Signal decompositions using trans-dimensional Bayesian methods.

Alireza Roodaki

N° D'ORDRE: 2012-05-TH

Thèse de doctorat

DOMAINE: STIC
Spécialité: Traitement du Signal

École Doctorale "Sciences et Technologies de l'Information,
des Télécommunications et des Systèmes"

Présentée par:

Alireza Roodaki

Décomposition de signaux dans un cadre bayésien trans-dimensionnel

Signal decompositions using trans-dimensional Bayesian methods

Soutenue le 14 Mai 2012 devant les membres du jury:

M.	Bect	Julien	Supélec	Encadrant
M.	Cappé	Olivier	Telecom ParisTech	Examinateur
M.	Djurić	Petar	Stony Brooks University	Rapporteur
M.	Fleury	Gilles	Supélec	Directeur de thèse
M.	Robert	Christian	Université Paris-Dauphine	Rapporteur
M.	Walter	Eric	Supélec	Président

Abstract

This thesis addresses the challenges encountered when dealing with signal decomposition problems with an unknown number of components in a Bayesian framework. Particularly, we focus on the issue of summarizing the variable-dimensional posterior distributions that typically arise in such problems. 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 Reversible-Jump MCMC (RJ-MCMC) sampler. No generic approach is available, however, to summarize the resulting variable-dimensional samples and extract from them component-specific parameters. One of the main challenges that needs to be addressed to this end is the label-switching issue, which is caused by the invariance of the posterior distribution to the permutation of the components.

We propose a novel approach to this problem, which consists in approximating the complex posterior of interest by a “simple”—but still variable-dimensional parametric distribution. We develop stochastic EM-type algorithms, driven by the RJ-MCMC sampler, to estimate the parameters of the model through the minimization of a divergence measure between the two distributions. Two signal decomposition problems are considered, to show the capability of the proposed approach 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.

Résumé

Cette thèse porte sur le problème de la décomposition de signaux contenant un nombre inconnu de composantes, envisagé dans un cadre bayésien. En particulier, nous nous concentrons sur la question de la description des lois a posteriori qui ont la spécificité, pour les problèmes de ce genre, d'être définies sur une union de sous-espaces de dimensions différentes. Ces lois peuvent être échantillonnées à l'aide de techniques de Monte Carlo récentes, telles que l'échantillonneur MCMC à sauts réversibles (RJ-MCMC), mais aucune approche générique n'existe à l'heure actuelle pour décrire les échantillons produits par un tel échantillonneur et en extraire les paramètres spécifiques des composantes. L'un des principaux obstacles est le problème de la commutation des étiquettes (label-switching), causé par l'invariance de la loi a posteriori vis-à-vis de permutations de ses composantes.

Nous proposons une nouvelle approche pour résoudre ce problème, qui consiste à approcher la loi a posteriori d'intérêt par une loi paramétrique plus “simple”, mais toujours définie sur un espace de dimension variable. Nous développons des algorithmes de type SEM (Stochastic Expectation-Maximization), s'appuyant sur la sortie d'un échantillonneur RJ-MCMC, afin d'estimer les paramètres du modèle par minimisation d'une divergence entre les deux lois. Deux problèmes de décomposition de signaux illustrent la capacité de la méthode proposée à résoudre le problème de commutation des étiquettes et à produire des résumés de lois a posteriori définies sur des espaces de dimension variable : le problème classique de détection et d'estimation de composantes sinusoïdales dans un bruit blanc d'une part, et un problème de comptage de particules motivé par le projet Pierre Auger en astrophysique d'autre part.

Acknowledgments

This work would have not been possible without the contribution of many people. First and foremost, I feel indebted to my supervisor, Prof. Julien Bect, for his guidance, effort and patience. I will never forget his continued presence, the way he taught me, our interesting discussions, and particularly his kindness. I am also grateful to him for reading this entire manuscript.

I am deeply grateful to my second supervisor, Prof. Gilles Fleury, for his support throughout this work.

I would also like to thank Prof. Petar M. Djurić and Prof. Christian P. Robert for reviewing this work, providing insightful comments and attending my defense. I would like to thank as well Prof. Olivier Cappé and Prof. Eric Walter for attending my defense as examiners and for their helpful comments.

I am grateful to Prof. Balázs Kégl for his collaboration and for providing me data for Chapter 4 and attending my defense.

I would also like to thank my colleagues in the department of SSE at Supélec: Prof. Stéphane Font, the head of department, Karine Bernard, for making living in France much easier for me and my wife, Prof. Emmanuel Vazquez, for keeping the SSEcalcul machines alive, Prof. Jérôme Juillard, for his useful comments, Prof. Arthur Tenenhaus, Prof. José Picheral, Prof. Elisabeth Lahalle and Luc Batalie. I am very grateful to the department for providing me with support for attending conferences. My special thank to my fellow PhD students: Kian Jafari, Arash Behboodi, Ling Li, Romain Benassi, Jean-Michel Akre and Rémi Bardenet.

I would like to express my gratitude to our French friends, Emile Urvoy and Madeleine Trebuchon, for their kindness.

I wish to thank my parents, Roghayeh Afrasiabian and Khalil Roodaki, for all their love, support and encouragement. I would also like to thank my sister, Elnaz.

Lastly, and most importantly, I would like to thank my beloved wife, Sepideh Razaghi. She loved me, encouraged me and nourished me in all these days (I am sure everybody will remember the reception after my defense!). She is a talented artist who decorates my life every day.

To Sepideh.

Contents

Introduction (en français)

Le problème de la décomposition des signaux

Décomposer un signal, une image, ou plus généralement des données observées, en un ensemble d'« atomes » ou de « composantes » est une tâche fondamentale dans le domaine du traitement du signal et des images. Il est important de faire la distinction entre deux types de problèmes de décomposition, selon que l'objectif est la prédiction ou le débruitage d'une part, ou l'inférence sur les composantes d'autre part.

Dans le premier type de problème, l'objectif est d'obtenir une représentation parcimonieuse du signal observé en utilisant dictionnaire (généralement redondant) de signaux élémentaires. Ce type de technique trouve principalement ses applications en débruitage, en compression, en séparation de sources, en segmentation et en déconvolution. Le dictionnaire peut être soit appris à partir des données (voir, par exemple, Lewicki and Sejnowski, 2000 ; Elad and Aharon, 2006), soit construit à l'aide de familles classiques de signaux élémentaires (ou « atomes ») tels que la base de Fourier, les bases d'ondelettes, ou encore les *curvelets* (voir, par exemple, Mallat, 2009). Parmi les algorithmes permettant d'effectuer de telles décompositions, on peut citer par exemple *Matching Pursuit* (Mallat and Zhang, 1993), *Basis Pursuit*/LASSO (Chen et al., 1999 ; Tibshirani, 1996), ou encore le « sélecteur de Danzig » (Candes and Tao, 2007). Le même type de problème a également été abordé dans un cadre bayésien, en utilisant des idées venant de la littérature de la sélection variables bayésienne (voir, par exemple, Wolfe et al., 2004 ; Fevotte and Godsill, 2006 ; Dobigeon et al., 2009).

Dans le deuxième type de problème, le signal observé est supposé être une superposition de plusieurs signaux élémentaires (ou « composantes ») d'intérêt. Dans ce cas, l'objectif est à la fois *la détection* du nombre réel de composantes et *l'estimation* de leurs paramètres (alors que, dans le premier point de vue, l'estimation précise du nombre d'atomes est généralement secondaire). On rencontre ce type de problème en analyse spectrale, en traitement d'antennes (traitement de signal pour les réseaux de capteurs), en spectrométrie, ou encore pour la détection d'objets dans des images et l'analyse de données hétérogènes par des modèles de mélange. Des critères de sélection du modèle tels que les critères AIC (*Akaike Information Criterion*) ou BIC (*Bayesian Information Criterion*) ont été largement utilisés dans des problèmes de détection et d'estimation jointes (voir Stoica and Selen, 2004, pour une analyse). L'approche bayésienne a également été utilisée pour analyser ce genre de problèmes (voir, par exemple, Richardson and Green, 1997 ; Andrieu and Doucet, 1999 ; Lacoste et al., 2005), afin de mieux prendre en compte les différentes sources d'incertitudes.

Dans cette thèse, nous nous concentrons sur les défis rencontrés dans la deuxième type de problèmes de décomposition de signaux, avec un nombre inconnu de composants, en particulier quand ils sont traités dans un cadre bayésien.

Exemple : détection et estimation des muons dans le projet Auger

A titre d'illustration du type de problème de décomposition de signaux qui nous intéresse, considérons maintenant un problème de détection et d'estimation qui a été porté à notre attention par M. Balázs Kégl (Laboratoire de l'Accélérateur Linéaire (LAL), Université Paris Sud 11) dans le cadre du projet Auger (voir, par exemple, Auger Collaboration, 1997, 2004). Dans ce projet, l'objectif est d'étudier les rayons cosmiques d'ultra-haute énergie (on entend par là des énergies de l'ordre de 10 ^{19} eV, c'est-à-dire les particules les plus énergétiques trouvées à ce jour dans l'Univers). Lorsque les particules contenues dans les rayons cosmiques entrent en collision avec celles de l'atmosphère terrestre, des gerbes atmosphériques contenant des particules secondaires appelées muons sont générées. Pour les détecter, l'Observatoire Pierre Auger (*Pierre Auger Cosmic Ray Observatory*) a été construit en Argentine. L'observatoire est composé de deux détecteurs indépendants : une matrice de détecteurs de surface et un certain nombre de détecteurs de fluorescence.

Quand un muon traverse un détecteur de surface, il génère le long de sa trajectoire des photo-électrons (PE) Cherenkov, dont le taux dépend de l'énergie du muon. Ces photo-électrons sont capturés par des détecteurs et créent un signal analogique, qui est ensuite discrétisé par un convertisseur analogique-numérique. Sachant le nombre k de muons, le signal observé peut être modélisé (Kégl, 2008 ; Bardenet et al., 2010) par un processus de Poisson non homogène d'intensité

$$h(t \,|\, \boldsymbol{a}_\mu, \boldsymbol{t}_\mu) \;=\; \sum_{j=1}^{k} a_{\mu,j}\, p_{\tau,t_d}(t - t_{\mu,j}),$$

où $p_{\tau,t_d}(t)$ est la loi des temps de réponse, paramétrée par son temps de montée t_d

et sa décroissance exponentielle τ. Les paramètres inconnus de ce modèle sont les amplitudes $\boldsymbol{a}_\mu = (a_{\mu,1}, \ldots, a_{\mu,k})$, les temps d'arrivée $\boldsymbol{t}_\mu = (t_{\mu,1}, \ldots, t_{\mu,k})$, ainsi que le nombre k de muons. La figure 0.1 montre le signal observé et l'intensité $h(t \,|\, \boldsymbol{a}_\mu, \boldsymbol{t}_\mu)$ pour une exemple simulé avec $k = 3$ muons.

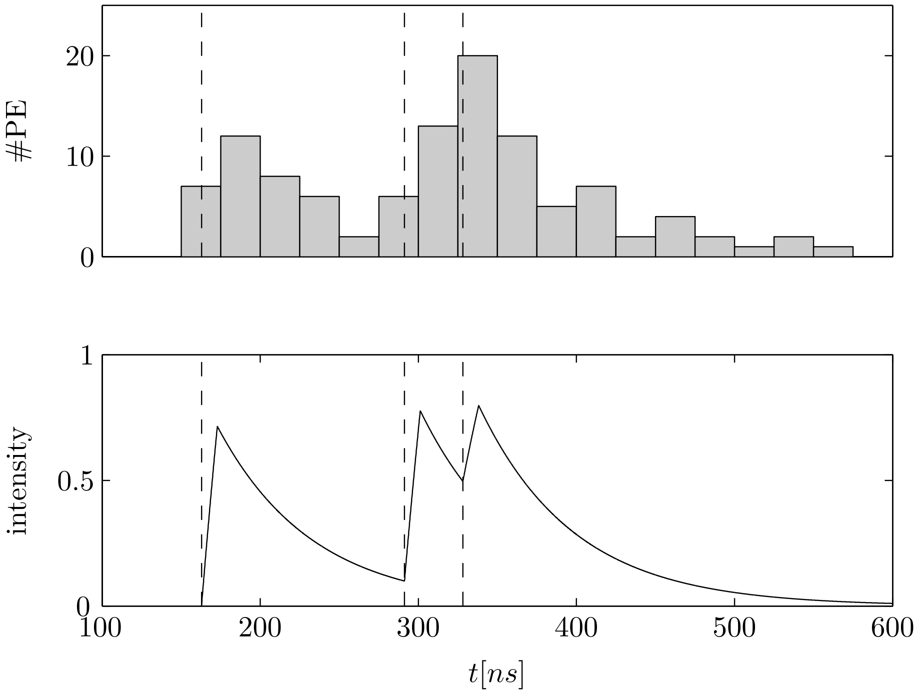


Figure 0.1 – *Signal observé (en haut) et intensité du modèle $h(t \,|\, \boldsymbol{a}_\mu, \boldsymbol{t}_\mu)$ (en bas) pour un exemple avec $k = 3$ muons. Les vrais temps d'arrivée, i.e., $\boldsymbol{t}_\mu =$* (162, 291, 328)*, sont indiqués par les lignes pointillées verticales.*

On s'attend à ce que le nombre de muons reçus par les détecteurs de surface, ainsi que leurs caractéristiques individuelles (particulièrement les temps d'arrivée), soient des informations utiles pour inférer la composition chimique de la particule qui était au l'origine de la gerbe observée. Il s'agit d'un problème de sélection du modèle et d'estimation des paramètres, aussi connu comme un problème "trans-dimensionnel" dans la littérature (voir, par exemple, Green, 2003), où un ensemble dénombrable de modèles concurrents, $\mathcal{M} = \{\mathcal{M}_1, \mathcal{M}_2, \cdots\}$, indexé par $k \in \mathcal{K} \subset \mathbb{N}$, est considéré pour décrire les données observées. Le modèle $\mathcal{M}_k$ suppose que le signal observé est une superposition de k signaux élémentaires (correspondant à k muons). Le vecteur correspondant aux paramètres spécifiques à chacune des composantes est $\boldsymbol{\theta}_k = ((t_{\mu,1}, a_{\mu,1}), \ldots, (t_{\mu,k}, a_{\mu,k})) \in \boldsymbol{\Theta}^k$, où $\boldsymbol{\Theta} = \mathbb{R}_+^2$. Par conséquent, le problème est défini sur l'espace $\mathbb{X} = \bigcup_{k \geq 0} \{k\} \times \boldsymbol{\Theta}^k$, qui est une union disjointe de sous-espaces de dimensions différentes.

Un état de l'art succinct des approches pour la sélection de modèle

Les problèmes trans-dimensionnels et le choix de modèle (bayésien) ont une longue histoire en science et en ingénierie. Lahiri (2001) et Burnham and Anderson (2002) fournissent des discussions exhaustives des méthodes existantes, tandis que Congdon (2006) et Robert (2007, Chapitre 7) se focalisent plus particulièrement sur les méthodes bayésiennes. Un principe commun à toutes les méthodes existantes est la recherche de la *parcimonie* du modèle sélectionné par rapport à la taille de l'échantillon observé.

