Mixtures of stochastic differential equations with random effects: application to data clustering

Maud Delattre[1], Valentine Genon-Catalot[2] and Adeline Samson[3*]

[1] AgroParisTech, France

[2] UMR CNRS 8145, Laboratoire MAP5, Université Paris Descartes, Sorbonne Paris Cité, France

[3] UMR CNRS 5224, Laboratoire Jean Kuntzmann, Université Grenoble Alpes, France

October 21, 2015

Abstract

We consider N independent stochastic processes $(X_i(t), t \in [0, T_i])$, $i = 1, \ldots, N$, defined by a stochastic differential equation with drift term depending on a random variable ϕ_i. The distribution of the random effect ϕ_i is a Gaussian mixture distribution, depending on unknown parameters which are to be estimated from the continuous observation of the processes X_i. The likelihood of the observation is explicit. When the number of components is known, we prove the consistency of the exact maximum likelihood estimators and use the EM algorithm to compute it. When the number of components is unknown, BIC (Bayesian Information Criterion) is applied to select it. To assign each individual to a class, we define a classification rule based on estimated posterior probabilities. A simulation study illustrates our estimation and classification method on various models. A real data analysis is performed on growth curves with convincing results.

Key Words: mixed-effects models; stochastic differential equations; BIC; classification; EM algorithm; mixture distribution; maximum likelihood estimator

1 Introduction

The goal of clustering methods is to discover structures among individuals: data are grouped into a few clusters such that the observations in the same cluster

*Corresponding author

Email addresses: maud.delattre@agroparistech.fr (Maud Delattre), valentine.genon-catalot@parisdescartes.fr (Valentine Genon-Catalot), adeline.leclercq-samson@imag.fr (Adeline Samson)

are more similar to each other than those from the other clusters. In this paper we focus on individuals described by longitudinal data or functional data: data is represented by curves and the random variable underlying data is a stochastic process. Some papers deal with the problem of classification of longitudinal data through mixed-effects models or models with random effects, assuming that the classes are known (see Arribas-Gil *et al.*, 2015, and references therein). Their purpose is to build a classification rule of longitudinal curves/profiles into a given number of different classes to be able to predict the class of a new individual. This is very different from the problem of classification when the classes and the number of classes are unknown. Here, we adopt the latter point of view. We consider functional data modeled by a stochastic differential equation (SDE) with random effects. This is a new approach which is very different from usual functional data analysis methods (see e.g. Jacques and Preda, 2014, for a recent review). The clustering of the trajectories is then obtained by modeling the distribution of the random effects as a mixture of distributions (with unknown number of components).

Mixture of linear regression models with random effects is considered in Celeux *et al.* (2005). Unknown parameters are estimated by maximum likelihood, with the EM algorithm and BIC (Bayesian Information Criterion) for selecting the number of components. Here, we consider functional data modeled by a stochastic differential equation with drift term depending on random effects and diffusion term without random effects. More precisely, we consider N real valued stochastic processes $(X_i(t), t \geq 0)$, $i = 1, \ldots, N$, with dynamics ruled by the following SDEs:

$$d\,X_i(t) = (\phi_i' \, b(X_i(t)) + a(X_i(t)))dt + \sigma(X_i(t))\, dW_i(t), \quad X_i(0) = x, \qquad (1)$$

where $(W_1, \ldots, W_N)$ are N independent Wiener processes, $\phi_1, \ldots, \phi_N$ are N *i.i.d.* $\mathbb{R}^d$-valued random variables, $(\phi_1, \ldots, \phi_N)$ and $(W_1, \ldots, W_N)$ are independent and x is a known real value. The functions $\sigma(.), a(.) : \mathbb{R} \to \mathbb{R}$ and $b(.) : \mathbb{R} \to \mathbb{R}^d$ are known. Each process $(X_i(t))$ represents an individual and the random variable ϕ_i represents the random effect of individual i.

We consider continuous observations $(X_i(t), t \in [0, T], i = 1, \ldots, N)$ with a given T. The estimation of unknown parameters in the distribution of ϕ_i from the (X_i)'s is not straightforward, as the exact likelihood is generally not explicit. Maximum likelihood estimation in SDEs with random effects has been studied in a few papers (Ditlevsen and De Gaetano, 2005; Donnet and Samson, 2008;

Picchini *et al.*, 2010). In Delattre *et al.* (2013), model (1) is considered with ϕ_i having a Gaussian distribution. This has the advantage of leading to an explicit formula for the exact likelihood.

In this paper, we assume that the random effects ϕ_i have distribution given by a mixture of Gaussian distributions, this mixture distribution modeling the classes. We want to estimate the number of components of the mixture, as well as the parameters and the proportions. More precisely, we assume that the random variables $\phi_1, \ldots, \phi_N$ have a common distribution with density $g(\varphi, \theta)$ on $\mathbb{R}^d$, which is given by a mixture of Gaussian distributions:

$$g(\varphi, \theta) = \sum_{\ell=1}^{M} \pi_\ell \, n_d(\varphi, \tau_\ell), \quad n_d(\varphi, \tau_\ell) d\varphi = \mathcal{N}_d(\boldsymbol{\mu}_\ell, \Omega_\ell), \quad \tau_\ell = (\boldsymbol{\mu}_\ell, \Omega_\ell)$$

with M the number of components in the mixture and π_ℓ the proportions of the mixture ($\sum_{\ell=1}^{M} \pi_\ell = 1$), $\boldsymbol{\mu}_\ell \in \mathbb{R}^d$ and Ω_ℓ a $d \times d$ invertible covariance matrix. Set $\theta = ((\pi_\ell, \tau_\ell), \ell = 1, \ldots, M)$ for the unknown parameters to be estimated when M is known. Below, we denote by θ_0 the true value of the parameter.

Our aim is to estimate the parameters θ of the density of the random effects from the observations $\{X_i(t), 0 \leq t \leq T, i = 1, \ldots, N\}$. We prove that the exact likelihood of observations is explicit. This allows to use the EM-algorithm to compute the maximum likelihood estimator when the number of components is known. We discuss the convergence of the algorithm. Then BIC is applied for selecting the number of mixture components. The EM algorithm also enables to define a classification rule of individuals. As a theoretical result, we prove the consistency of the exact maximum likelihood estimator when the number M of components is known. Our methods show good results on simulated data, both for the parameter estimation and the classification rule. An implemention on real data coming from growth chicken curves (Jaffrézic *et al.*, 2006) is performed.

In Section 2, we introduce notations, assumptions and give the formula of the exact likelihood. In Section 3, the EM algorithm and its properties are described. We present BIC to select the number of components and the classification rule. In Section 4, we prove the consistency of the exact maximum likelihood estimator when the number of components is known. Section 5 is devoted to a simulation study on various models. Section 6 concerns the implementation on real data. Some concluding remarks are given in Section 7. Theoretical proofs are gathered in the Appendix.

2 Model, assumptions and notations

Consider N real valued stochastic processes $(X_i(t), t \geq 0)$, $i = 1, \ldots, N$, with dynamics ruled by (1). The processes $(W_1, \ldots, W_N)$ and the r.v.'s $\phi_1, \ldots, \phi_N$ are defined on a common probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Consider the filtration $(\mathcal{F}_t = \sigma(\phi_i, W_i(s), s \leq t, i = 1, \ldots, N), t \geq 0)$. We introduce the following assumptions:

(H1) The functions $x \to a(x)$ and $x \to b(x) = (b_1(x), \ldots, b_d(x))'$ are Lipschitz continuous on $\mathbb{R}$ and $x \to \sigma(x)$ is Hölder continuous with exponent $\alpha \in [1/2, 1]$ on $\mathbb{R}$.

Under (H1), for $i = 1, \ldots, N$, for all $\varphi = (\varphi_1, \ldots, \varphi_d)' \in \mathbb{R}^d$, the stochastic differential equation (SDE)

$$d\, X_i^{\varphi}(t) = (\varphi'\, b(X_i^{\varphi}(t)) + a(X_i^{\varphi}(t)))dt + \sigma(X_i^{\varphi}(t))\, dW_i(t), \quad X_i^{\varphi}(0) = x \qquad (2)$$

admits a unique strong solution process $(X_i^{\varphi}(t), t \geq 0)$ adapted to the filtration $(\mathcal{F}_t)$. Moreover, the SDE (1) admits a unique strong solution adapted to $(\mathcal{F}_t)$ such that the joint process $(\phi_i, X_i(t))$ is strong Markov and the conditional distribution of $(X_i(t))$ given $\phi_i = \varphi$ is identical to the distribution of (2). The Markov property of $(\phi_i, X_i(t))$ is straightforward by looking at (1) as the two-dimensional SDE:

$$\begin{aligned} d\, \phi_i(t) &= 0, \quad \phi_i(0) = \phi_i, \\ d\, X_i(t) &= (\phi_i(t)'\, b(X_i(t)) + a(X_i(t)))dt + \sigma(X_i(t))\, dW(t), \quad X_i(0) = x. \end{aligned}$$

The processes $(\phi_i, X_i(t), t \geq 0), i = 1, \ldots, N$ are *i.i.d.*. (see *e.g.* Delattre *et al.*, 2013; Genon-Catalot and Larédo, 2015; Comte *et al.*, 2013).
To define the likelihood of the observations, let us introduce the associated canonical model. Let C_T denote the space of real continuous functions $(x(t), t \in [0, T])$ defined on $[0, T]$, endowed with the σ-field $\mathcal{C}_T$ associated with the topology of uniform convergence on $[0, T]$. Under (H1), we introduce the distribution $Q_{\varphi}^{x,T}$ on $(C_T, \mathcal{C}_T)$ of $(X_i^{\varphi}(t), t \in [0, T])$ given by (2). On $\mathbb{R}^d \times C_T$, let $\mathbb{P}_\theta = g(\varphi, \theta) d\varphi \otimes Q_{\varphi}^{x,T}$ denote the joint distribution of $(\phi_i, X_i(t), t \in [0, T])$ and let $\mathbb{Q}_\theta$ denote the marginal distribution of $(X_i(t), t \in [0, T])$ on $(C_T, \mathcal{C}_T)$.
We also denote $\tau_\ell = (\boldsymbol{\mu}_\ell, \Omega_\ell)$, P_{τ_ℓ} (resp. Q_{τ_ℓ}) the distribution $n_d(\varphi, \tau_\ell) d\varphi \otimes Q_{\varphi}^{x,T}$ of $(\phi_i, X_i(.))$ when ϕ_i has distribution $\mathcal{N}_d(\boldsymbol{\mu}_\ell, \Omega_\ell)$ (resp. of $(X_i(t), t \in [0, T])$

when ϕ_i has distribution $\mathcal{N}_d(\boldsymbol{\mu}_\ell, \Omega_\ell)$). With these notations,

$$\mathbb{P}_\theta = \sum_{\ell=1}^{M} \pi_\ell P_{\tau_\ell}, \quad \mathbb{Q}_\theta = \sum_{\ell=1}^{M} \pi_\ell Q_{\tau_\ell} \tag{3}$$

From now on, we denote by (ϕ, X), with $X = (X(t), t \in [0,T])$, the canonical process of $\mathbb{R}^d \times C_T$. We assume that,

(H2) for all $\varphi \in \mathbb{R}^d$, $Q_\varphi^{x,T}\left(\int_0^T \frac{b'(X(t))\, b(X(t)) + a^2(X(t))}{\sigma^2(X(t))} dt < +\infty\right) = 1$.

