An adaptive Monte-Carlo Markov chain algorithm for inference from mixture signals

Rémi Bardenet and Balázs Kégl

Linear Accelerator Laboratory (LAL) & Computer Science Laboratory (LRI)
IN2P3/CNRS & University of Paris-Sud, 91405 Orsay (France)

E-mail: bardenet@lri.fr, balazs.kegl@gmail.com

Abstract. Adaptive Metropolis (AM) is a powerful recent algorithmic tool in numerical Bayesian data analysis. AM builds on a well-known Markov Chain Monte Carlo algorithm but optimizes the rate of convergence to the target distribution by automatically tuning the design parameters of the algorithm on the fly. Label switching is a major problem in inference on mixture models because of the invariance to symmetries. The simplest (non-adaptive) solution is to modify the prior in order to make it select a single permutation of the variables, introducing an identifiability constraint. This solution is known to cause artificial biases by not respecting the topology of the posterior. In this paper we describe an online relabeling procedure which can be incorporated into the AM algorithm. We give elements of convergence of the algorithm and identify the link between its modified target measure and the original posterior distribution of interest. We illustrate the algorithm on a synthetic mixture model inspired by the muonic water Cherenkov signal of the surface detectors in the Pierre Auger Experiment.

1. Introduction

Inferring properties of individual particles in a mixture signal is a challenging problem. The goal of this paper is to describe an adaptive Monte-Carlo Markov chain (MCMC) algorithm that estimates the parameters of individual components from the mixture signal. Our objective is twofold: we aim at "popularizing" an existing technique [1] that eliminates the parameter-tuning step from MCMC, and we propose an extension of the algorithm to the case when the signal exhibits certain symmetries (such as in the case of mixtures). The main idea of the algorithm is to combine adaptive Metropolis [1] with online relabeling [2] with two goals in mind: 1) carrying out inference on individual components and 2) accelerating convergence. More details and further analysis of the algorithm can be found in [3].[1]

The paper is organized as follows. In section 2 we describe a synthetic model inspired by the muonic signal in water Cherenkov tanks of the Pierre Auger Experiment [4]. Section 3 contains the formal description of the Adaptive Metropolis with Online Relabeling (AMOR)

[1] The precise connection of the two papers is the following. The present paper concentrates on a concrete application (it describes the model and provides quantitative results), whereas [3] uses the physics example as an illustration and concentrates on the theoretical analysis of the algorithm (with a proof sketch available as supplementary material). The two papers are written to two almost mutually exclusive communities. The theoretical analysis of the algorithm is of little interest in the physics community whereas the concrete physics model we use in analyzing the Auger signal would draw equally little interest in the statistics community.

algorithm. We validate the algorithm on experiments conducted on synthetic data in section 4, and conclude the paper in section 5.

2. A motivating example

When a muon crosses a water tank, it generates Cherenkov photons and photons coming from other processes (e.g., delta rays) along its track at a rate depending on its energy. Some of these photons are captured by photomultipliers. The resulting photoelectrons (PEs) then generate analog signals that are discretized by an analog-to-digital converter.

In this section, we model the integer photoelectron (PE) count vector $\mathbf{n} = (n_1, \ldots, n_M) \in \mathbb{N}^M$ in the M bins of the signal, which means that we omit the model of the PMT electronics. Formally, n_i is the number of PEs in the i-th bin

$$[t_{i-1}, t_i) \triangleq [t_0 + (i-1)t_\Delta, t_0 + it_\Delta), \tag{1}$$

where t_0 is the absolute starting time of the signal, and $t_\Delta = 25\,\text{ns}$ is the signal resolution (size of one bin). The goal is to parametrize the likelihood $\mathcal{P}(\mathbf{n} \mid t_\mu, A_\mu)$ and the priors $\mathcal{P}(A_\mu)$ and $\mathcal{P}(t_\mu)$, where t_μ is the arrival time of the muon and A_μ is the integrated signal amplitude. A_μ is itself defined by $A_\mu = L_\mu \nu \phi_\mu$, where L_μ is the tracklength of the muon, ν is the mean number of PEs generated by a muon with kinetic energy of 1 GeV on a tracklength of 1 m, and ϕ_μ is the ratio between the number of photons produced by the muon and the expected number of photons generated by a muon with kinetic energy of 1 GeV. In this paper we use the dimensions of the Auger tank ($R = 1.8\,\text{m}$, $h = 1.2\,\text{m}$), and we choose to model signals of muons that arrive at $\theta = 45^\circ$. Furthermore, we use an energy spectrum measured for atmospheric muons [5] and assume that a muon with kinetic energy of 1 GeV generates, on average, $\nu = 228$ PEs per 1 m of tracklength. Putting it all together numerically, the distribution $\mathcal{P}(A_\mu)$ of the total number of PEs deposited by a muon is depicted by figure 1(a).

