Relabeling and Summarizing Posterior Distributions in Signal Decomposition Problems when the Number of Components is Unknown

Alireza Roodaki, Julien Bect and Gilles Fleury

***Abstract*—This paper addresses the problems of relabeling and summarizing posterior distributions that typically arise, in a Bayesian framework, when dealing with signal decomposition problems with an unknown number of components. Such posterior distributions are defined over union of subspaces of differing dimensionality and can be sampled from using modern Monte Carlo techniques, for instance the increasingly popular RJ-MCMC method. No generic approach is available, however, to summarize the resulting variable-dimensional samples and extract from them component-specific parameters.**

We propose a novel approach, named *Variable-dimensional Approximate Posterior for Relabeling and Summarizing* (VAPoRS), to this problem, which consists in approximating the posterior distribution of interest by a "simple"—but still variable-dimensional—parametric distribution. The distance between the two distributions is measured using the Kullback-Leibler divergence, and a Stochastic EM-type algorithm, driven by the RJ-MCMC sampler, is proposed to estimate the parameters. Two signal decomposition problems are considered, to show the capability of VAPoRS both for relabeling and for summarizing variable dimensional posterior distributions: the classical problem of detecting and estimating sinusoids in white Gaussian noise on the one hand, and a particle counting problem motivated by the Pierre Auger project in astrophysics on the other hand.

***Index Terms*—Bayesian inference; Signal decomposition; Trans-dimensional MCMC; Label-switching; Stochastic EM.**

I. Introduction

Nowadays, owing to the advent of Markov Chain Monte Carlo (MCMC) sampling methods [2–5], Bayesian data analysis is considered as a conventional approach in machine learning, signal and image processing, and data mining problems—to name but a few. Nevertheless, in many applications, practical challenges remain in the process of extracting, from the generated samples, quantities of interest to summarize the posterior distribution.

Summarization consists, loosely speaking, in providing a few simple yet interpretable parameters and/or graphics to the end-user of a statistical method. For instance, in the case of a scalar parameter with a unimodal posterior distribution, measures of location and dispersion (e.g., the empirical mean and the standard deviation, or the median and the interquartile range) are typically provided in addition to a graphical summary of the distribution (e.g., a histogram or a kernel density estimate). In the case of multimodal distributions, summarization becomes more difficult but can be carried out using, for instance, the approximation of the posterior by a Gaussian Mixture Model (GMM) [6]. Summarizing or approximating posterior distributions has also been used in designing proposal distributions of Metropolis-Hastings (MH) samplers in an adaptive MCMC framework; see, e.g., [7–9].

This paper addresses the problem of summarizing posterior distributions in the case of some trans-dimensional problems (i.e., "problems in which the number of things that we don't know is one of the things that we don't know" [10, 11]). More specifically, we concentrate on the problem of signal decomposition when the number of components is unknown, which is an important case of trans-dimensional problem. Examples of such problems include the detection and estimation of sinusoids in white Gaussian noise [12] and the related problem of estimating directions of arrival in array processing [13], the detection of objects in images [14, 15], and the detection of physical particles (neutrons, muons, ...) using noisy data from various types of sensors, for instance in spectroscopy [16] or astrophysics [17, 18].

Let $\mathbf{y} = (y_1, y_2, \ldots, y_N)^t$ be a vector of N observations, where the superscript t stands for vector transposition. As a generic description of a signal decomposition problem, we consider a countable family of models $(\mathscr{M}_k)_{k\in\mathbb{N}}$, where it is assumed that, under model $\mathscr{M}_k$, the observed signal $\mathbf{y}$ is made of k components. A "component" might be a sinusoid (see Section I-B) or a decaying exponential (see Section III-B) in a one-dimensional signal processing problem, for instance, or an elementary geometric form in an image processing problem [15]; we simply assume that each component is completely described by a vector of parameters $\boldsymbol{\theta}_j \in \Theta \subseteq \mathbb{R}^d$, $1 \le j \le k$. We denote by $\boldsymbol{\theta}_{1:k} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_k) \in \Theta^k$ the vector of all component-specific parameters, where $\Theta^0 = \{\varnothing\}$.

One feature that the problems we are considering have in common is the invariance of the likelihood $p(\mathbf{y} \,|\, k, \boldsymbol{\theta}_{1:k})$ with respect to permutations (relabeling) of the components, which is called the "label-switching" issue in the literature; see, e.g., [19–23]. We will discuss this issue further in Section I-A.

In a Bayesian framework, a joint posterior density

Copyright (c) 2014 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works.

Alireza Roodaki did this work during his Ph.D. in SUPELEC[1] and during his post-doctoral stay in LTCI[2]. The results presented here are also part of his Ph.D. dissertation [1]. Email: al.roodaki@gmail.com.

Julien Bect is Associate Professor in SUPELEC[1], Department of Signal Processing and Electronic System. Gilles Fleury is Head of Research in SUPELEC[1]. Email: firstname.lastname@supelec.fr.

[1] E3S—SUPELEC Systems Sciences, SUPELEC, Gif-sur-Yvette, France.

[2] LTCI, CNRS / Télécom ParisTech, Paris, France.

$f(k, \boldsymbol{\theta}_{1:k}) \triangleq p(k, \boldsymbol{\theta}_{1:k} | \mathbf{y})$ is obtained through Bayes' formula for the number k of components and the vector of component-specific parameters, after assigning prior distributions on them:

$$f(k, \boldsymbol{\theta}_{1:k}) \propto p(\mathbf{y} | k, \boldsymbol{\theta}_{1:k}) p(\boldsymbol{\theta}_{1:k} | k) p(k), \quad (1)$$

where $\propto$ indicates proportionality. This density is defined over a variable-dimensional space $\boldsymbol{\Theta}$, which is a union of subspaces of differing dimensionality, i.e., $\boldsymbol{\Theta} = \cup_{k \in \mathbb{N}} \{k\} \times \Theta^k$.

The posterior density (1) completely describes the information (and the associated uncertainty) provided by the data $\mathbf{y}$ about the candidate models and the vector of unknown parameters. Since it is only known up to a normalizing constant in most cases, and potentially multimodal, Monte Carlo simulation methods, such as Reversible Jump MCMC (RJ-MCMC) [10], have been widely used to approximate it.

A. The label-switching issue

One of the most challenging issues when attempting at summarizing posterior distributions, that even occurs in fixed-dimensional situations, is the label-switching phenomenon (see, e.g., [19–25]), which is caused by the invariance of both the likelihood and the prior distribution under permutations of the components. As a consequence, the component-specific marginal posterior distributions are all equal, and therefore useless for the purpose of summarizing the information contained in the posterior distribution about individual components. A symptom of this issue is the multimodality of marginal posterior distributions.

The simplest way of dealing with the label-switching issue is to introduce an Identifiability Constraint (IC), such as sorting the components with respect to one of their parameters; see [19] for more discussion concerning the use of ICs in the Bayesian analysis of GMMs. However, in most practical examples, choosing an appropriate IC manually is not feasible. Many relabeling algorithms have therefore been developed to "undo" the label-switching effect automatically—i.e., change sample labels to make the marginals as unimodal as possible—but all of them are restricted to the case of *fixed*-dimensional posterior distributions; see [23, 25–27] for recent advances and references.

In variable-dimensional posterior distributions, there is an extra uncertainty about the "presence" of components, as will be clarified in the following illustrative example. This additional difficulty has hindered previous attempts to undo label-switching in the variable-dimensional scenario, where, according to [28] "*the meaning of individual components is vacuous*".

B. Illustrative example: joint Bayesian detection and estimation of sinusoids in white Gaussian noise

In this example, it is assumed that under $\mathscr{M}_k$, the observed signal $\mathbf{y}$ is composed of k sinusoidal components observed in white Gaussian noise. That is, under $\mathscr{M}_k$,

$$y[i] = \sum_{j=1}^{k} \left(a_{c,j} \cos(\omega_j i) + a_{s,j} \sin(\omega_j i) \right) + n[i],$$

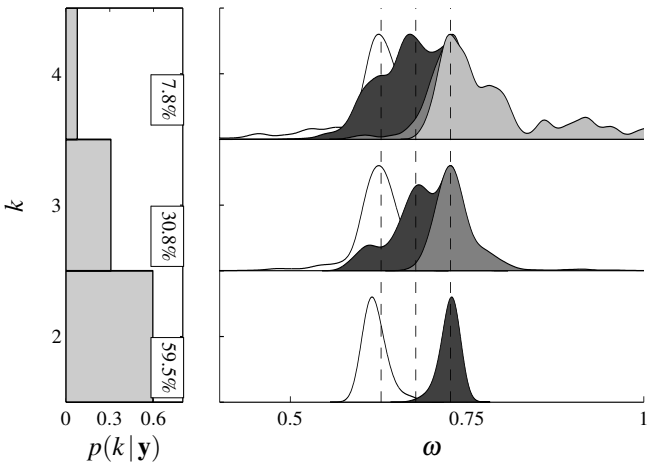


Figure 1: Posterior distributions of k (left) and sorted radial frequencies, $\boldsymbol{\omega}_{1:k}$, given k (right) from 100 000 output RJ-MCMC samples. The true number of components is three. The vertical dashed lines in the right figure locate the true radial frequencies. Not shown on the figure: $p(k \leq 1 \mid \mathbf{y}) \approx 0\%$ and $p(k \geq 5 \mid \mathbf{y}) \approx 1.9\%$.

where $a_{c,j}$ and $a_{s,j}$ are the cosine and sine amplitudes of the j^{th} sinusoidal component and ω_j its radial frequency. Moreover, n is a white Gaussian noise of variance σ^2.

The unknown parameters are the number k of sinusoidal components, the vectors $\boldsymbol{\theta}_j = (a_{c,j}, a_{s,j}, \omega_j)$ of component-specific parameters, $1 \leq j \leq k$, and the noise variance σ^2. Thus, $\Theta = \mathbb{R}^2 \times (0, \pi)$ and $\boldsymbol{\Theta} = \left(\cup_{k \in \mathbb{N}} \{k\} \times \Theta^k \right) \cup \mathbb{R}^+$. We use the hierarchical model, prior distributions, and the RJ-MCMC sampler proposed in [12] for this problem; the interested reader is thus referred to [10, 12] for more details[1].

Here, we consider an experiment in which the observed signal $\mathbf{y}$ of length $N = 64$ consists of three sinusoids with energies $\boldsymbol{A}_{1:k} = (20, 6.32, 20)^t$, where $A_j = a_{c,j}^2 + a_{s,j}^2$, phases $\boldsymbol{\phi}_{1:k} = (0, \pi/4, \pi/3)^t$, where $\phi_j = -\arctan(a_{s,j}/a_{c,j})$, and true radial frequencies $\boldsymbol{\omega}_{1:k} = (0.63, 0.68, 0.73)^t$. The signal-to-noise ratio[2] is set to the moderate value of 7 dB. Figure 1 represents the posterior distributions of both the number k of components and the sorted radial frequencies $\boldsymbol{\omega}_{1:k} = (\omega_1, \ldots, \omega_k)^t$ given k obtained using 100 000 samples generated by the RJ-MCMC sampler. Note that, here, we used sorting to mitigate the effect of label-switching for visualization. Each row is dedicated to one value of k, for $2 \leq k \leq 4$. Observe that other models have negligible posterior probabilities, since $p(2 \leq k \leq 4 \mid \mathbf{y}) = 0.981$.