Par exemple, les méthodes classiques de la sélection de modèle tels que AIC et BIC comprennent l'évaluation de deux termes : un terme d'attache aux données, qui vise à mesurer la proximité des données observées et le modèle supposé, et un terme de pénalisation qui vise à pénaliser la complexité du modèle. Plus précisément, les méthodes AIC / BIC sélectionnent

$$\hat{k} \;=\; \operatorname*{arg\,max}_{k\in\mathcal{K}} \bigl\{ -\mathcal{L}(\hat{\theta}_k) + F(k) \bigr\},$$

où $\mathbf{y}$ est le signal observé, $\mathcal{L}(\hat{\theta}_k)$ est la fonction de log-vraisemblance évaluée en l'estimateur du maximum de vraisemblance, et $\hat{\theta}_k$ et $F(k)$ est la pénalité associée au modèle $\mathcal{M}_k$.

Dans le paradigme bayésien, des lois a priori exprimant les croyances *a priori* sont attribués aux paramètres inconnus (i.e., dans l'exemple précédent, au nombre k de muons et à leur vecteur de paramètres $\boldsymbol{\theta}_k = (\boldsymbol{t}_\mu, \boldsymbol{a}_\mu))$. Puis la sélection de modèle et l'estimation des paramètres sont effectuées en utilisant la loi a posteriori, définie sur un espace de dimension variable,

$$p(k,\, \boldsymbol{\theta}_k \,|\, \mathbf{y}) \;=\; \frac{p(\mathbf{y} \,|\, k,\, \boldsymbol{\theta}_k)\, p(k,\, \boldsymbol{\theta}_k)}{\sum_{k'\in\mathcal{K}} \int_{\boldsymbol{\Theta}_k} p(\mathbf{y} \,|\, \boldsymbol{\theta}'_k,\, k')\, p(k',\, \boldsymbol{\theta}'_k)\, \mathrm{d}\boldsymbol{\theta}'_k},$$

où $\boldsymbol{\Theta}_k$ est l'espace des paramètres pour le modèle $\mathcal{M}_k$.

Avant l'introduction des échantillonneurs de Monte Carlo trans-dimensionnels, le choix de modèle bayésien a souvent été effectué par le calcul des facteurs de Bayes, qui peuvent être vus comme une généralisation des arguments de test d'hypothèse (voir Kass and Raftery (1995) pour une analyse exhaustive et des commentaires utiles). Par exemple, la comparaison des modèles $\mathcal{M}_k$ et $\mathcal{M}_{k'}$ peut se faire en calculant le facteur de Bayes dans les cas k de k'

$$B(k \,:\, k') \;=\; \frac{p(\mathbf{y} \,|\, k)}{p(\mathbf{y} \,|\, k')},$$

où

$$p(\mathbf{y} \,|\, k) \;=\; \int_{\boldsymbol{\Theta}_k} p(\mathbf{y} \,|\, \boldsymbol{\theta}_k, k)\, p(\boldsymbol{\theta}_k \,|\, k)\, \mathrm{d}\boldsymbol{\theta}_k.$$

Néanmoins, pour calculer cette intégrale pour chaque modèle, dans la plupart des cas, on doit utiliser des méthodes de simulation Monte Carlo, comme les méthodes de chaîne de Markov Monte Carlo (MCMC) ou les méthodes de Monte Carlo séquentielles (voir, par exemple, Robert and Casella, 2004), qui sont coûteuses en temps de calcul (voir, par exemple, Han and Carlin, 2001).

L'échantillonneur de MCMC à sauts réversibles (RJ-MCMC) proposé par Green (1995) a finalement permis d'approcher ce type de lois a posteriori $p(k, \boldsymbol{\theta}_k \,|\, \mathbf{y})$, définies sur une union de sous-espaces de dimensions différentes, en toute généralité. L'échantillonneur de RJ-MCMC peut être vu comme une généralisation du célèbre échantillonneur de Metropolis-Hastings (Metropolis et al., 1953 ; Hastings, 1970), qui est capable d'explorer non seulement l'espace des paramètres $\boldsymbol{\Theta}_k$, mais aussi l'espace $\mathcal{K}$ de tous les modèles considérés.

Ré-étiquetage et résumé des lois a posteriori trans-dimensionnelles

Résumer une loi a posteriori consiste à fournir quelques statistiques simples mais interprétables et/ou des graphiques à l'utilisateur final d'une méthode statistique. Par exemple, dans le cas d'un paramètre scalaire avec une loi a posteriori unimodale, des statistiques de positions et de dispersion (par exemple, la moyenne et l'écart-type, ou la médiane et l'intervalle interquartile) sont généralement fournies en plus d'un résumé graphique de la distribution (par exemple, un histogramme ou une estimation à noyau de la densité).

Dans la plupart des problèmes de décomposition de signaux, l'une des principales difficultés rencontrées lorsqu'on essaie de résumer la loi a posteriori est le problème de la commutation des étiquettes (*label-switching*), causé par l'invariance de la loi a posteriori vis-à-vis de permutations des composantes. Ce problème a surtout été étudié pour les modèles de mélange gaussien dans la littérature (voir, par exemple, Richardson and Green, 1997 ; Celeux et al., 2000 ; Stephens, 2000 ; Jasra et al., 2005). En raison de ce problème, toutes les lois marginales a posteriori des paramètres spécifiques des composantes sont identiques, rendant ainsi les moyennes a posteriori, habituellement utilisé pour le résumé, inexploitables. Toutes les méthodes proposées jusqu'à présent dans la littérature pour résoudre le problème de la commutation des étiquettes sont limitées aux modèles de dimension fixée. (Ces méthodes peuvent néanmoins êtres utilisées dans des problèmes trans-dimensionnels, en effectuant préalable un choix de modèle puis en résumant les lois posteriori sachant le modèle sélectionné.)

La figure 0.2 montre les lois marginales a posteriori du nombre k de muons (à gauche) et des temps d'arrivée triés sachant k (à droite), obtenues en utilisant un échantillonneur RJ-MCMC sur l'exemple représenté dans la section précédente. Chaque ligne correspond à un valeur de k, pour $2 \leq k \leq 4$. Trier les composantes — la plus simple des stratégies de ré-étiquetage (voir, par exemple, Richardson and Green, 1997) — en fonction de leur temps d'arrivée, i.e.,$t_{\mu,1} < \ldots < t_{\mu,k}$, permet de briser la symétrie de la loi a posteriori. On peut voir à partir de la figure que le modèle $\mathcal{M}_3$ a la plus grande probabilité a posteriori $p(k = 3 \,|\, \mathbf{y}) = 0,43$. En choisissant $\mathcal{M}_3$, ce qui serait le résultat de l'utilisation de l'approche que nous nommerons BMS (*Bayesian Model Selection*) dans la suite, tous les échantillons correspondant aux autres modèles seraient écartés (notez que $p(k = 2 \,|\, \mathbf{y}) = 0,38$ et $p(k = 4 \,|\, \mathbf{y}) = 0,16$). Ce faisant, nous perdrions l'incertitude concernant le nombre k de muons. Par ailleurs, résumer la composante centrale apparaissant sous le modèle $\mathcal{M}_3$ (représentée en bleu clair) par sa moyenne a posteriori n'aurait pas de sens, cette composante étant fortement bimodale en raison de l'effet de « commutation trans-dimensionnelle » des étiquettes. Remarquez que, sous le modèle $\mathcal{M}_4$, la densité de cette composante est divisée en deux parties relativement « compactes ». Ainsi, lors du déplacement entre les modèles de l'algorithme RJ-MCMC par la naissance (ajout) ou la mort (suppression) d'une composante, l'étiquette correspondante est insérée ou supprimée. Nous appelons ce problème "naissance, mort, et commutation des étiquettes".

La principale contribution de cette thèse est une nouvelle approche pour le ré-étiquetage et le résumé des lois a posteriori trans-dimensionnelles, qui consiste à approcher la loi a posteriori d'intérêt par un modèle paramétrique original, lui aussi trans-dimensionnel.

Plan de la thèse

Le chapitre 1 décrit brièvement des techniques de simulation de Monte Carlo avancées, telles que les méthodes de Monte Carlo par chaînes de Markov (MCMC) et les méthodes de Monte Carlo séquentielles (SMC), qui sont, de nos jours, couramment utilisée dans la littérature bayésienne. Ces méthodes seront utilisées tout au long de la thèse. Nous fournissons également dans ce chapitre des énoncés clairs et rigoureux de certains résultats mathématiques, probablement pas totalement nouveaux mais jamais vraiment explicités, qui permettent une justification propre du taux d'acceptation des mouvements de naissance ou de mort dans les problèmes de décomposition du signaux (entres autres). Nous corrigeons ainsi une erreur concernant ce type de mouvements qui s'est glissée dans le document fondateur de Andrieu and Doucet (1999, équation (20)) et s'est ensuite largement propagée dans la littérature du traitement du signal (Andrieu et al., 2000, 2001a, 2002 ;

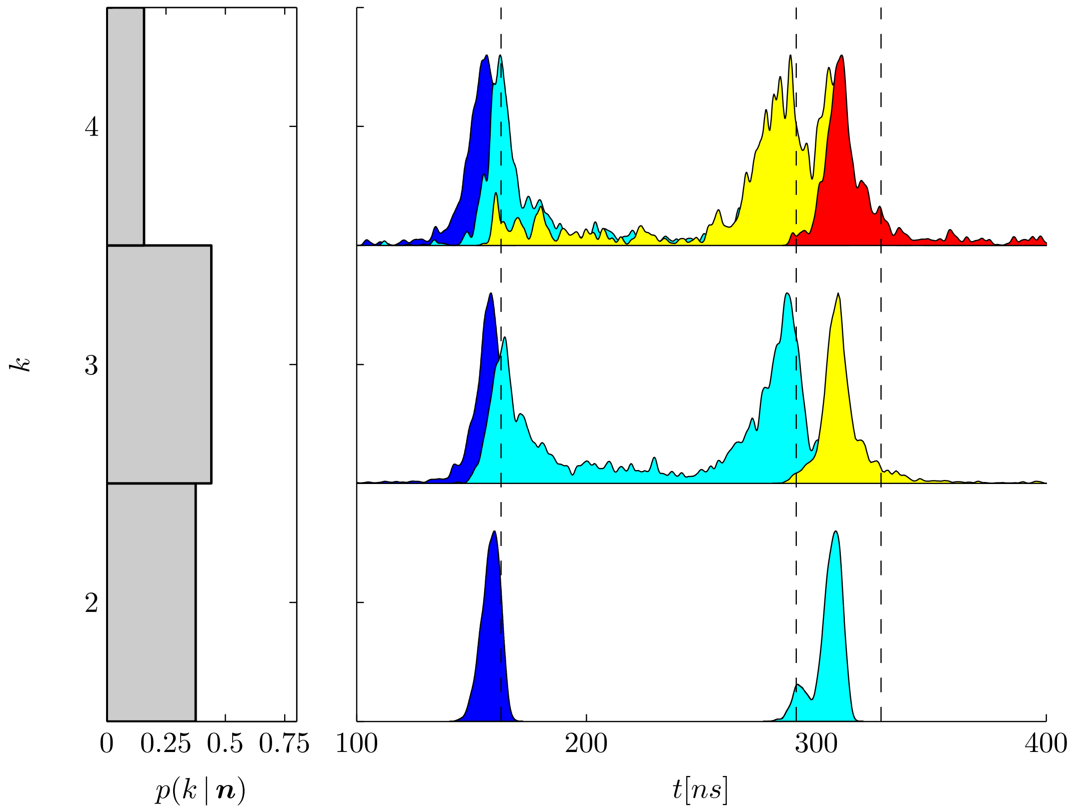


Figure 0.2 – *Des lois a posteriori du nombre k de muons (à gauche) et des temps d'arrivée triés, $\boldsymbol{t}_\mu$ sachant k (à droite) construit en utilisant 60 000 échantillons de fourni par RJ-MCMC. Le nombre réel de composants est trois. Les lignes verticales en pointillés dans la figure de droite localisent les temps d'arrivée, i.e., $\boldsymbol{t}_\mu = (163, 291, 328)$.*

Larocque and Reilly, 2002 ; Larocque et al., 2002 ; Ng et al., 2005 ; Davy et al., 2006 ; Rubtsov and Griffin, 2007 ; Shi et al., 2007 ; Melie-García et al., 2008 ; Ng et al., 2008 ; Hong et al., 2010 ; Schmidt and Mørup, 2010).

Le chapitre 2 introduit le problème de « naissance, mort, commutation des étiquettes » dans des lois a posteriori trans-dimensionnelles, et décrit la nouvelle approche que nous proposons pour le ré-étiquetage et le résumé de ces lois a posteriori. Cette approche consiste à approcher la loi a posteriori d'intérêt par un modèle paramétrique « simple »— mais toujours trans-dimensionnel — dans l'esprit de l'approche développée par Stephens (2000) pour le ré-étiquetage en dimension fixée. L'approximation est réalisée par minimisation d'une mesure de divergence entre les lois. Nous considérons, successivement, le divergence de Kullback-Leibler (KL) et une mesure de divergence plus robuste proposé par Basu et al. (1998). Des algorithmes de type EM stochastique (SEM), entraînés par l'échantillonneur RJ-MCMC, sont développés afin d'estimer les paramètres du modèle d'approximation.

Le chapitre 3 revisite le problème de la détection et l'estimation de composantes sinusoïdales observées dans un bruit blanc. Nous discutons brièvement le problème de la spécification des lois a priori pour l'hyperparamètre réglant le rapport signal sur bruit et l'analyse de sensibilité bayésienne. La partie principale de ce chapitre étudie la capacité de l'approche proposée dans le chapitre 2 à ré-étiquer et résumer les lois a posteriori trans-dimensionnelles rencontrées dans ce problème. Plus précisément, nous illustrons la convergence de l'algorithme et le ré-étiquetage à travers trois exemples de détection de composantes sinusoïdales. Nous discutons également de l'aide de simulations, un certain propriétés fréquentistes des résumés obtenus.

Le chapitre 4 aborde le problème de la détection et l'estimation des muons dans le projet Auger. Comme dans le chapitre 3, nous étudions en utilisant cette application la capacité de l'approche proposée à ré-étiqueter et résumer des lois a posteriori trans-dimensionnelles. Cette étude est menée sur des données simulées fournis par M. Balázs Kégl. En outre, dans ce chapitre, nous discutons des questions relatives à l'initialisation des algorithmes de type SEM que nous avons proposés et à l'interprétation des résumés obtenus.

Enfin, nous concluons la thèse et donnons des orientations possibles pour les travaux futurs.

Introduction (in English)

The problem of signal decomposition

Decomposing an observed signal, image, or data, into a set of “atoms” or “components” is an important task in the fields of signal/image processing and data analysis. Depending on whether the objective is prediction, denoising or inference about individual components, it is helpful at this point to distinguish between two kinds of signal decomposition problems.

In the first kind of problem, the objective is to obtain a sparse representation of the observed signal using a large (possibly over-complete) dictionary. It has applications in denoising, compression, source separation, deconvolution and segmentation, to name a few. The dictionary can either be learned from the data (see, e.g., Lewicki and Sejnowski (2000) ; Elad and Aharon (2006)) or constructed based on elementary bases or atoms such as Fourier basis, wavelet basis, or curvelets for instance (see, e.g., Mallat, 2009). Influential algorithms include the matching pursuit algorithm of Mallat and Zhang (1993), the basis pursuit or lasso algorithm of Chen et al. (1999) ; Tibshirani (1996), and the more recent Dantzig selector of Candes and Tao (2007). The problem has also been addressed in a Bayesian framework using ideas from the Bayesian variable selection literature (see, e.g., Wolfe et al., 2004 ; Fevotte and Godsill, 2006 ; Dobigeon et al., 2009).