To model the time response of a muon, we assume that the photon generation is uniform in a window of width $t_d = 10\,\text{ns}$. The decay phase is modeled by an exponential distribution with parameter $\tau = 60\,\text{ns}$. The distribution of PEs in time can be modeled by a convolution of the generative process and the decay process. In our case the convolution can be solved analytically to obtain the time response distribution

$$\mathcal{P}_{\tau,t_d}(t) = \frac{1}{t_d} \cdot \begin{cases} 0 & \text{if } t < 0, \\ 1 - \exp(-t/\tau) & \text{if } 0 \le t < t_d, \\ \exp\big(-(t-t_d)/\tau\big) - \exp(-t/\tau) & \text{if } t_d \le t. \end{cases} \tag{2}$$

depicted by figure 1(b). The time of arrival distribution $\mathcal{P}(t_\mu)$ of the muons in a tank depends on the energy and the geometry of the shower, on the distance of the tank from the shower core, and on other external parameters. In this paper we will use an inverse Gamma distribution, with parameters dependent on several factors (distance of the tank from the shower axis, zeinth angle and energy of the shower). An example is shown in figure 1(c) with parameters 2.5 and 350. All distributions and parameter values were picked to realistically model the Auger water detector signal in certain regimes.

Given the arrival time t_μ of a muon and the associated total number of PEs A_μ, the PE count in the ith bin is a Poisson variate with parameter

$$\bar{n}_i(A_\mu, t_\mu) = A_\mu \int_{t_{i-1}}^{t_i} p_{\tau,t_d}(t - t_\mu)\mathrm{d}t. \tag{3}$$