Roughly speaking, two approaches co-exist in the literature for summarizing variable-dimensional posterior distributions: Bayesian Model Selection (BMS) and Bayesian Model Averaging (BMA). The BMS approach ranks models according to their posterior probabilities $p(k \mid \mathbf{y})$, selects one model, denoted by k^{MAP} here, where MAP stands for Maximum A Posteriori, and then summarizes the posterior distribution of the component-specific parameters under the (fixed-dimensional)

[1] In fact, the "Birth-or-Death" moves' acceptance ratio provided in the seminal paper [12] is erroneous. See [1, Chapter 1] or [29] for justification and true expression of the acceptance ratio, which is used in this paper.

[2] defined here as $\|\mathbf{D}\mathbf{a}_{1:k}\|^2 / \left(N\sigma^2\right)$, where $\mathbf{D}$ is the $N \times 2k$ "design matrix" of sines and cosines associated to $\boldsymbol{\omega}_{1:k}$ and $\mathbf{a}_{1:k} = \left(a_{c,1}, a_{s,1}, \ldots, a_{c,k}, a_{s,k}\right)^t$.

selected model. This is at the price of losing valuable information provided by the other (discarded) models. For instance, in the example of Figure 1, all information about the small—and therefore harder to detect—middle component is lost by selecting the most *a posteriori* probable model $\mathscr{M}_2$. On the other hand, the BMA approach consists in reporting results that are averaged over all possible models. Although BMA is suitable for signal reconstruction and prediction, it is not appropriate for studying component-specific parameters, the number of which changes in each model[3]. More information concerning these two approaches can be found in [10, 30] and references therein.

To the best of our knowledge, no generic method is currently available that would allow to summarize the information that is so easily read on Figure 1 for this very simple example[4]: namely, that *there seem to be three sinusoidal components in the observed noisy signal, the middle one having a smaller "probability of presence" than the others*.

C. Outline of the paper

In this paper, we propose a novel approach, named *Variable-dimensional Approximate Posterior for Relabeling and Summarizing* (VAPoRS), for relabeling and summarizing posterior distributions defined over variable-dimensional subspaces that typically arise in signal decomposition problems when the number of components is unknown. It consists in approximating the true posterior distribution with a parametric model (of varying-dimensionality), by minimization of the Kullback-Leibler (KL) divergence between the two distributions. A Stochastic Expectation Maximization (SEM)-type algorithm [31–33], driven by the output of an RJ-MCMC sampler, is used to estimate the parameters of the approximate model.

VAPoRS shares some similarities with the relabeling algorithms proposed in [20, 26, 27] to solve the label switching problem, and also with the EM-type algorithm used in [8] in the context of adaptive MCMC algorithms (both in a *fixed*-dimensional setting). The main contribution of this paper is the introduction of an original variable-dimensional parametric model, which allows to tackle directly the difficult problem of approximating a distribution defined over a union of subspaces of differing dimensionality, and thus provides a first solution to the "trans-dimensional label-switching" problem, so to speak.

Perhaps, the algorithm that we propose can be seen as a realization of the idea that M. Stephens had in mind when he wrote [34, page 94]:

"*This raises the question of whether we might be able to obtain an alternative view of the [variable-dimensional] posterior by combining the results for all different k's, and grouping together components which are "similar", in that they have similar predictive density estimates. However, attempts to do this have failed to produce any easily interpretable results.*"

The paper is organized as follows. Section II introduces the proposed model and stochastic algorithm for relabeling and summarizing variable dimensional posterior distributions. Section III illustrates the performance of VAPoRS using two signal decomposition examples, namely, the problem of joint Bayesian detection and estimation of sinusoids in white Gaussian noise and the problem of joint Bayesian detection and estimation of particles in the Auger project (in astrophysics). Section IV confirms the performances of VAPoRS using a Monte Carlo experiment. Finally, Section V concludes the paper and gives directions for future work.

II. VAPoRS

A. Introduction: observed components and t-components

We assume that the target posterior distribution, defined on the variable-dimensional space $\boldsymbol{\Theta} = \bigcup_{k\in\mathbb{N}} \{k\} \times \Theta^k$, admits a probability density function (pdf) f with respect to the kd-dimensional Lebesgue measure on each $\{k\} \times \Theta^k$, $k \in \mathbb{N}^*$.

Our objective is to approximate the true posterior density f using a "simple" parametric model. This parametric model will *also* be defined on the variable-dimensional space $\boldsymbol{\Theta}$ (i.e., it is not a fixed-dimensional approximation as in BMS).

We assume that a Monte Carlo sampling method—e.g., an RJ-MCMC sampler [10]—is available to generate M samples from f, which we denote by $\boldsymbol{\theta}^{(i)} = \left(k^{(i)}, \boldsymbol{\theta}^{(i)}_{1:k^{(i)}}\right)$, for $i = 1, \ldots, M$. These M variable-dimensional samples, which will be used to fit the approximate model to the true posterior, will play the role of input data for our method; therefore, an element $\boldsymbol{\theta}^{(i)}_j \in \Theta$ $(1 \leq j \leq k^{(i)})$ of the sample $\boldsymbol{\theta}^{(i)}$ will be referred to as an "observed component" (remember that a "component" is a point in the set Θ).

As our main device to handle the variable dimensionality of $\boldsymbol{\Theta}$, our parametric model (described more precisely in the next section) introduces "transdimensional components", that we will refer to as *t-components* for the sake of brevity. Intuitively, in the example shown on Figure 1, these t-components will serve the purpose of aggregating bumps in parameter posterior probability across dimensions, thus identifying observed components occurring in different models as manifestations of the same true component in the signal.

B. Variable-dimensional parametric model

Instead of trying to describe the proposed parametric model directly, let us now adopt a generative point of view, i.e., let us describe how to sample an $\boldsymbol{\Theta}$-valued random variable $\boldsymbol{\theta} = (k, \boldsymbol{\theta}_{1:k})$ from the corresponding probability distribution. We assume that a positive integer L is given, which represents the number of t-components in the model. Each t-component can be thought of as a "virtual component" that can generate zero or one (or several, in the case of the Poisson point process component that will be introduced later) observed component(s), according to some prescribed distribution on Θ.

To generate a sample $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, we first generate, independently for each of the L t-components, a binary indicator variable $\xi_l \in \{0,1\}$ drawn from the Bernoulli distribution $\mathscr{B}er(\pi_l)$, where $\xi_l = 1$ indicates that an observed component corresponding to the l^{th} t-component is actually present in $\boldsymbol{\theta}$. The actual number k of components in the generated samples

[3] See, however, the intensity plot provided in Section III (Figure 10) as an example of a BMA summary related to a component-specific parameter.

[4] Reporting additional models would improve the situation for BMS, in this example, but would not directly provide a "probability of presence" for each component, as our approach does.

To generate a sample from the VAPoRS model, do:

1) For $l = 1, \ldots, L$,
   generate $\xi_l \sim \mathscr{B}er(\pi_l)$.
2) Set $k = \sum_{l=1}^{L} \xi_l$.
3) For each l such that $\xi_l = 1$,
   generate $\tilde{\boldsymbol{\theta}}_l \sim \mathscr{N}(\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)$.
4) Randomly arrange the samples $\tilde{\boldsymbol{\theta}}_l$ generated at step 3 in a vector $\boldsymbol{\theta}_{1:k} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_k)$.

Figure 2: Steps to generate samples from the proposed parametric model.

is thus $k = \sum_{l=1}^{L} \xi_l$. The parameter $\pi_l \in (0, 1]$ will be called the *probability of presence* of the l^{th} component.

Second, given the vector of indicator variables $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_L)$, a Θ-valued random vector is generated for each t-component that is present (i.e., for each l such that $\xi_l = 1$). This random vector is generated according to some probability distribution associated to the t-component, that will be assumed, in this paper[5], to be a d-dimensional Gaussian distribution with mean $\boldsymbol{\mu}_l$ and covariance matrix $\boldsymbol{\Sigma}_l$. In order to achieve the required invariance with respect to component relabeling, the generated vectors are *randomly* arranged in a vector $\boldsymbol{\theta}_{1:k} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_k)$—i.e., a permutation of the k components that are present is drawn uniformly in the set of all permutations. The above mentioned steps to generate samples from the proposed parametric model are summarized in Figure 2.

Example. Assume that there are $L = 3$ univariate Gaussian t-components in the model, with means $\boldsymbol{\mu} = (0.63, 0.68, 0.73)$, variances $\boldsymbol{s}^2 = (0.01, 0.02, 0.01)$ and probabilities of presence $\boldsymbol{\pi} = (0.8, 0.3, 0.8)$. Figure 3 shows the pdf's of the three Gaussian t-components along with six random samples generated from the parametric model. Moreover, the kernel density estimates of 10 000 random samples generated from the parametric model are depicted in Figure 4 (a). It can be seen from both figures that the dimension of the generated samples varies from $k = 0$ to $k = L = 3$.

Let us look at the marginal posterior distributions of the sorted radial frequencies, depicted in the right panel of Figure 1. An important fraction of the posterior distribution is concentrated in compact bumps[6]; as will be seen later (see Section III-A), our Gaussian t-component will be effective at describing this fraction of the posterior. However, it can be observed on the plots related to the models with three and four sinusoidal components that the posterior also contains "diffuse" parts, meaning that a small fraction of the total probability mass is spread across a wide region (resulting, after sorting at fixed k, in heavy asymmetric tails for some marginal distributions). Another small fraction of the posterior distribution, not shown on Figure 1, is similarly scattered over the frequency axis for $k = 4$ and $k = 5$. It is clear that a model made of Gaussian t-components only will not be able to provide a parsimonious representation of this feature of the posterior f. Moreover, the observed components from the "diffuse" part, which would behave as *outliers* with respect to a model built with a small number of Gaussian t-components, could adversely influence the process of fitting the approximate posterior to the true posterior distribution of interest.

To overcome this issue, we propose to include in the model a "noise-like" Poisson Point Process (PPP; see, e.g., [35]) to account for the presence of outliers in the observed samples. For simplicity[7], we assume in this paper that Θ is bounded and that the PPP is homogeneous on Θ, with intensity $\lambda/|\Theta|$. We denote by $\xi_{L+1} \in \mathbb{N}$ the number of components generated by the PPP, which follows a Poisson distribution with mean λ, and keep using the notation $\boldsymbol{\xi}$ for the extended vector $(\xi_1, \ldots, \xi_L, \xi_{L+1})$. Finally, ξ_{L+1} random samples are generated uniformly on Θ and inserted *randomly* among the samples drawn from the Gaussian components. The actual number k of components in the generated sample is thus, now, redefined as $k = \sum_{l=1}^{L+1} \xi_l$. Figure 5 provides the directed acyclic graph of the model.

Example (continued). Setting Θ to the interval $(0, \pi)$ (as is the case in the sinusoid detection problem) and $\lambda = 0.5$, Figure 4 (b) shows the intensities of generated samples from the toy example's parametric model equipped with the Poisson point process component. It can be observed that the model equipped with PPP is capable of generating diffuse samples and thus, provides a better approximation to the distribution of the observed samples in practice (see, for example, Figure 1). Another interesting point that can be seen in Figure 4 (b) is that the model with PPP is able to generate samples with dimensions greater than the given number L of Gaussian t-components. This allows the model to deal with vector of observed samples of dimension greater than L.

