FORWARD-REVERSE EM ALGORITHM FOR MARKOV CHAINS: CONVERGENCE AND NUMERICAL ANALYSIS

CHRISTIAN BAYER, HILMAR MAI, AND JOHN SCHOENMAKERS

Abstract. We develop a forward-reverse EM (FREM) algorithm for estimating parameters of a discrete time Markov chain evolving through a certain measurable state space. For the construction of the FREM method we develop forward-reverse representations for Markov chains conditioned on a certain terminal state. We proof almost sure convergence of our algorithm for a Markov chain model with curved exponential family structure. On the numerical side we give a complexity analysis of the forward-reverse algorithm by deriving its expected cost. Two application examples are discussed.

1. Introduction

The EM algorithm going back to the seminal paper [Dempster et al., 1977] is a very general method for iterative computation of maximum likelihood estimates in the setting of incomplete data. The algorithm consists of an expectation step (E-step) followed by a maximization step (M-step) which led to the name EM algorithm.

Due to its general applicability and relative simplicity it has nowadays found its way into a great number of applications. These include maximum likelihood estimates of hidden Markov models in [MacDonald and Zucchini, 1997], non-linear time series models in [Chan and Ledolter, 1995] and full information item factor models in [Meng and Schilling, 1996] to give just a very limited selection.

Despite the simplicity of the basic idea of the algorithm its implementation in more complex models can be rather challenging. The global maximization of the likelihood in the M-step has recently been addressed successfully (see e.g. [Meng and Rubin, 1993] and [Liu and Rubin, 1994]). On the other hand, when the expectation of the complete likelihood is not known in closed form only partial solutions have been given yet. One approach developed in [Wei and Tanner, 1990] uses Monte Carlo approximations of the unknown expectation and was therefore named Monte Carlo EM (MCEM) algorithm. As an alternative procedure the stochastic approximation EM algorithm was suggested in [Lavielle and Moulines, 1999].

In this paper we take a completely different route by using a forward-reverse algorithm (cf. [Bayer and Schoenmakers, 2014]) to approximate the conditional expectation of the complete data likelihood. In this respect we extend the idea from [Bayer and Schoenmakers, 2014] to a Markov chain setting, which is considered an interesting contribution on its own. Indeed, Markov chains are more general in a sense, since any diffusion monitored at discrete times yields canonically a Markov chain, but not every chain can be embedded (straightforwardly) into some continuous time diffusion the other way around.

The central issue is the identification of a parametric Markov chain model (X_n, $n = 0, 1, \ldots$) based on incomplete data, i.e. realizations of the model, given on a typically course grid of time points, let us say $n_1, n_2, \ldots n_N$. Let us assume that the chain runs through $\mathbb{R}^d$ and that the transition densities $p^{\theta}_{n,m}(x, y)$, $m \geq n$, of the chain exist (with $p^{\theta}_{n,n}(x, y) := \delta_x(y)$), where the unknown parameter θ has to be determined. The log-likelihood function

2010 *Mathematics Subject Classification.* 65C05, 65J20, 62F12, 62M05.

Key words and phrases. EM algorithm; Forward-reverse representations; Markov chain estimation; maximum likelihood estimation; Monte Carlo simulation.

based on the incomplete observations $(X_{n_1}, \dots, X_{n_N})$ is then given by

$$l(\theta, x_{n_1}, \dots, x_{n_N}) = \sum_{i=0}^{N-1} \ln p^{\theta}_{n_i, n_{i+1}}(x_{n_i}, x_{n_{i+1}}), \tag{1.1}$$

with $X_{n_0} = x_0$ being the initial state of the chain. Then the standard method of maximum likelihood estimation would suggest to evaluate

$$\arg\max_{\theta} l(\theta, X) = \arg\max_{\theta} \sum_{i=0}^{N-1} \ln p^{\theta}_{n_i, n_{i+1}}(X_{n_i}, X_{n_{i+1}}). \tag{1.2}$$

The problem in this approach lies in the fact that usually only the one-step transition densities $p^{\theta}_{n,n+1}(x, y)$ are explicitly known, while any multi-step density $p^{\theta}_{n,m}(x, y)$ for $m > n$ can be expressed as an $m - n - 1$ fold integral of one-step densities. In particular for larger $m - n$, these multiple integrals are numerically intractable however.

In the EM approach, we therefore consider the alternative problem

$$\arg\max_{\theta} \sum_{i=0}^{N-1} \sum_{j=n_i}^{n_{i+1}-1} \mathbb{E} \ln p^{\theta}_{j,j+1}(X_j, X_{j+1}), \tag{1.3}$$

given the "missing data" $X_{n_i+1}, ..., X_{n_{i+1}-1}$, $i = 0, ..., N - 1$. As such, between two such consecutive time points, n_i and n_{i+1} say, the chain may be considered as a bridge process starting in realization X_{n_i} and ending up in realization $X_{n_{i+1}}$ (under the unknown parameter θ though), and so each term in (1.3) may be considered as an expected functional of the "bridged" Markov chain starting at time n_i in (data point) X_{n_i}, conditional on reaching (data point) $X_{n_{i+1}}$ at time n_{i+1}.

We will therefore develop firstly an algorithm for estimating the terms in (1.3) for a given parameter θ. This algorithm will be of forward-reverse type in the spirit of the one in [Bayer and Schoenmakers, 2014] developed for diffusion bridges. It should be noted here that in the last years the problem of simulating diffusion bridges has attracted much attention. Without pretending to be complete, see for example, [Bladt and Sørensen, 2014, Delyon and Hu, 2006, Milstein and Tretyakov, 2004, Stinis, 2011, Stuart et al., 2004, Schauer et al., 2013].

Having the forward-reverse algorithm at hand, we may construct an approximate solution to (1.3) in a sequential way by the EM algorithm: Once a generic approximation θ_m is constructed after m steps, one estimates

$$\theta_{m+1} := \arg\max_{\theta} \sum_{i=0}^{N-1} \sum_{j=n_i}^{n_{i+1}-1} \widehat{\mathbb{E}} \ln p^{\theta}_{j,j+1}(X^{\theta_m}_j, X^{\theta_m}_{j+1}),$$

where X^{θ_m} denotes the Markov bridge process under the transition law due to parameter θ_m and each term

$$\widehat{\mathbb{E}} \ln p^{\theta}_{j,j+1}(X^{\theta_m}_j, X^{\theta_m}_{j+1})$$

represents a forward-reverse approximation of

$$\mathbb{E} \ln p^{\theta}_{j,j+1}(X^{\theta_m}_j, X^{\theta_m}_{j+1})$$

as a (known) function of θ.

Convergence properties of approximate EM algorithms have drawn considerable recent attention in the literature mainly driven by it's success in earlier intractable estimation problems. An overview of existing convergence results for the MCEM algorithm can be found in [Neath, 2013]. Starting from a convergence result for the forward-reverse representation for Markov chains we prove almost sure convergence of the FREM sequence in the setting of curved exponential families based on techniques developed in [Bayer and Schoenmakers, 2014] and [Fort and Moulines, 2003]. Essentially the only ingredient from the Markov chain model for this convergence to hold are exponential tails of the transition densities and their derivatives that are straightforward to check in many examples.

Since computational complexity is always an issue in estimation techniques that involve a simulation step, we also include a complexity analysis for the forward-reverse algorithm. We show that the algorithm achieves an expected cost of the order $O(N \log N)$ for sample size of N forward-reverse trajectories.

In order to demonstrate the scope of our method, two application examples ranging from a Markov chain obtained as a time discretized Ornstein-Uhlenbeck process to a discrete time Markov chain on $\mathbb{R}$, driven by heavy tailed transition probabilities, are discussed. We also mention a recent application paper [Bayer et al., 2016], which focuses on practical and implementation issues of a related forward-reverse EM algorithm in the setting of Stochastic Reaction Networks (SRNs), i.e., of continuous time Markov chains with discrete state space.

The structure of the paper is as follows. In Section 2 we recapitulate and adapt the concept of reversed Markov chains, initially developed in [Milstein et al., 2007] using the ideas in [Milstein et al., 2004] on reversed diffusions. A general stochastic representation — involving standard (unconditional) expectations only — for expected functionals of conditional Markov chains is constructed in Section 2.3. This representation allows for a forward reverse EM algorithm that is introduced and analyzed in Section 3. In Section 4 we proof almost sure convergence of the forward-reverse EM algorithm in the setting of curved exponential families. Implementation and Complexity of the FREM algorithm are addressed in Section 5. The paper is concluded with two application examples in Section 6.

2. Recap of forward and reverse representations for Markov chains

2.1. **Notation.** Consider a discrete-time Markov process $(X_n, \mathcal{F}_n)$, $n = 0, 1, 2, ...,$ on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ with phase space $\left(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d)\right)$, henceforth called Markov chain. Let P_n, $n \geq 0$, denote the one-step transition probabilities defined by

$$P_n(x, B) := \mathbb{P}(X_{n+1} \in B \mid X_n = x),\ n = 0, 1, 2, ...,\quad x \in S,\ B \in \mathcal{S}. \tag{2.1}$$

In the case of an autonomous Markov chain all the one-step transition probabilities coincide and are equal to $P := P_0$. The multi-step transition probabilities $P_{n,m}$ are defined by

$$P_{n,m}(x, B) := \mathbb{P}(X_m \in B \mid X_n = x), \quad m \geq n.$$

We will assume that the one-step (and, a-forteriori, the multi-step) transition probabilities have densities (w.r.t. the Lebesgue measure) denoted by $p_n = p_n(x, y)$ and $p_{n,m} = p_{n,m}(x, y)$, respectively.

Sometimes we denote by $X_m^{n,x}$, $m \geq n$, a trajectory of the Markov chain which is at step n in the point x, i.e., $X_n^{n,x} = x$.

2.2. **Reverse probabilistic representations.** First consider the integral $I(f)$ of a function f against the multi-step transition density $p_{n,N}(x, y)$ against the "forward variable" y. Clearly, we have the "forward stochastic representation"

$$I(f) := \int p_{n,N}(x, y) f(y) dy = \mathbb{E}\left[f(X_N^{n,x})\right]. \tag{2.2}$$

We want to construct a stochastic representation for the integral $J(g)$ of some functional against the "backward variable" x, i.e., for the expression

$$J(g) := \int g(x) p_{n,N}(x, y) dx, \tag{2.3}$$

which we call "reverse stochastic representations". Of course, $x \mapsto p_{n,N}(x, y)$ does not need to be a density. Still, we are going to construct a class of reverse Markov chains that allow for a probabilistic representation for the solution of (2.3).

Let us fix a number $N \in \mathbb{N}$ and consider for $0 \leq m < N$, functions $\psi_m : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}_+$ such that for each m and y the function

$$z \mapsto q_m(y, z) := \frac{p_{N-m-1}(z, y)}{\psi_m(y, z)} \tag{2.4}$$

is a density. One possible choice is

$$\psi_m(y) = \int p_{N-m-1}(z, y) dz$$

independent of z.[1]

Definition 2.1. The *reverse process* $(Y_m, \mathcal{Y}_m)_{m=0}^N$ is a pair of processes with values in $\mathbb{R}^d$ and $\mathbb{R}$, respectively, such that

- Y is a Markov chain with one-step transition densities q_m defined in (2.4);
- $\mathcal{Y}_{m+1} := \mathcal{Y}_m \psi_m(Y_m, Y_{m+1})$, with $\mathcal{Y}_0 := 1$.

When needed, we shall denote an instance of the reverse process for a given N started at $Y_0 = y$ by $\left(Y_\cdot^{y;N}, \mathcal{Y}_\cdot^{y;N}\right)$.