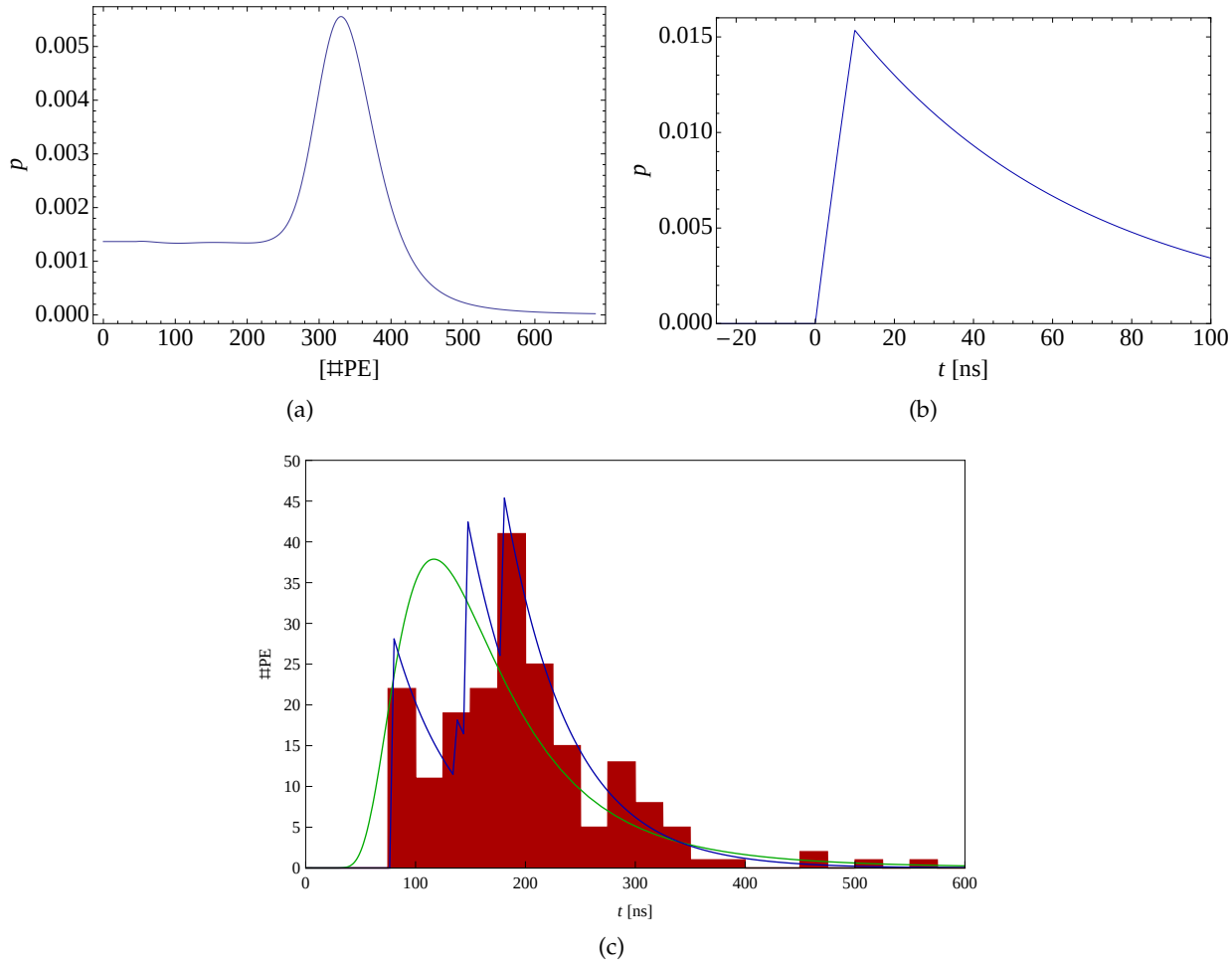


Figure 1. The generative model of the muonic signal. (a) The distribution $\mathcal{P}(A_\mu)$ of the muonic signal amplitude, used as a prior in the inference. (b) The muonic time response model $p_{\tau,t_d}(t)$. (c) An example signal. The green curve is the time-of-arrival distribution $\mathcal{P}(t_\mu)$, used as a prior in the inference. The blue curve is the "ideal" response $\sum_{j=1}^{N} A_{\mu_j} p_{\tau,t_d}(t - t_{\mu_j})$ of $N = 4$ muons drawn randomly from $t_\mu \sim p(t_\mu)$ and $A_\mu \sim \mathcal{P}(A_\mu)$, and the red histogram is the signal (PE count vector) $\mathbf{n}$.

Given N muons with signal amplitudes $\mathbf{A}_\mu = (A_{\mu_1}, \ldots, A_{\mu_N})$ and arrival times $\mathbf{t}_\mu = (t_{\mu_1}, \ldots, t_{\mu_N})$, the binwise signal expectation is

$$\bar{n}_i(\mathbf{A}_\mu, \mathbf{t}_\mu) = \sum_{j=1}^{N} \bar{n}_i(A_{\mu_j}, t_{\mu_j}), \tag{4}$$

and so our model is fully specified by the likelihood

$$\mathcal{P}(\bar{\mathbf{n}} \mid \mathbf{A}_\mu, \mathbf{t}_\mu) = \prod_{i=1}^{M} \mathrm{Poi}_{\bar{n}_i(\mathbf{A}_\mu, \mathbf{t}_\mu)}(n_i) \tag{5}$$

and the prior

$$\mathcal{P}(\mathbf{A}_\mu, \mathbf{t}_\mu) = \prod_{j=1}^{N} \mathcal{P}(A_{\mu,j}) \mathcal{P}(t_{\mu,j}), \tag{6}$$

where all bins and muons are assumed independent. Figure 1(c) depicts an example drawn from the model.