Under (H1)-(H2), by Theorem 7.19 p.294 in Lipster and Shiryaev (2001), the distributions $Q_\varphi^{x,T}$ and $Q_0^{x,T}$ are equivalent and

$$\frac{dQ_\varphi^{x,T}}{dQ_0^{x,T}}(X) := L_T(X, \varphi) = \exp\left(\varphi'\, U(X) - \frac{1}{2}\varphi'\, V(X)\varphi\right)$$

where $U(X)$ is the vector

$$U(X) = \int_0^T \frac{b(X(s))}{\sigma^2(X(s))}(dX(s) - a(X(s))ds) \tag{4}$$

and $V(X)$ is the $d \times d$ matrix

$$V(X) = \int_0^T \frac{b(X(s))\, b'(X(s))}{\sigma^2(X(s))} ds. \tag{5}$$

Therefore, the density of $\mathbb{Q}_\theta$ (the distribution of X_i on $\mathcal{C}_T$) w.r.t. $Q_0^{x,T}$ is obtained as follows:

$$\frac{d\mathbb{Q}_\theta}{dQ_0^{x,T}}(X) = \int_{\mathbb{R}^d} g(\varphi, \theta) \exp\left(\varphi'\, U(X) - \frac{1}{2}\varphi'\, V(X)\varphi\right) d\varphi := \Lambda(X, \theta). \tag{6}$$

The exact likelihood of $(X_i = (X_i(t), t \in [0,T]), i = 1, \ldots, N)$ is

$$L_N(\theta) = \prod_{i=1}^{N} \Lambda(X_i, \theta). \tag{7}$$

To get a tractable formula for the exact likelihood, we have to consider distributions for ϕ_i such that the integral (6) has a closed form expression. This is the case when ϕ_i has a Gaussian distribution as shown in Delattre *et al.* (2013). This is also the case for the larger class of Gaussian mixtures.
The following assumption is required.

(H3) The matrix $V(X)$ is positive definite $Q_0^{x,T}$-a.s. and $\mathbb{Q}_\theta$-a.s. for all θ.

If the functions (b_j/σ^2) are not linearly independent, (H3) is not true. Thus, (H3) can be interpreted as ensuring a well-defined dimension of the vector ϕ.

Proposition 1. *Assume that* $g(\varphi,\theta)d\varphi = \sum_{\ell=1}^M \pi_\ell \mathcal{N}_d(\boldsymbol{\mu}_\ell, \Omega_\ell)$ *and set* $\tau_\ell = (\boldsymbol{\mu}_\ell, \Omega_\ell)$, $U_i = U(X_i), V_i = V(X_i)$.
Under (H3), the matrices $V_i + \Omega_\ell^{-1}$, $I_d + V_i\Omega_\ell$, $I_d + \Omega_\ell V_i$ *are invertible* $Q_0^{x,T}$*-a.s. and* $\mathbb{Q}_\theta$*-a.s. for all* θ. *Set* $R_{i,\ell}^{-1} = (I_d + V_i\Omega_\ell)^{-1}V_i$, *we have,*

$$\Lambda(X_i,\theta) = \sum_{\ell=1}^M \pi_\ell\ \lambda(X_i,\tau_\ell) \tag{8}$$

where

$$\begin{aligned}\lambda(X_i,\tau_\ell) \quad &= \quad \frac{1}{\sqrt{\det(I_d + V_i\Omega_\ell)}} \exp\left[-\frac{1}{2}(\boldsymbol{\mu}_\ell - V_i^{-1}U_i)' R_{i,\ell}^{-1}(\boldsymbol{\mu}_\ell - V_i^{-1}U_i)\right] \\ &\quad \times \exp\left(\frac{1}{2}U_i' V_i^{-1} U_i\right) = \frac{dQ_{\tau_\ell}}{dQ_0^{x,T}}(X_i)\end{aligned}$$

Recall that $n_d(x,(\boldsymbol{\mu},\Omega))$ denotes the Gaussian density with mean $\boldsymbol{\mu}$ and covariance matrix Ω. Then, we have

$$\lambda(X_i,\tau_\ell) = \sqrt{2\pi\det(V_i)}\exp\left(\frac{1}{2}U_i'V_i^{-1}U_i\right)\ n_d(U_i,(V_i\boldsymbol{\mu}_\ell,(I_d+\Omega_\ell V_i)V_i)). \tag{9}$$

The formula for $\lambda(X_i,\tau_\ell)$ was obtained in Delattre *et al.* (2013) (Propositions 4, 9 and Lemma 2). The exact likelihood (7) is explicit. We can therefore study the asymptotic behaviour of the exact maximum likelihood estimator. To compute it, instead of maximizing the likelihood, we proceed using the EM-algorithm which performs well and rapidly for parameter estimation in mixture models.

3 Algorithm of estimation

In the case of mixtures distributions with a known number of components, instead of solving the likelihood equation, it is standard and much less cumbersome to use the EM algorithm for finding a stationary point of the log-likelihood. Below, we recall the EM algorithm and discuss its convergence. Then, a penalized criterion is used to estimate the number of components in the mixture.

3.1 EM algorithm

The modeling of ϕ_i by a mixture of distributions means that the population of individuals is divided in M clusters. More precisely, for the individual i, we may introduce a random variable $Y_i \in \{1, \ldots, M\}$, with $P_\theta(Y_i = \ell) = \pi_\ell$ and $P_\theta(\phi_i \in d\varphi | Y_i = \ell) = \mathcal{N}_d(\boldsymbol{\mu}_\ell, \Omega_\ell)$. We assume that (ϕ_i, Y_i) are *i.i.d.* and $(\phi_i, Y_i)_{i=1,\ldots,N}$ independent of $(W_1, \ldots, W_N)$.
The idea of the EM algorithm (Dempster *et al.*, 1977) is to consider the data (X_i) as incomplete and to introduce the unobserved variables $(Y_1, \ldots, Y_N)$. However in the algorithm, it is simpler to consider random variables $Z = (Z_i)_{i=1,\ldots,N}$, $Z_i = (Z_{i1}, \ldots, Z_{iM})$ whose values indicate which density component drives the equation of subject i, i.e. $Z_{i\ell} = \mathbf{1}_{(Y_i=\ell)}$, for $\ell = 1, \ldots, M$. The logarithm of the likelihood function of the complete data (X_i, Z_i) is explicitly given by

$$\mathcal{L}_N((X_i, Z_i); \theta) = \sum_{i=1}^{N} \sum_{\ell=1}^{M} Z_{il} \log(\pi_\ell \ \lambda(X_i, \tau_\ell)) \tag{10}$$

The EM algorithm is an iterative algorithm which alternates between the Expectation step (E-step) which is the computation of $Q(\theta, \theta') = \mathbb{E}(\mathcal{L}_N((X_i, Z_i); \theta) | (X_i); \theta')$ and the Maximization step (M-step) which is the maximization of $Q(\theta, \theta')$ with respect to θ. Here, $\mathbb{E}(.|(X_i); \theta')$ is the conditional expectation given (X_i) computed with the distribution of the complete data under the value θ' of the parameter (with the adequate augmented sample space).
For the E-step, we compute $Q(\theta, \theta') = \sum_{i=1}^{N} \sum_{\ell=1}^{M} \tilde{\pi}_\ell(X_i, \theta') \log(\pi_\ell \ \lambda(X_i, \tau_\ell))$ where $\tilde{\pi}_\ell(X_i, \theta')$ is the *posterior* probability:

$$\tilde{\pi}_\ell(X_i, \theta') := \mathbb{P}(Z_{i\ell} = 1 | X_i, \theta') = \frac{\pi'_\ell \lambda(X_i, \tau'_\ell)}{\Lambda(X_i, \theta')} \tag{11}$$

At iteration m of the EM algorithm, one wants to maximize $Q(\theta, \hat{\theta}^{(m)})$ w.r.t. to θ where $\hat{\theta}^{(m)}$ is the current value of θ. We can maximize the term containing π_ℓ and the term containing $\tau_\ell = (\boldsymbol{\mu}_\ell, \Omega_\ell)$ separately. To maximize w.r.t. π_ℓ, we introduce one Lagrange multiplier α with the constraint $\sum_{\ell=1}^{M} \pi_\ell = 1$ and solve the following equation

$$\frac{\partial}{\partial \pi_\ell} \left[\sum_{i=1}^{N} \sum_{\ell=1}^{M} \tilde{\pi}_\ell(X_i, \hat{\theta}^{(m)}) \log(\pi_\ell) + \alpha \left(\sum_{\ell} \pi_\ell - 1 \right) \right] = 0.$$

This yields the classical solution:

$$\hat{\pi}_\ell^{(m+1)} = \frac{1}{N}\sum_{i=1}^{N} \tilde{\pi}_\ell(X_i, \hat{\theta}^{(m)}).$$

Then we maximize $\sum_{i=1}^{N}\sum_{\ell=1}^{M} \tilde{\pi}_\ell(X_i, \hat{\theta}^{(m)}) \log \lambda(X_i, \tau_\ell)$. For this, we compute the derivatives w.r.t to the components of $\boldsymbol{\mu}_\ell$ and Ω_ℓ.
When the Ω_ℓ's are known, we obtain the explicit solutions for $\ell = 1, \ldots, M$:

$$\widehat{\mu}_\ell^{(m+1)} = \left(\sum_{i=1}^{N} \tilde{\pi}_\ell(X_i, \hat{\theta}^{(m)})(I_d + \Omega_\ell V_i)^{-1} V_i\right)^{-1} \sum_{i=1}^{N} \tilde{\pi}_\ell(X_i, \hat{\theta}^{(m)})(I_d + \Omega_\ell V_i)^{-1} U_i. \tag{12}$$

Otherwise, we have a system of up-dating parameters.