Notations. We denote by $\boldsymbol{\eta}_l = (\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l, \pi_l)$ the vector of parameters for the l^{th} Gaussian t-component, and by $\boldsymbol{\eta} = (\boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_L, \lambda)$ the full vector of parameters of the model. We denote by $q_{\boldsymbol{\eta}}$ the pdf of the random variable $\boldsymbol{\theta} = (k, \boldsymbol{\theta}_{1:k})$ generated by the above construction. As a convenient abuse of notation, we will also use $q_{\boldsymbol{\eta}}$ in Sections II-C and II-D, to denote all joint, marginal and conditional distributions involving $\boldsymbol{\theta}$ and the auxiliary variables used for its generation.

[5]Any parametric family of d-variate distributions could be used at this point. As often in the literature [8, 20, 26, 27], the Gaussian distribution is chosen as a convenient mean of describing a "compact" and unimodal d-dimensional distribution, nothing more. Note that, because the Gaussian distribution is supported by $\mathbb{R}^d$, our parametric model is actually defined on a variable-dimensional space bigger than $\boldsymbol{\Theta}$ if Θ is a strict subset of $\mathbb{R}^d$.

[6]which are multimodal, because sorting the frequencies at fixed k does not properly resolve the label switching problem

[7]Homogeneity of the PPP component has been assumed for the sake of simplicity, but more elaborate (non-homogeneous) models are easily accommodated by our approach, if needed. A non homogeneous PPP is needed, in particular, if $|\Theta| = +\infty$; indeed, an homogeneous PPP on a such a set would generate an almost-surely infinite number of components. As an example: if one of the component-specific parameters is positive—an amplitude parameter in a positive mixture, say—then a lognormal distribution, a gamma distribution or uniform distribution on a bounded segment can be used to build a (bounded) non-homogeneous intensity.

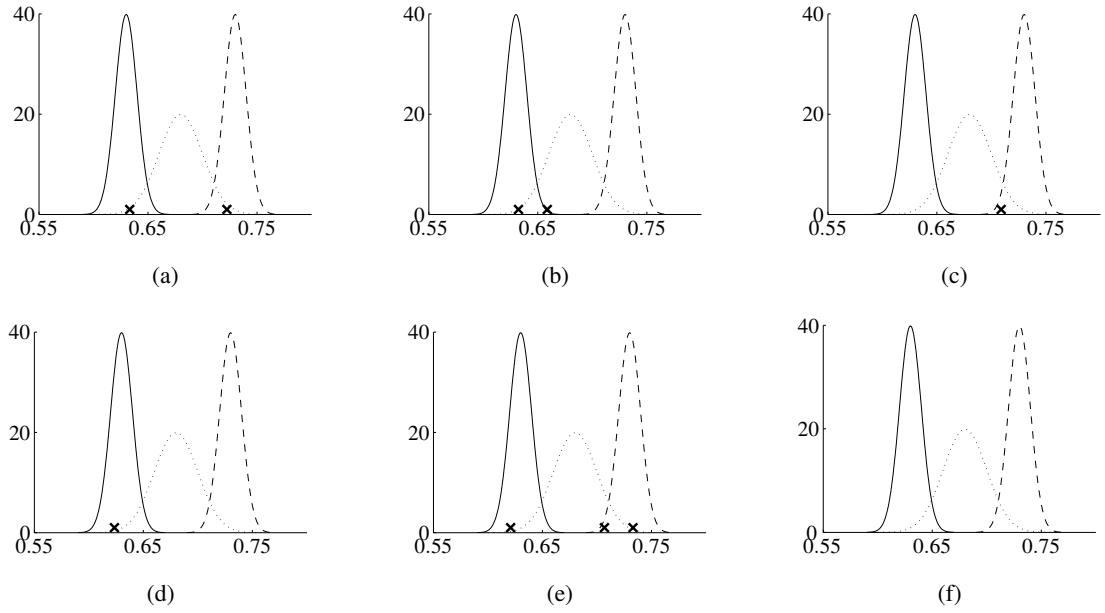


Figure 3: Generated samples from an example of the proposed variable-dimensional parametric model. There are $L = 3$ Gaussian t-components in the model with the means $\boldsymbol{\mu} = (0.63, 0.68, 0.73)^t$, the variances $\boldsymbol{s}^2 = (0.01, 0.02, 0.01)$ and the probabilities of presence $\boldsymbol{\pi} = (0.8, 0.3, 0.8)$. The $\times$ signs indicate the location of the generated random samples. (a) $\boldsymbol{\xi} = (1, 0, 1)$ and $\boldsymbol{\theta} = (2, (0.63, 0.72))$, (b) $\boldsymbol{\xi} = (1, 1, 0)$ and $\boldsymbol{\theta} = (2, (0.63, 0.66))$, (c) $\boldsymbol{\xi} = (0, 0, 1)$ and $\boldsymbol{\theta} = (1, (0.71))$, (d) $\boldsymbol{\xi} = (1, 0, 0)$ and $\boldsymbol{\theta} = (1, (0.62))$, (e) $\boldsymbol{\xi} = (1, 1, 1)$ and $\boldsymbol{\theta} = (3, (0.62, 0.70, 0.73))$, (f) $\boldsymbol{\xi} = (0, 0, 0)$ and $\boldsymbol{\theta} = (0, ())$.

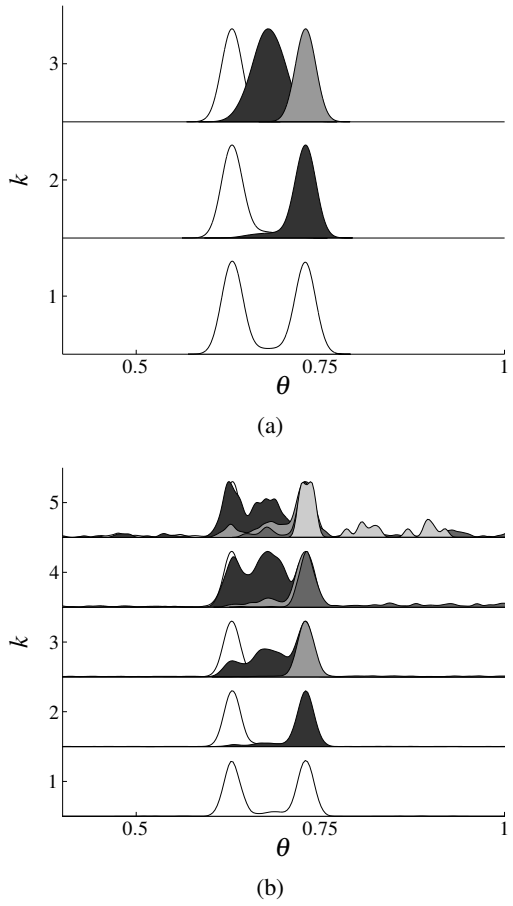


Figure 4: Kernel density estimates for 10 000 sorted random samples generated from the parametric model; (a) without PPP (b) with a Poisson point process component and setting $\lambda = 0.5$ and $\Theta = (0, \pi)$.

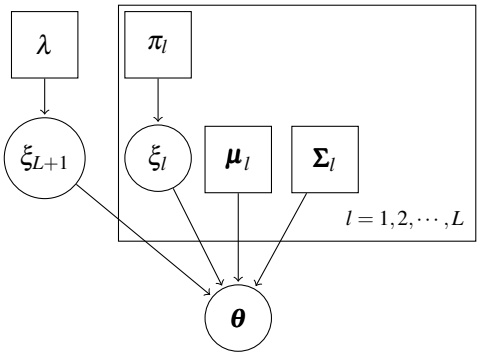


Figure 5: Proposed variable-dimensional parametric model in a generative viewpoint. Square nodes show fixed variables while circle nodes denote random variables. Note that $\boldsymbol{\eta} = (\boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_L, \lambda)$ and $\boldsymbol{\eta}_l = (\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l, \pi_l)$.

C. Distribution of the labeled samples

A random variable $\boldsymbol{\theta} = (k, (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_k))$ drawn from the density $q_{\boldsymbol{\eta}}$ can be thought of as an "unlabeled sample", since the label $l \in \mathscr{L} \triangleq \{1, \ldots, L+1\}$ of the component from which each $\boldsymbol{\theta}_j$ $(1 \leq j \leq k)$ originates cannot be recovered from $\boldsymbol{\theta}$ itself. Let us now introduce the (variable-dimensional) *allocation vector*

$$\mathbf{z} = (k, (z_1, \ldots, z_k)) \in \bigcup_{k \in \mathbb{N}} \{k\} \times \mathscr{L}^k,$$

which provides the missing piece of information: $z_j = l$ indicates that $\boldsymbol{\theta}_j$ originates from the l^{th} (Gaussian) t-component if $l \leq L$, while $z_j = L+1$ indicates that $\boldsymbol{\theta}_j$ originates from the point process t-component. We will refer to the pair $(\boldsymbol{\theta}, \mathbf{z})$ as a *labeled sample*. In the following, we will derive its joint

distribution $q_{\boldsymbol{\eta}}(\boldsymbol{\theta},\mathbf{z}) = q_{\boldsymbol{\eta}}(\boldsymbol{\theta} \mid \mathbf{z}) q_{\boldsymbol{\eta}}(\mathbf{z})$.

The distribution of the allocation vector $\mathbf{z}$ is

$$q_{\boldsymbol{\eta}}(\mathbf{z}) \;=\; q_{\boldsymbol{\eta}}(\mathbf{z} \mid \boldsymbol{\xi})\, q_{\boldsymbol{\eta}}(\boldsymbol{\xi}), \tag{2}$$

where $q_{\boldsymbol{\eta}}(\boldsymbol{\xi})$ is given by

$$q_{\boldsymbol{\eta}}(\boldsymbol{\xi}) \;=\; \frac{e^{-\lambda} \cdot \lambda^{\xi_{L+1}}}{\xi_{L+1}!} \prod_{l=1}^{L} \pi_l^{\xi_l} (1-\pi_l)^{(1-\xi_l)}. \tag{3}$$

Note that $\boldsymbol{\xi}$ is a deterministic function of $\mathbf{z}$: $\boldsymbol{\xi} = n(\mathbf{z})$, with $n_l(\mathbf{z}) = \sum_{j=1}^{k} \mathbb{1}_{z_j=l}$, for $1 \le l \le L+1$. To compute the first term of (2), remember that the points generated by the components of the parametric model are *randomly* arranged in $\boldsymbol{\theta}_{1:k}$. Therefore, for all $\boldsymbol{\xi} \in \{0,1\}^L \times \mathbb{N}$ such that $\sum_{l=1}^{L+1} \xi_l = k$,

$$q_{\boldsymbol{\eta}}(\mathbf{z} \mid \boldsymbol{\xi}) \;=\; \frac{\xi_{L+1}!}{k!}\, \mathbb{1}_{\boldsymbol{\xi}=n(\mathbf{z})}, \tag{4}$$

since two arrangements that differ only by the position of the points corresponding to the PPP give rise to the same allocation vector.