We note that the dynamics of the reverse process depend on N if the original chain X is time-inhomogeneous.

Theorem 2.2. *For any n, $0 \leq n \leq N$, (2.3) has the following probabilistic representation.*

$$\int g(x) p_{n,N}(x, y) dx = \mathbb{E}\left[g(Y_{N-n}^y)\mathcal{Y}_{N-n}^y\right],$$

where g is an arbitrary test function (a "density" $p_{m,m}$ has to be interpreted as a Dirac distribution or δ-function).

Proof. The proof consists of simply unwrapping Definition 2.1. Specifically, with $y_0 := y$,

$$\begin{aligned}
\mathbb{E}\left[g(Y_{N-n}^y)\mathcal{Y}_{N-n}^y\right] &= \int g(y_{N-n}) \prod_{l=0}^{N-n-1} \left(q_l(y_l, y_{l+1})\psi_l(y_l, y_{l+1})\right) dy_1 \cdots dy_{N-n} \\
&= \int g(y_{N-n}) \prod_{l=0}^{N-n-1} p_{N-l-1}(y_{l+1}, y_l) dy_1 \cdots dy_{N-n} \\
&= \int g(y_{N-n}) p_n(y_{N-n}, y_{N-n-1}) p_{n+1}(y_{N-n-1}, y_{N-n-2}) \cdots p_{N-1}(y_1, y_0) dy_1 \cdots dy_{N-n} \\
&= \int g(y_{N-n}) p_{n,N}(y_{N-n}, y_0) dy_{N-n},
\end{aligned}$$

which gives the claim. □

Remark 2.3. It should be stressed that, in contrast to a corresponding theorem in [Milstein et al., 2007], Theorem 2.2 provides a family of probabilistic representations indexed by $n = 1, \ldots, N$, that involves only one common reverse process Y^y independent of n. It turns out that this extension of the related result in [Milstein et al., 2007] is crucial for deriving probabilistic representations for conditional Markov chains below (cf. [Bayer and Schoenmakers, 2014]). If X is time-homogeneous, then the reverse process does not depend on N, either.

[1] This choice may be impractical if the corresponding integral cannot be expressed in closed form.

2.3. **Conditional expectations via forward-reverse representations.** In this section we describe for a Markov Chain (2.1) an efficient procedure for estimating the final distributions of a chain $X = (X_n)_{n=0,...,N}$ conditioned, or pinned, on a terminal state X_N. More specifically, for some given (unconditional) process X we aim at simulation of the functional

$$\mathbb{E}\left[g(X_0, \ldots, X_N) \,|\, X_0 = x,\, X_N = y\right]. \tag{2.5}$$

The procedure proposed below is in fact an extension of the method developed in [Bayer and Schoenmakers, 2014] to discrete time Markov chains. We note that similar techniques as in [Bayer and Schoenmakers, 2014] also allow us to treat the more general problem

$$\mathbb{E}\left[g(X_0, \ldots, X_N) \,|\, X_0 = x,\, X_N \in A\right]$$

for suitable sets $A \subset \mathbb{R}^d$.

Let us consider the problem (2.5) for fixed $x, y \in \mathbb{R}^d$ (i.e. $A = \{y\}$). A direct adaptation of Theorem 2.2 gives the following result:

Theorem 2.4. *Given $0 \le n < N$ and the reverse process $(Y, \mathcal{Y}) := (Y^{y;N}, \mathcal{Y}^{y;N})$. Then, for any (suitably integrable) f we have*

$$\mathbb{E}\left[f(Y_{N-n}, Y_{N-n-1}, \ldots, Y_0)\mathcal{Y}_{N-n}\right] = \int_{(\mathbb{R}^d)^{N-n}} f(y_0, y_1, \ldots, y_{N-n}) \prod_{i=0}^{N-n-1} p_{n+i}(y_i, y_{i+1}) dy_i$$

with $y_{N-n} := y$.

The basic idea of the forward-reverse representation for $\mathbb{E}\left[g(X_0, \ldots, X_N) \mid X_0 = x,\, X_N = y\right]$ is the following: we fix some intermediate point $0 < n^* < N$. We send out paths of the forward process started at $X_0 = x$ until time n^*. Independently, we send out paths of the reverse process $(Y, \mathcal{Y})$ started at $Y_0 = y$ until time $N - n^*$. While the dynamics of Y is defined in the standard way, i.e., forward in time, we should really think of the paths of Y as running backward in time, i.e., as starting in position y at time N and running backwards until their final time $N - (N - n^*) = n^*$. Let us denote these time-reversed paths of Y by $\widetilde{Y}$, i.e., $\widetilde{Y}_N = y$. If one of the forward paths X and one of the reverse paths $\widetilde{Y}$ happen to meet at time n^*, i.e., $X_{n^*} = \widetilde{Y}_{n^*}$, then we join the paths and consider the joint path as a sample from the law of the conditioned Markov chain—up to the re-weighting factor (measure change) $\mathcal{Y}$.

Even when checking for matches among all pairs of paths sampled from X and Y, such perfect matches happen extremely rarely (with probability zero). Therefore, we relax the perfect matching condition using a kernel K_ϵ of the form

$$K_\epsilon(u) := \epsilon^{-d} K(u/\epsilon), \quad y \in \mathbb{R}^d,$$

with K being integrable on $\mathbb{R}^d$ and $\int_{\mathbb{R}^d} K(u)du = 1$. Formally K_ϵ converges to the delta function δ_0 on $\mathbb{R}^d$ (in distribution sense) as $\epsilon \downarrow 0$. We obtain the following stochastic representation (involving standard expectations only) for (2.5):

Theorem 2.5. *Let the chain $(Y, \mathcal{Y}) := \left(Y^{y;N}, \mathcal{Y}^{y;N}\right)$ be given by Definition 2.1 and assumed independent of the forward chain $X := X^{0,x}$. Then, with $0 < n^* < N$,*

$$\begin{aligned} &\mathbb{E}\left[g(X_0, \ldots, X_N) \mid X_0 = x,\, X_N = y\right] = \\ &\qquad \lim_{\epsilon \to 0} \frac{\mathbb{E}\left[g\left(X_0, \ldots, X_{n^*}, Y_{N-n^*-1}, \ldots, Y_1, Y_0\right) K_\epsilon\left(Y_{N-n^*} - X_{n^*}\right) \mathcal{Y}_{N-n^*}\right]}{\mathbb{E}\left[K_\epsilon\left(Y_{N-n^*} - X_{n^*}\right) \mathcal{Y}_{N-n^*}\right]}. \end{aligned} \tag{2.6}$$

Proof. The proof is analogous to the corresponding one in [Bayer and Schoenmakers, 2014]. As a rough sketch, apply Theorem 2.4 to

$$f(y_0, y_1, \ldots, y_{N-n^*}) := g(X_0, \ldots, X_{n^*}, y_1, \ldots, y_{N-n^*}) K_\epsilon(y_0 - X_{n^*}),$$

conditional on $X_0, \dots, X_{n^*}$, send $\epsilon \to 0$, and divide the result by

$$p_{0,N}(x,y) = \lim_{\epsilon \to 0} \mathbb{E}\left[K_\epsilon\left(Y_{N-n^a st} - X_{n^*}\right)\mathcal{Y}_{N-n^*}\right]. \qquad \square$$

2.4. **Forward-Reverse algorithm.** Given Theorem 2.5 the corresponding forward-reverse Monte Carlo estimator for (2.6) suggests itself: Sample i.i.d. copies $X^{(1)}, ..., X^{(M)}$ of the process X and, independently, i.i.d. copies

$$\left(Y^{(1)}, \mathcal{Y}^{(1)}\right), \dots, \left(Y^{(M)}, \mathcal{Y}^{(M)}\right)$$

of the process $(Y, \mathcal{Y})$. Take for K a second order kernel (such as a Gaussian kernel) and choose a bandwidth $\epsilon_M \sim M^{-1/d}$ if $d \leq 4$, or $\epsilon_M \sim M^{-2/(4+d)}$ if $d \geq 4$. By next replacing the expectations in the numerator and denominator of (2.6) by their respective Monte Carlo estimates involving double sums, one ends up with an estimator with Root-Mean-Square error $O(M^{-1/2})$ in the case $d \leq 4$ and $O(M^{-4/(4+d)})$ in the case $d > 4$ (cf. [Bayer and Schoenmakers, 2014] for details).

3. The forward-reverse EM algorithm

Let us now formulate the forward-reverse EM (FREM) algorithm in the setting of the missing data problem. Suppose that the parameter $\theta \in \Theta \subset \mathbb{R}^s$ and that the Markov chain $X = (X_n, n \in \mathbb{N})$ has state space $\mathbb{R}^d$. Assuming that the transition densities $p_{n,k}$ of X exist for $n, k \in \mathbb{N}$ the full-data log-likelihood then reads

$$l_c(\theta, x) = \sum_{i=0}^{n_N - 1} \log p^\theta_{i,i+1}(x_i, x_{i+1}), \quad x \in \mathbb{R}^{n_N+1}. \tag{3.1}$$

In the missing data problem only partial observations $X_{n_0}, \dots, X_{n_N}$ are available for $0 = n_0 < n_1 < \dots < n_N$ with log-likelihood function l given in (1.1) that is intractable in most cases. Instead, the maximization of l in θ has to be replaced by a two step iterative procedure, the EM algorithm.

E-step: In this step, evaluate in the m-th iteration the conditional expectation of the complete data log-likelihood

$$Q(\theta, \theta_m, X_{n_\bullet}) := \mathbb{E}_{\theta_m}[l_c(\theta, X_0, \dots, X_{n_N}) | X_{n_0}, \dots, X_{n_N}].$$

M-step: Update the parameter by

$$\theta_{m+1} = \arg\max_\theta Q(\theta, \theta_m, X_{n_\bullet}),$$

using the shorthand $X_{n_\bullet} = X_{n_0}, \dots, X_{n_N}$. Since in many Markov chain models the E-step is intractable in this form, we propose a forward-reverse approximation for the expectation of the transition densities evaluated at the observations.

FR E-step: Evaluate

$$Q_m(\theta, \theta_m, X_{n_\bullet}) := \mathbb{E}^{FR}_{\theta_m}[l_c(\theta, X) | X_{n_\bullet}], \tag{3.2}$$

where $\mathbb{E}^{FR}_{\theta_m}$ denotes a forward-reverse approximation of the conditional expectation under the parameter θ_m.

It should be noted that, due to the dependence of the number of FR-trajectories on m, we now have an explicit dependence of Q on m. Due to the Markov property of the model the forward-reverse approximation can be performed independently in each interval between missing observations. This point will be further exploited for the implementation in Section 5.

After this FR E-step is computed the M-step remains unchanged. The FREM algorithm gives a random sequence $(\theta_m)_{m \geq 0}$ that under certain conditions given in the next section converges to stationary points of the likelihood function. To assure a.s. boundedness of this sequence we apply a stabilization technique as introduced in [Chen et al., 1988].