3. An adaptive MCMC algorithm with online relabeling

3.1. The symmetric random walk Metropolis algorithm

The goal of MCMC algorithms in the Bayesian context is to sample from the posterior distribution

$$\pi(X) \triangleq \pi(\mathbf{A}_\mu, \mathbf{t}_\mu) \propto \mathcal{P}(\bar{\mathbf{n}} \mid \mathbf{A}_\mu, \mathbf{t}_\mu)\mathcal{P}(\mathbf{A}_\mu, \mathbf{t}_\mu). \tag{7}$$

The posterior $\pi(X) = \pi(\mathbf{A}_\mu, \mathbf{t}_\mu)$, whose normalization constant is unknown, can be explored by running a Markov chain $X_t = (\mathbf{A}_\mu, \mathbf{t}_\mu)_t$ with stationary distribution π. In this context, π is also said to be the *target distribution* of the MCMC chain. The symmetric random walk Metropolis algorithm (SRWM [6], corresponding to the blue steps in figure 3) is one of the most popular techniques for simulating such a chain (X_t). In SRWM the user has to provide a symmetric proposal kernel that will be used to propose a new sample $\widetilde{X}$ given the previous sample X_{t-1}. This new sample is then accepted with probability $\min(1, \pi(\widetilde{X})/\pi(X_{t-1}))$. When the posterior is a distribution over a continuous space $\mathcal{X} = \mathbb{R}^d$ (with $d = 2N$ in our case), the most common proposal kernel is a multivariate Gaussian $\mathcal{N}\big(\,\cdot\,|X_{t-1}, \Sigma\big)$.

The goal of *adaptive Metropolis* (AM) (corresponding to the blue and green steps in figure 3) is to automatically calibrate the design parameter Σ of SRWM. When the target $\pi(x)$ is multivariate Gaussian with covariance Σ_π, the optimal choice of Σ is of the order of $(2.38)^2\Sigma_\pi/d$ [7, 8]. In practice, Σ_π is unknown thus motivating the use of an estimate of the covariance of the posterior based on samples $(X_1, \ldots, X_{t-1})$ generated so far. From a theoretical point of view, the conditional distribution of X_t given the past then depends on the whole past, rendering the analysis of AM more challenging. The convergence of AM has been recently addressed under quite general conditions (see, e.g., [9, 10] and references therein).

3.2. The challenge of permutation invariance

The optimal choice for Σ is appropriate only when the target distribution is strongly unimodal [8]. In our case both the signal likelihood $p(\bar{\mathbf{n}} \mid \mathbf{A}_\mu, \mathbf{t}_\mu)$ and the prior $\mathcal{P}(\mathbf{A}_\mu, \mathbf{t}_\mu)$ are invariant under any permutation of the muons which means that the posterior (7) has, in the general case, $N!$ modes. More precisely, let us assume that the number of muons is fixed to $N = 3$ and drop the μ index for the sake of simplicity, then any parameter vector $X = (A_1, A_2, A_3, t_1, t_2, t_3)^\intercal$ has the same posterior value as the relabeled vector $PX = (A_2, A_3, A_1, t_2, t_3, t_1)^\intercal$, where

$$P = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 \end{pmatrix}. \tag{8}$$

Actually, any other permutation $P \in \mathfrak{P}$ of the 3 muons would give the same posterior value. Throughout this paper, $\mathfrak{P}$ will denote the set of all block permutation matrices such as the one shown in (8). The described permutation invariance implies a risk for the MCMC chain to exhibit *label switching*, making componentwise averages meaningless. Figures 2(a)-2(b) illustrate the challenges when running vanilla AM on the example presented in figure 2(d). The red variable gets stuck in one of the mixture components, whereas the blue, green, and brown variables visit all the three remaining components. Marginal estimates computed for the blue,

green, and brown variables are then mostly identical as seen on figure 2(b). In addition, the shaded ellipses, depicting the marginal posterior covariances of the two parameters A_μ and t_μ of each muon, indicate that the resulting empirical covariance estimate is very broad, resulting in poor efficiency of the adaptive algorithm.