The conditional distribution $q_{\boldsymbol{\eta}}(\boldsymbol{\theta} \mid \mathbf{z})$ reads

$$q_{\boldsymbol{\eta}}(\boldsymbol{\theta} \mid \mathbf{z}) \;=\; \prod_{j=1}^{k} q_{\boldsymbol{\eta}}(\boldsymbol{\theta}_j \mid z_j), \tag{5}$$

where

$$q_{\boldsymbol{\eta}}(\boldsymbol{\theta}_j \mid z_j) \;=\; \begin{cases} \mathcal{N}\left(\boldsymbol{\theta}_j \mid \boldsymbol{\mu}_{z_j}, \boldsymbol{\Sigma}_{z_j}\right) & \text{if } z_j \le L, \\ \frac{1}{|\Theta|} & \text{if } z_j = L+1. \end{cases} \tag{6}$$

Therefore, from Equations (3) to (6), we have

$$\begin{aligned} q_{\boldsymbol{\eta}}(\boldsymbol{\theta},\mathbf{z}) \;=\; & \frac{e^{-\lambda}}{k!} \left(\frac{\lambda}{|\Theta|}\right)^{\xi_{L+1}} \prod_{\substack{1 \le j \le k \\ z_j \ne L+1}} \mathcal{N}\left(\boldsymbol{\theta}_j \mid \boldsymbol{\mu}_{z_j}, \boldsymbol{\Sigma}_{z_j}\right) \\ & \times \prod_{l=1}^{L} \pi_l^{\xi_l} (1-\pi_l)^{(1-\xi_l)} \quad \mathbb{1}_{\mathscr{Z}}(\mathbf{z}), \end{aligned} \tag{7}$$

where $(\xi_1, \ldots, \xi_{L+1}) = n(\mathbf{z})$ and $\mathscr{Z}$ is the set of all allocation vectors (i.e., the set of all $\mathbf{z} \in \cup_{k \in \mathbb{N}} \{k\} \times \mathscr{L}^k$ such that $\xi_l = n_l(\mathbf{z}) \in \{0,1\}$, for $1 \le l \le L$).

D. Estimating the model parameters

We propose to fit the parametric distribution $q_{\boldsymbol{\eta}}$ to the posterior f of interest by minimizing a divergence measure[8] from f to $q_{\boldsymbol{\eta}}$. We use the KL divergence as a divergence measure in this paper, though other divergence measures can be used as well (see, e.g., [1, Chapter 2]).

Denoting the KL divergence from f to $q_{\boldsymbol{\eta}}$ by $D_{KL}(f(\boldsymbol{\theta}) \| q_{\boldsymbol{\eta}}(\boldsymbol{\theta}))$, we define the criterion to be minimized as

$$\mathscr{J}(\boldsymbol{\eta}) \;\triangleq\; D_{KL}(f(\boldsymbol{\theta}) \| q_{\boldsymbol{\eta}}(\boldsymbol{\theta})) \;=\; \int_{\boldsymbol{\Theta}} f(\boldsymbol{\theta}) \log \frac{f(\boldsymbol{\theta})}{q_{\boldsymbol{\eta}}(\boldsymbol{\theta})}\, \mathrm{d}\boldsymbol{\theta}.$$

[8] It would also be possible to assign prior distributions over the unknown parameters $\boldsymbol{\eta}$ and study their posterior distributions (for example, using an MCMC sampler with the latent variable $\mathbf{z}$ added to the state of the chain, in the spirit of the "data augmentation" algorithm [36]). This would, however, leave the label-switching issue unsolved (because of the invariance of $q_{\boldsymbol{\eta}}$ to permutations of its components).

At the $(r+1)^{\text{th}}$ iteration, do:

(S-step) Draw allocation vectors $\mathbf{z}^{(i,r+1)}$, $1 \le i \le M$, using an IMH step with target $q_{\hat{\boldsymbol{\eta}}^{(r)}}(\cdot \mid \boldsymbol{\theta}^{(i)})$.

(E-step) Construct the pseudo-completed log-likelihood $\widehat{\mathscr{J}_M}(\boldsymbol{\eta}) \;=\; -\sum_{i=1}^{M} \log\big(q_{\boldsymbol{\eta}}(\boldsymbol{\theta}^{(i)}, \mathbf{z}^{(i,r+1)})\big)$.

(M-step) Estimate $\hat{\boldsymbol{\eta}}^{(r+1)}$ such that $\hat{\boldsymbol{\eta}}^{(r+1)} \;=\; \operatorname{argmin}_{\boldsymbol{\eta}}\, \widehat{\mathscr{J}_M}(\boldsymbol{\eta})$.

Figure 6: Pseudo-code for the proposed SEM-type algorithm

Using samples $\boldsymbol{\theta}^{(i)}$, $i = 1, \ldots, M$, generated by the RJ-MCMC sampler, this criterion can be approximated as

$$\mathscr{J}(\boldsymbol{\eta}) \;\simeq\; -\frac{1}{M} \sum_{i=1}^{M} \log\left(q_{\boldsymbol{\eta}}(\boldsymbol{\theta}^{(i)})\right) + C, \tag{8}$$

where $C = \int f(\boldsymbol{\theta}) \log f(\boldsymbol{\theta})\, \mathrm{d}\boldsymbol{\theta}$ is a constant that does not depend on $\boldsymbol{\eta}$. One should note that minimizing the right-hand side of (8) amounts to choosing

$$\hat{\boldsymbol{\eta}} \;=\; \operatorname{argmax}_{\boldsymbol{\eta}} \sum_{i=1}^{M} \log\left(q_{\boldsymbol{\eta}}(\boldsymbol{\theta}^{(i)})\right). \tag{9}$$

To estimate the model parameters $\boldsymbol{\eta}$, one of the extensively used algorithms for Maximum Likelihood (ML) parameter estimation in latent variable models is the EM algorithm proposed by [37]. However, it turns out that the EM algorithm, which has been used in similar works [8, 20, 26], is not appropriate for solving this problem, as computing the expectation in the E-step is intricate. More explicitly, in our problem the computational burden of the summation in the E-step over the set of all possible allocation vectors $\mathbf{z}$ increases very rapidly with both L and k. In fact, even for moderate values of L and k, say, $L = 15$ and $k = 10$, the summation is far too expensive to compute as it involves $\sum_{m=0}^{k} \frac{L!}{(L-k+m)!} \approx 1.3\,10^{10}$ terms.

In this paper, we propose to use the SEM algorithm [31–33], a variation of the EM algorithm in which the E-step is substituted with stochastic simulation of the latent variables from their conditional posterior distributions given the previous estimates of the unknown parameters. In other words, at the iteration $r+1$ of the SEM algorithm, denoting the estimated parameters at iteration r by $\hat{\boldsymbol{\eta}}^{(r)}$, for $i = 1, \ldots, M$, the allocation vectors $\mathbf{z}^{(i)}$ are drawn from $q_{\hat{\boldsymbol{\eta}}^{(r)}}(\cdot \mid \boldsymbol{\theta}^{(i)})$. This step is called the Stochastic (S)-step. Then, these random samples are used to construct the so-called pseudo-completed log-likelihood.

Exact sampling from $q_{\hat{\boldsymbol{\eta}}^{(r)}}(\cdot \mid \boldsymbol{\theta}^{(i)})$, as required by the S-step of the SEM-type algorithm, is unfortunately not feasible—not even using the accept-reject algorithm, due to the heavily combinatorial expression of the normalizing constant $q_{\hat{\boldsymbol{\eta}}^{(r)}}(\boldsymbol{\theta}^{(i)})$. Instead, since

$$q_{\hat{\boldsymbol{\eta}}^{(r)}}(\mathbf{z}^{(i)} \mid \boldsymbol{\theta}^{(i)}) \;\propto\; q_{\hat{\boldsymbol{\eta}}^{(r)}}(\boldsymbol{\theta}^{(i)}, \mathbf{z}^{(i)})$$

can be computed up to a normalizing constant, we choose to use an Independent Metropolis-Hasting (IMH) step with $q_{\hat{\boldsymbol{\eta}}^{(r)}}(\mathbf{z}^{(i)} \mid \boldsymbol{\theta}^{(i)})$ as its stationary distribution; see [1, Algorithm 2.2] for more details.

The proposed SEM-type algorithm is summarized in Figure 6. Explicit expressions for the M-step are easily obtained from Equation (7) (see [1, page 76]). The computational cost of one iteration is of the order $O\left(ML^2\right)$ for a given d, the most expensive part being the IMH algorithm used in the S-step.

Remark 1. Convergence results of the SEM algorithm in the general form are provided by [33] and, in the particular example of mixture analysis problems, by [38]. Unfortunately, the assumptions in [33, 38] do not hold in the problem we are dealing with as, 1) the observed samples $\boldsymbol{\theta}^{(i)}$ are correlated, owing to the fact that they are generated from the true posterior distribution using some MCMC methods, e.g., the RJ-MCMC sampler; 2) an IMH sampler is used to draw $\mathbf{z}^{(i)}$ from the conditional posterior distribution. Nevertheless, empirical evidence of the "good" convergence properties of the SEM-type algorithm we proposed will be provided in the next two sections.

E. Robustified algorithm

Preliminary experiments with the SEM-type algorithm described in Figure 6 were not satisfactory, because the sample mean and (co)variance estimates in the M-step, obtained from minimizing the KL divergence from the posterior distribution f to the parametric model $q_{\boldsymbol{\eta}}$, still suffer from sensitivity to the outliers in the observed samples, even after including the Poisson point process component. Similar robustness concerns are widespread in the clustering literature; see, e.g., [39] and references therein.

As a workaround, we propose to use robust estimates [40] of the means and (co)variances of Gaussian distributions instead of the empirical means and (co)variances in the M-step[9]. For example, in the case of univariate Gaussian distributions, we use the median and 1.349 times the interquartile range as robust estimators of the mean and standard deviation, respectively. See [1, Section 2.5] for more discussion of this robustness issue, including an alternative solution using the (robust) α-divergence of [41] instead of the KL divergence.

Remark 2. Note that the robustification introduced in this section does not render the PPP component useless. Indeed, removing it would result (experiments not shown) in the introduction of one or several large-variance small-probability Gaussian components in the model, to account for the diffuse part of the posterior discussed in Section II-B, and would also impose to choose L at least as large as the largest value of $k^{(i)}$ in the MCMC samples.

[9]As an alternative to the simple plug-in method proposed here, robustification could be achieved by using heavy-tailed distributions (e.g., Student distributions) instead of Gaussian distributions in the parametric model. Note that this approach, more elegant perhaps, makes the M-step more complicated since (at least a few steps of) an optimization algorithm must be used to find the new value of the parameters.

III. Illustrative examples

In this section, we will investigate the capability of VAPoRS for summarizing variable-dimensional posterior distributions using two signal decomposition examples; 1) joint Bayesian detection and estimation of sinusoids in white Gaussian noise [12] and 2) joint Bayesian detection and estimation of astrophysical particles in the Auger project [17, 18]; see [1, Chapters 3 and 4] for more results and discussion. We emphasize again that the output of the trans-dimensional Monte Carlo sampler, e.g, the. RJ-MCMC sampler in this paper, is considered as the observed data for VAPoRS.

A. Joint Bayesian detection and estimation of sinusoids in white Gaussian noise

Let us consider the problem of detection and estimation of sinusoidal components introduced in Section I-B where the unknown parameters are the number k of components, the component-specific parameters $(a_{c,j}, a_{s,j}, \omega_j)$, $1 \leq j \leq k$, and the noise variance σ^2. Since the amplitudes and the noise variance can be analytically integrated out, we focus on summarizing the joint posterior distribution $p(k, \boldsymbol{\omega}_{1:k} \mid \mathbf{y})$ of the form illustrated in Figure 1. Therefore, we assume that the proposed parametric model introduced in Section II-B consists of univariate Gaussian components, with means μ_l, variances s_l^2, and probabilities of presence π_l, $1 \leq l \leq L$, to be estimated. Moreover, the space of component-specific parameters is $\Theta = (0, \pi) \subset \mathbb{R}$.