Proposition 2. *The sequence* $(\hat{\theta}^{(m)})$ *generated by the EM algorithm converges to a stationary point of the likelihood.*

A crucial issue for EM algorithm is the initialization. If we had direct observations of the ϕ_i's, following Devijver (2014), we would initiate the EM algorithm with the k-means method on the ϕ_i's. As we do not have the direct observations, we propose to replace them by the estimator $\hat{\phi}_i = V_i^{-1} U_i$. This estimator is exactly the maximum likelihood estimator of φ when φ is fixed (Comte *et al.*, 2013; Dion and Genon-Catalot, 2015; Genon-Catalot and Larédo, 2015). Then we initialize the EM algorithm with the k-means method on the $\hat{\phi}_i$'s. Although these estimators of the ϕ_i's may be rough (they have no reason to be consistent for fixed T for example), this proposal provides good starting values (see Section 5). Finally, to stop the EM algorithm, we fix a maximum number of iterations and check that the sequence $(\hat{\theta}^{(m)})$ is stabilized.
Let us denote $\hat{\hat{\theta}}$ the estimator produced at the end of the EM iterations.
For an individual i, it is interesting to know to which cluster it belongs. This will be done by estimating the posterior probabilities $\tilde{\pi}_\ell(X_i, \theta)$ (11). Standardly, we decide that individual i belongs to cluster ℓ if

$$\tilde{\pi}_\ell(X_i, \hat{\hat{\theta}}) = \arg \max_{1 \le \ell' \le M} \tilde{\pi}_{\ell'}(X_i, \hat{\hat{\theta}}) \tag{13}$$

3.2 Selection of the component number

When M is unknown, an estimation procedure for M has to be implemented after having obtained an estimator of $\theta = \theta(M)$. We propose to use BIC. BIC is

a penalized criterion of the likelihood with a penalty proportional to the number of model parameters and the logarithm of the number of observations. BIC is defined as follows:

$$BIC(M, \theta(M)) = -2\log(L_N(\theta(M))) + \alpha(d, M)\log(N), \tag{14}$$

where $\alpha(d, M) = M(d+1)(d/2+1) - 1$ if all the matrices Ω_ℓ have non-null entries and $\alpha(d, M) = M(2d+1) - 1$ if all the matrices Ω_ℓ are diagonal. Then we select the number of mixture components as follows:

$$\hat{M} = \arg\min_{M \in \{1,2,\ldots,M_N\}} BIC(M, \hat{\hat{\theta}}(M)) \tag{15}$$

where $\hat{\hat{\theta}}(M)$ is the estimator obtained at the end of the iterations of the EM estimation algorithm for M components and $M_N \leq N$. The properties of BIC have been theoretically studied in finite mixture models. In particular, Leroux (1992) established that, asymptotically when $N \to +\infty$, it does not underestimate the number of components a.s.. Keribin (2000) proved that the selected number of components by BIC converges to the true number of mixture components a.s. when $N \to +\infty$.

4 Consistency of the maximum likelihood estimator with known number of mixture components

In this section, the number of components M is supposed to be known and our aim is to investigate theoretically the asymptotic properties of the exact maximum likelihood estimator of θ_0. To avoid cumbersome details, we only consider the case $d = 1$. The parameter set Θ is given by:

$$\Theta = \{(\pi_\ell, \tau_\ell), \ell = 1, \ldots, M, \forall \ell \in \{1, \ldots, M-1\}, 0 < \pi_\ell < 1, 0 < 1 - \sum_{\ell=1}^{M-1} \pi_\ell < 1,$$

$$\tau_\ell = (\mu_\ell, \omega_\ell^2) \in \mathbb{R} \times (0, +\infty), \ell \neq \ell' \Rightarrow \tau_\ell \neq \tau_{\ell'}\}$$

We set $\pi_M = 1 - \sum_{\ell=1}^{M-1} \pi_\ell$, but there are only $3M - 1$ parameters to be estimated. When necessary in notations, we set $\theta = (\theta_1, \ldots, \theta_{3M-1})$.

The maximum likelihood estimator is defined as any solution of

$$\hat{\theta}_N = \arg\max_{\theta\in\Theta} L_N(\theta)$$

where L_N is defined by (7)-(8). As in Delattre *et al.* (2013), the following assumption is required to prove the identifiability property.

(H4) Either, the function $b(.)/\sigma(.)$ is constant. Or, the function $b(.)/\sigma(.)$ is not constant and under $Q_0^{x,T}$, the random variable $(U(X), V(X))$ (see (4)) admits a density $f(u,v)$ w.r.t. the Lebesgue measure on $\mathbb{R}\times(0,+\infty)$ which is jointly continuous and positive on $\mathbb{R}\times(0,+\infty)$.

The case where $b(.)/\sigma(.)$ is constant is simple. For instance, let $b(.)/\sigma(.) \equiv 1$. Then, $V(X) = T$ is deterministic, and under $Q_0^{x,T}$, $U(X) = W_T$. Under Q_θ, $U(X)$ is a mixture of Gaussian distributions with means $(\mu_\ell T)$, variances $(T(1+\omega_\ell^2 T))$ and proportions (π_ℓ).

When $b(.)/\sigma(.)$ is not constant, under smoothness assumptions on functions b, σ, assumption (H4) will be fulfilled by application of Malliavin calculus tools (see Delattre *et al.* (2013)). As we deal with mixture distributions, the identifiability of the whole parameter θ can only be obtained in the following sense:

$$\theta \sim \theta_0 \Longleftrightarrow \{(\pi_\ell, \tau_\ell), \ell = 1,\dots,M\} = \{(\pi_{\ell,0}, \tau_{\ell,0}), \ell = 1,\dots,M\}. \tag{16}$$

We can prove:

Proposition 3. *Under* (H1)-(H2)-(H4), $\mathbb{Q}_\theta = \mathbb{Q}_{\theta_0}$ implies that $\theta \sim \theta_0$.

Proposition 4. *The function* $\theta \to \log\Lambda(X,\theta)$ *is* C^∞ *on* Θ *and*

- $\mathbb{E}_{\theta_0}\left(\frac{\partial \log\Lambda(X,\theta)}{\partial\theta_k}|_{\theta=\theta_0}\right)^2 < +\infty$ *and* $\mathbb{E}_{\theta_0}\left(\frac{\partial \log\Lambda(X,\theta)}{\partial\theta_k}|_{\theta=\theta_0}\right) = 0$ *for all* $1 \le k \le 3M-1$.
- $\mathbb{E}_{\theta_0}\left|\frac{\partial^2 \log\Lambda(X,\theta)}{\partial\theta_k\partial\theta_j}|_{\theta=\theta_0}\right| < +\infty$ *for all* $1 \le k,j \le 3M-1$
- $\mathbb{E}_{\theta_0}\left(\frac{\partial \log\Lambda(X,\theta)}{\partial\theta_k}|_{\theta=\theta_0}\frac{\partial \log\Lambda(X,\theta)}{\partial\theta_j}|_{\theta=\theta_0}\right) = -\mathbb{E}_{\theta_0}\left(\frac{\partial^2 \log\Lambda(X,\theta)}{\partial\theta_k\partial\theta_j}|_{\theta=\theta_0}\right)$ *for all* $1 \le k,j \le 3M-1$.

($\mathbb{E}_{\theta_0}$ denotes the expectation under $\mathbb{Q}_{\theta_0}$).

Let us define the Fisher information matrix

$$I(\theta_0) = \left[\mathbb{E}_{\theta_0}\left(\frac{\partial \log\Lambda(X,\theta)}{\partial\theta_k}|_{\theta=\theta_0}\frac{\partial \log\Lambda(X,\theta)}{\partial\theta_j}|_{\theta=\theta_0}\right)\right]_1 \le k,j \le 3M-1.$$

Theorem 1. *Assume* (H1)-(H2) *and that* $I(\theta_0)$ *is invertible. Then, an estimator* $\hat{\theta}_N$ *exists that solves the likelihood estimating equation* $\partial L_N(\theta)/\partial\theta = 0$ *with a probability tending to* 1 *and* $\hat{\theta}_N \to \theta_0$ *in probability.*

5 Simulation study

In this section, we implement our estimating methods on simulated data. First, we assume that the number of components is known ($M = 2$). We consider three models with univariate random effect. The simulated mixed Gaussian distributions of ϕ_i are:

1. $0.5\,\mathcal{N}(-0.5, 0.25^2) + 0.5\,\mathcal{N}(-1.8, 0.25^2)$ (well separated components)
2. $0.7\,\mathcal{N}(-0.5, 0.5^2) + 0.3\,\mathcal{N}(-1.8, 0.5^2)$ (not well separated components)