The stable FREM algorithm. Let $K_m \subset \Theta$ for $m \in \mathbb{N}$ be a sequence of compact sets such that

$$K_m \subsetneq K_{m+1} \quad \text{and} \quad \Theta = \bigcup_{m \in \mathbb{N}} K_m \tag{3.3}$$

for all $m \in \mathbb{N}$. We define the stable FREM algorithm by checking if θ_m after the m-th maximization step lies in K_m and resetting the algorithm otherwise. Choose a starting value $\theta_0 \in K_0$ and let r_m for $m \in \mathbb{N}$, $p_0 := 0$, count the number of resets.

stable M-step:

$$\theta_{m+1} = \arg\max_{\theta} Q_m(\theta, \theta_m, X_{n_\bullet}) \text{ and } r_{m+1} = r_m, \qquad \text{if } \arg\max_{\theta} Q_m(\theta, \theta_m, X_{n_\bullet}) \in K_m, \tag{3.4}$$

$$\theta_{m+1} = \theta_0 \text{ and } r_{m+1} = r_m + 1, \qquad \text{if } \arg\max_{\theta} Q_m(\theta, \theta_m, X_{n_\bullet}) \notin K_m. \tag{3.5}$$

We will show in the next section that under weak assumption the number of resets r_m stays a.s. finite. Our stable FREM algorithm consists now of iteratively repeating the FR E-step and the stable M-step.

4. Almost sure convergence of the FREM algorithm

In this section we prove almost sure convergence of the stable FREM algorithm under the assumption that the complete data likelihood is from a curved exponential family. Our proof is mainly based on results from [Bayer and Schoenmakers, 2014], [Fort and Moulines, 2003] and the classical framework for the EM algorithm introduced in [Dempster et al., 1977] and [Lange, 1995].

4.1. **Model setting.** Suppose that $\phi : \Theta \to \mathbb{R}$, $\psi : \Theta \to \mathbb{R}^q$ and $S : \mathbb{R}^{(n_N+1)d} \to \mathbb{R}^q$ are continuous functions. We make the structural assumption that the full data log-likelihood is of the form

$$l_c(\theta, x_0, \ldots, x_{n_N}) = \phi(\theta) + \langle S(x_0, \ldots, x_{n_N}), \psi(\theta) \rangle, \tag{4.1}$$

i.e. l_c is from a curved exponential family. In order to proof convergence we need the following properties to be fulfilled that naturally hold in many popular models. In Section 5 we give practical examples that fall into this setting. Set $\bar{l}_c := \phi(\theta) + \langle s, \psi(\theta) \rangle$ for all $s \in \mathbb{R}^q$.

Assumption 4.1.

1. There exists a continuous function $\bar{\theta} : \mathbb{R}^q \to \Theta$ such that $\bar{l}_c(\bar{\theta}(s), s) = \sup_{\theta \in \Theta} \bar{l}_c(\theta, s)$ for all $s \in \mathbb{R}^q$.
2. The incomplete data likelihood l (cf. (1.1)) is continuous in θ, and the level sets $\{\theta \in \Theta | l(\theta, x) \geq C\}$ are compact for any $C > 0$ and all x.
3. The conditional expectation $\mathbb{E}_\theta[S(X_0, \ldots, X_{n_N}) | X_{n_0} = x_{n_0}, \ldots, X_{n_N} = x_{n_N}]$ exists for all $(x_{n_0}, \ldots, x_{n_N}) \in \mathbb{R}^{N+1}$ and $\theta \in \Theta$ and is continuous on Θ.

To simplify our notation we will neglect in the following the dependence of l on s. Under these assumptions we can separate the E- and M-step. In order to do so we define

$$g(\theta) := \mathbb{E}_\theta[S(X_0, \ldots, X_{n_N}) | X_{n_0} = x_{n_0}, \ldots, X_{n_N} = x_{n_N}].$$

An iteration of the EM algorithm can now be written as $\theta_{m+1} = \bar{\theta}(g(\theta_m))$. Let us denote by Γ the set of stationary points of the EM algorithm, i.e.

$$\Gamma = \{\theta \in \Theta | \bar{\theta}(g(\theta)) = \theta\}.$$

It was shown in Theorem 2 in [Wu, 1983] that if Θ is open, ϕ and ψ are differentiable and Assumption 4.1 holds, then

$$\Gamma = \{\theta \in \Theta | \partial_\theta l(\theta) = 0\},$$

such that the fixed points of the EM algorithm coincide with the stationary points of the incomplete data log-likelihood l as defined in (1.1). In [Wu, 1983] it was proved that the set Γ contains all limit points of (θ_n) and that $(l(\theta_n))$ converges to $l(\theta_0)$ for some $\theta_0 \in \Gamma$. In the following theorem we extend these results to the FREM algorithm. Let $d(x, A)$ be the distance between a point x and a set A.

For the convergence of our forward-reverse based EM algorithm, we naturally also need to guarantee convergence of the corresponding forward-reverse estimators.

Assumption 4.2.

(1) For any multi-indices $\alpha, \beta \in \mathbb{N}_0^d$ with $|\alpha|+|\beta| \leq 2$ and any index i there are constants $C_1 = C_1(i, \alpha, \beta), C_2 = C_2(i, \alpha, \beta) > 0$ such that

$$|\partial_x^\alpha \partial_y^\beta p_i(x, y)| \leq C_1 \exp\left(-C_2 |x-y|\right).$$

(2) The kernel K has exponential decay.

(3) S is twice differentiable in its arguments and both S and its first and second derivatives are polynomially bounded.

Remark 4.3. In the related paper [Bayer and Schoenmakers, 2014, Condition 4.1, Condition 4.4], the authors even required Gaussian decay for both the transition densities and the kernel (i.e., in (1) and (2) above). These requirements–especially the first one–were somewhat natural in a diffusion context, see the discussion in [Bayer and Schoenmakers, 2014, Remark 4.3]. However, it is easy to see that the following results also hold under the weaker assumptions imposed above, see, in particular, the proof of [Bayer and Schoenmakers, 2014, Theorem 4.7]. The above conditions could be further relaxed: in fact, it is enough that both the kernel and the relevant derivatives of the transition densities are rapidly decaying functions–i.e., decaying faster than the reciprocal of any polynomial. Indeed, the proof of [Bayer and Schoenmakers, 2014, Theorem 4.7] is based on the fact that the convolution of the second derivative of p_i and the kernel K is itself (up to normalization) a density with finite moments of all orders. As is well-known that the convolution of two rapidly decaying functions is again rapidly decaying, this fact is still true under the weaker conditions.

Now we are ready to state as the main result of this section a general convergence theorem for the FREM algorithm.

Theorem 4.4. *Let $(K_m)_{m\in\mathbb{N}}$ satisfy* (3.3) *and choose $\theta_0 \in K_0$. Suppose that Assumptions 4.1 and 4.2 hold, $l(\Gamma)$ is compact and the number M of forward-reverse iterations satisfies $M > m$, then the stable FREM random sequence $(\theta_m)_{m\geq 0}$ has the following properties:*

(1) $\lim_m r_m < \infty$ *a.s. and (θ_m) is almost surely bounded.*

(2) $\lim_m d(l(\theta_m), l(\Gamma)) = 0$ *almost surely.*

(3) *If also l is s-times differentiable, then* $\lim_m d(\theta_m, \Gamma) = 0$ *almost surely.*

Proof. (1) Set

$$g^{FR}(\theta) := \mathbb{E}_\theta^{FR}[S(X_0, \ldots, X_{n_N}) | X_{n_0} = x_{n_0}, \ldots, X_{n_N} = x_{n_N}].$$

With the above notation an iteration of the FREM algorithm can be written as

$$\theta_{m+1} = \bar{\theta}(g^{FR}(\theta_m)).$$

It was shown in Lemma 2 in [Lavielle and Moulines, 1999] that the incomplete data log-likelihood l is a natural Lyapunov function relative to T and to the set of fixed points Γ. If for any $\epsilon > 0$ and compact $K \subset \Theta$ we have

$$\sum_m \mathbf{1}_{\{|l(\bar{\theta}(g^{FR}(\theta_m)))-l(\bar{\theta}(g(\theta_m)))|\geq\epsilon\}} \mathbf{1}_{\{\theta_m \in K\}} < \infty \quad \text{a.s.,} \tag{4.2}$$

then Proposition 11 in [Fort and Moulines, 2003] implies in our setting that $\limsup_n r_n < \infty$ almost surely and that (θ_n) is a compact sequence such that (1) follows. To obtain (4.2) it is sufficient by Borel-Cantelli to prove that

$$\sum_m P\left(|l(\bar{\theta}(g^{FR}(\theta_m))) - l(\bar{\theta}(g(\theta_m)))| \mathbf{1}_{\{\theta_m \in K\}} \geq \epsilon\right) < \infty.$$

Define for any $\delta > 0$ a neighborhood of K by

$$K_\delta := \{x \in \mathbb{R}^q | \inf_{z \in K} |z - x| \leq \delta\}.$$

By assumption l and $\bar{\theta}$ are continuous, such that for any $\delta > 0$ there exists $\eta > 0$ such that for any $x, y \in K_\eta$ we have $|l(\bar{\theta}(x)) - l(\bar{\theta}(y))| \leq \delta$.

Choosing now $\bar{\epsilon} = \delta \wedge \eta$ yields

$$\begin{aligned}
&P\left(|l(\bar{\theta}(g^{FR}(\theta_m))) - l(\bar{\theta}(g(\theta_m)))|\mathbf{1}_{\{\theta_m \in K\}} \geq \epsilon\right)\\
&= P\left(|l(\bar{\theta}(g^{FR}(\theta_m))) - l(\bar{\theta}(g(\theta_m)))|\mathbf{1}_{\{\theta_m \in K\}} \geq \epsilon, |g^{FR}(\theta_m) - g(\theta_m)|\mathbf{1}_{\{\theta_m \in K\}} \leq \delta\right)\\
&+ P\left(|l(\bar{\theta}(g^{FR}(\theta_m))) - l(\bar{\theta}(g(\theta_m)))|\mathbf{1}_{\{\theta_m \in K\}} \geq \epsilon, |g^{FR}(\theta_m) - g(\theta_m)|\mathbf{1}_{\{\theta_m \in K\}} \geq \delta\right)\\
&\leq 2P\left(|g^{FR}(\theta_m) - g(\theta_m)|\mathbf{1}_{\{\theta_m \in K\}} \geq \bar{\epsilon}\right).
\end{aligned}$$

Markov's inequality then gives

$$P\left(|l(\bar{\theta}(g^{FR}(\theta_m))) - l(\bar{\theta}(g(\theta_m)))|\mathbf{1}_{\{\theta_m \in K\}} \geq \epsilon\right) \leq 2\epsilon^{-k}\mathbb{E}\left[\left|g^{FR}(\theta_m) - g(\theta_m)\right|^k \mathbf{1}_{\{\theta_m \in K\}}\right]$$

for some $k > 0$.

By [Bayer and Schoenmakers, 2014, Theorem 4.18] (see also Remark 4.5 below), we can always choose a number M of samples for the forward-reverse algorithm and a corresponding bandwidth $\epsilon = \epsilon_M = M^{-\alpha}$ such that

$$\mathbb{E}\left[\left|g^{FR}(\theta_m) - g(\theta_m)\right|^2 \mathbf{1}_{\{\theta_m \in K\}}\right] \leq \frac{C}{M}$$

for some constant C. We note that the choice of α depends on the dimension d as well as on the *order* of the kernel. For instance, for $d \leq 4$ and a standard first order accurate kernel K, we can choose any $1/4 \leq \alpha \leq 1/d$.

In any case, if we choose $M > m$ then

$$\sum_m P\left(|l(\bar{\theta}(g^{FR}(\theta_m))) - l(\bar{\theta}(g(\theta_m)))|\mathbf{1}_{\{\theta_m \in K\}} \geq \epsilon\right) < \infty, \tag{4.3}$$

which proves (1).