Before launching VAPoRS, we need first to initialize the parametric model. It is natural to deduce the number L of Gaussian components from the posterior distribution of k. Here, we set it to the 90$^{\text{th}}$ percentile of $p(k \mid \mathbf{y})$ to keep all the probable models in the play. To initialize the Gaussian components' parameters, i.e., μ_l and s_l^2, $1 \leq l \leq L$, we used the robust estimates of the means and variances of the marginal posterior distributions of the sorted radial frequencies given $k = L$. Finally, we set $\pi_l = 0.9$, for $1 \leq l \leq L$, and $\lambda = 0.1$.

We ran the "robustified" stochastic algorithm introduced in Section II on the specific example shown in Figure 1, for 100 iterations, with $L = 3$ Gaussian components (note that the posterior probability of $\{k \leq 3\}$ is approximately 90.3%). To assess the convergence of VAPoRS, Figure 7 illustrates the evolution of the model parameters $\boldsymbol{\eta}$ together with the criterion $\mathscr{J}$. Two substantial facts showing the convergence of VAPoRS can be deduced from this figure: first, the decreasing behavior of the criterion $\widehat{\mathscr{J}_M}$, which is almost constant after the 10$^{\text{th}}$ iteration; second, the convergence of the parameters of the parametric model, particularly the means μ_l and probabilities of presence π_l, $1 \leq l \leq L$, even though we used a naive initialization procedure. Indeed after the 40$^{\text{th}}$ iteration there is no significant move in the parameter estimates.

As discussed in Section I, one of the main objectives of the algorithm we proposed is to solve the label-switching issue in a trans-dimensional setting. Figures 8 shows the histograms of the labeled samples, i.e., $(\boldsymbol{\theta}^{(i)}, \mathbf{z}^{(i)})$, with $i = 1, \ldots, M$, along with the pdf's of the estimated Gaussian components (black solid line). Moreover, the summaries provided by VAPoRS for each component are presented in its corresponding panel. We

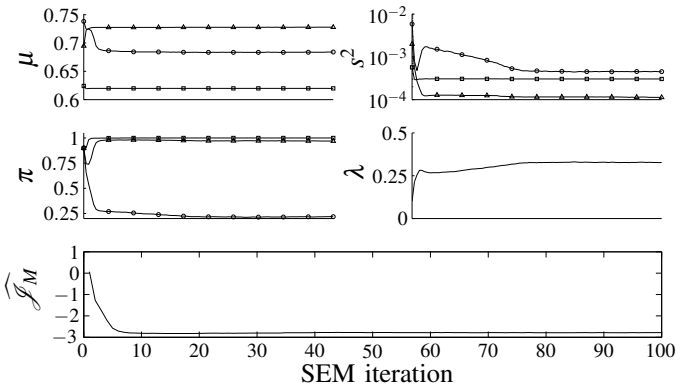


Figure 7: Evolution of the model parameters along with the criterion $\widehat{\mathscr{J}_M}$ defined in (8) using 100 iterations of VAPoRS with $L = 3$ on the RJ-MCMC output samples shown in Figure 1.

Comp.	μ	s	π	μ_{BMS}	s_{BMS}
1	0.62	0.017	1	0.62	0.016
2	0.68	0.021	0.22	—	—
3	0.73	0.011	0.97	0.73	0.012

Table I: Summaries of the variable-dimensional posterior distribution shown in Figure 1; VAPoRS vs. the BMS approach.

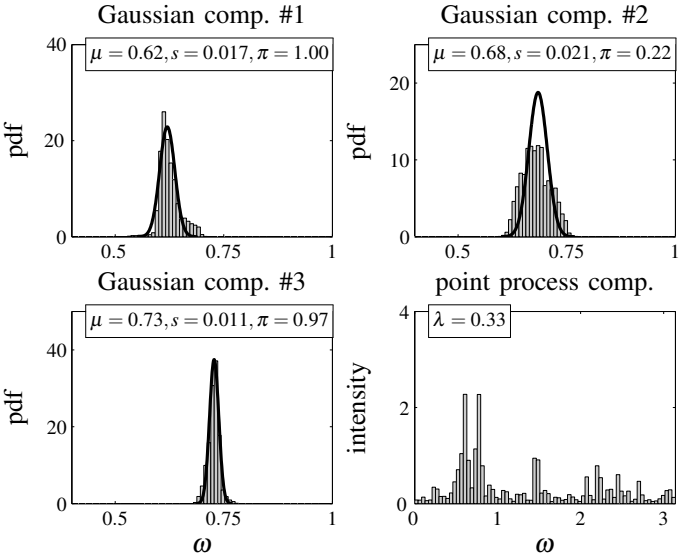


Figure 8: Normalized histogram of the labeled samples, that is, the samples allocated to the Gaussian and Poisson point process components, versus the pdf's of estimated Gaussian components in the model (black solid line) using VAPoRS on the sinusoid detection example. The estimated parameters of each component are presented in the corresponding panel.

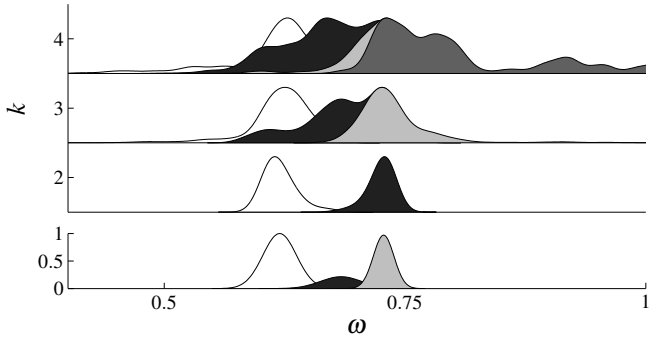


Figure 9: Posterior distribution of the sorted radial frequencies $\boldsymbol{\omega}_{1:k}$ given k (top) and normalized pdf of the fitted Gaussian components (bottom).

used the average of the last 50 SEM iterations as parameter estimates, as recommended in the SEM literature; see, for example, [32, 33]. Comparing the unimodal distributions of the labeled samples with the ones of the posterior distributions of the sorted radial frequencies given $k = 3$ shown in Figure 1, which are highly multimodal, reveals the capability of VAPoRS in solving label-switching in a variable-dimensional setting.

Looking at the bottom right panel of Figure 8, the role of the point process component in capturing the outliers in the observed samples, which cannot be described by the Gaussian components, becomes clearer[10]. Note that, without the point process component, these outliers would be allocated to the Gaussian components and would, consequently, induce a significant deterioration of the parameter estimates.

Table I presents the summaries provided using VAPoRS along with the ones obtained using the BMS approach. Contrary to the BMS approach, VAPoRS has enabled us to benefit from the information of all probable models to give summaries about the middle harder to detect component. Turning to the results of VAPoRS, it can be seen that the estimated means are compatible with the true radial frequencies. Furthermore, the estimated probabilities of presence are consistent with uncertainty of them in the variable-dimensional posterior shown in Figure 1.

To observe better the "goodness-of-fit" of the estimated Gaussian components, the bottom panel of Figure 9 depicts their normalized densities[11], under the posterior distributions of the sorted radial frequencies given k. This figure can be used to validate the coherency of the estimated summaries with the information in the variable-dimensional posterior distribution. It can be seen from the figures that the shape of the pdf's of the estimated Gaussian components are coherent in both the location and dispersion with the ones of the posterior of the sorted radial frequencies.

It is also useful, for the validation of the estimated parametric model, to compare the intensity [see, e.g., 35]

$$h = \sum_{l=1}^{L} \hat{\pi}_l \cdot \mathcal{N}(\cdot \,|\, \hat{\boldsymbol{\mu}}_l, \hat{\boldsymbol{\Sigma}}_l), \tag{10}$$

of the corresponding point process on Θ, where we ignore the point process component, with an histogram estimator of the

[10]The presence of two peaks on the left indicates that additional Gaussian t-components could be added here to get a more accurate approximation of the posterior. See Section III-B for a discussion of the interpretation of the estimated intensity as a residual.

[11]To obtain the normalized densities, first, we normalized the estimated pdf's to have their maximum equal to one. Then, we multiplied the estimated probability of presence of each Gaussian component to its corresponding normalized estimated pdf. Thus, the maximum of each normalized density is equal to the corresponding estimated probability of presence.

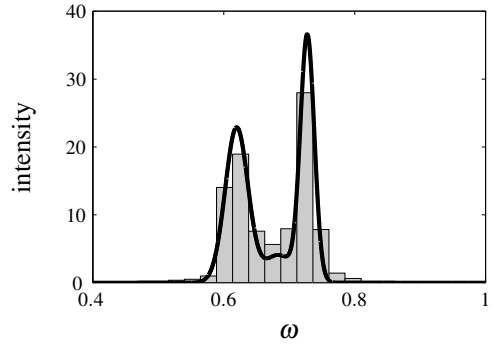


Figure 10: Comparison of the intensity of the fitted parametric model obtained with VAPoRS (black line) with an histogram estimator of the intensity (light gray).

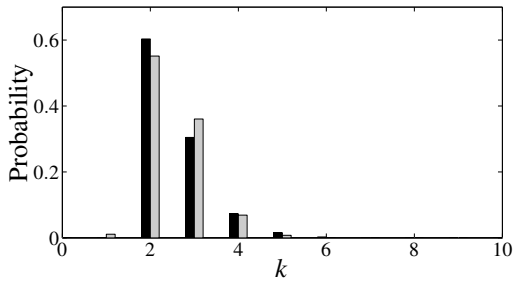


Figure 11: Posterior distribution of the number k of number of components (black) and its approximated version (gray) obtained from the fitted model.

intensity (which averages over MCMC samples from all model sizes, and is therefore an example of using the BMA approach; see [1, Chapter 2] for more information). Figure 10 shows such a figure for the specific example of this section where the solid black line indicates the intensity of the estimated parametric model. These figures also indicate the "goodness-of-fit" of the fitted approximate posterior and the true one.

Finally, to validate both the estimated probabilities of presence of the Gaussian components and the mean parameter λ of the Poisson point process component, Figure 11 illustrates the posterior distribution of the number k of components together with its approximated versions using VAPoRS. It can be seen from the figure that, in this example, VAPoRS successfully captured the information provided in the true posterior of the number k of components.

Similar results have been obtained for a wide range of configurations (with varying number of sinusoids, radial frequencies, signal-to-noise ratios, ...; not shown in the paper). In particular, an example with $k = 30$ sinusoids is discussed in [1, Section 3.3.4]; the computation time of our Matlab/C implementation, on this example where $L = 30$ and $M = 20\,000$, is approximately 15 seconds per SEM iteration (on a laptop with an Intel Core i5 M540 running at 2.53 GHz and 4 GB of RAM).

B. Joint Bayesian detection and estimation of astrophysical particles in the Auger project

As the second illustrative example, we show results on a signal decomposition problem encountered in the international astrophysics collaboration called Auger [17, 18]. The Auger project is aimed at studying ultra-high energy cosmic rays, with energies in order of 10^{19}eV, the most energetic particles found so far in the universe. The long-term objective of this project is to study the nature of those ultra-high energy particles and determine their origin in the universe.

These particles are not observed directly. When they collide the earth's atmosphere, a host of secondary particles are generated, some of which, mostly muons, finally reach the ground. To detect these muons, the Pierre Auger Cosmic Ray Observatory was built, which consists of two independent detectors: an array of Surface Detectors (SD) and a number of Fluorescence Detectors (FD). There are in total 1600 SD tanks, each separated from its neighbors by 1.5 kilometers, covering a surface of about 3000 km^2.