In the second kind of problem, the observed signal is assumed to be a superposition of number of fundamental elementary signals or components of interest. In this case, the objective is both to *detect* the true number of components and to *estimate* their parameters (whereas, in the first point of view, estimating accurately the number of included atoms is usually of minor importance). Applications include sensor array processing, spectral analysis, spectrometry, detection of objects in images, and mixture modeling of heterogeneous observed data. Model selection criteria such as the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC) have been extensively used in these joint detection and estimation problems (see Stoica and Selen, 2004, for a review). Bayesian approaches have also been used to analyze this kind of problems (see, e.g., Richardson and

Green, 1997 ; Andrieu and Doucet, 1999 ; Lacoste et al., 2005).

In this work, we concentrate on the challenges encountered in the second kind of signal decomposition problems with unknown number of components, specifically when they are treated in a Bayesian framework.

Example: Detection and estimation of muons in the Auger project

As an illustrative example of the kind of signal decomposition problems we are interested in, let us now describe a problem of detection and estimation that was brought to our attention by professor Balázs Kégl from the Laboratoire de l'Accélérateur Linéaire (LAL), Université Paris Sud 11, in connection with the Auger project (see, e.g., Auger Collaboration, 1997, 2004). In this project, the goal is to study ultra-high energy cosmic rays, with energies of the order of 10^{19}eV, the most energetic particles found so far in the universe. When these cosmic ray particles collide the earth's atmosphere, air showers containing secondary physical particles among which "muons" are of particular importance are generated. To detect the muons, the Pierre Auger Cosmic Ray Observatory was built in Argentina. The observatory consists of two independent detectors; an array of surface detectors and a number of fluorescence detectors.

When a muon crosses a surface detector, it generates "Cherenkov photons", the rate of which depends on the muon's energy, along its track. These photoelectrons (PE's) are then captured by detectors and create an analog signal which is consequently discretized using an analog-to-digital converter. Given the number k of muons, the observed signal is modeled (Kégl, 2008 ; Bardenet et al., 2010) by a non-homogeneous Poisson point process with intensity

$$h(t \,|\, \boldsymbol{a}_\mu, \boldsymbol{t}_\mu) \;=\; \sum_{j=1}^{k} a_{\mu,j}\, p_{\tau,t_d}(t - t_{\mu,j}),$$

where $p_{\tau,t_d}(t)$ is a known time response distribution, parametrized by its risetime t_d and its exponential decay τ (both measured in ns). The unknown parameters are the muons' amplitudes $\boldsymbol{a}_\mu = (a_{\mu,1}, \ldots, a_{\mu,k})$ and arrival times $\boldsymbol{t}_\mu = (t_{\mu,1}, \ldots, t_{\mu,k})$, along with the number k of muons. Figure 0.3 shows the observed signal and the intensity $h(t \,|\, \boldsymbol{a}_\mu, \boldsymbol{t}_\mu)$ for a simulated example with $k = 3$ muons.

The number of muons and their component-specific parameters, particularly the arrival times, are expected to be useful for making inference about the chemical composition of the particle that was at the origin of the observed shower. This is a joint model selection and parameter estimation problem, also known as a "trans-dimensional" problem

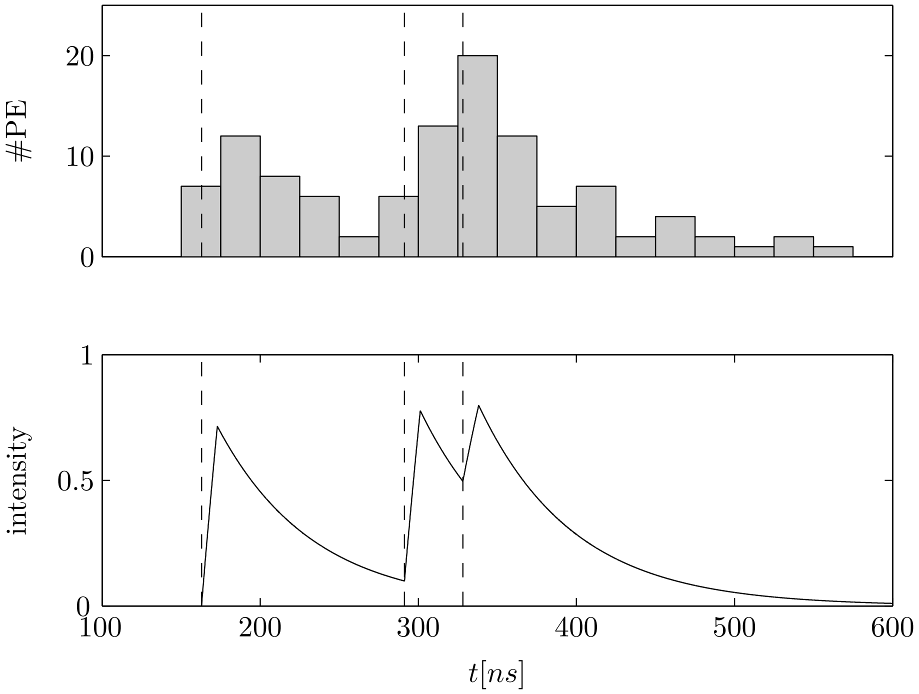


Figure 0.3 – *Observed signal (top) and intensity of the model* $h(t \,|\, \boldsymbol{a}_\mu,\, \boldsymbol{t}_\mu)$ *(bottom) for an example with* $k = 3$ *muons. The true arrival times, i.e.,* $\boldsymbol{t}_\mu = (162,\, 291,\, 328)$*, are indicated by the vertical dashed lines.*

in the literature (see, e.g., Green, 2003), where a countable set of competing models, $\mathcal{M} = \{\mathcal{M}_1,\, \mathcal{M}_2,\, \cdots\}$, indexed by $k \in \mathcal{K} \subset \mathbb{N}$, are considered to describe the observed data. The model $\mathcal{M}_k$ assumes that the observed signal is a superposition of k elementary signals (corresponding to k muons). The corresponding vector of component-specific parameters is $\boldsymbol{\theta}_k = ((t_{\mu,1}, a_{\mu,1}),\, \ldots,\, (t_{\mu,k}, a_{\mu,k})) \in \boldsymbol{\Theta}^k$, where $\boldsymbol{\Theta} = \mathbb{R}_+^2$ is the space of component-specific parameters. Hence, the problem is defined over the space $\mathbb{X} = \bigcup_{k \geq 0} \{k\} \times \boldsymbol{\Theta}^k$, which is a disjoint union of subspaces of differing dimension.

A short review of model selection approaches

Trans-dimensional problems and (Bayesian) model selection have a long history in science and engineering. Lahiri (2001) and Burnham and Anderson (2002) provide in-depth discussions of the existing methods for model choice and analysis; while Congdon (2006) and Robert (2007, Chapter 7) are more concentrated on the context of Bayesian model analysis. A common principle in all existing methods is *parsimony* of the selected model with respect to the sample size of the observed data. This idea is often referred to Occam's razor—"Shave away all but what is necessary".

For example, classical model selection methods such as AIC and BIC comprise evaluating two terms: a data term which aims at measuring the closeness of the observed data, denoted by $\mathbf{y}$, and the assumed model, and a penalty term which aims at penalizing the model complexity. More precisely, AIC/BIC select

$$\hat{k} \;=\; \arg\max_{k\in\mathcal{K}} \left\{ -\mathcal{L}(\hat{\theta}_k) + F(k) \right\},$$

where $\mathcal{L}(\hat{\theta}_k)$ is the log-likelihood function evaluated at the Maximum Likelihood (ML) estimate $\hat{\theta}_k$ and $F(k)$ is the so-called dimensionality penalty.

In the Bayesian paradigm, prior distributions expressing prior beliefs are assigned over the unknown parameters, i.e., in the previous example, the number k of muons and the component-specific parameters $\boldsymbol{\theta}_k = (\boldsymbol{t}_\mu, \boldsymbol{a}_\mu)$. Then, model selection and parameter estimation is carried out using the variable-dimensional posterior distribution

$$p(k,\, \boldsymbol{\theta}_k \,|\, \mathbf{y}) \;=\; \frac{p(\mathbf{y}\,|\,k,\, \boldsymbol{\theta}_k)\, p(k,\, \boldsymbol{\theta}_k)}{\sum_{k'\in\mathcal{K}} \int_{\boldsymbol{\Theta}_k} p(\mathbf{y}\,|\,\boldsymbol{\theta}'_k,\, k')\, p(k',\, \boldsymbol{\theta}'_k)\, \mathrm{d}\boldsymbol{\theta}'_k},$$

where $\boldsymbol{\Theta}_k$ is the parameter space for model $\mathcal{M}_k$.

Prior to the introduction of trans-dimensional Monte Carlo samplers, Bayesian model comparison has been, often, carried out by computing Bayes factors, which can be seen as a generalization of hypothesis testing arguments (see Kass and Raftery (1995) for a comprehensive review and useful comments). For example, comparing $\mathcal{M}_k$ against $\mathcal{M}_{k'}$ is achieved through computing the Bayes factor of k from k'

$$B(k\,:\,k') \;=\; \frac{p(\mathbf{y}\,|\,k)}{p(\mathbf{y}\,|\,k')},$$

where

$$p(\mathbf{y}\,|\,k) \;=\; \int_{\boldsymbol{\Theta}_k} p(\mathbf{y}\,|\,\boldsymbol{\theta}_k, k)\, p(\boldsymbol{\theta}_k\,|\,k)\, \mathrm{d}\boldsymbol{\theta}_k.$$

Nonetheless, to compute this integral for each model, in most cases, one should use Monte Carlo simulation methods, such as Markov Chain Monte Carlo (MCMC) and Sequential Monte Carlo (SMC) methods (see, e.g., Robert and Casella, 2004), that might be computationally expensive (see, e.g., Han and Carlin, 2001).

The Reversible Jump MCMC (RJ-MCMC) sampler proposed by Green (1995) make it possible to approximate the posterior distribution $p(k,\, \boldsymbol{\theta}_k \,|\, \mathbf{y})$ defined over a union of subspaces of differing dimensions. Green's RJ-MCMC sampler can be seen as a generalization of the well-known Metropolis-Hastings sampler (Metropolis et al., 1953 ; Hastings, 1970), which is capable of exploring not only the fixed-dimensional parameter spaces $\boldsymbol{\Theta}_k$, but also the space $\mathcal{K}$ of all models under consideration. However, in many applications, practical challenges remain in the process of making inference, from the generated samples, about the quantities of interest.

Relabeling and summarizing variable-dimensional posterior distributions

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 most signal decomposition problems, the main challenge encountered in the summarization process is switching of components' labels due to the invariance of the posterior distributions with respect to permutation of component labels. This issue is called "label-switching" in the literature and has mostly been investigated for Gaussian mixture models (see, e.g., Richardson and Green, 1997 ; Celeux et al., 2000 ; Stephens, 2000 ; Jasra et al., 2005). Because of this permutation invariance, all marginal posterior distributions of the component-specific parameters are equal, thus making posterior means, usually used for summarization, meaningless. To the best of our knowledge, all the methods proposed so far in the literature to solve the label-switching issue are restricted to the *fixed*-dimensional framework. (They have, however, been used in the trans-dimensional problems by, first, selecting a model, e.g., by the highest posterior probability, and then, summarizing the posterior distributions given the selected model.)

Figure 0.4 shows the marginal posterior distributions of the number k of muons (left) and sorted arrival times given k (right) obtained using an RJ-MCMC sampler on the example shown in the previous section. Each row corresponds to one value of k for $2 \leq k \leq 4$. Sorting the components—the simplest relabeling strategy (see, e.g., Richardson and Green, 1997)—based on their arrival times, i.e., $t_{\mu,1} < \ldots < t_{\mu,k}$, allows to break the symmetry in the posterior distribution. It can be seen from the figure that model $\mathcal{M}_3$ has the maximum posterior probability $p(k = 3 \,|\, \mathbf{y}) = 0.43$. However, by choosing $\mathcal{M}_3$, which is the result of using the Bayesian Model Selection (BMS) approach, all the samples corresponding to the other models would be discarded (note that $p(k = 2 \,|\, \mathbf{y}) = 0.38$ and $p(k = 4 \,|\, \mathbf{y}) = 0.16$). As a result, we would lose the uncertainty concerning the number k of muons. Moreover, summarizing the middle component under $\mathcal{M}_3$, (shown in light blue color), which is highly bimodal because of the effect of the trans-dimensional label-switching, by its posterior mean would be meaningless. Observe that, under $\mathcal{M}_4$, that component is split to two relatively compact components. Thus, when moving across models by the birth or death of a component, the corresponding label is either inserted or deleted. We call this issue “birth, death, and switching of labels”.

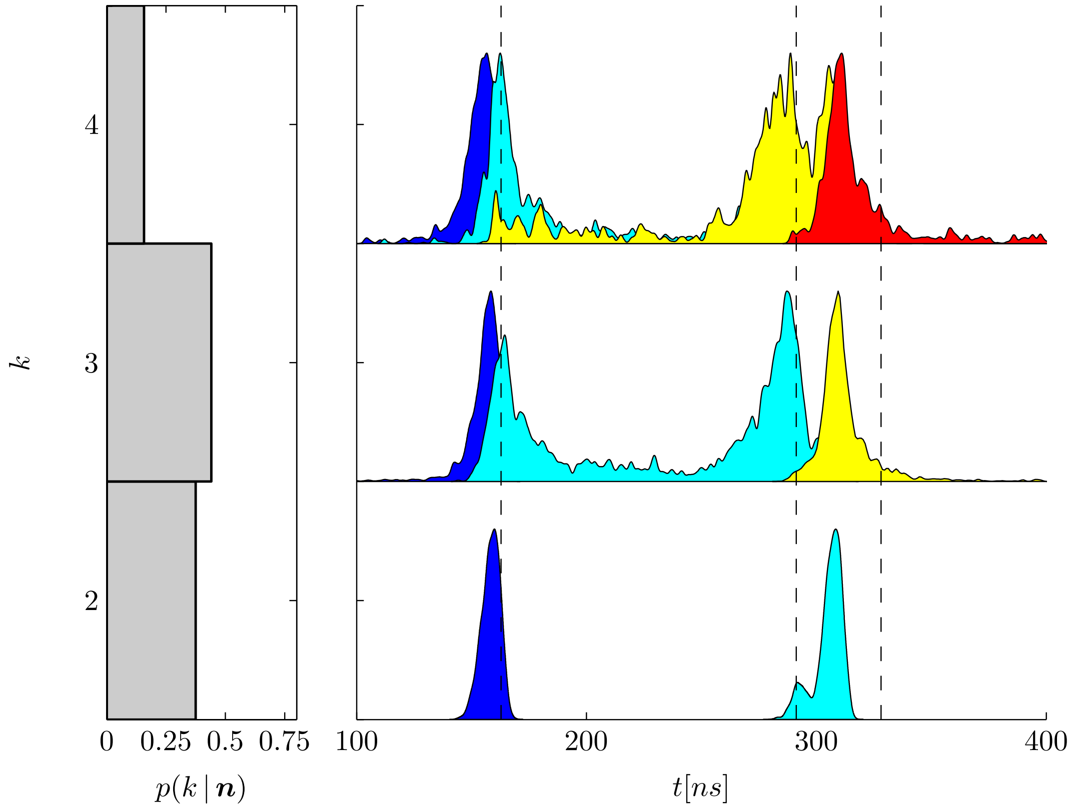


Figure 0.4 – *Posterior distributions of the number k of muons (left) and the sorted arrival times, $\boldsymbol{t}_{\mu}$, given k (right) constructed using 60 000 RJ-MCMC output samples. The true number of components is three. The vertical dashed lines in the right figure locate the arrival times, i.e., $\boldsymbol{t}_{\mu} = (163,\, 291,\, 328)$.*

The main contribution of this thesis is a novel approach for relabeling and summarizing variable-dimensional posterior distributions. It consists in fitting a new parametric model to the posterior distribution of interest encompassing all the uncertainties provided by the variable-dimensional samples generated, e.g., using the RJ-MCMC sampler.