To prove (2) and (3) observe that for every $K \subset \Theta$ we have

$$\lim_m |l(\theta_{m+1}) - l(\bar{\theta}(g(\theta_m)))|\mathbf{1}_{\{\theta_m \in K\}} = 0 \quad a.s.$$

By Borel-Cantelli it is sufficient to prove that

$$\sum_m P(|l(\theta_{m+1}) - l(\bar{\theta}(g(\theta_m)))|\mathbf{1}_{\{\theta_m \in K\}} \geq \epsilon) < \infty.$$

But since we have shown in (1) that r_n is finite a.s., we have in the above sum that $\theta_{m+1} = \bar{\theta}(g^{FR}(\theta_m))$ in almost all summands. Hence, it is sufficient to show that

$$\sum_m P(|l(\bar{\theta}(g^{FR}(\theta_m))) - l(\bar{\theta}(g(\theta_m)))|\mathbf{1}_{\{\theta_m \in K\}} \geq \epsilon) < \infty,$$

which is nothing else than (4.3). The statement of (2) and (3) follows now from Sard's theorem (cf. [Bröckner, 1975]) and Proposition 9 in [Fort and Moulines, 2003]. □

Remark 4.5. In the above convergence proof we need to rely on the convergence proof of the forward-reverse estimator when the bandwidth tends to zero and the number of simulated Monte Carlo samples tends to infinity. Such a proof is carried out for the diffusion case in [Bayer and Schoenmakers, 2014, Theorem 4.18], where also rates of convergence are given. We note that the proof only relies on the transition densities of (a discrete skeleton of) the underlying diffusion process. Hence, it immediately carries over to the present setting.

Theorem 4.4 is a general convergence statement that links the limiting points of the FREM sequence to the set of stationary points of l. In many concrete models the set Γ of stationary points consists of isolated points only such that an analysis of the Hessian of l gives conditions for local maxima. A more detailed discussion in this direction can be found in [Lavielle and Moulines, 1999] for example.

5. Implementation and complexity of the FREM algorithm

Before presenting two concrete numerical examples, we will first discuss general aspects of the implementation of the forward-reverse EM algorithm. For this purpose, let us, for simplicity, assume that the Markov chains X and $(Y, \mathcal{Y})$ are time-homogeneous, i.e., that $p \equiv p_k$ and $q \equiv q_k$ do not depend on time k. We assume that we observe the Markov process X at times $0 = n_0 < \cdots < n_N$, i.e., our data consist of the values $X_{n_k} = x_{n_k}$, $k = 0, \ldots, N$. For later use, we introduce the shortcut-notation $\mathbf{x} := (x_{n_j})_{j=0}^{N}$.

The law of X depends on an s-dimensional parameter $\theta \in \mathbb{R}^s$, which we are trying to estimate, i.e., $p = p^\theta$. To this end, let

$$\ell(\theta; x_0, \ldots, x_N) := \sum_{i=1}^{N} \log p^\theta(x_{i-1}, x_i)$$

denote the log-likelihood function for the estimation problem assuming full observation. As before, we make the structural assumption that

$$l_c(\theta; x_0, \ldots, x_{n_N}) = \phi(\theta) + \sum_{i=1}^{m} S_i(x_0, \ldots, x_{n_N}) \psi_i(\theta). \tag{5.1}$$

For simplicity, we further assume that there are functions S_i^j such that

$$S_i(x_0, \ldots, x_{n_N}) = \sum_{j=1}^{N} S_i^j(x_{n_{j-1}}, \ldots, x_{n_j}).$$

The structural assumption (5.1) allows us to effectively evaluate the conditional expectation of the log-likelihood l_c for different parameters θ, without having to re-compute the conditional expectations. More precisely, recall that for a given guess $\widetilde{\theta}$ the E step of the EM algorithm consists in calculating the function

$$\theta \mapsto Q(\theta; \widetilde{\theta}, \mathbf{x}) := \mathbb{E}_{\widetilde{\theta}}\left[\ell_c\left(\theta; X_0, \ldots, X_{n_N}\right) \middle| X_{n_j} = x_{n_j},\ j = 0, \ldots, N \right], \tag{5.2}$$

with $\mathbb{E}_{\widetilde{\theta}}$ denoting (conditional) expectation under the parameter $\widetilde{\theta}$. Inserting the structural assumption (5.1), we immediately obtain

$$\begin{aligned} Q(\theta; \widetilde{\theta}, \mathbf{x}) &= \phi(\theta) + \sum_{i=1}^{m} \psi_i(\theta) \mathbb{E}_{\widetilde{\theta}}\left[S_i(X_0, \ldots, X_{n_N}) \middle| X_{n_j} = x_{n_j},\ j = 0, \ldots, N \right] \\ &= \phi(\theta) + \sum_{i=1}^{m} S_i(\theta) z_i^{\widetilde{\theta}} \end{aligned}$$

with $z_i^{\widetilde{\theta}} := \mathbb{E}_{\widetilde{\theta}}\left[S_i(X_0, \ldots, X_{n_N}) \middle| X_{n_j} = x_{n_j},\ j = 0, \ldots, N \right]$, $i = 1, \ldots, m$. Note that the definition of $z_i^{\widetilde{\theta}}$ does not depend on the free parameter θ. Thus, only one (expensive) round of calculations of conditional expectations is needed for a given $\widetilde{\theta}$, producing a cheap-to-evaluate function in θ, which can then be fed into any maximization algorithm.

For any given $\widetilde{\theta}$, the calculation of the numbers $z_1^{\widetilde{\theta}}, \ldots, z_m^{\widetilde{\theta}}$ requires running the forward-reverse algorithm for conditional expectations. More precisely, using the Markov property we decompose

$$z_i^{\widetilde{\theta}} := \mathbb{E}_{\widetilde{\theta}}\Big[S_i(X_0,\ldots,X_{n_N})\Big|X_{n_j} = x_{n_j},\ j = 0,\ldots,N\Big]$$
$$= \sum_{j=1}^{N} \mathbb{E}_{\widetilde{\theta}}\Big[S_i^j(X_{n_{j-1}},\ldots,X_{n_j})\Big|X_{n_{j-1}} = x_{n_{j-1}}, X_{n_j} = x_{n_j}\Big].$$

All these conditional expectations are of the Markov-bridge type for which the forward-reverse algorithm is designed. Hence, for each iteration of the EM algorithm, we apply the forward-reverse algorithm N times, one for the time-intervals $n_{j-1},\ldots,n_j$, $j = 1,\ldots,N$, evaluating all the functionals $h_1^j,\ldots,h_m^j$ at one go.

5.1. **Choosing the reverse process.** Recall the defining equation for the one-step transition density q of the reverse process given in (2.4). For simplicity, we shall again assume that the forward and the reverse processes are time-homogeneous, implying that (2.4) can be re-expressed as

$$q(y,z) = \frac{p(z,y)}{\psi(y,z)}.$$

Notice that in this equation only p is given a-priori, i.e., the user is free to choose any re-normalization ψ provided that for any $y \in \mathbb{R}^d$ the resulting function $z \mapsto q(y,z)$ is non-negative and integrates to 1. In particular, we can turn the equation around, choose *any* transition density q and *define*

$$\psi(y,z) := \frac{p(z,y)}{q(y,z)}.$$

Note, however, that for the resulting forward-reverse process square integrability of the process $\mathcal{Y}$ is desirable. More precisely, only square integrability of the (numerator of the) complete estimator corresponding to (2.6) is required, but it seems far-fetched to hope for any cancelations giving square integrable estimators when $\mathcal{Y}$ itself is not square integrable. From a practical point of view, it therefore seems reasonable to aim for functions ψ satisfying

$$\psi \approx 1$$

in the sense that ψ is bounded from above by a number slightly smaller than 1 and bounded from below by a number slightly smaller than 1. Indeed, note that $\mathcal{Y}$ is obtained by multiplying terms of the form $\psi(Y_n, Y_{n+1})$ along the whole trajectory of the reverse process Y. Hence, if ψ is bounded by a large constant, $\mathcal{Y}$ could easily take extremely large values, to the extent that buffer-overflow might occur in the numerical implementation – think of multiplying 100 numbers of order 100. On the other hand, if ψ is considerably smaller than 1, $\mathcal{Y}$ might take very small values, which can cause problems in particular taking into account the division by the forward-reverse estimator for the transition density in the denominator of the forward-reverse estimator.

Heuristically, the following procedure seems promising.

- If $y \mapsto \int_{\mathbb{R}^d} p(z,y)dz$ can be computed in closed form (or so fast that one can think of a closed formula), then choose
$$\psi(y) := \psi(y,z) = \int_{\mathbb{R}^d} p(z,y)dz.$$
- Otherwise, assume that we can find a non-negative (measurable) function $\widetilde{p}(z,y)$ with closed form expression for $\int_{\mathbb{R}^d} \widetilde{p}(z,y)dz$ such that $p(z,y) \approx \widetilde{p}(z,y)$. Then define
$$q(y,z) := \frac{\widetilde{p}(z,y)}{\int_{\mathbb{R}^d} \widetilde{p}(z,y)dz},$$
which is a density in z. By construction, we have
$$\psi(y,z) = \frac{p(z,y)}{q(y,z)} = \int_{\mathbb{R}^d} \widetilde{p}(z,y)dz\,\frac{p(z,y)}{\widetilde{p}(z,y)},$$

implying that we are (almost) back in the first situation.

Remark 5.1. Even if we can, indeed, explicitly compute $\psi(y,z) = \int_{\mathbb{R}^d} p(z,y)dz$, there is generally no guarantee that $\mathcal{Y}$ has (non-exploding) finite second moments. However, in practice, this case seems to be much easier to control and analyze.

5.2. **Complexity of the forward-reverse algorithm.** We end this general discussion of the forward-reverse EM algorithm by a refined analysis of the complexity of the forward-reverse algorithm for conditional expectations as compared to [Bayer and Schoenmakers, 2014]. We start with an auxiliary lemma concerning the maximum product of numbers of two species of balls in bins, which is an easy consequence of a result by Gonnet [Gonnet, 1981], see also [Sedgewick and Flajolet, 1996, Section 8.4].

Lemma 5.2. *Let X be a random variable supported in a compact set $D \subset \mathbb{R}^d$ with a uniformly bounded density p. For any $K \in \mathbb{N}$ construct a partition $B_1^K, \ldots, B_K^K$ of D in measurable sets of equal Lebesgue measure $\lambda(B_i^K) = \lambda(D)/K$, $i = 1, \ldots, K$. Finally, for given $M \in \mathbb{N}$ let $X_1, \ldots, X_M$ be a sequence of independent copies of X and define*

$$M_k := \#\left\{ i \in \{1, \ldots, M\} \mid X_i \in B_k^K \right\}, \quad k = 1, \ldots, K.$$

For $M, K \to \infty$ such that $M = O(K)$ we have the asymptotic relation

$$\mathbb{E}\left[\max_{k=1,\ldots,K} M_k\right] = O\left(\frac{\log M}{\log\log M}\right).$$

Proof. W.l.o.g., we may assume that $\lambda(D) \le 1$. Let

$$p_k := P\left(X \in B_k^K\right) \le \|p\|_\infty / K$$

and observe that the random vector $(N_1, \ldots, N_K)$ satisfies a multi-nomial distribution with parameters K, N and $(p_1, \ldots, p_K)$.