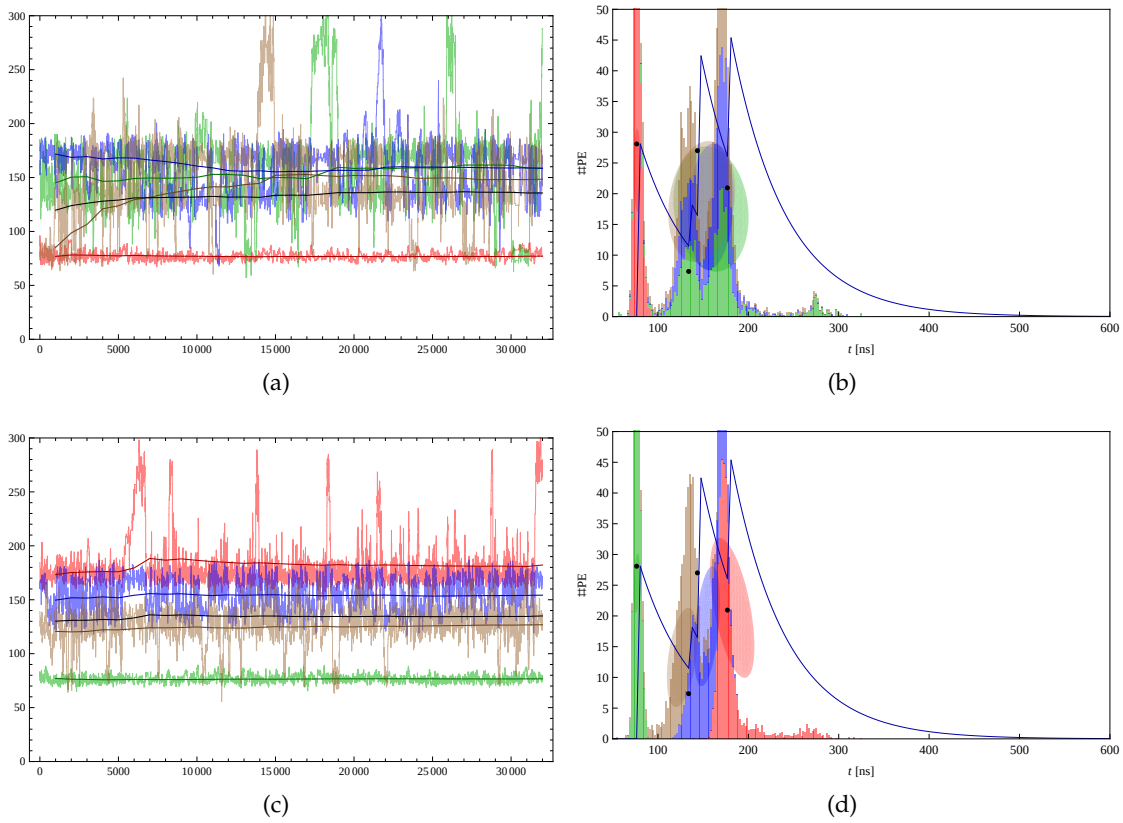


Figure 2. The results of AM (top row) and AMOR (bottom row) algorithms on the signal example of figure 1(c). The right panels show the parameters of the four muons. Black dots: the x-coordinates are the time-of-arrival parameters t_μ, and the y-coordinates are proportional to the amplitudes A_μ (more precisely, $A'_\mu = A_\mu t_\Delta (1 - e^{t_d/\tau})/t_d$ is the size of the "muon peak"). Colored ellipses are $\exp(1/2)$-level sets of Gaussian distributions: the means are the Bayesian estimates of (t_μ, A'_μ) for each muon, and the covariance is the marginal posterior covariance of each (t_μ, A'_μ) couple. The left panels show the four chains of the time-of-arrival parameters t_μ (light colors), the running means (dark colors), and the mean of the running means (black curve). The AM algorithm shows heavy label switching among the three rightmost muons whereas the AMOR algorithm separates the muons nicely.

Several approaches have been proposed to deal with the label switching problem. The first solution consists in modifying the prior in order to make it select a single permutation of the variables, introducing an *identifiability constraint* [11]. This solution is known to cause artificial biases by not respecting the topology of the posterior [12]. An effort has then been made to adapt to the estimated posterior surface through the design of relabeling algorithms [13, 12] that process the MCMC sample *after* the completion of the simulation run. These techniques look for a permutation P_t of each individual sample point X_t so as to minimize a posterior-based criterion depending on the *whole* chain history. [2] proposed an *online* version of the relabeling procedure in which the simulation of each X_t is followed by a permutation P_t of its

components. The permutation P_t is chosen to minimize a user-defined criterion that depends only on the *past* history of the chain up to time t. The major advantage of this online approach is that it is compatible with our objective of solving label switching "on the fly" in order to optimize AM for permutation-invariant models.