A hundred repeated datasets are simulated for $N = 100$, $N = 200$. Exact simulations of discretized sample paths are performed with $T = 1$ and $\delta = 0.0002$. Tables show the mean and standard deviation of the 100 estimators.
We compare three EM algorithms. First, we apply the EM algorithm to the direct observations of ϕ_i. This is the standard EM for mixture of Gaussian distributions. Second, we apply this standard EM algorithm to the estimators of the random variables ϕ_i given by $\hat{\phi}_i = V_i^{-1}U_i$. Third, we apply the EM algorithm described in Section 3 to the trajectories X_i. In the two first cases, the EM algorithm is initiated by the k-means method (using the kmeans function of R software). The third EM algorithm is initiated with the same initialization as the second one.
Model (1). Ornstein-Uhlenbeck process with multiplicative random effect: $dX_i(t) = \phi_i X_i(t)dt + \sigma dW_i(t), X_i(0) = x$.
Model (2). Ornstein-Uhlenbeck process with additive random effect: $dX_i(t) = (\phi_i - X_i(t))dt + \sigma dW_i(t), X_i(0) = x$
For both models, exact simulations relying on the explicit solutions are performed with $\sigma = 0.1$ and $x = 1$. Results are given in Tables 1 and 2.
Model (3). Square-root process:
$dX_i(t) = (\phi_i X_i(t) + \alpha)dt + c\sqrt{X_i(t)^+}dW_i(t), X_i(0) = x > 0$ with $a(x) = \alpha$ a constant. The study of the process $X^{\varphi,\alpha}(t)$ with fixed effect given by

$$d\,X^{\varphi,\alpha}(t) = (\varphi\, X^{\varphi,\alpha}(t) + \alpha)dt + c\sqrt{X^{\varphi,\alpha}(t)}\,dW(t), \quad X^{\varphi,\alpha}(0) = x > 0, \qquad (17)$$

Parameters	μ_1	μ_2	ω_1	ω_2	π_1	π_2
Well separated components						
True values	-0.5	-1.8	0.25	0.25	0.5	0.5
$N = 100$, $T = 1$						
EM on ϕ_i	-0.50 (0.04)	-1.80 (0.04)	0.25 (0.03)	0.25 (0.03)	0.50 (0.05)	0.50 (0.05)
EM on $\hat{\phi}_i$	-0.49 (0.04)	-1.76 (0.05)	0.27 (0.03)	0.31 (0.04)	0.49 (0.06)	0.51 (0.06)
EM on X_i	-0.48 (0.04)	-1.76 (0.05)	0.24 (0.04)	0.23 (0.05)	0.50 (0.06)	0.50 (0.06)
$N = 200$, $T = 1$						
EM on ϕ_i	-0.50 (0.03)	-1.80 (0.03)	0.25 (0.02)	0.25 (0.02)	0.50 (0.04)	0.50 (0.04)
EM on $\hat{\phi}_i$	-0.48 (0.03)	-1.77 (0.04)	0.27 (0.02)	0.32 (0.03)	0.49 (0.04)	0.51 (0.04)
EM on X_i	-0.48 (0.03)	-1.76 (0.04)	0.25 (0.03)	0.23 (0.03)	0.50 (0.04)	0.50 (0.04)
Not well separated components						
True values	-0.5	-1.8	0.5	0.5	0.7	0.3
$N = 100$, $T = 1$						
EM on ϕ_i	-0.46 (0.09)	-1.74 (0.21)	0.48 (0.06)	0.50 (0.11)	0.65 (0.09)	0.35 (0.09)
EM on $\hat{\phi}_i$	-0.44 (0.11)	-1.68 (0.23)	0.48 (0.07)	0.54 (0.12)	0.63 (0.11)	0.37 (0.11)
EM on X_i	-0.44 (0.11)	-1.70 (0.23)	0.47 (0.07)	0.48 (0.14)	0.65 (0.10)	0.35 (0.10)
$N = 200$, $T = 1$						
EM on ϕ_i	-0.47 (0.07)	-1.71 (0.18)	0.48 (0.05)	0.54 (0.08)	0.66 (0.07)	0.34 (0.07)
EM on $\hat{\phi}_i$	-0.44 (0.07)	-1.64 (0.17)	0.48 (0.05)	0.58 (0.07)	0.63 (0.07)	0.37 (0.07)
EM on X_i	-0.44 (0.07)	-1.66 (0.17)	0.47 (0.05)	0.53 (0.08)	0.65 (0.07)	0.35 (0.07)

Table 1: Model (1) with two different distributions of the random effects: well separated components (top of the table) and not well separated components (bottom of the table). Two designs are illustrated: $N = 100, T = 1$, $N = 200, T = 1$. For each design, mean and standard deviation (SD) are computed from 100 simulated datasets.

is detailed in Overbeck (1998). If $p \in \mathbb{N}^*$, the process $X(t) = \sum_{j=1}^{p} \xi_j^2(t), t \geq 0$ where $\xi_j, j = 1, \ldots, p$ are *i.i.d.* Ornstein-Uhlenbeck processes given by: $d\xi_j(t) = \theta\, \xi_j(t)dt + \sigma dB_j(t), \xi_j(0) = y_j$, satisfies

$$dX(t) = (2\theta X(t) + p\sigma^2)dt + 2\sigma\sqrt{X(t)}d\beta(t), \quad X(0) = x = \sum_{j=1}^{p} y_j^2$$

where $\beta(t)$ is a standard Brownian motion ($\varphi = 2\theta, \sigma = c/2, \alpha = pc^2/4$). For $p \geq 2$, the process $X(t)$ is always positive. Relying this property, we use exact simulations with $p = 2$, $\sigma = 0.05$, $\alpha = 1/40$ and $x = 1$. Results are given in Table 3.

Results. Whatever the simulated model, the parameters are well estimated overall for well and not well-separated mixture components. When the mixture components are not well separated, we see that the estimation is more difficult

Parameters	μ_1	μ_2	ω_1	ω_2	π_1	π_2
Well separated components						
True values	-0.5	-1.8	0.25	0.25	0.5	0.5
$N = 100$, $T = 1$						
EM on ϕ_i	-0.50 (0.04)	-1.80 (0.04)	0.25 (0.03)	0.24 (0.03)	0.50 (0.05)	0.50 (0.05)
EM on $\hat{\phi}_i$	-0.51 (0.04)	-1.80 (0.04)	0.27 (0.03)	0.26 (0.03)	0.50 (0.05)	0.50 (0.05)
EM on X_i	-0.51 (0.04)	-1.80 (0.04)	0.25 (0.04)	0.24 (0.03)	0.50 (0.05)	0.50 (0.05)
$N = 200$, $T = 1$						
EM on ϕ_i	-0.50 (0.03)	-1.80 (0.02)	0.25 (0.02)	0.25 (0.02)	0.50 (0.04)	0.50 (0.04)
EM on $\hat{\phi}_i$	-0.50 (0.03)	-1.80 (0.03)	0.27 (0.03)	0.27 (0.02)	0.50 (0.04)	0.50 (0.04)
EM on X_i	-0.50 (0.03)	-1.80 (0.03)	0.25 (0.03)	0.25 (0.02)	0.50 (0.04)	0.50 (0.04)
Not well separated components						
True values	-0.5	-1.8	0.5	0.5	0.7	0.3
$N = 100$, $T = 1$						
EM on ϕ_i	-0.45 (0.11)	-1.70 (0.21)	0.47 (0.07)	0.52 (0.12)	0.65 (0.10)	0.35 (0.10)
EM on $\hat{\phi}_i$	-0.44 (0.12)	-1.70 (0.21)	0.47 (0.07)	0.53 (0.12)	0.65 (0.11)	0.35 (0.11)
EM on X_i	-0.44 (0.11)	-1.70 (0.21)	0.46 (0.07)	0.52 (0.12)	0.65 (0.10)	0.35 (0.10)
$N = 200$, $T = 1$						
EM on ϕ_i	-0.47 (0.07)	-1.74 (0.16)	0.48 (0.04)	0.51 (0.08)	0.66 (0.07)	0.34 (0.07)
EM on $\hat{\phi}_i$	-0.46 (0.08)	-1.72 (0.16)	0.49 (0.04)	0.53 (0.08)	0.65 (0.07)	0.35 (0.07)
EM on X_i	-0.46 (0.08)	-1.72 (0.16)	0.47 (0.04)	0.52 (0.08)	0.65 (0.07)	0.35 (0.07)

Table 2: Model (2) with two different distributions of the random effects: well separated components (top of the table) and not well separated components (bottom of the table). Two designs are illustrated: $N = 100, T = 1$, $N = 200, T = 1$. For each design, mean and standard deviation (SD) are computed from 100 simulated datasets.

Parameters	μ_1	μ_2	ω_1	ω_2	π_1	π_2
Well separated components						
True values	-0.5	-1.8	0.25	0.25	0.5	0.5
$N = 100$, $T = 1$						
EM on ϕ_i	-0.50 (0.04)	-1.81 (0.04)	0.25 (0.03)	0.25 (0.03)	0.50 (0.05)	0.50 (0.05)
EM on $\hat{\phi}_i$	-0.51 (0.04)	-1.81 (0.04)	0.28 (0.03)	0.28 (0.03)	0.50 (0.05)	0.50 (0.05)
EM on X_i	-0.52 (0.04)	-1.81 (0.04)	0.25 (0.04)	0.24 (0.03)	0.50 (0.05)	0.50 (0.05)
$N = 200$, $T = 1$						
EM on ϕ_i	-0.50 (0.03)	-1.80 (0.02)	0.25 (0.02)	0.25 (0.02)	0.50 (0.03)	0.51 (0.03)
EM on $\hat{\phi}_i$	-0.50 (0.03)	-1.80 (0.03)	0.28 (0.02)	0.27 (0.02)	0.50 (0.03)	0.51 (0.03)
EM on X_i	-0.51 (0.03)	-1.80 (0.03)	0.25 (0.03)	0.25 (0.02)	0.49 (0.03)	0.51 (0.03)
Not well separated components						
True values	-0.5	-1.8	0.5	0.5	0.7	0.3
$N = 100$, $T = 1$						
EM on ϕ_i	-0.46 (0.11)	-1.70 (0.24)	0.48 (0.07)	0.50 (0.12)	0.65 (0.10)	0.35 (0.10)
EM on $\hat{\phi}_i$	-0.47 (0.11)	-1.70 (0.24)	0.50 (0.07)	0.51 (0.12)	0.65 (0.10)	0.35 (0.10)
EM on X_i	-0.47 (0.11)	-1.70 (0.24)	0.48 (0.07)	0.49 (0.13)	0.65 (0.10)	0.35 (0.10)
$N = 200$, $T = 1$						
EM on ϕ_i	-0.47 (0.07)	-1.73 (0.15)	0.49 (0.05)	0.51 (0.08)	0.66 (0.06)	0.34 (0.06)
EM on $\hat{\phi}_i$	-0.47 (0.08)	-1.72 (0.16)	0.50 (0.05)	0.52 (0.09)	0.66 (0.07)	0.34 (0.07)
EM on X_i	-0.47 (0.07)	-1.72 (0.16)	0.48 (0.05)	0.51 (0.09)	0.66 (0.07)	0.34 (0.07)

Table 3: Model (3) with two different distributions of the random effects: well separated components (top of the table) and not well separated components (bottom of the table). Two designs are illustrated: $N = 100, T = 1$, $N = 200, T = 1$. For each design, mean and standard deviation (SD) are computed from 100 simulated datasets.