The proof for the statement in the special case of $p_1 = \ldots = p_K = 1/K$ is given in Gonnet [Gonnet, 1981], so we only need to argue that the relation extends to the non-uniform case. To this end, let $K' := \lfloor K / \|p\|_\infty \rfloor$ and let $(N_1, \ldots, N_{K'})$ denote a multi-nomial random variable with parameters M, K' and $(1/K', \ldots, 1/K')$. Observe that $\|p\|_\infty \ge 1$ by the assumption that $\lambda(D) \le 1$. Hence, $K' \le K$ and $1/K' \ge p_k$, $\forall k = 1, \ldots, K$. As $M = O(K')$ we have by Gonnet's result that

$$\mathbb{E}\left[\max_{k=1,\ldots,K'} N_k\right] = O\left(\frac{\log M}{\log\log M}\right).$$

Moreover, it is clear that $\mathbb{E}\left[\max_{k=1,\ldots,K} M_k\right] \le \mathbb{E}\left[\max_{k=1,\ldots,K'} N_k\right]$—in the latter case, the same number of samples is put into fewer boxes with uniformly higher probabilities each—and we have proved the assertion. □

Theorem 5.3. *Assume that the transition densities p and q have compact support in $\mathbb{R}^d$.*[2] *Moreover, assume that the kernel K is supported in a ball of radius $R > 0$. Then the forward-reverse algorithm for M forward and reverse trajectories based on a bandwidth proportional to $M^{-1/d}$ can be implemented in such a way that its expected cost is $O(M \log M)$ as $M \to \infty$.*

Proof. In order to increase the clarity of the argument, we re-write the double sum in the forward-reverse algorithm to a simpler form, which highlights the computational issues. Indeed, we are trying to compute a double sum of the form

$$\sum_{i=1}^{M} \sum_{j=1}^{M} F_{i,j} K_\epsilon \left(X_{n^*}^i - Y_{\hat{n}_l}^j \right), \tag{5.3}$$

[2] Obviously, this assumption can be weakened.

where $F_{i,j}$ obviously depends on the whole ith sample of the forward process X and on the whole jth sample of the reverse process $(Y, \mathcal{Y})$.

We may assume that the end points $X^i_{n^*}$ and $Y^j_{\hat{n}_l}$ of the M samples of the forward and reverse trajectories are contained in a compact set $[-L, L]^d$. (Indeed, the necessary rescaling operation can obviously be done with $O(M)$ operations.) In fact, for ease of notation we shall assume that the points are actually contained in $[0,1]^d$. We sub-divide $[0,1]^d$ in boxes with side-length $S\epsilon$, where $S > R$ is chosen such that $1/(S\epsilon) \in \mathbb{N}$. Note that there are $K := (S\epsilon)^{-d}$ boxes which we order lexicographically and associate with the numbers $1, \dots, K$ accordingly.

In the next step, we shall order the points $X^i_{n^*}$ and $Y^j_{\hat{n}_l}$ into these boxes. First, let us define a function $f_1 : [0,1]^d \to \{1, \dots, 1/(S\epsilon)\}^d$ by setting

$$f_1(x) := (\lceil x_1/(S\epsilon)\rceil, \dots, \lceil x_d/(S\epsilon)\rceil),$$

with $\lceil\cdot\rceil$ denoting the smallest integer larger or equal than a number. Moreover, define $f_2 : \{1, \dots, 1/(S\epsilon)\}^d \to \{1, \dots, K\}$ by

$$f_2(i_1, \dots, i_d) := (i_1 - 1)(S\epsilon)^{-d+1} + (i_2 - 1)(S\epsilon)^{-d+2} + \cdots + (i_d - 1) + 1.$$

Obviously, a point $x \in [0,1]^d$ is contained in the box number k if and only if $f_2(f_1(x)) = k$.[3] Now we apply a sorting algorithm like quick-sort to both sets of points $\left(X^1_{n^*}, \dots, X^M_{n^*}\right)$ and $\left(Y^1_{\hat{n}_l}, \dots, Y^M_{\hat{n}_l}\right)$ using the ordering relation defined on $[0,1]^d \times [0,1]^d$ by

$$x < y : \Longleftrightarrow f_2(f_1(x)) < f_2(f_1(y)).$$

Sorting both sets incurs a computational cost of $O(M \log M)$, so that we can now assume that the vectors $X^i_{n^*}$ and $Y^i_{\hat{n}_l}$ are ordered.

Notice that $K_\epsilon(x - y) \neq 0$ if and only if x and y are situated in neighboring boxes, i.e., if $|f_1(x) - f_1(y)|_\infty \leq 1$, where we define $|\alpha|_\infty := \max_{i=1,\dots,d} |\alpha_i|$ for multi-indices α. Moreover, there are 3^d such neighboring boxes, whose indices can be easily identified, in the sense that there is a simple set-valued function f_3 which maps an index k to the set of all the indices $f_3(k)$ of the at most 3^d neighboring boxes. Moreover, for any $k \in \{1, \dots, K\}$ let $X^{i(k)}_{n^*}$ be the first element of the ordered sequence of $X^i_{n^*}$ lying in the box k. Likewise, let $Y^{j(k)}_{\hat{n}_l}$ be the first element in the ordered sequence $Y^j_{\hat{n}_l}$ lying in the box with index k. Note that identifying these $2K$ indices $i(1), \dots, i(K)$ and $j(1), \dots, j(K)$ can be achieved at computational costs of order $O(K \log M) = O(M \log M)$.

After all these preparations, we can finally express the double sum (5.3) as

$$\sum_{i=1}^{M}\sum_{j=1}^{M} F_{i,j} K_\epsilon\left(X^i_{n^*} - Y^j_{\hat{n}_l}\right) = \sum_{k=1}^{K}\sum_{r\in f_3(k)}\sum_{i=i(k)}^{i(k+1)-1}\sum_{j=j(r)}^{j(r+1)-1} F_{i,j} K_\epsilon\left(X^i_{n^*} - Y^j_{\hat{n}_l}\right). \tag{5.4}$$

Regarding the computational complexity of the right hand side, note that we have the deterministic bounds

$$K = O(M),$$
$$|f_3(k)| \leq 3^d.$$

Moreover, regarding the stochastic contributions, the expected maximum number of samples $X^i_{n^*}$ ($Y^j_{\hat{n}_l}$, respectively) contained in any of the boxes is bounded by $O(\log M)$ by Lemma 5.2, i.e.,

$$\mathbb{E}\left[\max_{r=1,\dots,K}(j(r+1) - j(r))\right] = O(\log M),$$

which then needs to be multiplied by the total number M of points $X^i_{n^*}$ to get the complexity of the double summation step. □

[3] To make this construction fully rigorous, we would have to make the boxes half-open and exclude the boundary of $[0,1]^d$.

Remark 5.4. As becomes apparent in the proof of Theorem 5.3, the constant in front of the asymptotic complexity bound does depend exponentially on the dimension d.

Remark 5.5. Notice that the box-ordering step can be omitted by maintaining a list of all the indices of trajectories whose end-points lie in every single box. The asymptotic rate of complexity of the total algorithm does not change by omitting the ordering, though.

6. Applications of the FREM algorithm

The forward reverse EM algorithm is a versatile tool for parameter estimation in dynamic stochastic models. It can be applied in discrete time Markov models, but also in the the setting of discrete observations of time-continuous Markov processes such as diffusions for example.

In this section we give examples from both worlds: we start by a discretized Ornstein-Uhlenbeck process that serves as a benchmark model, since the likelihood function can be treated analytically. Then we give an example of a discrete time Markov chain with heavy tailed transition densities given by a two-parameter Z-distribution, see [Barndorff-Nielsen et al., 1982] for details. For a complex real data application of our method we refer the interested reader to the paper by Bayer et al. [Bayer et al., 2016] that was developed in parallel to an earlier version of the present one.

6.1. **Ornstein-Uhlenbeck dynamics.** In this section we apply the forward-reverse EM algorithm to simulated data from a discretized Ornstein-Uhlenbeck process. The corresponding Markov chain is thus given by

$$X_{n+1} = X_n + \lambda X_n \Delta t + \Delta W_{n+1}, \quad n \geq 0 \tag{6.1}$$

where W_n are independent random variables distributed according to $\mathcal{N}(0, \Delta t)$. The drift parameter $\lambda \in \mathbb{R}$ is unknown and we will employ the forward reverse EM algorithm to estimate it from simulated data. The Ornstein-Uhlenbeck model has the advantage that the likelihood estimator is available in closed form and we can thus compare it to the results of the EM algorithm.

In each simulation run we suppose that we have known observations

$$X_0, X_{10\Delta t}, \ldots, X_{N\Delta t}$$

for varying step size Δt and use the EM methodology to approximate the likelihood function in between. We perform six iteration of the algorithm with increasing number of data points N.

In Table 1 we summarize the results of two runs for the discrete Ornstein-Uhlenbeck chain. The mean and standard deviation are estimated from 1000 Monte Carlo iterations. We find that already after three steps the mean is very close to the corresponding estimate of the true MLE. This indicates a surprisingly fast convergence for this example. Note also that the approximated value of the likelihood function stabilizes extremely fast at the maximum.

Table 2 gives results for the same setup as in Table 1 but with initial guess $\lambda = 2$ such that the forward-reverse EM algorithm converges from above to the true maximum of the likelihood function. We observe that the smaller step size $\Delta t = 0.05$ results in a more accurate approximation of the likelihood and also of the true MLE. It seems that the step size has crucial influence on the convergence rate of the algorithm, since for $\Delta t = 0.05$ the likelihood stabilizes already from the second iteration.

In Figure 1 the empirical distribution of 1000 estimates for λ is plotted. The initial value was 0.5 and the true maximum of the likelihood function is at 1.161. The step size between observations was chosen to be $\Delta t = 0.1$. The histogram on the left shows the estimates after only one iteration and on the right the estimates were obtained from five iterations of the forward-reverse EM algorithm.

Δt	N	bandwidth	mean $\hat{\lambda}$	std dev $\hat{\lambda}$	likel.	std dev likel.
0.1	2000	0.0005	0.972	0.0135	-3.402	0.00290
	8000	0.000125	1.098	0.00841	-3.383	0.00062
	32000	3.125e-05	1.132	0.00476	-3.381	0.000123
	128000	7.812e-06	1.151	0.00236	-3.381	2.783e-05
	512000	1.953e-06	1.157	0.00117	-3.381	4.745e-06
	2048000	4.882e-07	1.159	0.000581	-3.381	1.005e-06
0.05	2000	0.0005	1.160	0.0141	-3.107	0.000854
	8000	0.000125	1.247	0.00872	-3.103	9.867e-05
	32000	3.125e-05	1.253	0.00468	-3.103	1.329e-05
	128000	7.812e-06	1.265	0.00225	-3.103	3.772e-06
	512000	1.953e-06	1.265	0.00111	-3.103	6.005e-07

Table 1. Behavior of the forward-reverse EM algorithm for a discretized Ornstein-Uhlenbeck model for different step sizes Δt, initial guess $\lambda = 0.5$ and true MLE $\hat{\lambda}_{\text{MLE}} = 1.161$ and 1.266, respectively.

Δt	N	bandwidth	mean $\hat{\lambda}$	std dev $\hat{\lambda}$	likel.	std dev likel.
0.1	2000	0.0005	1.554	0.0353	-3.457	0.0134
	8000	0.000125	1.312	0.0127	-3.393	0.00221
	32000	3.125e-05	1.217	0.00544	-3.382	0.000351
	128000	7.812e-06	1.185	0.00245	-3.381	5.817e-05
	512000	1.953e-06	1.168	0.00121	-3.381	1.227e-05
0.05	2000	0.0005	1.390	0.0248	-3.108	0.00238
	8000	0.000125	1.289	0.00925	-3.103	0.000130
	32000	3.125e-05	1.261	0.00471	-3.103	1.451e-05
	128000	7.812e-06	1.266	0.00221	-3.103	2.538e-06
	512000	1.953e-06	1.266	0.00113	-3.103	5.855e-07

Table 2. Behavior of the forward-reverse EM algorithm for a discretized Ornstein-Uhlenbeck model for different step sizes Δt, initial guess $\lambda = 2$ and true MLE $\hat{\lambda}_{\text{MLE}} = 1.161$ and 1.266, respectively.