3.3. The AMOR algorithm

To prevent AM from label-switching, we go one step further and introduce online adaptive relabeling, the red steps in figure 3, resulting in the *adaptive Metropolis with online relabeling* algorithm (AMOR). The key step is the selection, after each proposal $\widetilde{X} \sim \mathcal{N}(.|X_{t-1}, \Sigma_{t-1}))$, of a permutation (relabeling) of the proposed vector by minimizing a quadratic cost function $L_{\mu_{t-1},\Sigma_{t-1}}(x)$ over all permutations of the proposed vector, where L is defined by

$$L_{\mu,\Sigma}(x) = (x-\mu)^\intercal \Sigma^{-1}(x-\mu). \tag{9}$$

This step forces the posterior sample to look as unimodal as possible. A side effect is the modification of the acceptance ratio (line 7 of the algorithm in figure 3), which now includes sums over $\mathfrak{P}$. Computing this correction factor for all the permutations in $\mathfrak{P}$ can be prohibitively slow if $|\mathfrak{P}|$ is large. In practice, in this case we can use an approximate term which sums over all single inversions, i.e. permutations of only two components/muons, which decreases the computational cost from $\mathcal{O}(|\mathfrak{P}|!)$ to $\mathcal{O}(|\mathfrak{P}|^2)$. Although this modification does not change the practical performance of the algorithm, the price we pay is the lack of theoretical convergence guarantee.

To illustrate the algorithm, we ran AMOR on the same example we used for AM. The mode selected by AMOR (figure 2(c)) corresponds roughly to ordering the arrival times (i.e., multiplying the target π by the indicator of the set $\{t_{\mu_1} < t_{\mu_2} < t_{\mu_3} < t_{\mu_4}\}$), although AMOR does allow the eventual mixing of neighboring components which is important to avoid the well-known bias of the ordering-based relabeling.

AMOR, similarly to identifiability constraints, aims at approximately selecting one of the many repeated modes of π. However, AMOR has the advantage of 1) selecting its mode automatically, avoiding the need for human intervention and possible bad choices of the constraint, and 2) providing good conditions to the AM algorithm since the sample looks as Gaussian as possible among its relabelings. The double adaptivity of AMOR, both in its proposal and in its selection mechanisms (and hence in its target π), makes its analysis particularly challenging. The convergence proof of AMOR is out of the scope of this paper and has been recently addressed in [3].

4. Experimental validation of AMOR

To validate AMOR and to illustrate the pitfalls of using AM without removing the permutation invariance of π, we ran both algorithms on 1200 tank signals with 20 bins and $N = 4$ muons each, simulated from the model presented in section 2. To create difficult and realistic situations, we set the arrival time distribution to be inverse Gamma with parameters 2 and 100. This had the effect of reducing the variance of the arrival time distribution, thus making simulations exhibit a reasonable number of overlapping muons. Examples of such simulated signals are depicted in figures 4(a), 4(b) and 4(c).

To quantify the performance after T iterations, we first selected the permutation of the running posterior mean components $(\hat{t}^{(T)}_{\mu_i})_{i=1,2,3,4}$ which minimized the sum of the ℓ_2 errors on the four estimates of the times of arrival $t_{\mu_i}, i = 1,2,3,4$, and we considered the sum of squared errors taken at this best permutation τ of the posterior mean to compute an error per

AMOR$(\pi(x), X_0, T, \mu_0, \Sigma_0, c)$

1 $\quad \mathcal{S} \leftarrow \emptyset$

2 $\quad$ **for** $t \leftarrow 1$ **to** T

3 $\quad\quad \Sigma \leftarrow c\Sigma_{t-1}$ ▷ *scaled adaptive covariance*

4 $\quad\quad \widetilde{X} \sim \mathcal{N}(\cdot | X_{t-1}, \Sigma)$ ▷ *proposal*

5 $\quad\quad \widetilde{P} \sim \underset{P \in \mathfrak{P}}{\arg\min}\, L_{(\mu_{t-1}, \Sigma_{t-1})}(P\widetilde{X})$ ▷ *pick an optimal permutation*

6 $\quad\quad \widetilde{X} \leftarrow \widetilde{P}\widetilde{X}$ ▷ *permute*

7 $\quad\quad$ **if** $\dfrac{\pi(X)\sum_{P \in \mathfrak{P}} \mathcal{N}(PX_{t-1} | X, \Sigma)}{\pi(X_{t-1})\sum_{P \in \mathfrak{P}} \mathcal{N}(PX | X_{t-1}, \Sigma)} > \mathcal{U}[0,1]$ **then**