	Well-separated		Not well-separated	
	$N = 100$	$N = 200$	$N = 100$	$N = 200$
Model (1)	98.32	98.57	87.64	88.51
Model (2)	99.03	99.19	88.73	89.31
Model (3)	99.06	99.09	87.77	89.54

Table 4: Classification rate for two different distributions of the random effects: well separated components (top of the table) and not well separated components (bottom of the table). Two designs are illustrated: $N = 100, T = 1$, $N = 200, T = 1$. For each design and each diffusion model, the correct classification rate is computed from 100 simulated datasets.

(larger bias and SD). The estimation based on the data X_i is close to the ideal case of direct observation of the ϕ_i's. The estimation based on the $\hat{\phi}_i$'s is worse than the two others. This is not surprising as $\hat{\phi}_i$ has no reason to be a consistent estimator of the ϕ_i's for fixed T. Here, the interest of the $\hat{\phi}_i$'s is that they allow to produce a good initialisation for the EM-algorithm based on the X_i's.
We use formula (13) to classify individuals in two groups. By the simulation, we know exactly which group the individuals belong to. Thus we can compare the exact and the estimated classifications. Table 4 shows the rate of correct classification. For well separated groups, the classification rate is almost perfect. Even for not well separated groups, the classification rate is very satisfactory.

Second, we assume that the number of components is unknown. We consider model (2). Model (2) was simulated with the true number of components $M_0 = 3$. Three different distributions for ϕ_i were simulated:

1. $0.2\,\mathcal{N}(-0.5, 0.25^2) + 0.3\,\mathcal{N}(-3.5, 0.25^2) + 0.5\,\mathcal{N}(-5.5, 0.25^2)$
2. $0.2\,\mathcal{N}(-0.5, 0.25^2) + 0.3\,\mathcal{N}(-1.8, 0.25^2) + 0.5\,\mathcal{N}(-2.5, 0.25^2)$
3. $0.2\,\mathcal{N}(-0.5, 0.5^2) + 0.3\,\mathcal{N}(-1.8, 0.5^2) + 0.5\,\mathcal{N}(-2.5, 0.5^2)$

Figure 5 plots the densities of these three Gaussian mixtures, the first one with well-separated components, the two others with not well separated components.

The selection of $\hat{M}$ according to (15) was applied to a hundred repetitions with $M_N = 4$. Design is $N = 100, T = 1$. A similar procedure is applied on the $\hat{\phi}_i$'s and when the ϕ_i's are directly observed. Results are presented in Table 5.
Table 5 shows that the selections given by the SDEs and the ϕ_i's are quite close. The true number of components is well selected when the mixture components

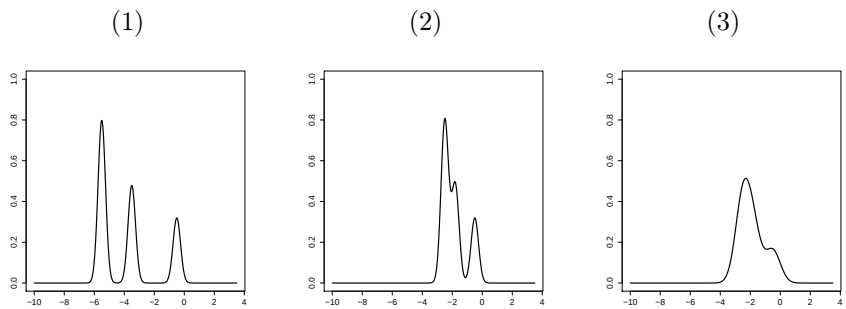

Figure 1: Density of three distributions of the random effects: (1)$0.2\ \mathcal{N}(-0.5, 0.25^2) + 0.3\ \mathcal{N}(-3.5, 0.25^2) + 0.5\ \mathcal{N}(-5.5, 0.25^2)$ (2) $0.2\ \mathcal{N}(-0.5, 0.25^2) + 0.3\ \mathcal{N}(-1.8, 0.25^2) + 0.5\ \mathcal{N}(-2.5, 0.25^2)$ and (3) $0.2\,\mathcal{N}(-0.5, 0.5^2) + 0.3\,\mathcal{N}(-1.8, 0.5^2) + 0.5\,\mathcal{N}(-2.5, 0.5^2)$

Selected number	Distribution (1)			Distribution (2)			Distribution (3)		
of components	SDE	$\hat{\phi}$	ϕ	SDE	$\hat{\phi}$	ϕ	SDE	$\hat{\phi}$	ϕ
1	0	0	0	0	0	0	49	47	46
2	0	0	0	76	70	69	51	53	54
3	89	88	84	24	30	31	0	0	0
4	11	12	16	0	0	0	0	0	0

Table 5: Model (2) with three different distributions of the random effects, $N = 100, T = 1$. True number of components $M_0 = 3$. From a hundred simulated datasets, the frequency of selected numbers of components with the BIC procedure is given. BIC is applied with estimation based on the SDE or the $\hat{\phi}_i$'s or the ϕ_i's.

are well separated (distribution (1)). When the mixture components become confused (distributions (2) and (3)), the performances of the criterion deteriorate. For distribution (3), BIC selects one or two components rather than three. Such a result is expected given the plot of the mixture distribution (Figure 5). In Table 6, we illustrate the validity of the classification rule (13) on model (2) with distributions (1) and (2). For each distribution, we have chosen one dataset for which BIC has selected the true number of components. Table 6 shows that the classification rule performs very well.

6 Clustering of real growth data

We study growth curve data, i.e. repeated measurements of a continuous growth process over time in a population of individuals. Data analyzed in this paper are

	EM classification of individuals					
	Distribution 1			Distribution 2		
True classes	1	2	3	1	2	3
1	22	0	0	22	1	0
2	0	27	0	0	29	0
3	0	0	51	0	9	39

Table 6: Model (2) with well separated components (distribution (1)) on the left and not well separated components (distribution 2) on the right, $N = 100, T = 1$, $M_0 = 3$. Classification of the individuals among the estimated components on two simulated datasets for which BIC selected 3 components.

chicken growth data described in Jaffrézic *et al.* (2006); Donnet *et al.* (2010). The aim is to differentiate animal phenotypes by characterizing their growth dynamics. The individuals are from four different genetic lines, with different expected juvenile and adult body weights. Line LH was selected for low juvenile weight at 8 weeks and high adult weight at 36 weeks (31 individuals), line HL was selected for high juvenile weight and low adult weight (35 individuals), line LL (55 individuals) was selected for low weights at both ages, and line HH for high weights (32 individuals). The data set comprised $N = 153$ chickens and 12 measurements for each animal at ages $0, 4, 6, 8, 12, 16, 20, 24, 32, 36$ and 40 weeks.

Donnet *et al.* (2010) have proved that a deterministic modeling of these growth curves is not appropriate and that SDE fits better the data (see also Filipe *et al.* (2010, 2013)). In Donnet *et al.* (2010), the logarithm of the weight was modeled by a SDE with a bivariate random effects in the drift. Only Gaussian distribution for the random effects was considered and mixture of distributions was not tested. The objective of this new analysis is to test the existence of a mixture of distribution and to classify individuals into groups. Then we will compare the estimated clusters to their genetic lines.

Let us denote $y_i(t)$ the weight at time t of individual i, $i = 1, \ldots, N$. Set $X_i(t) = \log(y_i(t))$. See Figure 2 where only 15 individuals curves of each genetic lines are plotted. Two models are considered and compared on the data. Donnet *et al.* (2010)'s model, called bilinear, is the following

$$dX_i(t) = (\phi_{i1} - \phi_{i2} X_i(t))dt + \sigma X_i(t) dW_i(t) \tag{18}$$

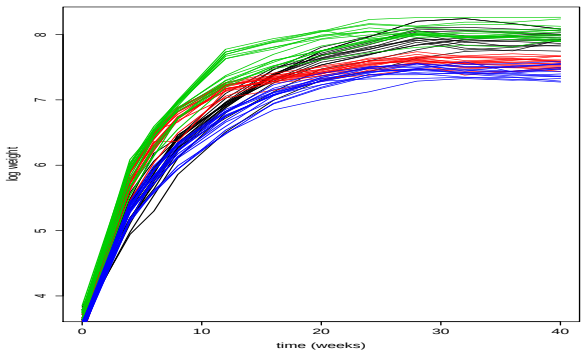

Figure 2: Real growth data: 15 individual trajectories of each genetic lines represented by four different colors (black, red, green, blue).