Figure 2 depicts the distribution of 1000 Monte Carlo samples of the likelihood values that led to the estimates in Figure 1. It is interesting to see that after one iteration of the algorithm the likelihood values are approximately bell shaped (left histogram) whereas after five iterations the distributions becomes more and more one-sided as would be expected, since the EM algorithm only increase the likelihood from step to step towards the maximum.

Figure 3 shows the convergence of the forward reverse EM algorithm when the number of iterations increases. We find that already after 4 iterations the estimate is very close to the true MLE for λ. After six iterations the algorithm has almost perfectly stabilized at the value of the true MLE $\lambda = 1.16$.

6.2. **Z-distribution models.** As a second application we consider the EM Algorithm in the context of an autonomous time series model where the one step probability is given by a Z-distribution [Barndorff-Nielsen et al., 1982]. The density of the z-distribution is given by

$$p(x;\alpha,\beta,\sigma) = \frac{\exp\left[\alpha x/\sigma\right]}{\sigma B(\alpha,\beta)\left(1+\exp\left[x/\sigma\right]\right)^{\alpha+\beta}}, \qquad x \in \mathbb{R};\ \alpha,\beta,\sigma > 0 \tag{6.2}$$

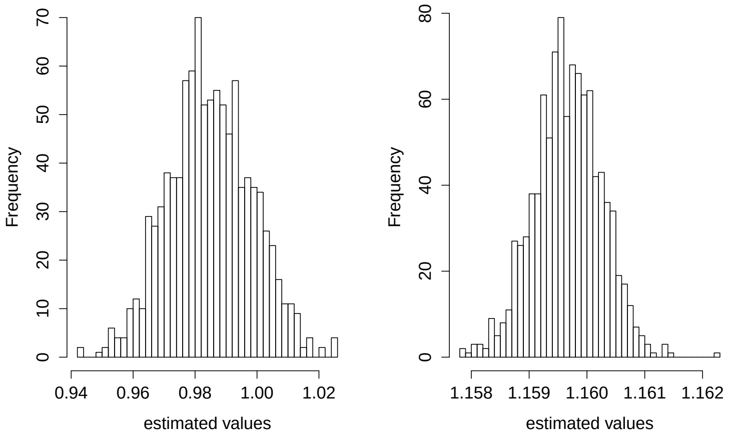

Figure 1. Empirical distribution of 1000 estimates after one iteration (right) and after five iteration (left) of the forward-reverse EM algorithm.

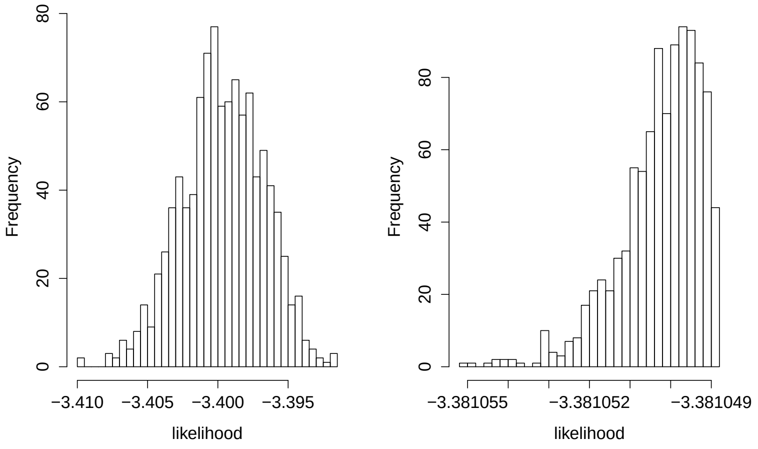

Figure 2. Empirical distribution of the likelihood values of 1000 Monte Carlo samples after one iteration (right) and after five iteration (left) of the forward-reverse EM algorithm.

with $B(\cdot,\cdot)$ being the beta-function. The density (6.2) is heavy tailed in the sense that the upper and lower tail of $\log p(x;\alpha,\beta,\sigma)$ tend to a straight line asymptotically. Further, if $\alpha = \beta$ it is symmetric and for $\alpha > \beta$ ($\beta > \alpha$) negatively (positively) skewed. The autonomous one-step transition density is thus given by,

$$p_{i,i+1}\,(x_i, x_{i+1}) := p_X(x_i, x_{i+1}) := p(x_{i+1} - x_i;\alpha,\beta,\sigma),$$

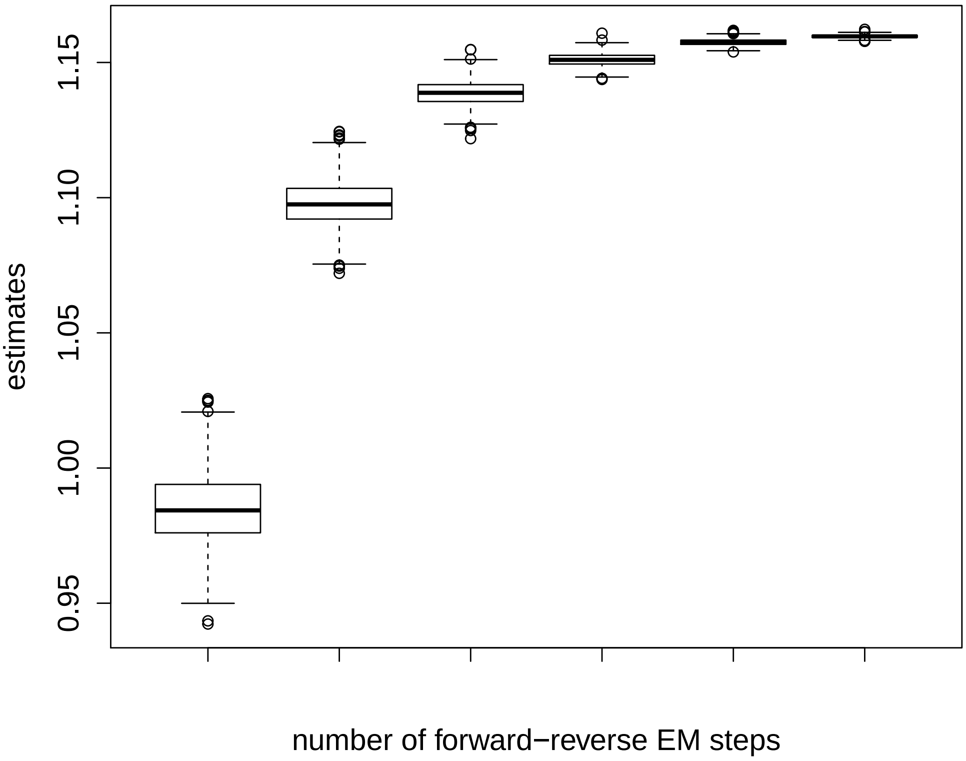

Figure 3. Convergence of the forward-reverse EM algorithm from one to six iterations for each 1000 estimates of λ. The value of the true MLE is $\hat{\lambda} = 1.161$.

and the full-data log-likelihood expression (3.1) becomes

$$
\begin{aligned}
l_c(\alpha,\beta,\sigma,x) &= \sum_{i=0}^{n_N-1} \log\left[\frac{\exp\left[\alpha\left(x_{i+1}-x_i\right)/\sigma\right]}{\sigma B(\alpha,\beta)\left(1+\exp\left[\left(x_{i+1}-x_i\right)/\sigma\right]\right)^{\alpha+\beta}}\right] \\
&= \frac{\alpha\left(x_{n_N}-x_0\right)}{\sigma} - n_N \log\left[\sigma B(\alpha,\beta)\right] \\
&\quad - (\alpha+\beta)\sum_{i=0}^{n_N-1} \log\left[1+\exp\left[\left(x_{i+1}-x_i\right)/\sigma\right]\right]..
\end{aligned} \tag{6.3}
$$

If we fix the choice of σ, i.e. $\sigma := \sigma_0$ is assumed to be known, then we are in the setting of an exponential family as in Section 4.1. Indeed, we have with $\theta = (\alpha,\beta)$ (while abusing notation a bit)

$$
l_c(\theta,x) = l_c(\alpha,\beta,x) = \phi(\alpha,\beta) + S_1\left(x_0,...,x_{n_N}\right)\psi_1(\alpha,\beta) + S_2\left(x_0,...,x_{n_N}\right)\psi_2(\alpha,\beta)
$$

with

$$\phi(\alpha,\beta) = -n_N \log\left[\sigma_0 B(\alpha,\beta)\right],$$
$$\psi_1(\alpha,\beta) = \frac{\alpha}{\sigma_0}, \quad S_1\left(x_0,...,x_{n_N}\right) = x_{n_N} - x_0,$$
$$\psi_2(\alpha,\beta) = \alpha+\beta, \quad S_2\left(x_0,...,x_{n_N}\right) = -\sum_{i=0}^{n_N-1} \log\left[1+\exp\left[(x_{i+1}-x_i)/\sigma_0\right]\right],$$

cf. (4.1). For a fixed appropriate choice σ_0 we first consider the EM algorithm for the symmetric one parameter case $\alpha = \beta$, and then, subsequently, we consider the two parameter case.

6.2.1. *The case $\theta = \alpha = \beta$ with fixed σ_0.* In this case the one step probability reads

$$p^{\alpha}_{i,i+1}\left(x_i, x_{i+1}\right) = \frac{\exp\left[\alpha(x_{i+1}-x_i)/\sigma_0\right]}{\sigma_0 B(\alpha,\alpha)\left(1+\exp\left[(x_{i+1}-x_i)/\sigma_0\right]\right)^{2\alpha}} \tag{6.4}$$

and the full data log-likelihood becomes,

$$\begin{aligned} l_c(\alpha, X) &= l_c(\alpha, X_0, X_1, ..., X_{n_N}) \\ &= \sum_{i=0}^{n_N-1} \log p^{\alpha}_{i,i+1}\left(X_i, X_{i+1}\right) \\ &= \sum_{i=0}^{n_N-1} \log\left[\frac{\exp\left[\alpha\left(X_{i+1}-X_i\right)/\sigma_0\right]}{\sigma_0 B(\alpha,\alpha)\left(1+\exp\left[\left(X_{i+1}-X_i\right)/\sigma_0\right]\right)^{2\alpha}}\right] \\ &= -n_N \log\sigma_0 - 2n_N \log\Gamma(\alpha) + n_N \log\Gamma(2\alpha) \\ &\quad - 2\alpha \sum_{i=0}^{n_N-1}\left(\log\left[1+\exp\left[\left(X_{i+1}-X_i\right)/\sigma_0\right]\right] - \frac{X_{i+1}-X_i}{2\sigma_0}\right). \end{aligned}$$