Outline of the thesis

Chapter 1 briefly describes advanced Monte Carlo simulation techniques, such as MCMC and SMC methods, that are, nowadays, routinely used in the Bayesian literature, since they will be used to analyze the problems encountered throughout the thesis. Moreover, due to the existence of a lasting mistake in the computation of the acceptance ratio of “Birth-or-Death” moves, the most elementary type of trans-dimensional move, in the seminal paper of Andrieu and Doucet (1999, Equation(20)) and its followers in the signal processing literature (Andrieu et al., 2000, 2001a, 2002 ; Larocque and Reilly, 2002 ; Larocque et al., 2002 ; Ng et al., 2005 ; Davy et al., 2006 ; Rubtsov and Griffin, 2007 ; Shi et al., 2007 ; Melie-García et al., 2008 ; Ng et al., 2008 ; Hong et al., 2010 ; Schmidt and Mørup, 2010), we provide, in this chapter, clear statements of some mathematical results, certainly not completely new but never stated explicitly, which can be used for a clean justification of the acceptance ratio of Birth-or-Death moves in signal decomposition (and similar) problems.

Chapter 2 introduces the issue of birth, death, and switching of labels in variable-dimensional posterior distributions and describes the novel approach that we propose for relabeling and summarizing variable-dimensional posterior distributions. This approach consists in approximating the posterior of interest by a "simple"—but still variable-dimensional—parametric distribution, in the spirit of the relabeling approach developed by Stephens (2000). We fit this parametric model to the posterior distributions of interest through the minimization of a divergence measure between them. We consider, successively, the Kullback-Leibler (KL) divergence and a more robust divergence measure proposed by Basu et al. (1998). Stochastic EM-type algorithms, driven by the RJ-MCMC sampler, are developed to estimate the parameters of the approximate model.

Chapter 3 revisits the problem of joint Bayesian detection and estimation of sinusoidal components observed in white Gaussian noise. We first show the effect of using the erroneous Birth-or-Death acceptance ratio provided by Andrieu and Doucet (1999). We also briefly discuss the issue of prior specification for the signal-to-noise ratio hyperparameter and Bayesian sensitivity analysis. The main part of the chapter is devoted to investigating the capability of the summarizing approach we proposed in Chapter 2 for relabeling and summarization of the variable-dimensional posterior distributions encountered in this problem. More precisely, we illustrate the convergence and relabeling properties the proposed algorithms along with the goodness-of-fit of the fitted approximate model on three specific sinusoid detection examples. We also discuss, using simulations, some frequentist properties of the summaries obtained using the proposed approach.

Chapter 4 discusses the problem of joint detection and estimation of muons in the Auger project. As in Chapter 3, we investigate using this application the capability of the proposed summarizing approach in relabeling and summarizing the variable-dimensional posterior distributions. This study is conducted on simulated data kindly provided by Prof. Balázs Kégl. Moreover, in this chapter, we discuss issues concerning the initialization of the proposed SEM-type algorithms and the interpretation of the obtained summaries.

Finally, we conclude the thesis and give possible directions for future work.

List of publications

The publications that resulted from this work are as follows:

i) Alireza Roodaki, Julien Bect, and Gilles Fleury. Summarizing posterior distributions in signal decomposition problems when the number of components is unknown In 37th *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'12), Kyoto, Japan, March 25-30*, 2012.

   This paper briefly describes the proposed approach for relabeling and summarizing variable-dimensional posterior distributions with preliminary results on the problem of joint Bayesian detection and estimation of sinusoidal components in white Gaussian noise. The content of this paper is presented, in much more detail, in Chapters 2 and 3.

ii) Alireza Roodaki, Julien Bect, and Gilles Fleury. Note on the computation of the Metropolis-Hastings ratio for Birth-or-Death moves in trans-dimensional MCMC algorithms for signal decomposition problems. *Technical report, École Supérieur d'Électricité (Supélec), Gif-sur-Yvette, France, 2012.*

   This note provides results concerning the computation of the acceptance ratio in the Metropolis-Hastings algorithm, with a focus on the Birth-or-Death moves used in trans-dimensional MCMC samplers. The theoretical results provided in this note are expressed in Section 1.4.

iii) Alireza Roodaki, Julien Bect, and Gilles Fleury. An empirical Bayes approach for joint Bayesian model selection and estimation of sinusoids via reversible jump MCMC. In: *European signal Processing Conference (EUSIPCO'10), Aalborg , Denmark*, 2010.

iv) Alireza Roodaki, Julien Bect, and Gilles Fleury. On the joint Bayesian model selection and estimation of sinusoids via reversible jump MCMC in low SNR situations. In: 10th *International Conference on Information Sciences, Signal Processing and their Applications (ISSPA'10) Kuala Lumpur, Malaysia*, 2010.

   The last two papers address the issue of the prior specification over the signal-to-noise ratio hyperparameter in the problem of joint Bayesian detection and estimation of sinusoidal components in white Gaussian noise. Assigning a weakly-informative conjugate Inverse Gamma prior over it, as recommended in Andrieu and Doucet

(1999), the results provided in the above papers reveal that the value of its scale parameter has a significant influence on 1) the mixing rate of the Markov chain and 2) the posterior distribution of the number k of components. In iii), we investigated an Empirical Bayes approach to select an appropriate value for this hyperparameter in a data-driven way. In iv), we took a different approach and used a truncated Jeffreys prior. However, both approaches failed in low SNR situations, while in high SNR situations the sensitivity to β_{δ^2} is negligible.

This problem is briefly discussed in Section 3.2.5 of this thesis, where we propose to use an SMC sampler to study the sensitivity of the posterior distribution to the variations of this hyperparameter (following an idea of Bornn et al. (2010)). The papers are provided in Appendix B.

Chapter 1

Monte Carlo Sampling Methods

1.1 Introduction

1.1.1 Why are advanced sampling methods required?

Let π denotes the posterior distribution of interest defined over a measurable space $(\mathbb{X}, \mathcal{B})$ with vector $\boldsymbol{x} \in \mathbb{X}$. The space $\mathbb{X}$ might be quite general, and, in particular, it might contain some discrete and some continuous components, as in variable-dimensional problems discussed in Section 1.4. Given the observed data $\mathbf{y}$, suppose we are interested in computing the posterior expectation of a π-integrable function h written as

$$\begin{aligned} \mathbb{E}\{h(\boldsymbol{x}) \,|\, \mathbf{y}\} &= \int_{\mathbb{X}} h(\boldsymbol{x})\pi(\boldsymbol{x} \,|\, \mathbf{y})\mathrm{d}\boldsymbol{x} \\ &= \frac{\int_{\mathbb{X}} h(\boldsymbol{x})\mathbb{P}(\mathbf{y} \,|\, \boldsymbol{x})\pi_0(\boldsymbol{x})\mathrm{d}\boldsymbol{x}}{\int_{\mathbb{X}} \mathbb{P}(\mathbf{y} \,|\, \boldsymbol{x})\pi_0(\boldsymbol{x})\mathrm{d}\boldsymbol{x}}, \end{aligned} \tag{1.1}$$

where $\mathbb{P}(\mathbf{y} \,|\, \boldsymbol{x})$ and π_0 denote, respectively, the likelihood function and the assigned prior distribution.

Bayesian data analysis often involves high dimensional and/or intractable integrals when studying the posterior distributions' quantities of interest, such as the ones shown in (1.1)—making thus the inference infeasible. This has indeed been the main obstacle for Bayesian statisticians to use the Bayes approach for treating their problems. However, advanced computational methods developed in the previous decades—in parallel with developments in computing machines—have lead to major breakthroughs in Bayesian data analysis.

Assuming that the integrals in (1.1) are intractable, using "classical" Monte Carlo sampling methods (see, e.g., Robert and Casella, 2004), the posterior expectation (1.1) can be approximated by the empirical average

$$\widehat{\mathbb{E}}\{h(\boldsymbol{x}) \,|\, \mathbf{y}\} = \frac{1}{M}\sum_{m=1}^{M} h(\boldsymbol{x}^{(m)}), \tag{1.2}$$

where $\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(M)}$ are independent random variables generated from the target posterior distribution π. However, drawing random samples directly from π—which can be complex and/or known up to a normalizing constant—is not possible in most problems.

In this chapter, we will present advanced Monte Carlo sampling techniques extensively used in statistics, signal and image processing and machine learning to generate random samples from complex distributions.

1.1.2 Metropolis-Hastings algorithms for variable-dimensional problems

In many practical problems, the posterior distribution is of varying-dimensions where use of simple MCMC and IS methods is inappropriate. Green (1995) proposed a trans-dimensional MCMC sampler named Reversible Jump MCMC (RJ-MCMC) for generating samples from variable-dimensional posterior distributions. Green's RJ-MCMC sampler can be seen as a generalization of the well-known Metropolis-Hastings sampler (Metropolis et al., 1953 ; Hastings, 1970), which is capable of exploring not only the fixed-dimensional parameter spaces but also the space of all models under consideration. At the heart of this algorithm lies an accept/reject mechanism, with an acceptance ratio calibrated in such a way that the invariant distribution of the chain is the target distribution π. The computation of this acceptance ratio for trans-dimensional moves is in general a delicate issue, involving measure theoretic considerations. (Fortunately, the simple and powerful "dimension matching" argument of Green (1995) allows to bypass this difficulty for a large class of proposal distributions.)

Andrieu and Doucet (1999) pioneered the use of RJ-MCMC sampling in "signal decomposition" problems, by tackling joint model selection and parameter estimation for an unknown number of sinusoidal signals observed in white Gaussian noise. (At the same period, RJ-MCMC also became popular for image processing tasks such as segmentation and object recognition; see, e.g., (Hurn and Rue, 1997 ; Nicholls, 1998 ; Pievatolo and Green, 1998 ; Rue and Hurn, 1999 ; Descombes et al., 2001).) This seminal paper was followed by many others in the signal processing literature (Andrieu et al., 2000, 2001a, 2002 ; Larocque and Reilly, 2002 ; Larocque et al., 2002 ; Ng et al., 2005 ; Davy et al., 2006 ; Rubtsov and Griffin, 2007 ; Shi et al., 2007 ; Melie-García et al., 2008 ; Ng et al., 2008 ; Hong et al., 2010 ; Schmidt and Mørup, 2010), relying systematically on the original paper Andrieu and Doucet (1999) for the computation of the acceptance ratio of "Birth-or-Death" moves—the most elementary type of trans-dimensional move, which either adds or removes a component from the signal decomposition. Unfortunately, the expression of the acceptance ratio for Birth-or-Death moves provided by (Andrieu and Doucet, 1999, Equation (20)) turns out to be erroneous, as will be explained later. Worse, the exact same mistake has been reproduced in most of the following papers, referred to above.

1.1.3 Outline of the chapter

This chapter is organized as follows. In Section 1.2, Markov Chain Monte Carlo (MCMC) methods are introduced with a brief description of the properties of Markov chains that are essential for the study of MCMC methods. Moreover, two well-known fixed-dimensional MCMC samplers, namely, the Metropolis-Hastings (MH) and Gibbs samplers, are described. Next, Importance Sampling (IS) based methods are described in Section 1.3 where we specifically explain Sequential Monte Carlo (SMC) samplers that have found many applications in scenarios that the distribution of interest is evolving over "time". A toy example is provided to illustrate how MCMC and SMC sampler work in practice. Owing to the existence of the lasting mistake in the computation of birth-or-death move's acceptance ratio in Andrieu and Doucet (1999), Section 1.4 is devoted to explain elaborately the procedure of between-models moves in trans-dimensional MCMC samplers. Finally, Section 1.5 summarizes the arguments discussed in this Chapter.

1.2 Markov Chain Monte Carlo methods

1.2.1 Basic principles of MCMC methods

A Markov chain in discrete time is a sequence of random variables $\left(\boldsymbol{x}^{(n)}\right)_{n\geq 0} = \left(\boldsymbol{x}^{(0)}, \boldsymbol{x}^{(1)}, \ldots\right)$, with $\boldsymbol{x}^{(n)} \in \mathbb{X}$, respecting the *Markovian property*, that is, conditional on the current state $\boldsymbol{x}^{(n)}$, the distribution of the next state $\boldsymbol{x}^{(n+1)}$ is independent of the previous states $\left(\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}, \ldots, \boldsymbol{x}^{(n-1)}\right)$. A time-homogeneous Markov chain can be more formally specified using its *transition kernel* defined as

Definition 1.1. *A transition kernel is a function P on $\mathbb{X} \times \mathcal{B}$ such that*

i) $\forall \boldsymbol{x} \in \mathbb{X}, P(\boldsymbol{x}, \cdot)$ *is a probability measure;*

ii) $\forall A \in \mathcal{B}, P(\cdot, A)$ *is measurable.*

Note that in the time-inhomogeneous case the transition kernel itself depends on the index n of the current state. However, throughout this thesis, we only consider the time-homogeneous case unless otherwise stated. Then, the conditional distribution of $\boldsymbol{x}^{(n+1)}$

given the previous states is

$$\mathbb{P}\left(\boldsymbol{x}^{(n+1)} \in A \mid \boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}, \ldots, \boldsymbol{x}^{(n)}\right) = \mathbb{P}\left(\boldsymbol{x}^{(n+1)} \in A \mid \boldsymbol{x}^{(n)}\right) = P\left(\boldsymbol{x}^{(n)}, A\right).$$

In what follows, the properties of Markov chains that are essential for the study of MCMC methods are briefly described. For further information see Meyn and Tweedie (1993) ; Tierney (1994, 1998) ; Robert and Casella (2004) ; Roberts and Rosenthal (2004) ; Liu (2001) ; Roberts and Rosenthal (2006). Let ν be a positive measure on $(\mathbb{X}, \mathcal{B})$.

Definition 1.2. *A Markov chain is said to have invariant or stationary distribution ν if $\nu = \nu P$, where $(\nu P)(A) \triangleq \int \nu(\mathrm{d}\boldsymbol{x})P(\boldsymbol{x}, A)$, for all measurable sets $A \in \mathcal{B}$.*

Definition 1.3. *A Markov chain is ν-irreducible if for all $A \in \mathcal{B}$, $\nu(A) > 0$ induces $P(\boldsymbol{x}, A) > 0$ for all $x \in \mathbb{X}$.*

In other words, a ν-irreducible Markov chain, starting from any state, is able to visit any $A \in \mathcal{B}$, such that $\nu(A) > 0$, in a finite number of steps.

Definition 1.4. *Let $\left(\boldsymbol{x}^{(n)}\right)_{n\geq 0}$ be a ν-irreducible Markov chain on $\mathbb{X}$. Then, the transition kernel P is periodic if there exist an integer $d \geq 2$ and a sequence $\{A_1, \ldots, A_d\}$ of d nonempty disjoint sets in $\mathcal{B}$ (a "d-cycle") such that*

i) for $\boldsymbol{x} \in A_i$, $P(\boldsymbol{x}, A_{i+1}) = 1$, $i = 0, \ldots, d-1$ (mod d);

ii) the set $\left(\bigcup_{i=1}^{d} A_i\right)^c$ is ν-null.

Otherwise, the kernel is aperiodic.

