^*}\tilde{r}_{n,j}(|\lambda_j^*| - |\tilde{\mu}_j|) \\
&= \frac{2}{n}\sum_{i=1}^n b_i (y_i - \mathbb{E}(y_i))(f_{\hat{\mu}}(x_i) - f_{\lambda^*}(x_i)) + 2\sum_{j\in S^*}\tilde{r}_{n,j}(|\lambda_j^*| - |\tilde{\mu}_j|) \\
&= \frac{2}{n}\sum_{i=1}^n b_i \varepsilon_i (f_{\hat{\mu}}(x_i) - f_{\lambda^*}(x_i)) + 2\sum_{j\in S^*}\tilde{r}_{n,j}(|\lambda_j^*| - |\tilde{\mu}_j|) \\
&= 2\sum_{j=1}^M V_j(\hat{\mu}_j - \lambda_j^*) + 2\sum_{j\in S^*}\tilde{r}_{n,j}(|\lambda_j^*| - |\tilde{\mu}_j|).
\end{aligned}$$

Now let us assume that for any $j \in S^*$, $V_j < r_{n,j}$. Then,

$$\begin{aligned}
\|f_{\lambda^*} - f_{\hat{\mu}}\|_n^2 &< 2\sum_{j\in S^*}(r_{n,j} + \tilde{r}_{n,j})|\hat{\mu}_j - \lambda_j^*| \\
&< 2\sigma\sqrt{\frac{\log M}{n}}(\sqrt{\gamma} + \sqrt{\tilde{\gamma}})\sum_{j\in S^*}\|\varphi_j\|_n|\hat{\mu}_j - \lambda_j^*|.
\end{aligned}$$

So,

$$\sum_{k\in S^*}|\lambda_k^* - \hat{\mu}_k|\|\varphi_k\|_n < 2\sigma\sqrt{\frac{\log M}{n}}(\sqrt{\gamma} + \sqrt{\tilde{\gamma}})\frac{s^*}{1 - \rho(S^*)s^*}$$

and for any $j \notin S^*$,

$$\begin{aligned}
\rho(S^*)\|\varphi_j\|_n\sum_{k\in S^*}|\lambda_k^* - \hat{\mu}_k|\|\varphi_k\|_n &< 2\sigma\sqrt{\frac{\log M}{n}}\|\varphi_j\|_n(\sqrt{\gamma} + \sqrt{\tilde{\gamma}})\frac{\rho(S^*)s^*}{1 - \rho(S^*)s^*} \\
&< \frac{2\sigma c(\sqrt{\gamma} + \sqrt{\tilde{\gamma}})}{1 - c}\sqrt{\frac{\log M}{n}}\|\varphi_j\|_n \\
&< (\sqrt{\tilde{\gamma}} - \sqrt{\gamma})\sigma\sqrt{\frac{\log M}{n}}\|\varphi_j\|_n \\
&< \tilde{r}_{n,j} - r_{n,j}.
\end{aligned}$$

Therefore,

$$B \leq \sum_{j\in S^*}\mathbb{P}\{|V_j| \geq r_{n,j}\}$$

and using Lemma 1, since $\mathbb{P}\{\mathcal{S}^c\} \leq A + B$,

$$\mathbb{P}\{\mathcal{S}\} \geq 1 - 2M^{1-\gamma/2}.$$

Proof of Corollary 1

First note that λ^* satisfies the Dantzig constraint (11) where $r_{n,j}$ is replaced by $\tilde{r}_{n,j}$ with probability larger than $1 - M^{1-\tilde{\gamma}/2}$. On the event $\hat{S} \subset S^*$, we have $\lambda^*_{(S^*)^C} = \hat{\lambda}_{(S^*)^C} = 0$, then applying Theorem 2, we obtain that for any $\alpha > 0$

$$\|\hat{f} - f\|_n^2 \leq 32 s^* \left(\frac{1}{\alpha} + \frac{1}{\kappa_{s^*}^2} \right) \tilde{r}_n^2,$$

which implies the result of the theorem.

References

Bertin, K., Le Pennec, E., and Rivoirard, V. (2011). Adaptive Dantzig density estimation. *Annales de l'Institut Henri Poincaré*, 47:43–74.

Bickel, P. J., Ritov, Y., and Tsybakov, A. B. (2009). Simultaneous analysis of lasso and Dantzig selector. *Ann. Statist.*, 37(4):1705–1732.

Bühlmann, P. and van de Geer, S. (2011). *Statistics for high-dimensional data*. Springer Series in Statistics. Springer, Heidelberg. Methods, theory and applications.

Bunea, F. (2008). Consistent selection via the Lasso for high dimensional approximating regression models. In *Pushing the limits of contemporary statistics: contributions in honor of Jayanta K. Ghosh*, volume 3 of *Inst. Math. Stat. Collect.*, pages 122–137. Inst. Math. Statist., Beachwood, OH.

Bunea, F., Tsybakov, A., and Wegkamp, M. (2007a). Sparsity oracle inequalities for the Lasso. *Electronic Journal of Statistics*, 1:169–194.

Bunea, F., Tsybakov, A. B., and Wegkamp, M. H. (2006). Aggregation and sparsity via l_1 penalized least squares. In *Learning theory*, volume 4005 of *Lecture Notes in Comput. Sci.*, pages 379–391. Springer, Berlin.

Bunea, F., Tsybakov, A. B., and Wegkamp, M. H. (2007b). Aggregation for Gaussian regression. *The Annals of Statistics*, 35(4):1674–1697.