8 $\quad\quad\quad X_t \leftarrow X$ ▷ *accept*

9 $\quad\quad$ **else**

10 $\quad\quad\quad X_t \leftarrow X_{t-1}$ ▷ *reject*

11 $\quad\quad \mathcal{S} \leftarrow \mathcal{S} \cup \{X_t\}$ ▷ *update posterior sample*

12 $\quad\quad \mu_t \leftarrow \mu_{t-1} + \dfrac{1}{t}(X_t - \mu_{t-1})$ ▷ *update running mean and covariance*

13 $\quad\quad \Sigma_t \leftarrow \Sigma_{t-1} + \dfrac{1}{t}\left((X_t - \mu_t)(X_t - \mu_t)^\intercal - \Sigma_{t-1}\right)$

14 $\quad$ **return** $\mathcal{S}$

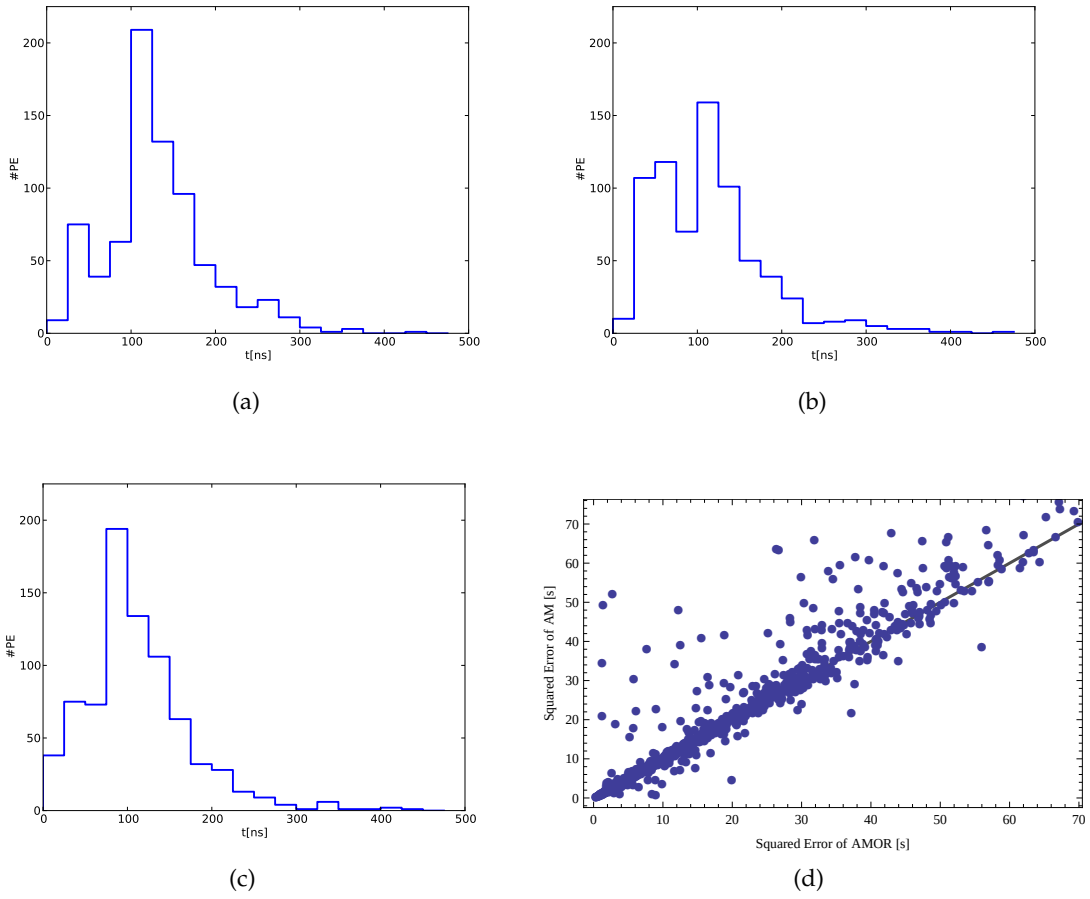
Figure 3. The pseudocode of the AMOR algorithm. The steps of the classical SRWM algorithm are in blue, the AM algorithm adds the green steps, and the new online relabeling steps are in red. Notice the adaptation of both the proposal (line 4) and the selection mechanism through the dependence of $L_{(\mu,\Sigma)}$ on (μ, Σ). Note that for practical reasons, a small εI_d is often added to the covariance matrix in line 3, but [14] recently confirmed that core AM does not lead to degenerate covariances. Note also that line 5 is usually a simple assignment of the optimal permutation. In case of ties, we draw uniformly from the finite set of optimal permutations.