Definition 1.5. *A π-irreducible Markov chain $\left(\boldsymbol{x}^{(n)}\right)_{n\geq 0}$ with the invariant distribution π is recurrent if, for any $A \in \mathcal{B}$ with $\pi(A) > 0$, the probability of visiting A infinitely often, denoted by $\mathbb{P}(A \text{ i.o.} \,|\, \boldsymbol{x}^{(0)} = \boldsymbol{x})$, is positive for all $\boldsymbol{x}$ and equals to one for π-almost all $\boldsymbol{x}$. The chain is Harris recurrent if $\mathbb{P}(A \text{ i.o.} \,|\, \boldsymbol{x}^{(0)} = \boldsymbol{x}) = 1$ for all $\boldsymbol{x}$. The chain is called positive recurrent if π is a proper measure.*

We can now state the following theorem (taken from (Tierney, 1994, Theorem 1) with appropriate notational modifications):

Theorem 1.6. *If the transition kernel P is π-irreducible and $\pi = \pi P$, then P is positive recurrent and π is its unique invariant distribution. If, in addition, P is aperiodic, then, the chain converges in total variation to π for π-almost all starting states $\boldsymbol{x}$, that is,*

$$\| P^n(\boldsymbol{x}, \cdot) - \pi \|_{TV} \to 0,$$

where $\| \cdot \|_{TV}$ denotes the total variation distance[1]*. If P is also Harris recurrent, then convergence occurs for all initial distributions (see also Robert and Casella, 2004, Theorem 6.51).*

Definition 1.7. *A Markov chain is ergodic if it is positive Harris recurrent and aperiodic.*

Now, we can state the following theorem concerning the *ergodic* Markov chains (Tierney, 1994, Theorem 3):

Theorem 1.8. *If $\left(\boldsymbol{x}^{(n)}\right)_{n\geq 0}$ is an ergodic Markov chain with invariant distribution π, and assuming that h is a real-valued function such that $\int_{\mathbb{X}} |h(\boldsymbol{x})|\pi(\mathrm{d}\boldsymbol{x}) < \infty$, then,*

$$\frac{1}{M}\sum_{m=1}^{M} h(\boldsymbol{x}^{(m)}) \xrightarrow[M\to+\infty]{} \int_{\mathbb{X}} h(\boldsymbol{x})\pi(\mathrm{d}\boldsymbol{x}).$$

This theorem asserts that under regularity assumptions on h, the sample path average (1.2) will converge almost surely to the integral (1.1) if the samples $\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(M)}$ are generated according to an ergodic Markov chain with invariant distribution π.

MCMC sampling methods, generally, proceed by constructing a time-homogeneous Markov chain $\left(\boldsymbol{x}^{(n)}\right)_{n\geq 0}$ with invariant distribution π, using a transition kernel P fulfilling the conditions of Theorem 1.6. One sufficient, but not necessary, condition to ensure that π is the invariant distribution of the transition kernel P, is the *reversibility* of P with respect to π. A kernel that satisfies the detailed balance condition

$$\pi\left(\mathrm{d}\boldsymbol{x}\right) P\left(\boldsymbol{x}, \mathrm{d}\boldsymbol{x}'\right) \;=\; \pi\left(\mathrm{d}\boldsymbol{x}'\right) P\left(\boldsymbol{x}', \mathrm{d}\boldsymbol{x}\right), \tag{1.3}$$

is reversible. For all measurable sets $A \in \mathcal{B}$, integrating (1.3) on $\mathbb{X} \times A$ yields

$$\int_{\mathbb{X}} \pi\left(\mathrm{d}\boldsymbol{x}\right)\, P\left(\boldsymbol{x},\, A\right) \;=\; \pi\left(A\right),$$

which means that π is an invariant distribution for the kernel P (it is also said that "P leaves π invariant").

Remark 1.1. Some of the above requirements on the chain $\left(\boldsymbol{x}^{(n)}\right)_{n\geq 0}$ can be relaxed. Most notably, time-inhomogeneous chains are used in the context of "adaptive MCMC" algorithms; see, e.g., Atchadé and Rosenthal (2005) ; Andrieu and Moulines (2006) ; Roberts and Rosenthal (2009) and the references therein . It is also possible to depart

[1] Assume that μ_1 and μ_2 are two probability measures on the measurable space $(\mathbb{X}, \mathcal{B})$. Then, for $A \in \mathcal{B}$, the total variation norm is

$$\| \mu_1 - \mu_2 \|_{TV} \;=\; \sup_A \,|\, \mu_1(A) - \mu_2(A) \,|.$$

from the reversibility assumption, which a sufficient but not necessary condition for π to be an invariant distribution (see, e.g., Diaconis et al. (2000)), though the vast majority of MCMC algorithms considered in the literature are based on reversible kernels.

Practical considerations: burn-in and convergence monitoring

Despite the fact that theoretical results prove the convergence of MCMC methods provided satisfying the conditions of Theorem 1.6, in practice, with naturally finite number of simulated samples, one should care about the properties of the MCMC sampler. In fact, Theorems 1.6 and 1.8 state asymptotic results advocating validity of MCMC algorithms in theory. However, they do not provide adequate information to answer the following questions concerning the Markov chain under study: how to choose the starting point, i.e., $\boldsymbol{x}^{(0)}$? When to stop the algorithm? What is the rate of convergence? Has the chain visited the entire support—or even likely regions—of the target distribution π ?

The first point is the initialization of the sampler. Although the Markov chains used in MCMC methods are assumed to be π-irreducible by construction and, consequently, it is unnecessary to worry about starting points, it turns out that in practice starting points become quite influential. Gelman et al. (2004, Chapter 12) discuss techniques for approximating the target distribution to assess appropriate starting points for the MCMC algorithms. Nonetheless, often, in high dimensional complex problems, there is no generic approach to choose the initial state of the Markov chain $\boldsymbol{x}^{(0)}$. Therefore, in order to reduce the dependence of the Markov chain to the initial points, it is conventional to discard a portion of the whole generated samples from the beginning of the chain. These discarded samples are called *burn-in* period in the literature.

Moreover, there is no standard stopping rule in MCMC algorithms in the case of complex problems. However, there are several methods in the literature to monitor the convergence of the Markov chains; see for example Cowles and Carlin (1996) and Mengersen and Robert (1999) for comprehensive reviews of the existing methods. Despite these methods need usually problem-specific analytical work and programming, which can be difficult, intricate, or even impossible in certain cases, none of them is foolproof. It is indeed concluded by Cowles and Carlin (1996, Section 5) that "... although many of the diagnostics often succeed at detecting the sort of convergence failure they were designed to identify, they can also fail in this role—even in low-dimensional idealized problems far simpler than those typically encountered in statistical practice." Therefore, in this document, as many other work, the behavior of the Markov chain is empirically assessed by means of monitoring the graphical plots of the evolution of the parameters and the corresponding autocorrelation function (see, e.g., Roberts and Rosenthal, 2001 ; Thompson, 2010).

1.2.2 Two fixed-dimensional MCMC samplers

This Section presents two well-known fixed-dimensional MCMC samplers, namely, the Metropolis-Hastings and Gibbs sampler, since they are used throughout this work and, furthermore, they can be considered as basis for trans-dimensional MCMC algorithms described later in Section 1.4.

The Metropolis-Hastings sampler

The very popular Metropolis-Hastings kernels proposed by Metropolis et al. (1953) and Hastings (1970) correspond to the following two-stage sampling procedure: first, given that the current state of the Markov chain is $\boldsymbol{x} \in \mathbb{X}$, a new state $\boldsymbol{x}' \in \mathbb{X}$ is proposed from a proposal transition kernel $Q\left(\boldsymbol{x},\, \mathrm{d}\boldsymbol{x}'\right)$; second, this move is accepted with probability $\alpha\left(\boldsymbol{x},\, \boldsymbol{x}'\right)$ and rejected otherwise—in which case the new state is equal to $\boldsymbol{x}$. More formally, for all $\boldsymbol{x} \in \mathbb{X}$ and $B \in \mathcal{B}$, the transition kernel is given by

$$P\left(\boldsymbol{x},\, B\right) \;=\; \int_B Q\left(\boldsymbol{x},\, \mathrm{d}\boldsymbol{x}'\right)\, \alpha\left(\boldsymbol{x}, \boldsymbol{x}'\right) \;+\; s\left(\boldsymbol{x}\right) \mathbb{1}_B\left(\boldsymbol{x}\right), \tag{1.4}$$

where $\mathbb{1}_B$ denotes the indicator function of B, and

$$s\left(\boldsymbol{x}\right) \;=\; \int_{\mathbb{X}} Q\left(\boldsymbol{x},\, \mathrm{d}\boldsymbol{x}'\right)\, \left(1 - \alpha\left(\boldsymbol{x}, \boldsymbol{x}'\right)\right)$$

is the probability of rejection at $\boldsymbol{x}$. It is easily seen that the detailed balance condition (1.3) holds if and only if (Tierney, 1994 ; Green, 1995 ; Tierney, 1998)

$$\pi\left(\mathrm{d}\boldsymbol{x}\right) Q\left(\boldsymbol{x}, \mathrm{d}\boldsymbol{x}'\right) \alpha\left(\boldsymbol{x}, \boldsymbol{x}'\right) \;=\; \pi\left(\mathrm{d}\boldsymbol{x}'\right) Q\left(\boldsymbol{x}', \mathrm{d}\boldsymbol{x}\right) \alpha\left(\boldsymbol{x}', \boldsymbol{x}\right). \tag{1.5}$$

This is achieved, for instance, by the acceptance probability

$$\alpha\left(\boldsymbol{x}, \boldsymbol{x}'\right) \;=\; \min\left\{1,\, r\left(\boldsymbol{x}, \boldsymbol{x}'\right)\right\}, \tag{1.6}$$

where $r(\boldsymbol{x}, \boldsymbol{x}')$ denotes the Metropolis-Hastings-Green (MHG) ratio

$$r\left(\boldsymbol{x}, \boldsymbol{x}'\right) \;=\; \frac{\pi\left(\mathrm{d}\boldsymbol{x}'\right) Q\left(\boldsymbol{x}', \mathrm{d}\boldsymbol{x}\right)}{\pi\left(\mathrm{d}\boldsymbol{x}\right) Q\left(\boldsymbol{x}, \mathrm{d}\boldsymbol{x}'\right)}\cdot \tag{1.7}$$

The right-hand side of (1.7) is the Radon-Nykodim derivative of $\pi\left(\mathrm{d}\boldsymbol{x}'\right) Q\left(\boldsymbol{x}', \mathrm{d}\boldsymbol{x}\right)$ with respect to $\pi\left(\mathrm{d}\boldsymbol{x}\right) Q\left(\boldsymbol{x}, \mathrm{d}\boldsymbol{x}'\right)$; see Tierney (1998, Section 2) for technical details. In fact, the general form of the acceptance ratio presented is (1.7) is also valid in the trans-dimensional case (Green, 1995) which we will discuss later in Section 1.4. Assuming further that the posterior distribution π and the proposal transition kernel $Q\left(\boldsymbol{x},\, \mathrm{d}\boldsymbol{x}'\right)$ admit densities denoted by f and $q(\boldsymbol{x},\, \boldsymbol{x}')$, respectively, with respect to a σ-finite measure denoted $\mathrm{d}\boldsymbol{x}$, the ratio can be written as

$$r\left(\boldsymbol{x}, \boldsymbol{x}'\right) \;=\; \frac{f\left(\boldsymbol{x}'\right) q\left(\boldsymbol{x}',\, \boldsymbol{x}\right)}{f\left(\boldsymbol{x}\right) q\left(\boldsymbol{x}, \boldsymbol{x}'\right)}. \tag{1.8}$$

Remark 1.2. It is proved in Tierney (1998, Section 4) that the acceptance probability (1.6) is optimal in the sense of minimizing the asymptotic variance of sample path averages among all acceptance rates satisfying (1.5).

The general MH sampler is presented in the following pseudo-code:

```
Algorithm 1.1. Metropolis-Hastings sampler

Initialization  Select randomly or deterministically $\boldsymbol{x}^{(0)}$.

For $n \geq 0$ iterate  Given $\boldsymbol{x}^{(n)}$,

        i) Generate $\boldsymbol{x}' \sim Q(\boldsymbol{x}^{(n)},\, \cdot\,)$.
        ii) Generate an auxiliary uniform variable $u \sim \mathcal{U}\left(0,1\right)$.
        iii) $\boldsymbol{x}^{(n+1)} = \begin{cases} \boldsymbol{x}' & \text{if } \alpha\left(\boldsymbol{x}^{(n)}, \boldsymbol{x}'\right) > u\,, \\ \boldsymbol{x}^{(n)} & \text{otherwise,} \end{cases}$
           where $\alpha\left(\boldsymbol{x}^{(n)}, \boldsymbol{x}'\right)$ is the acceptance probability defined in (1.6).
```

Several MH samplers can be derived by using proposal distributions $q\left(\boldsymbol{x},\, \cdot\,\right)$ of different natures. For instance, provided that the proposal distribution is symmetric, that is, $q\left(\boldsymbol{x}, \boldsymbol{x}'\right) = q\left(\boldsymbol{x}', \boldsymbol{x}\right)$, such as symmetric random walk proposal, then, we recover the Metropolis sampler with the simplified acceptance ratio $r\left(\boldsymbol{x}, \boldsymbol{x}'\right) \;=\; f\left(\boldsymbol{x}'\right) / f\left(\boldsymbol{x}\right)$. Moreover, the Independent MH (I-MH) sampler is achieved by using a proposal distribution which does not depend on the current state $\boldsymbol{x}$, i.e., $q\left(\boldsymbol{x}, \boldsymbol{x}'\right) = q\left(\boldsymbol{x}'\right)$. Then, the I-MH acceptance ratio reads $r\left(\boldsymbol{x}, \boldsymbol{x}'\right) \;=\; \frac{f(\boldsymbol{x}')q(\boldsymbol{x})}{f(\boldsymbol{x})q(\boldsymbol{x}')}$.

Sufficient conditions for the Markov chain constructed by Algorithm 1.1 to satisfy the conditions of Theorem 1.6 are given in the following proposition (see Robert and Casella, 2004, Section 7.3.2):

Proposition 1.9. *i) It enjoys the aperiodicity property if the algorithm allows rejection of the proposed moves with non zero probability, i.e., $s(\boldsymbol{x}) > 0$ π-almost everywhere.*

ii) It is π-irreducible provided that the proposal distribution is positive on the support of the target distribution π, that is, for every $(\boldsymbol{x}, \boldsymbol{x}') \in \mathbb{X} \times \mathbb{X}$ such that $\pi(\boldsymbol{x}') > 0$, then, $q(\boldsymbol{x}, \boldsymbol{x}') > 0$.

The Gibbs sampler

The Gibbs sampler is one of the most famous MCMC samplers proposed in the seminal paper Geman and Geman (1984); see Gelfand and Smith (1990) and Casella and George (1992) for statistical discussion. Because of its simplicity, it has been used in many Bayesian data analysis problems, specifically, when conditionally conjugate prior distributions are used.