Now let us consider the EM maximization step based on incomplete data $X_{(n)} := (X_{n_0}, X_{n_1}, ..., X_{n_N})$. Note that $X_{n_0} = X_0 = x_0$ is the initial state of the chain. Consider with $\theta_m = \alpha_m = \beta_m$,

$$\begin{aligned} Q(\alpha, \alpha_m, X_{n.}) &= \mathbb{E}_{\alpha_m}\left[l_c(\alpha, X_0, X_1, ..., X_{n_N}) \middle| X_{n_0}, X_{n_1} ..., X_{n_N}\right] \\ &= -n_N \log\sigma_0 - 2n_N \log\Gamma(\alpha) + n_N \log\Gamma(2\alpha) \\ &\quad - 2\alpha \sum_{j=0}^{N-1} \mathbb{E}_{\alpha_m}\left[\sum_{i=n_j}^{n_{j+1}-1}\left(\log\left[1+\exp\left[\left(X_{i+1}-X_i\right)/\sigma_0\right]\right] - \frac{X_{i+1}-X_i}{2\sigma_0}\right)\middle| X_{n_j}, X_{n_{j+1}}\right] \\ &= -n_N \log\sigma_0 - 2n_N \log\Gamma(\alpha) + n_N \log\Gamma(2\alpha) - 2\alpha\zeta\left(\alpha_m, x, \sigma_0, N\right) n_N, \end{aligned}$$

where we note that

$$\begin{aligned} &\zeta\left(\alpha_m, x, \sigma_0, N\right) := \\ &\frac{1}{n_N}\sum_{j=0}^{N-1} \mathbb{E}_{\alpha_m}\left[\sum_{i=n_j}^{n_{j+1}-1}\left(\log\left[1+\exp\left[\left(X_{i+1}-X_i\right)/\sigma_0\right]\right] - \frac{X_{i+1}-X_i}{2\sigma_0}\right)\middle| X_{n_j}, X_{n_{j+1}}\right] > \log 2 \end{aligned} \tag{6.5}$$

almost surely, since the function $y \to \log\left[1+e^y\right] - y/2$ has a global minimum $\log 2$ at $y = 0$. It is easy to check that

$$\alpha^\circ := \operatorname*{arg\,max}_{0<\alpha<\infty} Q(\alpha, \alpha_m, X_{(n)}) \tag{6.6}$$

necessarily satisfies

$$\frac{\Gamma'\left(2\alpha^\circ\right)}{\Gamma\left(2\alpha^\circ\right)} - \frac{\Gamma'\left(\alpha^\circ\right)}{\Gamma\left(\alpha^\circ\right)} = \zeta\left(\alpha_m, x, \sigma_0, N\right). \tag{6.7}$$

On the other hand, by properties of the polygamma function [Abramowitz and Stegun, 1964] the left-hand-side in (6.7) decays monotonically from $+\infty$ to $\log 2$ if $0 < \alpha^\circ < \infty$. Then, by using (6.5) and standard asymptotic behavior of the log-gamma function, it easily follows that (6.6) always has a unique global maximum.

In the EM algorithm the conditional expectations in (6.5) are of course not exactly known and therefore computed by the forward-reverse algorithm.

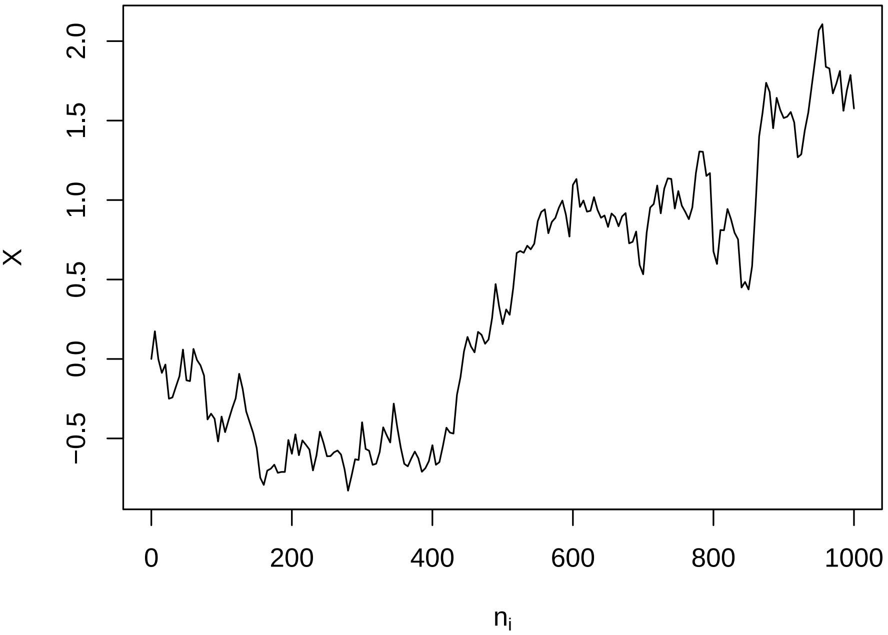

FIGURE 4. Synthetic data generated by a z-process with parameters $\alpha = \beta = 1.1$. The sample has size $N = 200$ and $n_{i+1} - n_i \equiv 5$, $i = 0, \ldots, N - 1$.

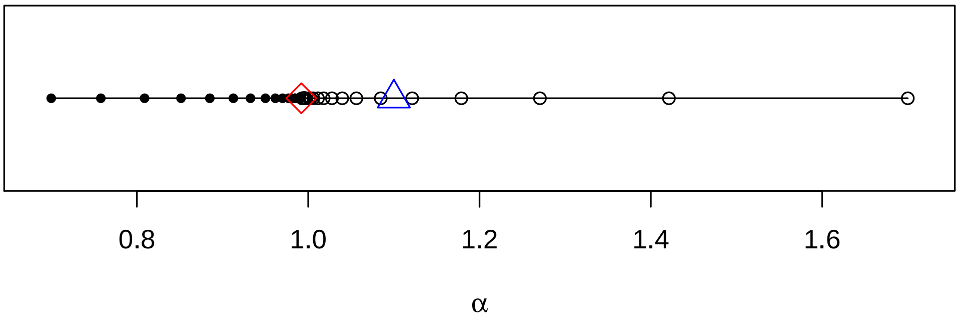

FIGURE 5. Convergence of the FREM algorithm for two different initial guesses. For each initial guess, the values obtained from the first 25 iterations are plotted. The upward-pointing triangle indicates the parameter used to generate the data, the diamond indicates the approximate limiting value of the FREM procedure. As the parameters are restricted to $\alpha = \beta$, we only plot the α coordinate.

We present one example for the estimation of the parameter $\alpha = \beta$ of a Z-process. In Figure 4 we show one trajectory of a Z-process generated with parameter $\alpha = \beta = 0.7$ and $\sigma = 0.0316$. [4]

[4] $\sigma = 0.0316 = 1/\sqrt{1000}$ essentially normalizes the variance of the terminal value X_{1000}.

Figure 5 shows the convergence of the FREM algorithm for two different initial guesses $\alpha = \beta = 0.7$ and $\alpha = \beta = 1.7$, respectively, for the data in Figure 4. We observe convergence to a parameter $\alpha = \beta \approx 0.992$.

6.2.2. *The case $\theta = (\alpha,\beta)$ with fixed σ_0.* Analogue to the one parameter case we now consider the two parameter case. Based on the incomplete data $X_{(n)} := (X_{n_0}, X_{n_1}, \dots, X_{n_N})$ we may write with $\theta_m = (\alpha_m,\beta_m)$ (cf. (6.3))

$$\begin{aligned}
Q(\alpha,\beta,\alpha_m,\beta_m,x) &= \mathbb{E}_{\theta_m}\left[l_c(\alpha,\beta,X_0,...,X_{n_N}) \middle| X_{n_0} = x_{n_0},...,X_{n_N} = x_{n_N}\right] \\
&= -n_N \log\left[\sigma_0 B(\alpha,\beta)\right] + \frac{\alpha-\beta}{2\sigma_0}\left(x_{n_N} - x_0\right) \\
&\quad - (\alpha+\beta)\,\mathbb{E}_{\alpha_m}\left[\sum_{i=n_j}^{n_{j+1}-1} \left(\log\left[1+\exp\left[\left(X_{i+1}-X_i\right)/\sigma_0\right]\right] - \frac{X_{i+1}-X_i}{2\sigma_0} \right) \middle| X_{n_j}, X_{n_{j+1}} \right] \\
&= -n_N \log\left[\sigma_0 B(\alpha,\beta)\right] + \frac{\alpha-\beta}{2\sigma_0}\left(x_{n_N} - x_0\right) - (\alpha+\beta)\,\zeta\left(\theta_m, x, \sigma_0, N\right) n_N
\end{aligned}$$

with $\zeta\left(\theta_m, x, \sigma_0, N\right)$ defined in (6.5). Similar to (6.6), it follows by solving for stationary points that

$$(\alpha^\circ,\beta^\circ) := \arg\max_{\alpha,\beta>0} Q(\alpha,\beta,\alpha_m,\beta_m,x)$$

satisfies the system

$$\frac{\Gamma'\left(\alpha^\circ+\beta^\circ\right)}{\Gamma\left(\alpha^\circ+\beta^\circ\right)} - \frac{\Gamma'\left(\alpha^\circ\right)}{2\Gamma\left(\alpha^\circ\right)} - \frac{\Gamma'\left(\beta^\circ\right)}{2\Gamma\left(\beta\right)} = \zeta\left(\theta_m, x, \sigma_0, N\right) > \log 2, \tag{6.8a}$$

$$\frac{\Gamma'\left(\alpha^\circ\right)}{\Gamma\left(\alpha^\circ\right)} - \frac{\Gamma'\left(\beta^\circ\right)}{\Gamma\left(\beta^\circ\right)} = \frac{1}{n_N\sigma_0}\left(x_{n_N} - x_0\right). \tag{6.8b}$$

It is possible to show (e.g. by using results in [Abramowitz and Stegun, 1964]) that

$$\operatorname*{range}_{\alpha,\beta>0}\left\{ \frac{\Gamma'\left(\alpha+\beta\right)}{\Gamma\left(\alpha+\beta\right)} - \frac{\Gamma'\left(\alpha\right)}{2\Gamma\left(\alpha\right)} - \frac{\Gamma'\left(\beta\right)}{2\Gamma\left(\beta\right)} \right\} = \left(\log 2, \infty\right),$$

and that the system (6.8) has a unique solution that represents a global maximum. In the corresponding EM algorithm the conditional expectations (6.5) need to be computed accurate enough by the forward-reverse algorithm.

We end this section by a numerical example based on a z-process with known parameter σ and two unknown parameters α and β that are not restricted to be equal. The sample is depicted in Figure 6. Note that $\alpha \neq \beta$ introduces a monotonous "drift" in the process.

Figure 7 shows the convergence of the FREM algorithm for the synthetic data shown in Figure 6 for the four different initial guesses $\theta_0 = (2,2)$, $\theta_0 = (0.2,2)$, $\theta_0 = (2,0.2)$ and $\theta_0 = (0.2,0.2)$. The results are based on $M = 5000$ simulations of the "unobserved" values of the process X and a bandwidth $\epsilon = 10^{-3}$, and 25 iterations of the FREM algorithm. No stabilization was used for the algorithm.

For all the initial guesses, we observe convergence to the same parameter values $\overline{\theta} \approx (0.981, 0.804)$. The iterations seem to rapidly jump to a one-dimensional sub-mainifold (almost, but not quite flat) in just one step of the FREM, and then rather slowly, but monotonously move towards their suggested limit.

Acknowledgements

The third author was partially supported by the Research Center Matheon "Mathematics for Key Technologies" in Berlin. The second author was partially supported by the "Finance and Sustainable Growth Laboratory" of the Louis Bachelier LABEX in Paris. The first author is grateful to Pedro Villanova for enlightening discussions regarding the implementation of the algorithm leading to substantial improvements in Theorem 5.3.