We also consider a Ornstein-Uhlenbeck process with a bivariate random effects:

$$dX_i(t) = (\phi_{i1} - \phi_{i2} X_i(t))dt + \sigma dW_i(t) \tag{19}$$

For both models, $X_i(0)$ is assumed known and observed. The random effects $\phi_i = (\phi_{i1}, \phi_{i2})$ have a common distribution given by a mixture of Gaussian distributions. Parameter σ is estimated as $\hat{\sigma}^2 = \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{11}\frac{(X_i(t_j)-X_i(t_{j-1}))^2}{(t_j-t_{j-1})X_i(t_{j-1})^2}$ and $\hat{\sigma}^2 = \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{11}\frac{(X_i(t_j)-X_i(t_{j-1}))^2}{t_j-t_{j-1}}$, respectively.
The estimation procedure described previously is applied to the dataset with $M_N = 8$. The two models (18) and (19) are compared with BIC. BIC is always lower for the Ornstein-Uhlenbeck process than for the bilinear model. This is why in the following, we focus on the Ornstein-Uhlenbeck model. Figure 3 shows BIC with respect to M for the Ornstein-Uhlenbeck process: the selected number of components is $\hat{M} = 4$. Estimated distributions of (ϕ_{i1}, ϕ_{i2}) are presented in Figure 3 (middle and right plots). The plot shows that they are very far from a uni-modal Gaussian distribution and the four components are well separated. It is worth stressing that BIC selects exactly the number of genetic lines of the studied population, without any prior knowledge.
We use the classification rule (13) to assign individuals in the $\hat{M} = 4$ groups. Comparison between these clusters and the true genetic lines is presented in Table 7. Ideally, we would like to have only non null numbers on the diagonal. This is not exactly the case but the four first clusters fit reasonably well the four genetic lines. The two first clusters mix lines LH and LL, which differ only

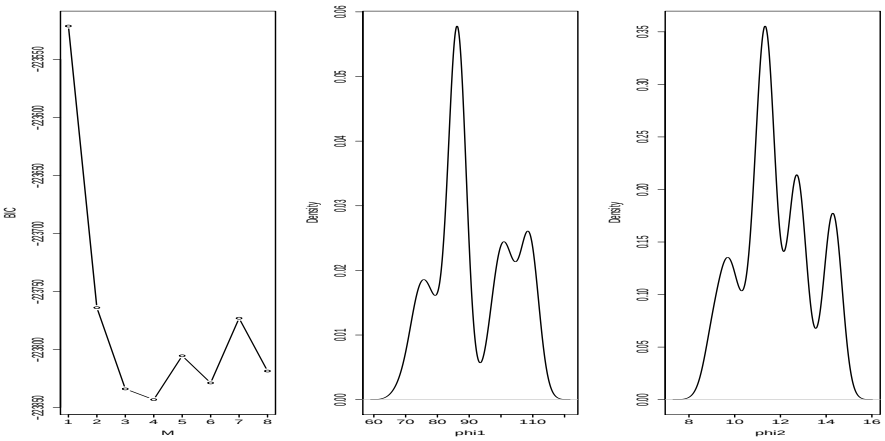

Figure 3: Real growth data. Left: BIC estimated for the Ornstein-Uhlenbeck process with two random effects distributed with a Gaussian mixture of M components. Middle: marginal distribution of ϕ_1 estimated from the model with $\hat{M} = 4$ components. Right: marginal estimated distribution of ϕ_2 estimated from the model with $\hat{M} = 4$ components.

through the expected adult weight. Similarly, clusters 3 and 4 mix lines HH and HL, which differ only through the expected adult weight.

7 Concluding remarks

In this paper, we study mixed-effects SDEs continuously observed with a multivariate linear random effect in the drift. The distribution of the random effect is a mixture of Gaussian distributions. When the number of components is known, the exact likelihood is explicit and we prove that the maximum likeli-

Estimated	Genetic classes				
classes	LH	LL	HL	HH	Total
1	17	13	0	0	30
2	14	40	5	1	60
3	0	2	26	9	37
4	0	0	4	22	26
Total	31	55	35	32	153

Table 7: Real growth data: Comparison of the estimated clustering with the genetic lines.

hood estimator is consistent. This estimator is computed via the EM algorithm. Afterwards, the number of components is estimated by BIC and a classification rule is proposed to assign individuals to the classes.
The method is implemented on simulated data with a univariate random effect and shows good performances. Then, it is applied to study growth curve data in a population of chickens coming from four genetic lines. Using an Ornstein-Uhlenbeck model with a bivariate random effect, the results are very convincing as the estimation method recovers four groups. Classification of individuals fits reasonably well the genetic lines.
An interesting direction would be to consider other criteria than BIC to select the number of components. For instance, one could investigate the criterion developed in Maugis and Michel (2011) which is proposed for direct observations of Gaussian mixtures.
We assume that the trajectories X_i's are continuously observed throughout a time interval, which is not realistic in practice. As it is explained in Delattre *et al.* (2013) (Section 6), to build estimators, we can replace the stochastic and deterministic integrals by their discretized versions. This discretisation is used in the simulations and real data analysis.

Acknowledgments

This work has been partially supported by the LabEx PERSYVAL-Lab (ANR-11-LABX-0025-01).

References

Arribas-Gil, A., De la Cruz, R., Lebarbier, E. and Meza, C. (2015). Classification of longitudinal data through a semiparametric mixed-effects model based on lasso-type estimators. *Biometrics* **71**, 333–343.

Celeux, G., Martin, O. and Lavergne, C. (2005). Mixture of linear mixed models - application to repeated data clustering. *Statistical Modelling* **5**, 243–267.

Comte, F., Genon-Catalot, V. and Samson, A. (2013). Nonparametric estimation for stochastic differential equations with random effects. *Stoch. Proc. Appl.* **123**, 2522–2551.

Delattre, M., Genon-Catalot, V. and Samson, A. (2013). Maximum likelihood estimation for stochastic differential equations with random effects. *Scandinavian Journal of Statistics* **40(2)**, 322–343.

Dempster, A., Laird, N. and Rubin, D. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Jr. R. Stat. Soc. B* **39**, 1–38.

Devijver, E. (2014). Model-based clustering for high-dimensional data. application to functional data. *arXiv:1409.1333* .

Dion, C. and Genon-Catalot, V. (2015). Bidimensional random effect estimation in mixed stochastic differential model. *Statistical Inference for Stochastic processes* **18**.

Ditlevsen, S. and De Gaetano, A. (2005). Stochastic vs. deterministic uptake of dodecanedioic acid by isolated rat livers. *Bull. Math. Biol.* **67**, 547–561.

Donnet, S., Foulley, J. and Samson, A. (2010). Bayesian analysis of growth curves using mixed models defined by stochastic differential equations. *Biometrics* **66**, 733ï£¡741.

Donnet, S. and Samson, A. (2008). Parametric inference for mixed models defined by stochastic differential equations. *ESAIM P&S* **12**, 196–218.

Filipe, P., Braumann, C., Brites, N. and Roquete, C. (2010). Modelling animal growth in random environments: an application using nonparametric estimation. *Biometrical Journal* **52**, 653–666.

Filipe, P., Braumann, C., Brites, N. and Roquete, C. (2013). *Recent development in Modeling and applications in Statistics*. Springer.

Genon-Catalot, V. and Larédo, C. (2015). Estimation for stochastic differential equations with mixed effects. *Hal-00807258* **V2**.

Jacques, J. and Preda, C. (2014). Functional data clustering: a survey. *Advances in Data Analysis and Classification* **8**, 231–255.

Jaffrézic, F., Meza, C., Lavielle, M. and Foulley, J. (2006). Genetic analysis of growth curves using the SAEM algorithm. *Genet. Sel. Evol.* **38**, 583–600.

Keribin, C. (2000). Consistent estimation of the order of mixture models. *Sankhya: The Indian Journal of Statistics* **62**, 49–66.

Leroux, B. (1992). Maximum penalized likelihood estimation for independent and markov-dependent mixture models. *Biometrics* **48**, 545–558.

Lipster, R. and Shiryaev, A. (2001). *Statistics of random processes I : general theory.* Springer.

Maugis, C. and Michel, B. (2011). A non asymptotic penalized criterion for gaussian mixture model selection. *ESAIM : P&S* **15**, 41–68.

McLachlan, G. and Krishnan, T. (2008). *The EM Algorithm and Extensions.*

Overbeck, L. (1998). Estimation for continuous branching processes. *Scandinavian Journal of Statistics* **25**, 111–126.

Picchini, U., De Gaetano, A. and Ditlevsen, S. (2010). Stochastic differential mixed-effects models. *Scand. J. Statist.* **37**, 67–90.

8 Appendix

8.1 Proof of Proposition 2

To simplify notations, we prove the convergence for $d = 1$. We use the results recalled in McLachlan and Krishnan (2008). The conditions are the following:

1. $\Theta \subset \mathbb{R}^{3M-1}$
2. $\Theta_{\theta_0} = \{\theta \in \Theta, L_N(\theta) \geq L_N(\theta_0)\}$ is a compact set if $L_N(X, \theta_0) > -\infty$
3. $L_N(\theta)$ is continuous on Θ and differentiable on the interior of Θ.
4. $Q(\theta, \theta')$ is continuous with respect to both θ and θ'.
5. $\frac{\partial Q(\theta, \hat{\theta}^{(m)})}{\partial \theta}\Big|_{\theta = \hat{\theta}^{(m+1)}} = 0$
6. $\frac{\partial Q(\theta, \theta')}{\partial \theta}$ is continuous in both θ and θ'.