Suppose that, for some $r > 1$, the vector of unknown parameters $\boldsymbol{x} \in \mathbb{X}$ can be partitioned into (possibly multidimensional) blocks $(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_r) \in \mathbb{X}_1 \times \ldots \times \mathbb{X}_r$. For convenience, we introduce the notation

$$\boldsymbol{x}_{-i} \triangleq (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{i-1}, \boldsymbol{x}_{i+1}, \ldots, \boldsymbol{x}_r)$$

to indicate the vector of parameters $\boldsymbol{x}$ without the i^{th} block. Then, the conditional distribution $f(\boldsymbol{x}_i \mid \boldsymbol{x}_{-i})$ is called the full conditional distribution of the block $\boldsymbol{x}_i$. Assume further that it is possible to sample (directly) form the full conditional distributions

$$f(\boldsymbol{x}_1 \mid \boldsymbol{x}_{-1}), \ldots, f(\boldsymbol{x}_i \mid \boldsymbol{x}_{-i}), \ldots, f(\boldsymbol{x}_r \mid \boldsymbol{x}_{-r}) \cdot$$

Then, the Gibbs sampler algorithm is as follows:

Algorithm 1.2. *One iteration of the Gibbs sampler*

```
    Given $\boldsymbol{x}^{(n)} = \left(\boldsymbol{x}_1^{(n)}, \ldots, \boldsymbol{x}_r^{(n)}\right)$, generate $\boldsymbol{x}^{(n+1)}$ in the following $r$ steps
Step 1. $\boldsymbol{x}_1^{(n+1)} \sim f\left(\boldsymbol{x}_1 \mid \boldsymbol{x}_2^{(n)}, \ldots, \boldsymbol{x}_r^{(n)}\right)$,
Step 2. $\boldsymbol{x}_2^{(n+1)} \sim f\left(\boldsymbol{x}_2 \mid \boldsymbol{x}_1^{(n+1)}, \boldsymbol{x}_3^{(n)}, \ldots, \boldsymbol{x}_r^{(n)}\right)$,
  ⋮
Step r. $\boldsymbol{x}_r^{(n+1)} \sim f\left(\boldsymbol{x}_r \mid \boldsymbol{x}_1^{(n+1)}, \ldots, \boldsymbol{x}_{r-1}^{(n+1)}\right) \cdot$
```

Remark 1.3. Note that each step of the Gibbs sampler presented in Algorithm 1.2 can be regarded as a MH step where both the target and proposal distributions are equal to the corresponding full conditional distribution. Thus, the acceptance ratio (1.7) is one. Conversely, the MH sampler can be used in any step of Algorithm 1.2 when direct sampling is impossible. The resulting algorithm is then called Metropolis-within-Gibbs sampler; see, e.g., Roberts and Rosenthal (2006) or Robert and Casella (2004, Section 10.3).

1.2.3 A toy example

We present a toy Bayesian example in order to illustrate the aforementioned MCMC algorithms and terminology. The objective of this example is to estimate the mean of a normal distribution from an observed sample $\mathbf{y}$ of length N. So, $y_1, \ldots, y_N \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, 1)$, where $\mathcal{N}(a, b)$ denotes a normal distribution with a and b as its mean and variance parameters, respectively. From the Bayes formula, we obtain the posterior distribution of the mean μ given the observed data $\mathbf{y}$

$$p(\mu \,|\, \mathbf{y}) \;=\; \frac{p(\mathbf{y} \,|\, \mu) \cdot p(\mu)}{p(\mathbf{y})},$$

where $p(\mu)$ is the prior distribution assigned over μ and $p(\mathbf{y})$ is the marginal distribution of $\mathbf{y}$

$$p(\mathbf{y}) \;=\; \int p(\mathbf{y} \,|\, \mu) \cdot p(\mu) \mathrm{d}\mu.$$

Though it is natural to put a conjugate prior over μ, in this example, for instructive reasons, a non-conjugate standard Cauchy distribution, that is,

$$p(\mu) \;=\; \frac{1}{\pi\,(1 + \mu^2)}, \tag{1.9}$$

is assigned as a prior distribution over the mean parameter μ. Note that in this case the marginal distribution $p(\mathbf{y})$ cannot be computed analytically. Thus, the posterior distribution of the mean μ is available only up to a normalizing constant which reads

$$p(\mu \,|\, \mathbf{y}) \;\propto\; \frac{\exp\left(-\frac{\|\mathbf{y} - \mu \mathbb{1}_N\|^2}{2}\right)}{1 + \mu^2}, \tag{1.10}$$

where $\propto$ denotes proportionality.

It is not possible to directly draw samples from this posterior distribution, but, correlated samples can be generated from the target distribution (1.10) by constructing a Markov chain $(\mu^{(n)})_{n \geq 0}$ which leaves the posterior distribution (1.10) invariant. For this purpose, a proposal distribution $q(\mu,\, \mu')$ has to be designed, first. Here, we use a normal random walk proposal distribution centered at the current state of the Markov chain, that is, $\mu' \sim \mathcal{N}(\mu,\, \sigma^2)$. Therefore, owing to symmetricity of the proposal distribution q, the acceptance ratio becomes

$$r\,(\mu,\, \mu') \;=\; \frac{p(\mu' \,|\, \mathbf{y})}{p(\mu \,|\, \mathbf{y})}.$$

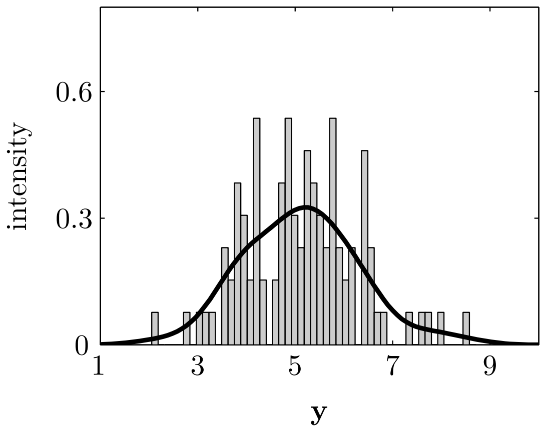


Figure 1.1 – *Histogram and kernel density estimate of the observed data.*

In the experiment, we set the true mean $\mu = 5$ and generated $N = 100$ i.i.d. samples from $\mathcal{N}(5, 1)$ which serve as the observed signal $\mathbf{y}$. Figure 1.1 illustrates the distribution of the observed data. The Markov chain was initialized, deliberately, to the value of $\mu^{(0)} = 2$ (far away from the true value), to clearly see how the algorithm approaches to the true value. Indeed, it is well-known that the standard deviation of the proposal distribution σ controls the rate of convergence and mixing of the chain (see for example Roberts and Rosenthal, 2001). For very "small" values of σ, the proposed jumps, which are mostly accepted, will be too short to explore rapidly the space and, thus, the convergence time will be so long. On the other hand, when σ is set to an extremely "large" value, the sampler will propose large jumps, even to regions of low posterior density. Thus, the acceptance probability will be low and the sampler will stand still for many iterations. Here, we used three different values for σ to show this fact. The length of the chain were set to $M = 5000$ and the first 1000 samples were discarded as burn-in period.

Figures 1.2, 1.3, and 1.4 illustrate the performance of the normal random walk sampler for the cases of "small" $\sigma = 0.01$, "large" $\sigma = 10$, and "good" $\sigma = 0.25$, respectively. From Theorem 1.8 and descriptions in Section 1.2.1, we chose to monitor the sample path average and the 25^{th} and 75^{th} percentiles of the chain, which are shown with red and green lines, respectively, to assess the behavior of the chain. One way to determine that the Markov chain has converged to its stationary distribution is to look for the locations where the sample path average and the percentiles become constant. This can be more easily detected in the zoomed figures demonstrated in the middle left panels. Moreover, the mean acceptance probability are shown in the right panel of the middle row to highlight the effect of the scale parameter σ. Finally, the bottom right panel in the figures demonstrate the autocorrelation function indicating the "mixing" of the Markov chains.

It can be inferred from the figures that the proposal distribution's scale parameter σ has a significant influence on the behavior of the Markov chain. This can be observed, for example, from the AFC plots shown in the bottom right panels of the figures; for either small or large σ, the ACF decayed very slowly. Whereas, for the case of "good" scale parameter $\sigma = 0.25$, the chain mixes and explores the support of the target distribution rapidly and, thus, the autocorrelation fades to zeros after a few lags.

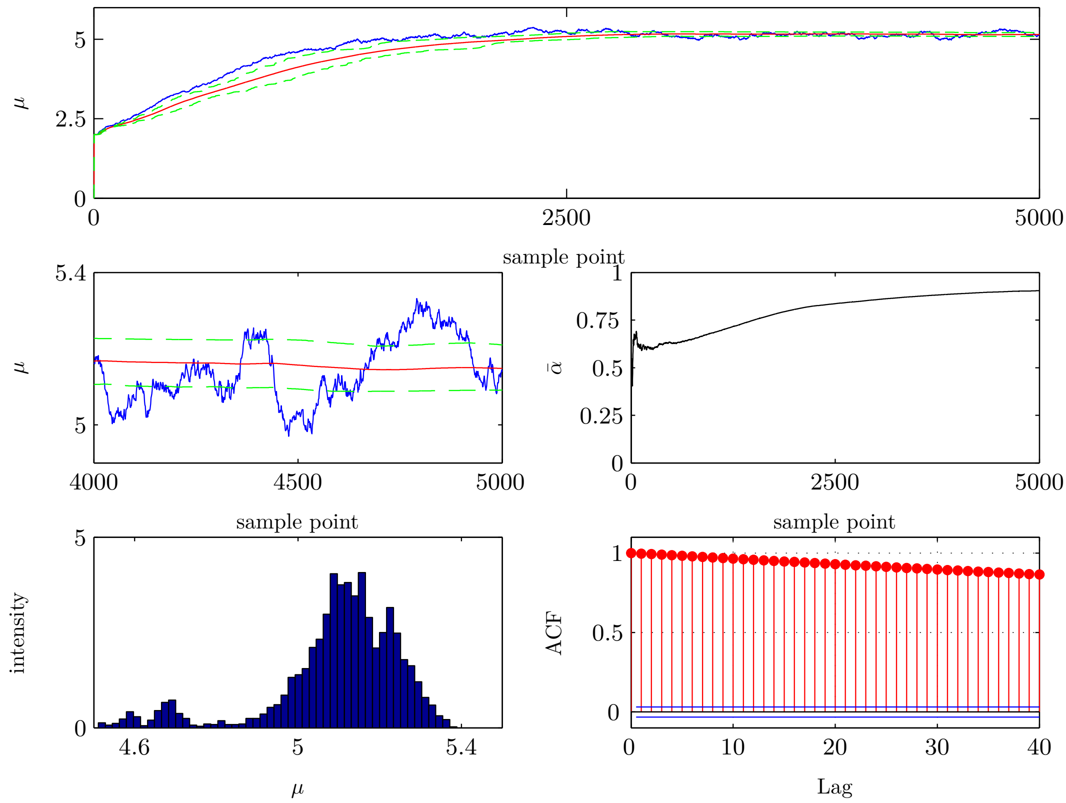


Figure 1.2 – *Performance of the normal random walk sampler with* $\sigma = 0.01$ *on the toy example. The top figure show the Markov chain* $(\mu^{(n)})_{n\geq 0}$ *(blue line), its average (red line),* 25^{th}*, and* 75^{th} *percentiles (green dashed line) (the left panel of middle row is a zoomed version of the last 1000 iterations of the top figure). The middle row right panel illustrates the mean acceptance probability. The bottom figures show the histogram intensity (on the left) and the ACF (on the right) of the output chain after discarding the first 1000 samples as burn-in period. The estimated mean is* $\hat{\mu} = 5.15$ *with the mean acceptance probability of* $\bar{\alpha} = 0.9$*.*

The sensitivity issue of the MCMC samplers to the parameters of the proposal distributions has motivated many researchers for developing adaptive MCMC methods; see, e.g., Haario et al. (2001) ; Atchadé and Rosenthal (2005) ; Andrieu and Moulines (2006) ; Roberts and Rosenthal (2009) and references therein for more information. In the spe-

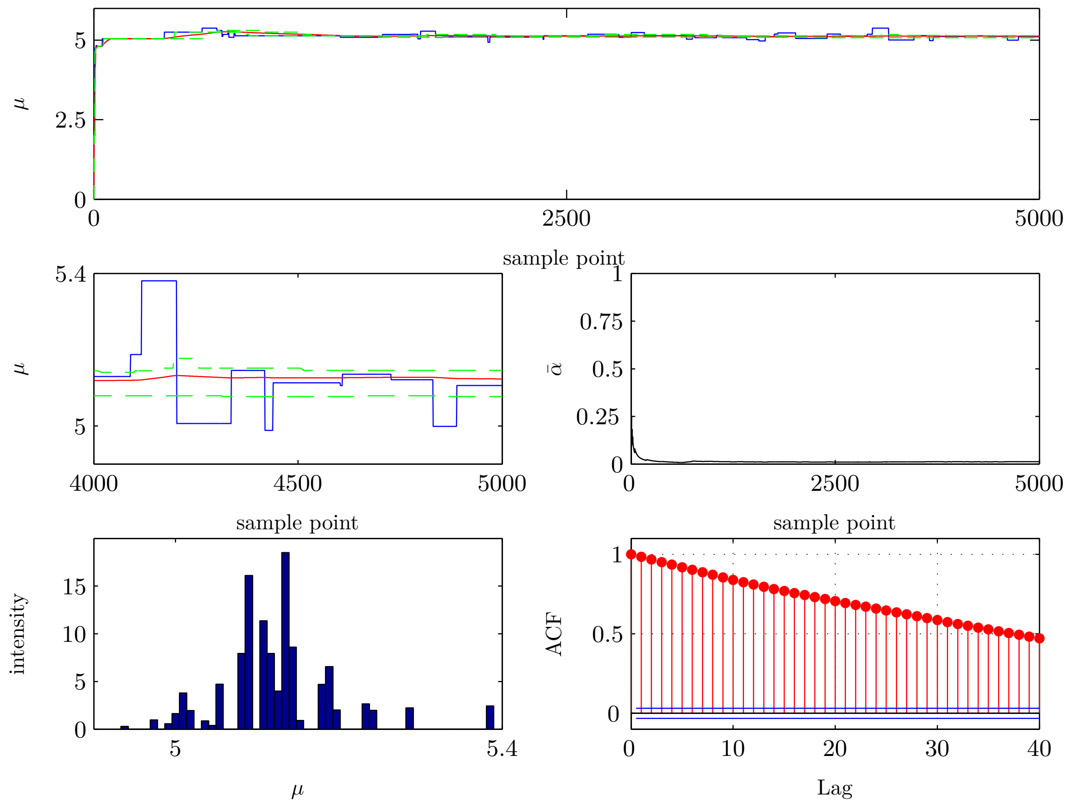


Figure 1.3 – *Performance of the normal random walk sampler with $\sigma = 10$ on the toy example. The top figure show the Markov chain $(\mu^{(n)})_{n \geq 0}$ (blue line), its average (red line), 25^{th}, and 75^{th} percentiles (green dashed line) (the left panel of middle row is a zoomed version of the last 1000 iterations of the top figure). The middle row right panel illustrates the mean acceptance probability. The bottom figures show the histogram intensity (on the left) and the ACF (on the right) of the output chain after discarding the first 1000 samples as burn-in period. The estimated mean is $\hat{\mu} = 5.13$ with the mean acceptance probability of $\bar{\alpha} = 0.05$.*

cial case of normal random walk Metropolis sampler, Gelman et al. (1996) ; Roberts and Rosenthal (2001) provided results for optimal scaling of proposal distribution. The mean acceptance rate corresponding to the case of $\sigma = 0.25$ is close to 0.44 which is the optimal value for one dimensional normal random walk sampler (note that its optimal value in higher dimensions becomes 0.234 (see, e.g., Roberts and Rosenthal, 2001)).