The number of muons and their arrival times can be used as indications of both the chemical composition and the origin of the primary particles (see [17, 18] for more information). Here, we concentrate on the signal decomposition problem, where the goal is to count the number of muons and estimate their individual parameters from the signals observed by SD detectors; while noting that, to investigate the characteristics of the primary particles, one needs to use information obtained from a few tens of SD's.

This problem has been addressed by [9, 42, 43] in a Bayesian framework, in which they developed an RJ-MCMC sampler to jointly count the muons and estimate their parameters. They run thousands of iterations of the RJ-MCMC sampler on the signals captured by each individual SD tank and, then, aggregate all the samples in a secondary analysis step to make inference on the characteristics of the primary particle [44, Chapter 7]. To make the whole process more efficient and easier to interpret, it would be preferable to "digest" the MCMC samples associated to each SD tank, and to convey only the resulting summaries to the next inference level. This would also make it much cheaper to store the data required for the second stage of inference, e.g., for traceability or future studies. Therefore, an algorithm that faithfully summarizes the RJ-MCMC samples is needed. In this section, we first briefly describe the problem and then use VAPoRS to relabel and summarize variable-dimensional output samples of the RJ-MCMC sampler developed by [9, 42, 43].

When a muon crosses a SD tank, it generates photoelectrons (PE's) along its track that are, then, captured by detectors and create a discrete observed signal. We denote the vector of observed signal by $\mathbf{n} = (n_1, \ldots, n_N) \in \mathbb{N}^N$, where the element n_i indicates the number of PE's deposited by the muons in the time interval

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

where t_0 is the absolute starting time of the signal and $t_\Delta = 25$ ns is the signal resolution (length of one bin).

Each muon has two component-specific parameters, namely, the arrival time t_μ and the signal amplitude a_μ. The absorption process of the photons generated by a muon is modeled by a non-homogeneous Poisson point process with intensity [43, Section 2.2]

$$h(t \,|\, a_\mu, t_\mu) = a_\mu \, p_{\tau, t_d}(t - t_\mu), \tag{11}$$

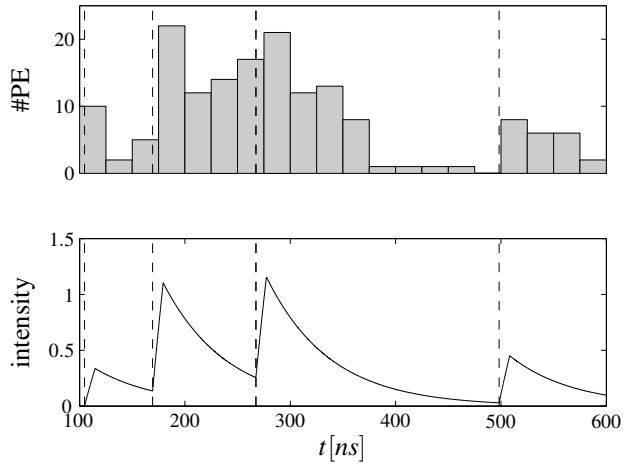


Figure 12: (top) Simulated observed signal $\mathbf{n}$. (bottom) Intensity of the model $h(t\,|\,\boldsymbol{a}_\mu, \boldsymbol{t}_\mu)$ defined in (11). There are $k = 5$ muons in the signal with the true arrival times, i.e., $\boldsymbol{t}_\mu = (105, 169, 267, 268, 498)$, indicated by vertical dashed lines.

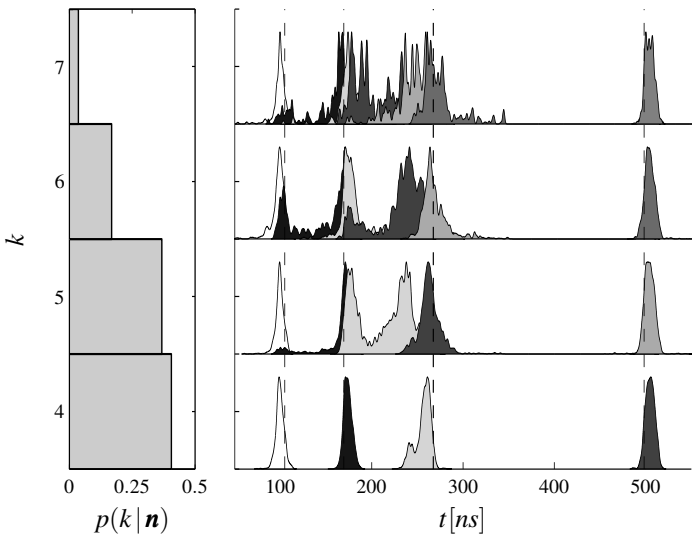


Figure 13: Posterior distributions of the number k of muons (left) and the sorted arrival times, $\boldsymbol{t}_\mu$, given k (right) constructed using 60 000 RJ-MCMC output samples after discarding the burn-in period. The true number of components is five. The vertical dashed lines in the right figure locate the arrival times.

where $p_{\tau,t_d}(t)$ is the time response distribution, t_d is the rise-time and τ is the exponential decay (both measured in ns); see Figure 12 (bottom) for such exponential shape intensities. Then, the expected number of PE's in the bin i is obtained by integrating the intensity (11) in the corresponding bin:

$$\bar{n}_i(a_\mu, t_\mu) \;=\; a_\mu \int_{t_{i-1}}^{t_i} p_{\tau,t_d}(t - t_\mu)\mathrm{d}t. \tag{12}$$

Conditioning on the number k of muons and the vector of parameters $\boldsymbol{t}_\mu = (t_{\mu,1}, \ldots, t_{\mu,k})$ and $\boldsymbol{a}_\mu = (a_{\mu,1}, \ldots, a_{\mu,k})$, and assuming that the number of PE's in each bin are independent, the likelihood is written as

$$p(\mathbf{n}\,|\,k, \boldsymbol{t}_\mu, \boldsymbol{a}_\mu) \;=\; \prod_{i=1}^{N} p(n_i\,|\,\bar{n}_i(k, \boldsymbol{a}_\mu, \boldsymbol{t}_\mu)), \tag{13}$$

where $p(n_i\,|\,\bar{n}_i(k, \boldsymbol{a}_\mu, \boldsymbol{t}_\mu))$ is a Poisson distribution with the mean $\bar{n}_i(k, \boldsymbol{a}_\mu, \boldsymbol{t}_\mu)$. Then, assuming independence of the muons, the expected number of PE's in the i^{th} bin, i.e., $\bar{n}_i(k, \boldsymbol{a}_\mu, \boldsymbol{t}_\mu)$, given k, $\boldsymbol{t}_\mu$, and $\boldsymbol{a}_\mu$ becomes

$$\bar{n}_i(k, \boldsymbol{a}_\mu, \boldsymbol{t}_\mu) \;=\; \sum_{j=1}^{k} \bar{n}_i(a_{\mu,j}, t_{\mu,j}). \tag{14}$$

We will now illustrate the performance of VAPoRS on a simulated PE counting signal (see [1, Chapter 4] for results on two other simulated experiments). The observed signal of the illustrative example considered here consists of five muons located at $\boldsymbol{t}_\mu = (105, 169, 267, 268, 498)$ (see Figure 12). The posterior distributions of the number k of muons and sorted arrival times are shown in Figure 13. Note that, in this example, there are two muons with almost equal arrival times, i.e., the third and fourth muons.

Using the BMS approach, the model with four muons would be selected ($p(k = 4\,|\,\mathbf{n}) = 0.4$), although $\mathscr{M}_5$ has an almost identical posterior probability of 0.38. Moreover, observe that the marginal posterior of the arrival time of the third component is bimodal under both $\mathscr{M}_4$ and, more significantly so, $\mathscr{M}_5$. We ran VAPoRS with $L = 6$ Gaussian components on the RJ-MCMC output samples shown in Figure 13 (note that $p(k \leq 6\,|\,\mathbf{n}) = 0.94$).

Figure 14 shows the histogram of the labeled samples and the estimated parameters of the components. From the figure, it can be seen that the bimodality effects caused by label-switching exhibited in Figure 13 is removed completely and the estimated Gaussian components are unimodal and enjoy reasonable variances. In the presented summary, there are four muons with high probabilities of presence corresponding to the ones shown in the bottom row of Figure 13. There are also two other muons with comparatively low probabilities of presence.

In fact, the samples allocated to the point process component shown the bottom row of Figure 14 can be regarded as the residuals of the fitted model, that is, the observed samples which the L Gaussian components in $q_{\boldsymbol{\eta}}$ have not been able to describe. These residuals can be used, as usual in statistics, as a tool for goodness-of-fit diagnostics and model choice.

Figure 15 illustrates the histograms of the residuals of the fitted model for different values of $L \in \{3, 4, 6, 8\}$. It can be seen from the top left panel of Figure 15 that the distribution of the residuals corresponding to the case where $L = 3$ contains a few "significant" peaks. The peaks are gradually removed by adding Gaussian components. When $L = 4$, a component is added at $t_\mu = 261$ that captures samples distributed around the most significant peak of the top left panel of Figure 15. However, there still exist a few peaks, particularly around $t_\mu = 173$ which are captured when $L \geq 6$. However, the distribution of residuals for the case of $L = 6$ and $L = 8$ do not differ significantly. Note the decrease of value of $\hat{\lambda}$ by increasing L.

Figure 16 compares the normalized intensities of the estimated Gaussian components for $6 \leq L \leq 9$. When moving from $L = 6$ to $L = 9$, the six Gaussian components that are estimated in the case with $L = 6$ always exist, but additional Gaussian components with very low probabilities of presence are added to the summary, which improve the fit but does not

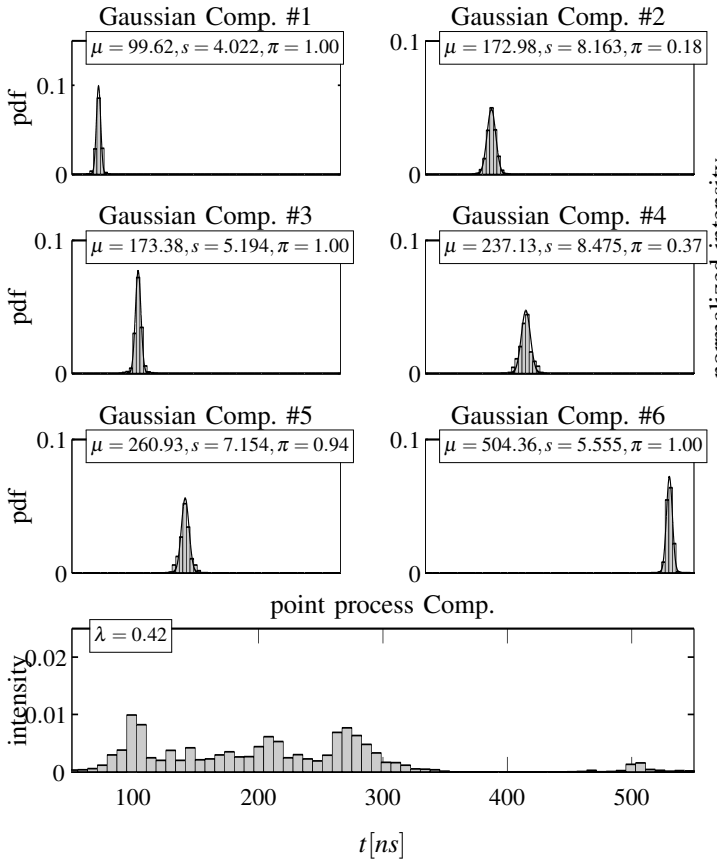


Figure 14: Normalized histogram of the labeled samples along with the pdf's of estimated Gaussian components in the model (black solid line) using VAPoRS with $L = 6$ on the variable-dimensional posterior shown in Figure 13. The estimated parameters of each component are presented in the corresponding panel.