Conditions 3, 4, 5 and 6 are verified by regularity of the likelihood (see Proposition 4). Condition 2 is usually not verified by a standard Gaussian mixture (see McLachlan and Krishnan (2008)). However, here, one has (see (9)):

$$\Lambda(X, \theta) \propto \sum_{\ell=1}^{M} \pi_\ell \, n(U, (\mu_\ell V, \sigma_\ell^2(V)))$$

where $\sigma_\ell^2(V) = V(1+\omega_\ell^2 V) \geq V > 0$. So, $\Lambda(X,\theta)$ has the form of a mixture of Gaussian distributions with variances all bounded from below. This implies condition 2.□

8.2 Proof of Proposition 3

We first treat the case of $b(.)/\sigma(.)$ non constant. Let us consider two parameters θ and θ_0. We want to prove that $\mathbb{Q}_\theta = \mathbb{Q}_{\theta_0}$ implies $\theta \sim \theta_0$. As $\Lambda(X,\theta)$ and $\lambda(X,\tau_\ell)$ depend on X only through the statistics $U(X) := U, V(X) := V$, with a slight abuse of notation, we set:

$$\Lambda(X,\theta) = \Lambda(U,V,\theta), \qquad \lambda(X,\tau_\ell) = \lambda(U,V,\theta_\ell). \tag{20}$$

Under (H4), $\Lambda(u,v,\theta)$ is the density of the distribution of (U,V) under $\mathbb{Q}_\theta$ w.r.t. the density of (U,V) under $Q_0^{x,T}$ and $\mathbb{Q}_\theta = \mathbb{Q}_{\theta_0}$ implies $\Lambda(u,v,\theta) = \Lambda(u,v,\theta_0)$ almost everywhere, hence everywhere on $\mathbb{R}\times(0,+\infty)$ by the continuity assumption. We deduce that the following equality holds, for all $u \in \mathbb{R}, v > 0$:

$$\sum_{\ell=1}^M \pi_\ell \frac{1}{\sqrt{1+\omega_\ell^2 v}} \exp\left[-\frac{v(u/v-\mu_\ell)^2}{2(1+\omega_\ell^2 v)}\right] = \sum_{\ell=1}^M \pi_{\ell,0} \frac{1}{\sqrt{1+\omega_{\ell,0}^2 v}} \exp\left[-\frac{v(u/v-\mu_{\ell,0})^2}{2(1+\omega_{\ell,0}^2 v)}\right].$$

Let us set

$$p(v) = \prod_{1\leq\ell\leq M} \sqrt{1+\omega_\ell^2 v}, \quad q_\ell(v) = \prod_{1\leq\ell'\leq M, \ell'\neq\ell} \sqrt{1+\omega_{\ell'}^2 v}$$

and $p_0(v), q_{\ell,0}(v)$ the same quantities with $(\omega_{\ell,0})$ instead of (ω_ℓ). Note that $q(v)\sqrt{1+\omega_\ell^2 v} = p(v), q_{\ell,0}(v)\sqrt{1+\omega_{\ell,0}^2 v} = p_0(v)$ so that these quantities do not depend on ℓ. By reducing to the same denominator, we get:

$$\frac{p_0(v)}{p(v)} = \frac{\sum_{\ell=1}^M \pi_{\ell,0} q_{\ell,0}(v) \exp\left[-\frac{v(u/v-\mu_{\ell,0})^2}{2(1+\omega_{\ell,0}^2 v)}\right]}{\sum_{\ell=1}^M \pi_\ell q_\ell(v) \exp\left[-\frac{v(u/v-\mu_\ell)^2}{2(1+\omega_\ell^2 v)}\right]}.$$

The left-hand side is a function of v only while the right-hand side is a function of (u,v). This is only possible if $p(v) = p_0(v)$ for all $v > 0$. Thus

$$\{\omega_1^2, \ldots, \omega_M^2\} = \{\omega_{1,0}^2, \ldots, \omega_{M,0}^2\} \tag{21}$$

and the equality of the variances is obtained with reordering the terms if needed.

Then, we have for a fixed v and $\sigma_\ell^2(v) = v(1+\omega_\ell^2 v)$,

$$\sum_{\ell=1}^{M} \pi_\ell q_\ell(v) \exp\left[-\frac{v(u/v-\mu_\ell)^2}{2(1+\omega_\ell^2 v)}\right] = p(v)\sqrt{2\pi v} \sum_{\ell=1}^{M} \pi_\ell \; n(u, (\mu_\ell v, \sigma_\ell^2(v))) \qquad (22)$$

where $n(u,(m,\sigma^2))$ denotes the Gaussian density with mean m and variance σ^2. Analogously, using the equality (21),

$$\sum_{\ell=1}^{M} \pi_{\ell,0} q_{\ell,0}(v) \exp\left[-\frac{v(u/v-\mu_{\ell,0})^2}{2(1+\omega_{\ell,0}^2 v)}\right] = p(v)\sqrt{2\pi v} \sum_{\ell=1}^{M} \pi_{\ell,0} \; n(u, (\mu_{\ell,0} v, \sigma_\ell^2(v))).$$

For all fixed $v > 0$, we thus have for all $u \in \mathbb{R}$,

$$\sum_{\ell=1}^{M} \pi_\ell \; n(u, (\mu_\ell v, \sigma_\ell^2(v))) = \sum_{\ell=1}^{M} \pi_{\ell,0} \; n(u, (\mu_{\ell,0} v, \sigma_\ell^2(v))).$$

This is the equality of two mixtures of Gaussian distributions with expectations $(\mu_\ell v)$ and $(\mu_{\ell,0} v)$, proportions (π_ℓ) and $(\pi_{\ell,0})$ respectively, and same set of known variances $(v(1+\omega_\ell^2 v))$. By identifiability of Gaussian mixtures, we obtain the equality

$$\{(\pi_\ell, \mu_\ell), l = 1, \ldots, M\} = \{(\pi_{\ell,0}, \mu_{\ell,0}), l = 1, \ldots, M\},$$

and thus $\theta \sim \theta_0$.
Now, suppose that $b(.)/\sigma(.) \equiv 1$. As noted above, under $Q_0^{x,T}$, $U = W_T$ is $\mathcal{N}(0,T)$. Under $\mathbb{Q}_\theta$, U has density $\Lambda(u,\theta)\exp(-u^2/2T)$ with respect to the Lebesgue measure on $\mathbb{R}$. This is exactly a mixture of Gaussian densities and we can conclude by the identifiability property of Gaussian mixtures.□

8.3 Proof of Proposition 4

For this proposition, we use results proved in Delattre *et al.* (2013) (section 4.2, Lemma 1, Proposition 5) that we recall now. For all $\tau = (\mu, \omega^2)$, for all $u \in \mathbb{R}$,

$$E_{Q_\tau} \exp\left(u \frac{U}{1+\omega^2 V}\right) < +\infty.$$

(Recall that Q_τ is the distribution of X_i when ϕ_i has Gaussian distribution with parameters $\tau = (\mu, \omega^2)$). This implies that $E_{Q_\tau} |\frac{U}{1+\omega^2 V}|^m < +\infty$ for all $m \geq 1$.

Let

$$\gamma(\tau) = \frac{U - \mu V}{1 + \omega^2 V}, \qquad I(\omega^2) = \frac{V}{1 + \omega^2 V}.$$

The r.v. $\gamma(\tau)$ has moments of any order under Q_τ, $I(\omega^2)$ is bounded and the following relations hold:

$$E_{Q_\tau}\gamma(\tau) = 0,\ E_{Q_\tau}(\gamma^2(\tau) - I(\omega^2)) = 0, \tag{23}$$

$$E_{Q_\tau}\left[\left(\frac{1}{2}(\gamma^2(\tau_\ell) - I(\omega_\ell^2))\right)^2 - \gamma^2(\tau_\ell)I(\omega_\ell^2) - \frac{1}{2}I^2(\omega_\ell^2)\right] = 0. \tag{24}$$

$$E_{Q_\tau}(\frac{1}{2}\gamma^3(\tau_\ell) - \frac{3}{2}\gamma(\tau_\ell)I(\omega_\ell^2)) = 0. \tag{25}$$

All derivatives of $\Lambda(X, \theta)$ are well defined. For $\ell = 1, \ldots, M-1$, we have

$$\frac{\partial \Lambda(X, \theta)}{\partial \pi_\ell} = \lambda(X, \tau_\ell) - \lambda(X, \tau_M).$$

As for all $\tau = (\mu, \omega^2)$, $Q_\tau = \lambda(X, \tau)Q_0^{x,T}$, the r.v. above are $Q_0^{x,T}$-integrable and

$$\int_{C_T} \frac{\partial \Lambda(X, \theta)}{\partial \pi_\ell} dQ_0^{x,T} = \int_{C_T} \frac{\partial \log \Lambda(X, \theta)}{\partial \pi_\ell} d\mathbb{Q}_\theta = 0.$$

Moreover, as $\lambda(X, \tau_\ell)/\Lambda(X, \theta) \leq \pi_\ell^{-1}$, we have

$$\left(\frac{\partial \log \Lambda(X, \theta)}{\partial \pi_\ell}\right)^2 \Lambda(X, \theta) = \frac{\left(\frac{\partial \Lambda(X, \theta)}{\partial \pi_\ell}\right)^2}{\Lambda(X, \theta)} \leq \frac{2}{\pi_\ell}\lambda(X, \tau_\ell) + \frac{2}{\pi_M}\lambda(X, \tau_M).$$

Therefore,

$$\mathbb{E}_{\mathbb{Q}_\theta}\left(\frac{\partial \log \Lambda(X, \theta)}{\partial \pi_\ell}\right)^2 \leq \int_{C_T}\left(\frac{2}{\pi_\ell}\lambda(X, \tau_\ell) + \frac{2}{\pi_M}\lambda(X, \tau_M)\right) dQ_0^{x,T} = \frac{2}{\pi_\ell} + \frac{2}{\pi_M}.$$

Higher order derivatives of $\Lambda(X, \theta)$ w.r.t. the π_ℓ's are nul:

$$\frac{\partial^2 \Lambda(X, \theta)}{\partial \pi_\ell \partial \pi_{\ell'}} = 0.$$

We next study the derivatives w.r.t. the parameters μ_ℓ, ω_ℓ^2. We have:

$$\frac{\partial \Lambda(X, \theta)}{\partial \mu_\ell} = \pi_\ell \frac{\partial \lambda(X, \tau_\ell)}{\partial \mu_\ell} = \pi_\ell\, \gamma(\tau_\ell)\lambda(X, \tau_\ell).$$

We know that:

$$E_{Q_{\tau_\ell}}|\gamma(\tau_\ell)| = \int_{C_T} |\gamma(\tau_\ell)|\lambda(X,\tau_\ell)dQ_0^{x,T} < +\infty,\ E_{Q_{\tau_\ell}}\gamma(\tau_\ell) = \int_{C_T} \gamma(\tau_\ell)\lambda(X,\tau_\ell)dQ_0^{x,T} = 0.$$

Consequently,

$$\int_{C_T} \left|\frac{\partial\Lambda(X,\theta)}{\partial\mu_\ell}\right| dQ_0^{x,T} < +\infty,\ \int_{C_T} \frac{\partial\Lambda(X,\theta)}{\partial\mu_\ell} dQ_0^{x,T} = \mathbb{E}_{\mathbb{Q}_\theta}\frac{\partial\log\Lambda(X,\theta)}{\partial\mu_\ell} = 0.$$