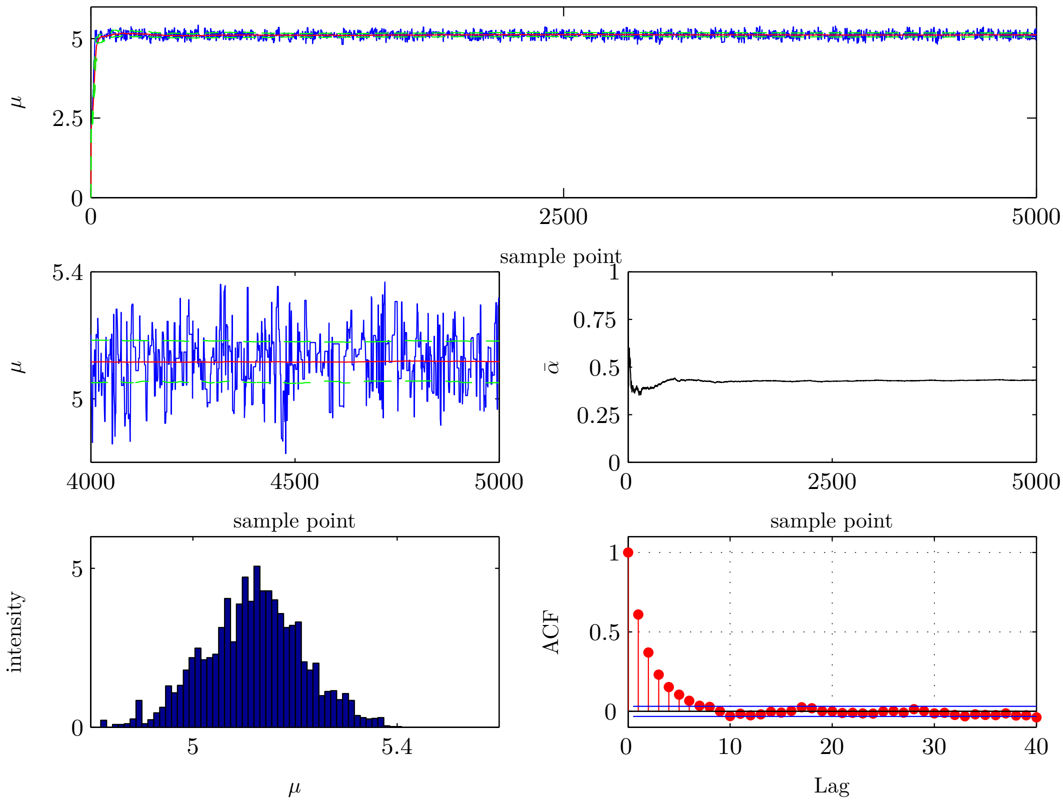


Figure 1.4 – *Performance of the normal random walk sampler with* $\sigma = 0.25$ *on the toy example. The top figure show the Markov chain* $(\mu^{(n)})_{n\geq 0}$ *(blue line), its average (red line),* 25^{th}*, and* 75^{th} *percentiles (green dashed line) (the left panel of middle row is a zoomed version of the last 1000 iterations of the top figure). The middle row right panel illustrates the mean acceptance probability. The bottom figures show the histogram intensity (on the left) and the ACF (on the right) of the output chain after discarding the first 1000 samples as burn-in period. The estimated mean is* $\hat{\mu} = 5.11$ *with the mean acceptance probability of* $\bar{\alpha} = 0.44$*.*

1.3 Importance sampling and sequential Monte Carlo methods

1.3.1 Importance sampling

Assume that we are interested in computing the integral given in (1.1) and, as before, it is not possible to compute it analytically. When generating samples directly from the distribution π is expensive or impossible, Importance Sampling (IS) is another Monte Carlo sampling strategy to resort to (Liu, 2001 ; Robert and Casella, 2004). Again we denote by f the density of π with respect to a dominating measure on $(\mathbb{X}, \mathcal{B})$. Let us rewrite the integral (1.1) as below

$$\mathbb{E}_\pi\{h(\boldsymbol{x})\,|\,\mathbf{y}\} \;=\; \mathbb{E}_g\left\{h(\boldsymbol{x})\frac{f(\boldsymbol{x})}{g(\boldsymbol{x})}\right\} \;=\; \int_{\mathbb{X}} h(\boldsymbol{x})\frac{f(\boldsymbol{x})}{g(\boldsymbol{x})}g(\boldsymbol{x})\mathrm{d}\boldsymbol{x} \tag{1.11}$$

where g is an easy-to-sample instrumental distribution. In the IS method, the integral is approximated by first generating M samples $\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(M)}$ from g, and then compensating for the discrepancy between the target and instrumental distributions by using *importance weights*

$$w^{(m)} \;=\; \frac{f(\boldsymbol{x}^{(m)})}{g(\boldsymbol{x}^{(m)})},$$

for $m = 1, \ldots, M$. We will call the pair $(\boldsymbol{x}^{(m)}, w^{(m)})$ the *weighted samples*, hereafter. Then, using the weighted samples, an unbiased and consistent approximation of the integral (1.1) is

$$\widehat{\mathbb{E}}_\pi^{\mathrm{IS}}[h(\boldsymbol{x})] \;=\; \frac{1}{M}\sum_{m=1}^{M} w^{(m)}h(\boldsymbol{x}^{(m)}).$$

If f is only known up to a normalizing constant, and consequently, the importance weights become $w^{(m)} \propto f(\boldsymbol{x}^{(m)})/g(\boldsymbol{x}^{(m)})$, for $m = 1, \ldots, M$, the following approximation called “self-normalized” IS, which is biased but yet consistent, is usually used

$$\widetilde{\mathbb{E}}_\pi^{\mathrm{IS}}[h(\boldsymbol{x})] \;=\; \sum_{m=1}^{M} W^{(m)}h(\boldsymbol{x}^{(m)}), \tag{1.12}$$

where $W^{(m)} = w^{(m)}/\sum_{i=1}^{M} w^{(i)}$ is the *normalized* weights. Note that when in addition to f, sampling directly from g is impossible, for example when g is only available up to a multiplicative constant, the “self-normalized” IS estimator (1.12) can be used through constructing a Markov chain—using one of the MCMC methods—which has g as its stationary distribution. This has applications for example in using Monte Carlo Expectation Maximization (MCEM) method for estimating model's hyperparameters in the Empirical Bayes (EB) approach (see for example Quintana et al., 1999 ; Levine and Casella, 2001).

Remark 1.4. We stress here that the application of IS method is not limited to the cases that generating samples directly from π is impossible. Indeed it is also a variance reduction technique (see, e.g., Robert and Casella, 2004, Chapter 4). Moreover, it can be useful in studying rare events such as sampling from the tail of a heavy-tailed distribution; for more information see, e.g., Asmussen et al. (2000).

In order to have finite importance weights and thus preventing degeneracy, f should be absolutely continuous with respect to g, that is, for all $\boldsymbol{x} \in \mathbb{X}$, if $g(\boldsymbol{x}) = 0$ then $f(\boldsymbol{x}) = 0$. In other words, this condition $\mathrm{supp}(f) \subset \mathrm{supp}(g)$ must always hold. Moreover, to ensure that inference made by means of the IS estimator is reliable, it is recommended in the literature to monitor several criteria. For example, a “rule of thumb” proposed by Liu (2001, page 34) is to monitor Efficient Sample Size (ESS) defined as

$$\mathrm{ESS} \triangleq \frac{M}{1+V_g(w(\boldsymbol{x}))}, \tag{1.13}$$

where $V_g\left(w(\boldsymbol{x})\right)$ is the variance of the unnormalized importance weights with respect to the distribution g. When M is large, the variance $V_g\left(w(\boldsymbol{x})\right)$ can be approximated by

$$V_g(w(\boldsymbol{x})) \simeq \frac{1}{M}\sum_{m=1}^{M}(w^{(m)})^2 - 1 \simeq M\sum_{m=1}^{M}(W^{(m)})^2 - 1. \tag{1.14}$$

Then, plugging (1.14) into (1.13), the estimated value of ESS becomes $1/\sum_{m=1}^{M}\left(W^{(m)}\right)^2$. This criterion tells us that the M i.i.d. weighted samples generated from g are worth of $M/(1+V_g\left(w(\boldsymbol{x})\right))$ i.i.d. samples drawn from the target distribution π.

Remark 1.5. An interesting property of the IS based methods is that the weighted samples generated from the instrumental distribution can be reused when the target distribution is slightly changed by just updating the weights. This concept is one of the principles of the Sequential Monte Carlo (SMC) samplers introduced in the following section (see, e.g., Bornn et al., 2010).

1.3.2 Sequential Monte Carlo samplers

In many applications, the goal is to generate samples from a sequence of distributions $\{\pi_t\}_{t\in\mathbb{T}}$, where π_t is defined on some $\mathbb{X}_t$ and $t \in \mathbb{T} = \{1, 2, \ldots, T\}$. In a Bayesian setting, this sequence of (posterior) distributions might arise either by observing *sequentially* input data (see, for example, Gordon et al., 1993 ; Liu and Chen, 1998 ; Doucet et al., 2001 ; Liu, 2001), that is,

$$f_1(\boldsymbol{x}) = p(\boldsymbol{x}\,|\,y_1),\; f_2(\boldsymbol{x}) = p(\boldsymbol{x}\,|\,y_{1:2}), \ldots,\; f_T(\boldsymbol{x}) = p(\boldsymbol{x}\,|\,y_{1:T}),$$

where $y_{1:T}$ simply denotes $(y_1, \ldots, y_T)$, for example, in target tracking, or by the fact that a certain hyperparameter θ of the model is evolving over “time”, that is,

$$f_1(\boldsymbol{x}) = p(\boldsymbol{x}\,|\,\theta_1, \mathbf{y}),\; f_2(\boldsymbol{x}) = p(\boldsymbol{x}\,|\,\theta_2, \mathbf{y}), \ldots,\; f_T(\boldsymbol{x}) = p(\boldsymbol{x}\,|\,\theta_T, \mathbf{y}),$$

for example in Bayesian sensitivity analysis (Bornn et al., 2010). In addition, one may be interested in *artificially* partitioning a huge set of observed data into several batches to reduce the computational complexity (Chopin, 2002).

Although ergodic Markov kernels could be designed to generate samples from each of the distributions π_t, for $t \in \mathbb{T}$, separately using MCMC methods, this would be very time demanding when T is large and considering further the fact that one should generate a large number of samples from each distribution π_t to have an acceptable approximation. However, importance sampling concept can be used, in these situations, to efficiently draw samples from the sequence of distributions $\{\pi_t\}_{t\in\mathbb{T}}$. Sequential Monte Carlo (SMC) samplers, also known as Particle filters, (elaborated mainly in Gordon et al., 1993 ; Liu and Chen, 1998 ; Doucet et al., 2000, 2001 ; Liu, 2001 ; Gilks and Berzuini, 2001 ; Chopin, 2002 ; Del Moral et al., 2006, among others) are particular algorithms developed for this purpose by generalizing the idea of importance sampling.

In fact, considering the state space that defines the model and parameter space, the sequential problems can be divided into two main groups: “dynamic” and “static” models. In the former case the target distribution, at time t, is defined on $\mathbb{X}_t$, where, often, $\mathbb{X}_{t-1} \subset \mathbb{X}_t$, such as target tracking problem while in the latter one the target distributions are all defined on the same space $\mathbb{X}$ such as Bayesian sensitivity analysis. In this section and, thus, throughout this thesis we will concentrate on the static case. Refer to (Gordon et al., 1993 ; Liu and Chen, 1998 ; Doucet et al., 2000, 2001 ; Liu, 2001) for information concerning “dynamic” models.

We briefly explain the Sequential Importance Sampling (SIS) technique (see, e.g., Del Moral et al., 2006, Section 2), as it will be helpful in presenting the principles of the SMC samplers. For $t \in \mathbb{T}$, let

$$f_t(\boldsymbol{x}) \;=\; \frac{\gamma_t(\boldsymbol{x})}{\mathbf{z}_t}, \tag{1.15}$$

where $\gamma_t(\boldsymbol{x})$ is the unnormalized density assumed to be known and $\mathbf{z}_t$ is the unknown normalizing constant. Assume further that there is a density $g_t(\boldsymbol{x})$ on $\mathbb{X}$ which will be used as the instrumental distribution. Thus, for example, $f_{t-1}(\boldsymbol{x})$ can be approximated by the set of weighted samples $(\boldsymbol{x}_{t-1}^{(m)},\, w_{t-1}^{(m)})$, or simply particles, for $m = 1, \ldots, M$, generated from $g_{t-1}(\boldsymbol{x})$, as described in Section 1.3.1. The M particles $(\boldsymbol{x}_{t-1}^{(m)},\, w_{t-1}^{(m)})$, then, can be reused to approximate $f_t(\boldsymbol{x})$ in two steps; first the particles $\boldsymbol{x}_{t-1}^{(m)}$—currently distributed according to $f_{t-1}(\boldsymbol{x})$—are moved using a Markov kernel $P_t(\boldsymbol{x}, \boldsymbol{x}')$ to $\boldsymbol{x}_t^{(m)}$ which are distributed according to

$$g_t(\boldsymbol{x}') \;=\; \int_{\mathbb{X}} g_{t-1}(\boldsymbol{x}) P_t(\boldsymbol{x}, \boldsymbol{x}')\mathrm{d}\boldsymbol{x},$$

and, subsequently, *reweighted* to be distributed according to $f_t(\boldsymbol{x})$. The corresponding instrumental distribution is

$$g_t(\boldsymbol{x}_t) \;=\; \int g_1(\boldsymbol{x}_1) \prod_{i=2}^{t} P_i(\boldsymbol{x}_{i-1}, \boldsymbol{x}_i) \mathrm{d}\boldsymbol{x}_{1:n-1}. \tag{1.16}$$

This is, in fact, a major drawback of SIS method, as the integral (1.16) is, in most practical problems, impossible to compute; furthermore, the proposed methods to approximate it are intricate (see Del Moral et al., 2006, for more details).

The idea proposed in the seminal paper by Del Moral et al. (2006) to side-step this difficulty is to employ an auxiliary backward Markov kernel $L_{t-1}\left(\boldsymbol{x}_t, \mathrm{d}\boldsymbol{x}_{t-1}\right)$ that allows us to carry out proper reweighting by approximating "surrogate" joint target distributions

$$\tilde{\pi}_t\left(\boldsymbol{x}_{1:t}\right) \;=\; \frac{\tilde{\gamma}_t\left(\boldsymbol{x}_{1:t}\right)}{\mathbf{z}_t},$$

where

$$\tilde{\gamma}_t\left(\boldsymbol{x}_{1:t}\right) \;=\; \gamma_t\left(\boldsymbol{x}_t\right) \prod_{i=1}^{t-1} L_i\left(\boldsymbol{x}_{i+1}, \mathrm{d}\boldsymbol{x}_i\right) \cdot$$

The key observation here is that that the surrogate joint distribution $\tilde{\pi}_t\left(\boldsymbol{x}_{1:t}\right)$ admits $\pi_t\left(\boldsymbol{x}\right)$ as a marginal distribution. The general expression for the updated unnormalized importance weights reads

$$w_t(\boldsymbol{x}_{1:t}) \;=\; w_{t-1}(\boldsymbol{x}_{1:t-1})\, \tilde{w}_t(\boldsymbol{x}_{t-1},\, \boldsymbol{x}_t), \tag{1.17}$$

with the unnormalized *incremental* weight

$$\tilde{w}_t(\boldsymbol{x}_{t-1},\, \boldsymbol{x}_t) \;=\; \frac{\gamma_t(\boldsymbol{x}_t)\, L_{t-1}(\boldsymbol{x}_t,\, \mathrm{d}\boldsymbol{x}_{t-1})}{\gamma_{t-1}(\boldsymbol{x}_{t-1})\, P_t(\boldsymbol{x}_{t-1},\, \mathrm{d}\boldsymbol{x}_t)}. \tag{1.18}$$

As in importance sampling, a routine procedure in the SMC samplers is to monitor the ESS criterion (1.13) to avoid *degeneracy* of particles. Then, when it becomes less than a certain threshold, say, $M/2$, the particles are *resampled.* The aim of this step is to duplicate the particles with significant importance weights and discard the ones with negligible weights. In the resampling procedure, the particles $\boldsymbol{x}_t^{(m)}$, for $m = 1, \ldots, M$, are copied $\mathbf{m}_t^m$ times—$\mathbf{m}_t^{(m)}$ can be even zero—such that $\sum_{m=1}^{M} \mathbf{m}_t^{(m)} = M$, depending on the corresponding normalized weights $W_t^{(m)}$. Then, all the unnormalized weights are set to one. There are various resampling algorithms in the literature (see for example Liu, 2001 ; Doucet et al., 2000, pages 72–75), though, we use the one consisting of generating random numbers $\mathbf{m}_t^{(m)}$, for $m = 1, \ldots, M$, from a multinomial distribution of parameters $W_t^{(m)}$. A general SMC sampler is described in Algorithm 1.3.