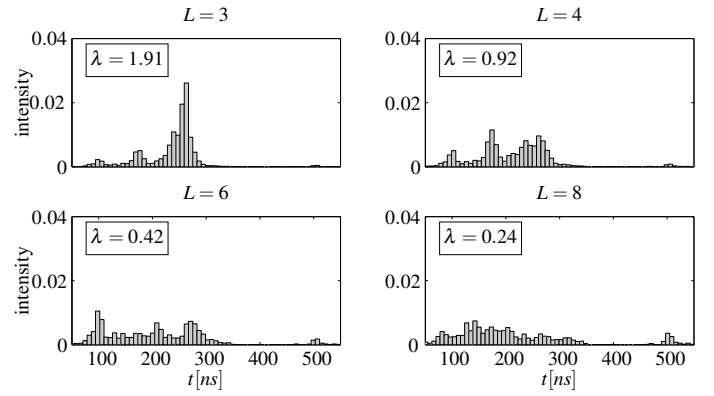


Figure 15: Normalized histograms of the residuals of the fitted model using VAPoRS with different values of $L = \{3, 4, 6, 8\}$.

change much the final inference.

IV. Monte Carlo experiment

The examples of Section III have illustrated the capability of VAPoRS to relabel and summarize variable-dimensional posterior distributions encountered in two signal decomposition problems. In order to confirm these findings, we will now investigate more systematically, by means of a Monte Carlo simulation experiment, how faithfully the approximate posterior distribution preserves certain features of the true posterior distribution.

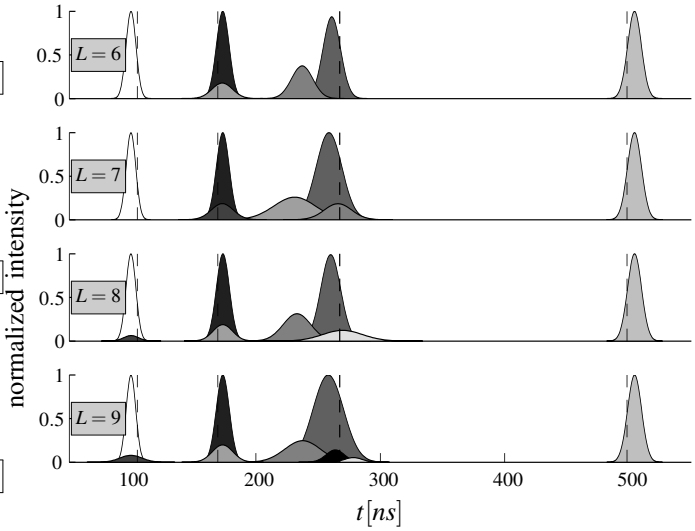


Figure 16: Normalized pdf's of the fitted Gaussian components using VAPoRS with different values of $6 \leq L \leq 9$.

One hundred realizations of the sinusoid detection experiment described in Section I-B (see Figure 1) were simulated and analyzed using the same RJ-MCMC sampler as before. The number of RJ-MCMC iterations was set to 100 000 and the first 20 000 samples were discarded as the burn-in period. Then, the samples were thinned to one every fifth. To initialize the parametric model $q_{\boldsymbol{\eta}}$ in a systematic fashion, we set L to the largest k such that its posterior probability is not less than 0.05. Then, during the process of the SEM-type algorithms, if sufficient number of samples, say, 10, is not allocated to a Gaussian component (or, equivalently, its probability of presence fades to zero), we will remove it from the parametric model and decrease L by one. Using this approach generally results in approximate posterior distributions which are "richer" than those provided by the BMS approach[12], in the sense that $L \geq k^{\mathrm{MAP}}$, where $k^{\mathrm{MAP}} = \mathrm{argmax}_k \; p(k|\mathbf{y})$. To initialize the Gaussian components' parameters, i.e., the means μ_l and variances s_l^2, we used as previously robust estimates of the mean and variances of the posterior distributions of sorted radial frequencies given $k = L$.

Figure 17 compares various features of the fitted approximate posterior distribution $q_{\hat{\boldsymbol{\eta}}}$, obtained using 100 iterations of VAPoRS, with the corresponding features of the true variable-dimensional posterior distribution. These features are described in the rest of this section.

The scatter plots shown in panels (a), (b), and (c) compare the posterior distribution of the number k of components, i.e., $p(k|\mathbf{y})$, with its approximated version, denoted here by $\hat{p}(k|\mathbf{y})$, in 100 runs. We only show the posterior probabilities of $k = 2$ and $k = 3$ in this comparison, as the other probabilities were

[12]Later, in a post-processing step, since each Gaussian component has been endowed with a probability of presence π_l, with $1 \leq l \leq L$, one can decide to discard the ones with π_l smaller than a certain threshold; see [1, Section 3.4.3] for more discussion about this idea.

close to zero. The digits situated on the right of the points in the panel (a) indicate the number of occurrence of the corresponding event in 100 runs and $\hat{k}^{\mathrm{MAP}} = \mathrm{argmax}_k\, \hat{p}(k|\mathbf{y})$. It can be seen from these three panels that the information in $p(k|\mathbf{y})$ was well preserved by the approximated posterior distributions.

Next we compare the performance of VAPoRS with the one of the "direct" BMA approach (i.e., using the RJ-MCMC samples directly, and not the VAPoRS posterior) in reconstructing the noiseless signal $\mathbf{y}_0 = \mathbf{D}\,\mathbf{a}_{1:k}$. To this end, the estimated reconstructed noiseless signal is defined as

$$\begin{aligned}\hat{\mathbf{y}}_0 \;&=\; \mathbb{E}(\mathbf{y}_0\,|\,\mathbf{y}) \\ &=\; \sum_{k\in\mathbb{N}} \int_{\Theta^k} \mathbb{E}(\mathbf{y}_0\,|\,k,\boldsymbol{\theta}_{1:k},\mathbf{y})\, p(k,\boldsymbol{\theta}_{1:k}\,|\,\mathbf{y})\,\mathrm{d}\boldsymbol{\theta}_{1:k}. \qquad (15)\end{aligned}$$

In the direct BMA approach, using the samples generated with the RJ-MCMC sampler, the above integral is approximated by

$$\hat{\mathbf{y}}_0^{\mathrm{BMA}} \;=\; \frac{1}{M}\sum_{i=1}^{M} \mathbf{D}^{(i)}\,\hat{\boldsymbol{a}}^{(i)}_{1:k^{(i)}},$$

where $\mathbf{D}^{(i)}$ is the design matrix of the i^{th} vector of the sampled radial frequencies $\boldsymbol{\omega}^{(i)}_{1:k^{(i)}}$ and $\hat{\boldsymbol{a}}^{(i)}_{1:k^{(i)}}$ is the posterior mean of the amplitudes given $\boldsymbol{\omega}^{(i)}_{1:k^{(i)}}$ and its hyperparameters. To reconstruct the noiseless signal from the approximate posterior $q_{\hat{\boldsymbol{\eta}}}$ using VAPoRS, one can generate R pairs of samples $(k^{(r)}, \boldsymbol{\omega}^{(r)}_{1:k^{(r)}})$ as explained in Section II-B and set

$$\hat{\mathbf{y}}_0^{\mathrm{VAPoRS}} = \frac{1}{R}\sum_{r=1}^{R} \mathbf{D}^{(r)}\,\hat{\boldsymbol{a}}^{(r)}_{1:k^{(r)}}.$$

Panel (d) compares the normalized reconstruction errors when using VAPoRS with the ones of the direct BMA approach in dB, defined as

$$10\log_{10}\left(\frac{\|\hat{\mathbf{y}}_0-\mathbf{y}_0\|^2}{\|\mathbf{y}_0\|^2}\right), \qquad (16)$$

where $\|\cdot\|$ is the L_2-norm and we set $\hat{\mathbf{y}}_0 = \hat{\mathbf{y}}_0^{\mathrm{BMA}}$ and $\hat{\mathbf{y}}_0 = \hat{\mathbf{y}}_0^{\mathrm{VAPoRS}}$, when using the BMA approach and VAPoRS, respectively. It can be seen from the figure that the normalized errors of the reconstructed noiseless signals using the compact summary obtained by VAPoRS are quite comparable with the ones obtained using the BMA approach.

Finally, the scatter plots in the last two panels compare the expected number of components in the intervals $(0,\pi/4)$ and $(\pi/4,\pi/2)$ using VAPoRS with, again, the ones obtained using the direct BMA approach. For the BMA approach, the expected number of components in an interval $T \subset (0;\pi)$ is

$$\mathbb{E}(N(T)\,|\,\mathbf{y}) \;=\; \sum_{k\in\mathbb{N}} \mathbb{E}(N(T)\,|\,k,\mathbf{y})\,p(k\,|\,\mathbf{y}) \;\approx\; \frac{1}{M}\sum_{i=1}^{M} N^{(i)}(T),$$

where $N^{(i)}(T)$ is the number of radial frequencies observed in T on the i^{th} sample. On the other hand, from the summary provided by VAPoRS, the expected number of components in interval T is

$$\mathbb{E}_{\hat{\boldsymbol{\eta}}}\,(N(T)\,|\,\mathbf{y}) \;=\; \sum_{l=1}^{L} \hat{\pi}_l\,\mathscr{N}(T;\hat{\boldsymbol{\eta}}_l) \;+\; \hat{\lambda}\,\frac{|T|}{|\Theta|},$$

where $\mathscr{N}(T;\hat{\boldsymbol{\eta}}_l)$ denotes the probability of T under the Gaussian distribution with parameters $\hat{\boldsymbol{\eta}}_l$. The figures confirm that the expected number of components in the chosen intervals computed using both approaches are very similar.

The results shown in this section confirmed that the approximate posterior distribution $q_{\hat{\boldsymbol{\eta}}}$ obtained using VAPoRS preserves faithfully several important features of the true posterior distribution; see [1, Section 3.4] for more results in this vein, including a numerical investigation comparison of the properties of estimators derived from VAPoRS.

V. Conclusion

In this paper, we have proposed a novel algorithm to relabel and summarize variable dimensional posterior distributions encountered in signal decomposition problems when the number of component is unknown. For this purpose, a variable-dimensional parametric model has been designed to approximate the posterior of interest. The parameters of the approximate model have been estimated by means of an SEM-type algorithm, using samples from the true posterior distribution f generated by a trans-dimensional (e.g., RJ-MCMC) Monte Carlo sampler. Modifications of our initial SEM-type algorithm have been proposed, in order to cope with the lack of robustness of maximum likelihood-type estimates.

The relevance of the proposed algorithm, both for summarizing and for relabeling variable-dimensional posterior distributions, has been illustrated on two signal decomposition examples, namely, the problem of detection and estimation of sinusoids in Gaussian white noise and a particle counting problem motivated by the astrophysics project Auger. Most notably, VAPoRS has been shown to be the first approach in the literature capable of solving the label-switching issue in trans-dimensional problems. We have shown that the proposed parametric model provides a good approximation for the posteriors encountered in both applications. Moreover, VAPoRS can provide the user with more insight concerning not only the component-specific parameters but also the uncertainties about their presence.