Now,

$$\left(\frac{\partial\log\Lambda(X,\theta)}{\partial\mu_\ell}\right)^2 \Lambda(X,\theta) = \frac{\left(\frac{\partial\Lambda(X,\theta)}{\partial\mu_\ell}\right)^2}{\Lambda(X,\theta)} = \frac{\pi_\ell^2\gamma^2(\tau_\ell)\lambda^2(X,\tau_\ell)}{\Lambda(X,\theta)} \leq \pi_\ell\gamma^2(\tau_\ell)\lambda(X,\tau_\ell).$$

Hence,

$$\mathbb{E}_{\mathbb{Q}_\theta}\left(\frac{\partial\log\Lambda(X,\theta)}{\partial\mu_\ell}\right)^2 \leq \pi_\ell E_{Q_{\tau_\ell}}(\gamma^2(\tau_\ell)) = \pi_\ell\, I(\omega_\ell^2).$$

Next, we have

$$\frac{\partial\Lambda(X,\theta)}{\partial\omega_\ell^2} = \pi_\ell\frac{\partial\lambda(X,\tau_\ell)}{\partial\omega_\ell^2} = \pi_\ell\,\frac{1}{2}(\gamma^2(\tau_\ell) - I(\omega_\ell^2))\lambda(X,\tau_\ell).$$

Again, we know that this r.v. is $Q_0^{x,T}$-integrable with nul integral. This yields:

$$\mathbb{E}_{\mathbb{Q}_\theta}\left|\frac{\partial\log\Lambda(X,\theta)}{\partial\omega_\ell^2}\right| < +\infty, \quad \text{and } \mathbb{E}_{\mathbb{Q}_\theta}\frac{\partial\log\Lambda(X,\theta)}{\partial\omega_\ell^2} = 0.$$

Moreover,

$$\left(\frac{\partial\log\Lambda(X,\theta)}{\partial\omega_\ell^2}\right)^2 \Lambda(X,\theta) = \frac{\left(\frac{\partial\Lambda(X,\theta)}{\partial\omega_\ell^2}\right)^2}{\Lambda(X,\theta)} \leq \pi_\ell\left[\frac{1}{2}(\gamma^2(\tau_\ell) - I(\omega_\ell^2))\right]^2\lambda(X,\tau_\ell).$$

This implies:

$$\mathbb{E}_{\mathbb{Q}_\theta}\left(\frac{\partial\log\Lambda(X,\theta)}{\partial\omega_\ell^2}\right)^2 < +\infty.$$

Now, we look at second order derivatives. The successive derivatives w.r.t. $\mu_\ell, \mu_{\ell'}$, $\omega_\ell, \omega_{\ell'}$ with $\ell \neq \ell'$ are nul. We have

$$\frac{\partial^2\Lambda(X,\theta)}{\partial\mu_\ell^2} = \pi_\ell\frac{\partial^2\lambda(X,\tau_\ell)}{\partial\mu_\ell^2} = \pi_\ell\,(\gamma^2(\tau_\ell) - I(\omega_\ell^2))\lambda(X,\tau_\ell).$$

This r.v. is integrable w.r.t. $Q_0^{x,T}$ with nul integral. Consequently, noting that

$$\frac{\partial^2 \log \Lambda(X,\theta)}{\partial \mu_\ell^2}\Lambda(X,\theta) = \frac{\partial^2 \Lambda(X,\theta)}{\partial \mu_\ell^2} - \frac{\left(\frac{\partial \Lambda(X,\theta)}{\partial \mu_\ell}\right)^2}{\Lambda(X,\theta)},$$

we get that this r.v. is integrable w.r.t. $Q_0^{x,T}$ and computing the integral yields

$$\mathbb{E}_{\mathbb{Q}_\theta} \frac{\partial^2 \log \Lambda(X,\theta)}{\partial \mu_\ell^2} = -\mathbb{E}_{\mathbb{Q}_\theta} \left(\frac{\partial \log \Lambda(X,\theta)}{\partial \mu_\ell}\right)^2.$$

Next,

$$\frac{\partial^2 \Lambda(X,\theta)}{\partial \mu_\ell \partial \omega_\ell^2} = \pi_\ell \frac{\partial^2 \lambda(X,\tau_\ell)}{\partial \mu_\ell \partial \omega_\ell^2} = \pi_\ell \,(\frac{1}{2}\gamma^3(\tau_\ell) - \frac{3}{2}\gamma(\tau_\ell) I(\omega_\ell^2))\lambda(X,\tau_\ell).$$

$$\frac{\partial^2 \Lambda(X,\theta)}{(\partial \omega_\ell^2)^2} = \pi_\ell \left[\left(\frac{1}{2}(\gamma^2(\tau_\ell) - I(\omega_\ell^2))\right)^2 - \gamma^2(\tau_\ell) I(\omega_\ell^2) - \frac{1}{2} I^2(\omega_\ell^2)\right] \lambda(X,\tau_\ell).$$

Therefore, we conclude the proof analogously using (24) and (25). □

8.4 Proof of Theorem 1

Thus, following the standard steps for weak consistency, it remains to prove a uniformity condition. We prove that

- there exists an open convex subset S of Θ, containing θ_0 and functions $G_{k,j,\ell}(X)$ such that, on S, $|\frac{\partial^3 \log \Lambda(X,\theta)}{\partial \theta_k \theta_j \theta_l}| \leq G_{k,j,\ell}(X)$ and $\mathbb{E}_{\theta_0}|G_{k,j,\ell}(X))| < +\infty$ for all $1 \leq k, j, l \leq 3M-1$.

Let $K, \alpha, \beta, c_0, c_1$ be positive numbers such that $0 < \alpha < \beta < 1$, $0 < c_0 < c_1$ and assume that θ_0 belongs to

$$S = \{(\pi_\ell, \mu_\ell, \omega_\ell^2)_{1 \leq \ell \leq M}, \alpha < \pi_\ell < \beta, |\mu_\ell| < K, c_0 < \omega_\ell^2 < c_1, 1 \leq \ell \leq M\} \quad (26)$$

where $\pi_M = 1 - \sum_{\ell=1}^{M-1} \pi_\ell$. We have to study

$$\begin{aligned}
\frac{\partial^3 \log \Lambda(X,\theta)}{\partial \theta_j \partial \theta_k \partial \theta_r} \quad = \quad & \frac{1}{\Lambda(X,\theta)} \frac{\partial^3 \Lambda(X,\theta)}{\partial \theta_j \partial \theta_k \partial \theta_r} - \frac{1}{\Lambda^2(X,\theta)} \frac{\partial \Lambda(X,\theta)}{\partial \theta_r} \frac{\partial^2 \Lambda(X,\theta)}{\partial \theta_j \partial \theta_k} \\
& - \frac{1}{\Lambda^2(X,\theta)} \left(\frac{\partial \Lambda(X,\theta)}{\partial \theta_k} \frac{\partial^2 \Lambda(X,\theta)}{\partial \theta_j \partial \theta_r} - \frac{\partial \Lambda(X,\theta)}{\partial \theta_j} \frac{\partial^2 \Lambda(X,\theta)}{\partial \theta_r \partial \theta_k}\right) \\
& + \frac{2}{\Lambda^3(X,\theta)} \frac{\partial \Lambda(X,\theta)}{\partial \theta_j} \frac{\partial \Lambda(X,\theta)}{\partial \theta_k} \frac{\partial \Lambda(X,\theta)}{\partial \theta_r}.
\end{aligned}$$

Thus, we have, for j, k, r distinct indexes:

$$\frac{\partial^3 \log \Lambda(X,\theta)}{\partial \pi_j \partial \pi_k \partial \pi_r} = \frac{2}{\Lambda^3(X,\theta)}(\lambda(X,\tau_j)-\lambda(X,\tau_M))(\lambda(X,\tau_k)-\lambda(X,\tau_M))(\lambda(X,\tau_r)-\lambda(X,\tau_M)).$$

As $\lambda(X,\tau_j)/\Lambda(X,\theta) \leq \pi_j^{-1} < \alpha^{-1}$,

$$\frac{\partial^3 \log \Lambda(X,\theta)}{\partial \pi_j \partial \pi_k \partial \pi_r} \leq 2^4 \alpha^{-3}.$$

To bound the other third order derivatives, we use again that $\lambda(X,\tau_j)/\Lambda(X,\theta) \leq \pi_j^{-1} < \alpha^{-1}$. Then, in the derivatives, there appears the random variables

$$\gamma^m(\tau) = \left(\frac{U-\mu V}{1+\omega^2 V}\right)^m$$

for different values of m. We now bound $\gamma(\tau)$ by a random variable indepedendent of τ and having moments of any order under $\mathbb{Q}_{\theta_0}$. Observe that:

$$\frac{U}{1+\omega^2 V} = \frac{U}{1+c_1 V}\left(1+\frac{(c_1-\omega^2)V}{1+\omega^2 V}\right).$$

Therefore,

$$|\gamma(\tau)| \leq \frac{c_1}{c_0}\left|\frac{U}{1+c_1 V}\right| + \frac{K}{c_0}.$$

Now, analogously,

$$\left|\frac{U}{1+c_1 V}\right| = \left|\frac{U}{1+\omega^2 V}\left(1+\frac{(\omega^2-c_1)V}{1+c_1 V}\right)\right| \leq 3\left|\frac{U}{1+\omega^2 V}\right|.$$

This implies that, for all τ,

$$\mathbb{E}_{Q_\tau}\left|\frac{U}{1+c_1 V}\right|^m < +\infty.$$

Consequently,

$$\mathbb{E}_{\mathbb{Q}_{\theta_0}}\left|\frac{U}{1+c_1 V}\right|^m = \sum_{\ell=1}^{M} \pi_{\ell,0}\mathbb{E}_{Q_{\tau_0}}\left|\frac{U}{1+c_1 V}\right|^m < +\infty.$$

The proof of Theorem 1 is complete.□