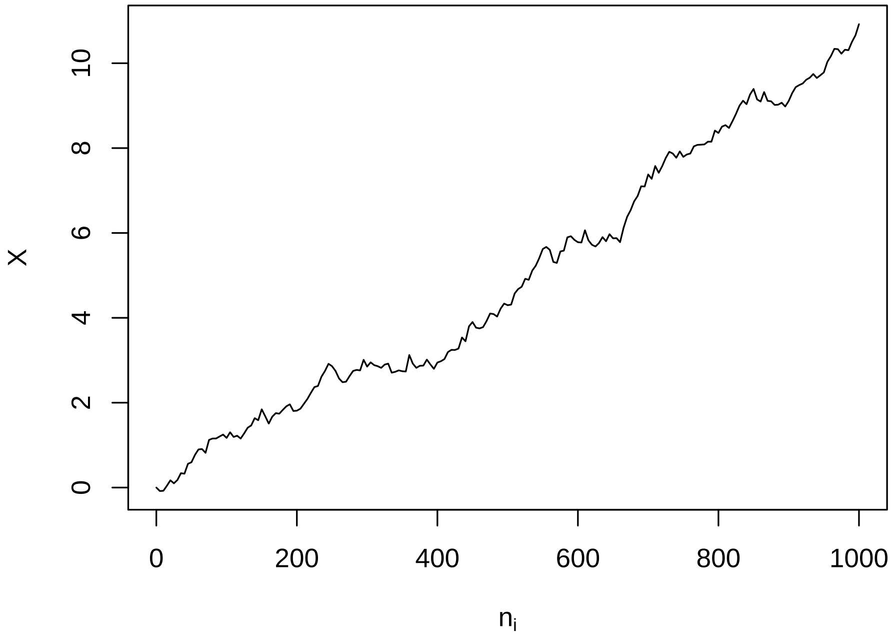

Figure 6. Synthetic data generated by a z-process with parameters $\alpha = 1.1$ and $\beta = 0.9$. The sample has size $N = 200$ and $n_{i+1} - n_i \equiv 5$, $i = 0, \dots, N - 1$.

References

[Abramowitz and Stegun, 1964] Abramowitz, M. and Stegun, I. A. (1964). *Handbook of mathematical functions with formulas, graphs, and mathematical tables*, volume 55 of *National Bureau of Standards Applied Mathematics Series*. For sale by the Superintendent of Documents, U.S. Government Printing Office, Washington, D.C.

[Barndorff-Nielsen et al., 1982] Barndorff-Nielsen, O., Kent, J., and Sørensen, M. (1982). Normal variance-mean mixtures and *z* distributions. *Internat. Statist. Rev.*, 50(2):145–159.

[Bayer et al., 2016] Bayer, C., Moraes, A., Tempone, R., and Vilanova, P. (2016). An efficient forward-reverse expectation-maximization algorithm for statistical inference in stochastic reaction networks. *Stoch. Anal. Appl.*, 34(2):193–231.

[Bayer and Schoenmakers, 2014] Bayer, C. and Schoenmakers, J. (2014). Simulation of forward-reverse stochastic representations for conditional diffusions. *Ann. Appl. Probab.*, 24(5):1994–2032.

[Bladt and Sørensen, 2014] Bladt, M. and Sørensen, M. (2014). Simple simulation of diffusion bridges with application to likelihood inference for diffusions. *Bernoulli*, 20(2):645–675.

[Bröckner, 1975] Bröckner, T. (1975). *Differential Germs and Catastrophes*. Cambridge University Press, Cambridge.

[Chan and Ledolter, 1995] Chan, K. and Ledolter, J. (1995). Monte Carlo EM estimation for time series models involving counts. *J. Am. Stat. Assoc.*, 90(429):242–252.

[Chen et al., 1988] Chen, H.-F., Guo, L., and Gao, A.-J. (1988). Convergence and robustness of the Robbins-Monro algorithm truncated at randomly varying bounds. *Stochastic Processes Appl.*, 27(2):217–231.

[Delyon and Hu, 2006] Delyon, B. and Hu, Y. (2006). Simulation of conditioned diffusion and application to parameter estimation. *Stochastic Process. Appl.*, 116(11):1660–1675.

[Dempster et al., 1977] Dempster, A., Laird, N., and Rubin, D. (1977). Maximum likelihood from incomplete data via the EM algorithm. Discussion. *J. R. Stat. Soc., Ser. B*, 39:1–38.

[Fort and Moulines, 2003] Fort, G. and Moulines, E. (2003). Convergence of the Monte Carlo expectation maximization for curved exponential families. *Ann. Stat.*, 31(4):1220–1259.

[Gonnet, 1981] Gonnet, G. H. (1981). Expected length of the longest probe sequence in hash code searching. *J. Assoc. Comput. Mach.*, 28(2):289–304.

[Lange, 1995] Lange, K. (1995). A gradient algorithm locally equivalent to the EM algorithm. *J. R. Stat. Soc., Ser. B*, 57(2):425–437.

[Lavielle and Moulines, 1999] Lavielle, B. D. M. and Moulines, E. (1999). Convergence of a stochastic approximation version of the EM algorithm. *Ann. Stat.*, 27(1):94–128.

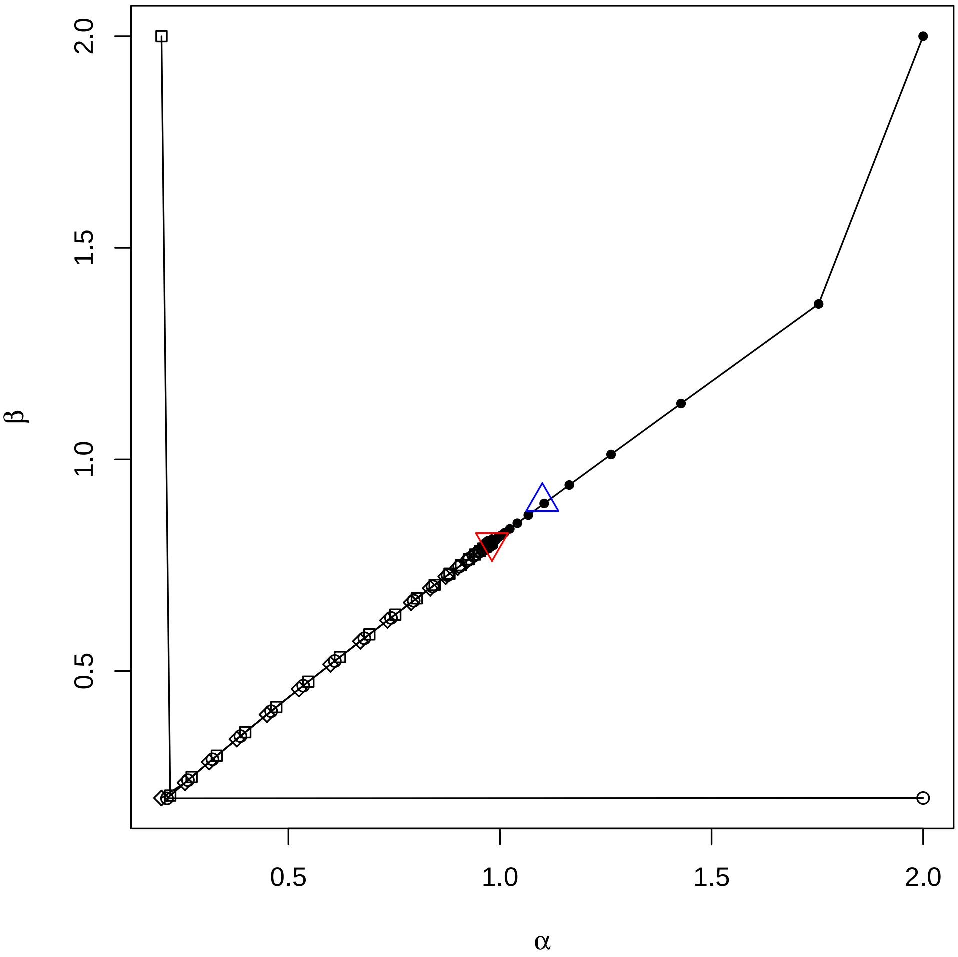

FIGURE 7. Convergence of the FREM algorithm for four different initial guesses. For each initial guess, the values obtained from the first 25 iterations are plotted. The upward pointing triangle indicates the parameters used to generate the data, the downward pointing triangle indicates the approximate limiting value of the FREM procedure.

[Liu and Rubin, 1994] Liu, C. and Rubin, D. B. (1994). The ECME algorithm: A simple extension of EM and ECM with faster monotone convergence. *Biometrika*, 81(4):633–648.

[MacDonald and Zucchini, 1997] MacDonald, I. L. and Zucchini, W. (1997). *Hidden Markov and other models for discrete-valued time series.* London: Chapman & Hall.

[Meng and Rubin, 1993] Meng, X.-L. and Rubin, D. B. (1993). Maximum likelihood estimation via the ECM algorithm: A general framework. *Biometrika*, 80(2):267–278.

[Meng and Schilling, 1996] Meng, X.-L. and Schilling, S. (1996). Fitting full-information item factor models and an empirical investigation of bridge sampling. *J. Am. Stat. Assoc.*, 91(435):1254–1267.

[Milstein et al., 2004] Milstein, G. N., Schoenmakers, J., and Spokoiny, V. (2004). Transition density estimation for stochastic differential equations via forward-reverse representations. *Bernoulli*, 10(2):281–312.

[Milstein et al., 2007] Milstein, G. N., Schoenmakers, J., and Spokoiny, V. (2007). Forward and reverse representations for Markov chains. *Stochastic Process. Appl.*, 117(8):1052–1075.

[Milstein and Tretyakov, 2004] Milstein, G. N. and Tretyakov, M. V. (2004). Evaluation of conditional Wiener integrals by numerical integration of stochastic differential equations. *J. Comput. Phys.*, 197(1):275–298.

[Neath, 2013] Neath, R. C. (2013). *On Convergence Properties of the Monte Carlo EM Algorithm*, volume 10. Institute of Mathematical Statistics.

[Schauer et al., 2013] Schauer, M., van der Meulen, F., and van Zanten, H. (2013). Guided proposals for simulating multi-dimensional diffusion bridges. Preprint.

[Sedgewick and Flajolet, 1996] Sedgewick, R. and Flajolet, P. (1996). *An Introduction to the Analysis of Algorithms*. Addison-Wesley.

[Stinis, 2011] Stinis, P. (2011). Conditional path sampling for stochastic differential equations through drift relaxation. *Commun. Appl. Math. Comput. Sci.*, 6(1):63–78.

[Stuart et al., 2004] Stuart, A. M., Voss, J., and Wiberg, P. (2004). Fast communication conditional path sampling of SDEs and the Langevin MCMC method. *Commun. Math. Sci.*, 2(4):685–697.

[Wei and Tanner, 1990] Wei, G. and Tanner, M. (1990). A Monte Carlo Implementation of the EM Algorithm and the Poor Man's Data Augmentation Algorithm. *J. Am. Stat. Assoc.*, 85:699–704.

[Wu, 1983] Wu, C. (1983). On the convergence properties of the EM algorithm. *Ann. Stat.*, 11:95–103.

E-mail address: `christian.bayer@wias-berlin.de`

Weierstrass Institute, Mohrenstr. 39, 10117 Berlin, Germany

E-mail address: `hilmar.mai@gmail.com`

Deutsche Bank, Berlin, Germany

E-mail address: `john.schoenmakers@wias-berlin.de`

Weierstrass Institute, Mohrenstr. 39, 10117 Berlin, Germany