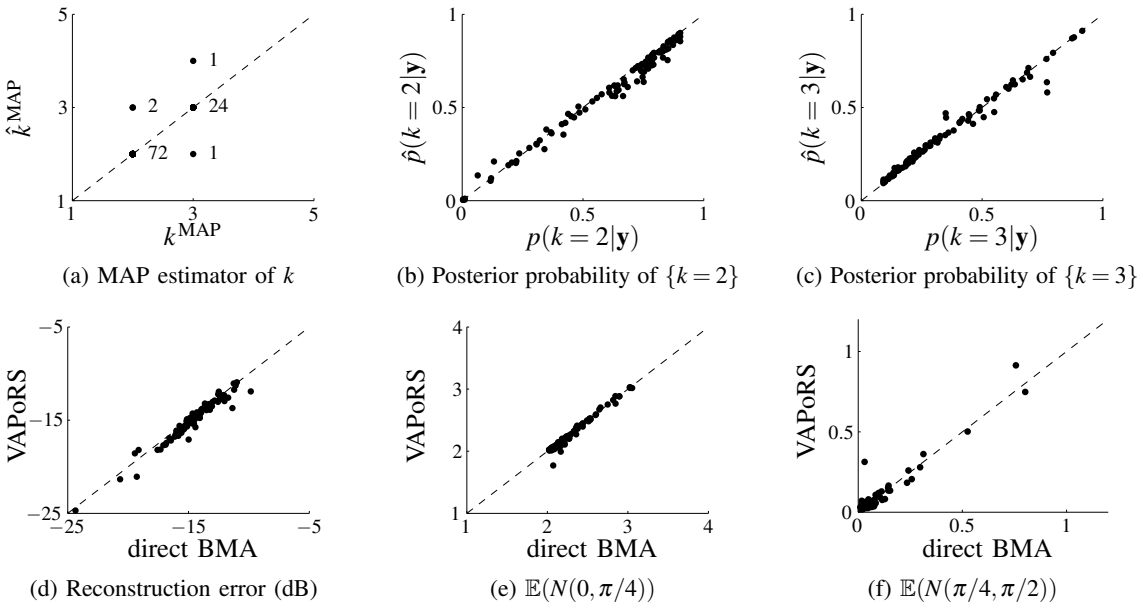
We believe that this algorithm can be useful in the vast domain of signal decomposition and mixture model analysis to enhance inference in trans-dimensional problems. In particular, it is useful in large-scale applications, such as the Auger project, where storing all samples can be problematic. Theoretical investigations are required in order to extend available existing convergence results to the SEM-type algorithm used in this paper (with correlated input data and Metropolis-Hastings updates). Future work will focus on using VAPoRS to design more efficient adaptive trans-dimensional MCMC methods, as a continuation of the ideas presented in [8, 9].

Acknowledgment

The authors would like to express their gratitude to B. Kégl for his collaboration and providing data for the Auger example.

References

[1] A. Roodaki, *Signal decompositions using trans-dimensional Bayesian methods*, Ph.D. thesis, École Supérieure d'Électricité (Supélec), Gif-sur-Yvette, France, 2012.



(a) MAP estimator of k (b) Posterior probability of $\{k=2\}$ (c) Posterior probability of $\{k=3\}$

(d) Reconstruction error (dB) (e) $\mathbb{E}(N(0,\pi/4))$ (f) $\mathbb{E}(N(\pi/4,\pi/2))$

Figure 17: Comparison of (some features of) the true posterior distribution with its VAPoRS approximation.

[2] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller, "Equation of state calculations by fast computing machines," *The Journal of Chemical Physics*, vol. 21, no. 6, pp. 1087, 1953.

[3] W. K. Hastings, "Monte Carlo sampling methods using Markov chains and their applications," *Biometrika*, vol. 57, no. 1, pp. 97–109, 1970.

[4] J. S. Liu, *Monte Carlo Strategies in Scientific Computing*, Springer Verlag, 2001.

[5] C. P. Robert and G. Casella, *Monte Carlo Statistical Methods (second edition)*, Springer Verlag, 2004.

[6] M. West, "Approximating posterior distributions by mixture," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 55, no. 2, pp. 409–422, 1993.

[7] H. Haario, E. Saksman, and J. Tamminen, "An adaptive metropolis algorithm," *Bernoulli*, pp. 223–242, 2001.

[8] Y. Bai, R. V. Craiu, and A. F. Di Narzo, "Divide and conquer: a mixture-based approach to regional adaptation for MCMC," *Journal of Computational and Graphical Statistics*, vol. 20, no. 11, pp. 63–79, 2011.

[9] R. Bardenet, O. Cappé, G. Fort, and B. Kégl, "An adaptive Metropolis algorithm with online relabeling," in *the proceeding of the 15th International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2012.

[10] P. J. Green, "Reversible jump MCMC computation and Bayesian model determination," *Biometrika*, vol. 82, no. 4, pp. 711–732, 1995.

[11] P. J. Green, "Trans-dimensional Markov chain Monte Carlo," in *Highly Structured Stochastic Systems*, P. J. Green, N. L. Hjort, and S. Richardson, Eds., pp. 179–198. O.U.P., 2003.

[12] C. Andrieu and A. Doucet, "Joint Bayesian model selection and estimation of noisy sinusoids via reversible jump MCMC," *IEEE Transactions on Signal Processing*, vol. 47, no. 10, pp. 2667–2676, 1999.

[13] J. R. Larocque and J. P. Reilly, "Reversible jump MCMC for joint detection and estimation of sources in coloured noise," *IEEE Transactions on Signal Processing*, vol. 50, pp. 231–240, 2002.

[14] H. Rue and M.A. Hurn, "Bayesian object identification," *Biometrika*, vol. 86, no. 3, pp. 649–660, 1999.

[15] M. Ortner, X. Descombes, and J. Zerubia, "Building outline extraction from digital elevation models using marked point processes," *International Journal of Computer Vision*, vol. 72, pp. 107–132, 2007.

[16] C. Andrieu, E. Barat, and A. Doucet, "Bayesian deconvolution of noisy filtered point processes," *IEEE Transactions on Signal Processing*, vol. 49, no. 1, pp. 134–146, 2002.

[17] Auger Collaboration, "The Pierre Auger Project Design Report (Second Edition)," http://www.auger.org/technical_info/design_report.html, 1997.

[18] Auger Collaboration, "Properties and performance of the prototype instrument for the Pierre Auger Observatory," *Nuclear Instruments and Methods in Physics Research A*, vol. 523, pp. 50–95, 2004.

[19] S. Richardson and P. J. Green, "On Bayesian analysis of mixtures with an unknown number of components," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 59, no. 4, pp. 731–792, 1997.

[20] M. Stephens, "Dealing with label switching in mixture models," *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, pp. 795–809, 2000.

[21] A. Jasra, C. C. Holmes, and D. A. Stephens, "Markov chain Monte Carlo methods and the label switching problem in Bayesian mixture modeling," *Statistical Science*, vol. 20, no. 1, pp. 50–67, 2005.

[22] G. Celeux, M. Hurn, and C. P. Robert, "Computational and inferential difficulties with mixture posterior distributions," *Journal of the American Statistical Association*, pp. 957–970, 2000.

[23] S. Frühwirth-Schnatter, "Dealing with label switching under model uncertainty," in *Mixtures: estimation and applications*, K. Mengersen, C. P. Robert, and D. Titterington, Eds., pp. 213–239. Wiley Online Library, 2011.

[24] P. Papastamoulis and G. Iliopoulos, "An artificial allocations based solution to the label switching problem in Bayesian analysis of mixtures of distributions," *Journal of Computational and Graphical Statistics*, vol. 19, no. 2, pp. 313–331, 2010.

[25] C. E. Rodríguez and S. G. Walker, "Label switching in bayesian mixture models: Deterministic relabeling strategies," *Journal of Computational and Graphical Statistics*, 2013.

[26] M. Sperrin, T. Jaki, and E. Wit, "Probabilistic relabelling strategies for the label switching problem in bayesian mixture models," *Statistics and Computing*, vol. 20, pp. 357–366, 2010.

[27] W. Yao, "Model based labeling for mixture models," *Statistics and Computing*, pp. 1–11, 2011.

[28] C. P. Robert, "Discussion of "On Bayesian analysis of mixtures with an unknown number of components," by S. Richardson and P. J. Green," *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, vol. 59, no. 4, pp. 758–764, 1997.

[29] A. Roodaki, J. Bect, and G. Fleury, "Comments on "Joint Bayesian model selection and estimation of noisy sinusoids via reversible jump MCMC"," *IEEE Transactions on Signal Processing*, vol. 61, no. 14, pp. 3653–3655, 2013.

[30] M. A. Clyde and E. I. George, "Model uncertainty," *Statistical Science*, vol. 19, no. 1, pp. 81–94, 2004.

[31] G. Celeux and J. Diebolt, "The SEM algorithm : a probabilistic teacher algorithm derived from the EM algorithm for the mixture problem," *Computational Statistics Quaterly*, vol. 2, pp. 73–82, 1985.

[32] G. Celeux and J. Diebolt, "A stochastic approximation type EM algorithm for the mixture problem," *Stochastics and Stochastics Reports*, vol. 41, no. 1, pp. 119–134, 1992.

[33] S. F. Nielsen, "The stochastic EM algorithm: estimation and asymptotic results," *Bernoulli*, vol. 6, no. 3, pp. 457–489, 2000.

[34] M. Stephens, *Bayesian methods for mixture of normal distributions*, Ph.D. thesis, D Phill Thesis. University of Oxford, Oxford., 1997.

[35] A. F. Karr, *Point Processes and their Statistical Inference (2nd ed.)*, CRC, 1991.

[36] M. A. Tanner and W. H. Wong, "The calculation of posterior distributions by data augmentation," *Journal of the American statistical Association*, vol. 82, no. 398, pp. 528–540, 1987.

[37] A. P. Dempster, N. B. Laird, and D. B. Rubin, "Maximum Likelihood from Incomplete Data via the EM Algorithm," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 39, no. 1, pp. 1–38, 1977.

[38] J. Diebolt and G. Celeux, "Asymptotic properties of a stochastic EM algorithm for estimating mixing proportions," *Stochastic Models*, vol. 9, no. 4, pp. 599–613, 1993.

[39] R. N. Davé and R. Krishnapuram, "Robust clustering methods: a unified view," *IEEE Transactions on Fuzzy Systems*, vol. 5, no. 2, pp. 270–293, 1997.

[40] P. J. Huber and E. M. Ronchetti, *Robust statistics (2nd Edition)*, Wiley., 2009.

[41] A. Basu, I. R. Harris, N. L. Hjort, and M. C. Jones, "Robust and efficient estimation by minimising a density power divergence," *Biometrika*, vol. 85, no. 3, pp. 549–559, 1998.

[42] B. Kégl, "Bayesian estimation, the Metropolis-Hastings algorithm, and a simple example," Technical report, LAL, University of Paris-Sud / CNRS, France, 2008.

[43] R. Bardenet, B. Kégl, and D. Veberic, "Single muon response: The signal model," Technical report, LAL, University of Paris-Sud / CNRS, France, 2010.

[44] R. Bardenet, *Towards adaptive learning and inference : Applications to hyper-parameter tuning and astroparticle physics*, Ph.D. thesis, University of Paris 11, Orsay, France, 2012.