Candès, E. J. and Tao, T. (2007). The Dantzig selector: statistical estimation when p is much larger than n. *The Annals of Statistics*, 35(6):2313–2351.

Comte, F. and Samson, A. (2012). Nonparametric estimation of random effects densities in linear mixed-effects model. Unpublished manuscript. Available at http://hal.archives-ouvertes.fr/hal-00657052/fr/.

Delyon, B., Lavielle, M., and Moulines, E. (1999). Convergence of a stochastic approximation version of the em algorithm. *The Annals of Statistics*, 27:94–128.

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

Ding, A. A. and Wu, H. (2001). Assessing antiviral potency of anti-HIV therapies in vivo by comparing viral decay rates in viral dynamic models. *Biostatistics*, 2:13–29.

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

Harville, D. (1974). Bayesian inference for variance components using only error contrasts. *Biometrika*, 61:383–385.

Jank, W. and Shmueli, G. (2006). Functional data analysis in electronic commerce research. *Statistical Science*, 21:155–166.

Ke, C. and Wang, Y. (2001). Semiparametric nonlinear mixed-effects models and their applications (with discussion). *Journal of the American Statistical Association*, 96(456):1272–1298.

Ke, C. and Wang, Y. (2004). Smoothing spline nonlinear nonparametric regression models. *Journal of the American Statistical Association*, 99(468):1166–1175.

Kuhn, E. and Lavielle, M. (2004). Coupling a stochastic approximation version of EM with an MCMC procedure. *ESAIM: P&S*, 8:115–131.

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

Liu, B. and Müller, H. G. (2008). Functional data analysis for sparse auction data. In Jank, W. and Shmueli, G., editors, *Statistical Methods in E-commerce research*, pages 269–290. Wiley, New York.

Liu, W. and Wu, L. (2007). Simultaneous inference for semiparametric nonlinear mixed-effects models with covariate measurement errors and missing responses. *Biometrics*, 63:342–350.

Liu, W. and Wu, L. (2008). A semiparametric nonlinear mixed-effects model with non-ignorable missing data and measurement errors for HIV viral data. *Computational Statistics & Data Analysis*, 53:112–122.

Liu, W. and Wu, L. (2009). Some asymptotic results for semiparametric nonlinear mixed-effects models with incomplete data. *Journal of Statistical Planning and Inference*. doi:10.1016j.jspi.2009.06.006.

Luan, Y. and Li, H. (2004). Model-based methods for identifying periodically expressed genes based on time course microarray gene expression data. *Bioinformatics*, 20(3):332–339.

Meza, C., Jaffrézic, F., and Foulley, J.-L. (2007). Reml estimation of variance parameters in nonlinear mixed effects models using the SAEM algorithm. *Biometrical Journal*, 49(6):876–888.

Patterson, H. D. and Thompson, R. (1971). Recovery of inter-block information when block sizes are unequal. *Biometrika*, 58:545–554.

Pinheiro, J. and Bates, D. (2000). *Mixed-Effects Models in S and S-PLUS*. Springer-Verlag, New York.

Ramos, R. and Pantula, S. (1995). Estimation of nonlinear random coefficient models. *Statistics & Probability Letters*, 24:49–56.

Reithinger, F., Jank, W., Tutz, G., and Shmueli, G. (2008). Modelling price paths in on-line auctions: smoothing sparse and unevenly sampled curves by using semiparametric mixed models. *Applied Statistics*, 57:127–148.

Schelldorfer, J., Bühlmann, P., and van de Geer, S. (2011). Estimation for high-dimensional linear mixed-effects models using l1-penalization. *Scandinavian Journal of Statistics*, 38:197–214.

Shmueli, G. and Jank, W. (2005). Visualizing online auctions. *Journal of Computational and Graphical Statistics*, 14:299–319.

Shmueli, G., Russo, R. P., and Jank, W. (2007). The BARISTA: a model for bid arrivals in online auctions. *The Annals of Applied Statistics*, 1:412–441.

Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society, Series B*, 58:267–288.

van de Geer, S. (2010). ℓ_1-regularization in high-dimensional statistical models. In *Proceedings of the International Congress of Mathematicians. Volume IV*, pages 2351–2369, New Delhi. Hindustan Book Agency.

van de Geer, S. A. and Bühlmann, P. (2009). On the conditions used to prove oracle results for the Lasso. *Electronic Journal of Statistics*, 3:1360–1392.

Vonesh, E. F. (1996). A note on the use of Laplace's approximation for nonlinear mixed-effects models. *Biometrika*, 83:447–452.

Wahba, G. (1990). *Spline models for observational data*, volume 59 of *CBMS-NSF Regional Conference Series in Applied Mathematics*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA.

Wang, Y. (1998). Smoothing spline models with correlated random errors. *Journal of the American Statistical Association*, 93:341–348.

Wang, Y. and Brown, M. B. (1996). A flexible model for human circadian rhythms. *Biometrics*, 52:588–596.

Wang, Y., Ke, C., and Brown, M. B. (2003). Shape-invariant modeling of circadian rhythms with random effects and smoothing spline anova decompositions. *Biometrics*, 59:804–812.

Wei, G. C. and Tanner, M. A. (1990). A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithm. *Journal of the American Statistical Association*, 85:699–704.

Wu, H. and Zhang, J. (2002). The study of longterm HIV dynamics using semi-parametric non-linear mixed-effects models. *Statistics in Medicine*, 21:3655–3675.

Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society, Series B*, 68(1):49–67.