muon:

$$S_T = \frac{1}{4}\left(\underset{\tau \in \mathfrak{S}_4}{\arg\min} \sum_{i=1}^{4} \left(\hat{t}^{(T)}_{\mu_{\tau(i)}} - t_{\mu_i}\right)^2\right)^{1/2}. \tag{10}$$

Figure 4(d) shows a scatterplot of S_T after $T = 3 \times 10^6$ iterations. On each signal, both AM and AMOR started with the same initial point, that is all points in figure 4(d) were lying on the diagonal. AMOR clearly outperformed AM on cases where label switching appears, leading to an estimate of the average error per muon of 17.0 ± 0.1ns versus 18.3 ± 0.1ns on these difficult cases.

The computational complexity of AMOR depends on several factors. The most costly step is the evaluation of the posterior ratio (line 7 in figure 3). In mixture models the evaluation of the posterior ratio is proportional to the number of components. Since it has to be done in every iteration, the number of iterations is also a linear factor. The execution time of our current unoptimized Mathematica implementation with a fixed number of iterations $T = 32000$ was in the order of minutes to tens of minutes, depending on these factors. An optimized C++ implementation would reduce this time by one or two orders of magnitudes.



(a) (b) (c) (d)

Figure 4. (a),(b),(c) Examples of generated signals, with 20 bins and 4 muons each. Arrival time distribution is $\mathcal{IG}(2, 100)$. (d) S_{3M}^{AM} vs. S_{3M}^{AMOR} results (see text).

5. Conclusion

We presented AMOR, a generic adaptive MCMC algorithm designed to tackle problems with permutation invariance, such as inferring the parameters of the muons of an Auger tank signal, and we demonstrated its capacities on simulated signals. The main advantages of AMOR are that 1) it selects a constraint on the posterior in a fully automatic fashion, avoiding the need for external expert knowledge, and 2) it allows for adaptive proposals which are known to be efficient when the target is unimodal [1]. The first advantage is a key point when it comes to models as complex as the generative model of a cosmic ray shower in Auger, since inferring the parameters of the muons is only meant to be a basic building block of a much larger scale inference algorithm.

Acknowledgments

This work was supported by the ANR-2010-COSI-002 grant of the French National Research Agency.

References

[1] Haario H, Saksman E and Tamminen J 2001 *Bernoulli* **7** 223–242
[2] Celeux G 1998 *COMPSTAT 98* ed Payne R and Green P (Physica-Verlag)

[3] Bardenet R, Cappé O, Fort G and Kégl B 2012 *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS)* vol 22 pp 91–99
[4] Pierre Auger Collaboration 1997 Pierre Auger project design report Tech. rep. Pierre Auger Observatory
[5] Boezio et al M 2003 *Phys. Rev. D* **67**
[6] Metropolis N, Rosenbluth A, Rosenbluth M, Teller A and Teller E 1953 *Journal of Chemical Physics* **21** 1087–1092
[7] Roberts G, Gelman A and Gilks W 1997 *The Annals of Applied Probability* **7** 110–120
[8] Roberts G and Rosenthal J 2001 *Statistical Science* **16** 351–367
[9] Saksman E and Vihola M 2010 *Annals of Applied Probability* **20** 2178–2203
[10] Fort G, Moulines E and Priouret P 2012 *Annals of Statistics* **39** 3262–3289
[11] Richardson S and Green P J 1997 *Journal of the Royal Statistical Society, Series B* **59** 731–792
[12] Marin J, Mengersen K and Robert C 2004 *Handbook of Statisics* **25**
[13] Stephens M 2000 *Journal of the Royal Statistical Society, Series B* **62** 795–809
[14] Vihola M 2011 *Electronic Journal of Probability* **16** 45–75