An important case of the SMC samplers extensively used in the literature is the one in which the Markov kernel P_t is a MCMC kernel of invariant distribution π_t. A suboptimal

Algorithm 1.3. *Sequential Monte Carlo sampler:*

```
initialization :

        • set $t = 1$;
        • for $m = 1, \dots, M$ draw particles $\boldsymbol{x}_1^{(m)}$ from $g_1(\boldsymbol{x})$;
        • compute importance weights $w_1^{(m)} \propto f_1(\boldsymbol{x}_1^{(m)})/g_1(\boldsymbol{x}_1^{(m)})$ and normalize them
          to obtain normalized weights $W_1^{(m)}$;

 for $t = 2 : T$ do

    resampling :
        If ESS$= 1/\sum W_{t-1}^{(m)}$ is less than a certain threshold, resample the particles
        and set all normalized weights $w_t^{(m)} = 1$;

    move :
            • for $m = 1, \dots, M$ move particles $\boldsymbol{x}_t^{(m)} \sim P_t(\boldsymbol{x}_{t-1}^{(m)}, \cdot)$;
            • compute the unnormalized weights using expressions (1.17) and (1.18);
```

backward kernel often used in this case is the so-called “reverse Markov kernel” (Del Moral et al., 2006, page 422)

$$L_{t-1}(\boldsymbol{x}_t, \mathrm{d}\boldsymbol{x}_{t-1}) \;=\; \frac{\pi_t(\mathrm{d}\boldsymbol{x}_{t-1})\, P_t(\boldsymbol{x}_{t-1}, \mathrm{d}\boldsymbol{x}_t)}{\pi_t(\mathrm{d}\boldsymbol{x}_t)}. \tag{1.19}$$

In this case, the expression of the unnormalized incremental weight (1.18) boils down to

$$\tilde{w}_t(\boldsymbol{x}_{t-1},\, \boldsymbol{x}_t) \;=\; \frac{\gamma_t(\boldsymbol{x}_{t-1})}{\gamma_{t-1}(\boldsymbol{x}_{t-1})} \propto \frac{f_t(\boldsymbol{x}_{t-1})}{f_{t-1}(\boldsymbol{x}_{t-1})}. \tag{1.20}$$

Remark 1.6. The asymptotic variance of the SMC sampler, under regularity assumptions, is computed in Del Moral et al. (2006, Proposition 2) and it is stated that the variance is upper bounded while the one of IS method goes to ∞ with t. Moreover, the mixing behavior of the Markov kernel has a direct influence on the variance of SMC sampler.

Toy example revisited

As an illustration of the SMC sampler, we revisit the toy example of Section 1.2.3 using the SMC sampler developed by Chopin (2002). Recall that the posterior distribution of interest is given by

$$p(\mu \,|\, \mathbf{y}) \;\propto\; \frac{\exp\left(-\frac{\|\mathbf{y}-\mu\mathbb{1}_N\|^2}{2}\right)}{1+\mu^2}.$$

Let us construct, as in Chopin (2002), a sequence of posterior distributions by partitioning the observed data $\mathbf{y}$ into T sections. More precisely, the sequence of posterior distributions is $\{f_t\,(\mu)\}_{t\in\mathbb{T}}$, where

$$f_t(\mu) \;=\; p\left(\mu \mid \mathbf{y}_{1:\frac{tN}{T}}\right).$$

Observe that, at each step of the algorithm, N/T observed samples are added to the model.

To initialize the SMC sampler, we need to generate M particles denoted by $\mu_1^{(m)}$ from an instrumental distribution denoted by g_1 in Algorithm 1.3. It is possible to draw directly samples from the prior distribution, i.e., standard Cauchy prior distribution (1.9), using the *inverse transform* approach (see, e.g., Robert and Casella, 2004, Section 2.1.2). It turns out that, however, due to the heavy-tailedness of the Cauchy distribution, many of the generated samples would be far away from the region of interest. Therefore, we opt for using a normal distribution $\mathcal{N}(\mu_{\mathbf{y}}, \sigma_{\mathbf{y}}^2)$, where $\mu_{\mathbf{y}}$ and $\sigma_{\mathbf{y}}^2$ are, respectively, the empirical mean and variance estimates of the observed data $\mathbf{y}$. Note that this is much faster than applying the MCMC sampler to generate M particles from $p\left(\mu \mid \mathbf{y}_{1:\frac{N}{T}}\right)$. Then, for $m = 1, \dots, M$, the unnormalized importance weights becomes

$$w_1^{(m)} = \frac{\exp\left(-\frac{\left\|\mathbf{y}_{1:\frac{N}{T}} - \mu_1^{(m)}\mathbb{1}_N\right\|^2}{2}\right)}{\left(1+(\mu_1^{(m)})^2\right)\cdot \mathcal{N}(\mu_1^{(m)} \mid \mu_{\mathbf{y}}, \sigma_{\mathbf{y}}^2)}.$$

Next the unnormalized weights are updated using the unnormalized incremental weight given by, for $m = 1, \dots, M$ and $t = 1, \dots, T$,

$$\tilde{w}_t^{(m)} \propto \frac{f_t\left(\mu_{t-1}^{(m)}\right)}{f_{t-1}\left(\mu_{t-1}^{(m)}\right)} = \exp\left(-\frac{1}{2}\left\|y_{\frac{(t-1)N}{T}+1:\frac{tN}{T}} - \mu_{t-1}^{(m)}\right\|^2\right).$$

We use ESS defined in (1.13) as a criterion to monitor the efficiency of the sampler. When ESS is smaller than $M/2$, to avoid degeneracy, the particles are resampled according to a multinomial distribution with the normalized weights as its probabilities. Next, the particles are moved according to a symmetric normal random walk kernel, as in Section 1.2.3.

We use the same generated data of length $N = 100$. The number of partitions, and thus the number of SMC iterations, T is set to 20, to ensure that $f_{t-1} \approx f_t$, while the number of particles M is chosen to be 1000. Thus, at each SMC step, five observation are added to the model and only when ESS is less than 500 we resample the particles. The variation of ESS is shown in Figure 1.5. It can be seen that in only four out of 20 iterations the resampling procedure is used (shown by red squares). Figure 1.6 illustrates the evolution of the particles through 20 iterations of SMC sampler. It can be observed from the depicted densities of the weighted samples that in the beginning the samples generated from the instrumental distribution $\mathcal{N}(\mu_{\mathbf{y}}, \sigma_{\mathbf{y}}^2)$ are quite spread; then, as more observations are added the particles are more concentrated around the true mean, i.e, $\mu = 5$. The final estimated mean $\hat{\mu} = \sum_{m=1}^{M} W_T^{(m)} \mu_T^{(m)} = 5.13$.

Remark 1.7. Note that we didn't aim at comparing the performance of the MCMC and the SMC samplers for the toy example. If so, repeated simulations are needed to give statistics such as bias, variance, and MSE of the final estimated values. Rather, this toy example was intended to illustrate how Monte Carlo sampling strategies work in practice.

We will use the SMC sampler for Bayesian sensitivity analysis of the posterior distribution to a certain hyperparameter (similar to the algorithm developed by Bornn et al. (2010)) in Chapter 3.

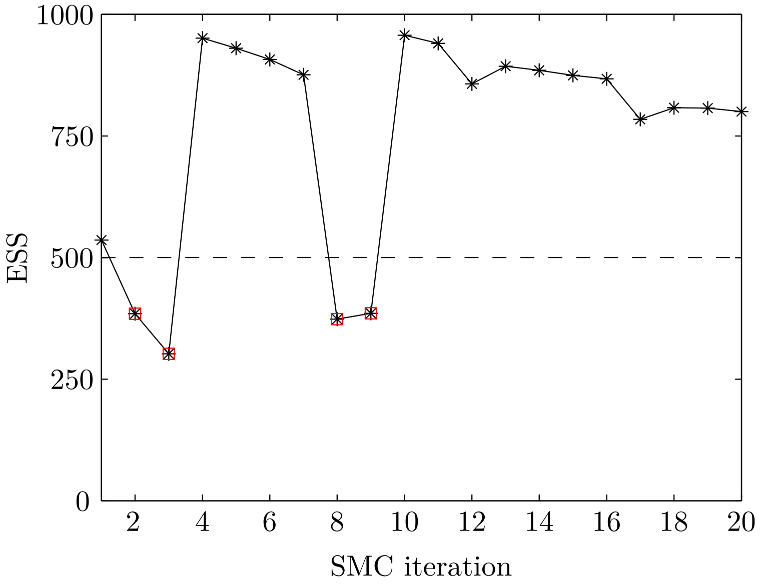


Figure 1.5 – *Efficient sample size (ESS) of SMC sampler applied for the toy example. The dark dashed line shows the threshold, that is, 500, used for deciding where to resample and the red squares highlight the resampled iterations.*

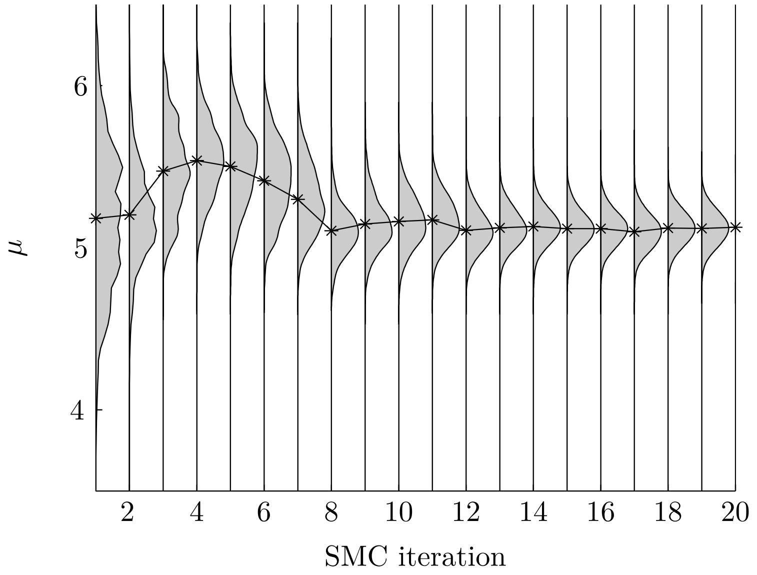


Figure 1.6 – *Density of the weighted samples (gray) along with the estimated mean of the normal distribution (black) for each SMC iteration. The final estimated value is* $\hat{\mu} = 5.13$.

1.4 Trans-dimensional MCMC sampler

In many problems of science and engineering "the number of things that we don't know is one of the things that we don't know"(Green, 1995, 2003). These variable-dimensional problems are also called "trans-dimensional" problems in the literature. In these problems, we are interested in making inference about a countable set of models, $\mathcal{M} = \{\mathcal{M}_1, \mathcal{M}_2, \cdots, \mathcal{M}_{k_{\max}}\}$, indexed by $k \in \mathcal{K} \subset \mathbb{N}$. In the general case, the model k has associated vector of parameters $\boldsymbol{\theta}_k \in \mathbb{X}_k$ of length n_k, where $\mathbb{X}_k$ is some n_k-dimensional space. Then, the objective is to make inference about the joint posterior distribution $\pi(k, \boldsymbol{\theta}_k)$ defined on the union of subspaces with different dimension $\mathbb{X} = \bigcup_{k \in \mathcal{K}} \{k\} \times \mathbb{X}_k$, of course after defining prior distributions over both k and $\boldsymbol{\theta}_k$.

In this Section, we restrict ourselves to signal decomposition problems in which the observed signal is assumed to be the superposition of a number of fundamental elementary signals or components of interest. To indicate that this restriction holds and to be consistent with the existing literature, we replace $\mathbb{X}_k$ with $\Theta_k \subseteq \mathbb{R}^{n_k}$. Thus, $\mathbb{X} = \bigcup_{k \in \mathcal{K}} \{k\} \times \Theta_k$ with pairs $\boldsymbol{x} = (k, \boldsymbol{\theta}_k) \in \mathbb{X}$. Then, the objective is, in addition to exploring the model space $\mathcal{M}$, assessing the vector of unknown component-specific parameters[2], denoted by $\boldsymbol{\theta}_k \in \Theta_k$ under $\mathcal{M}_k$, given the observed data (signal), $\mathbf{y}$. The variable-dimensional posterior distribution of interest is

$$\pi(k, \boldsymbol{\theta}_k) = \frac{p(\mathbf{y} \mid k, \boldsymbol{\theta}_k)\, p(k, \boldsymbol{\theta}_k)}{p(\mathbf{y})} \propto p(\mathbf{y} \mid k, \boldsymbol{\theta}_k)\, p(k, \boldsymbol{\theta}_k)\,. \tag{1.21}$$

This joint posterior distribution, then, can be used to express uncertainty about the candidate models and the vector of unknown parameters (see for example Clyde and George, 2004).

Simultaneous inference on both the model and parameter spaces through analyzing the joint posterior $\pi(k, \boldsymbol{\theta}_k)$ requires exploring the space $\mathbb{X}$ defined over the union of subspaces of varying-dimensionalities. Studying efficiently such posteriors has not been feasible until the introduction of Reversible Jump MCMC (RJ-MCMC) sampler by Green (1995). This sampler can be seen as a generalized version of Metropolis-Hastings sampler introduced in Section 1.2.2. In effect, it is not only capable of exploring the parameter space under $\mathcal{M}_k$ but also designed to span the model space by jumping between plausible models. To this end, in addition to fixed-dimensional (within-model) moves, as in standard MCMC methods, the RJ-MCMC sampler is equipped with trans-dimensional (between models) moves, which, under certain conditions, leaves the joint posterior distribution of interest, i.e. $\pi(k, \boldsymbol{\theta}_k)$, invariant.

[2] Note that, in addition to the component-specific parameters $\boldsymbol{\theta}_k$, one can further consider parameters that are common to all models. However, these parameters enjoy a fixed-dimensional space and usually are easily sampled using simple MH or Gibbs samplers. Thus, in this section, for clarity, we only concentrate on the variable-dimensional part and consider the case where there is no common parameter.

In this Section, due to the existence of erroneous Birth-or-Death moves acceptance ratio in signal processing society (see, e.g., Andrieu and Doucet, 1999), we aim at providing clear statements of some mathematical results, certainly not completely new but never stated explicitly, which can be used for a clean justification of the acceptance ratio of Birth-or-Death moves in signal decomposition (and similar) problems. For further information, see Green (1995) ; Richardson and Green (1997, 1998) ; Sisson (2005) ; Green (2003) ; Hastie and Green (2011).

Remark 1.8. Note that in some signal decomposition problems in which the goal is decomposing the observed signals into atoms, such as wavelet basis, the situation is different (see, e.g., George and Foster, 2000 ; Wolfe et al., 2004 ; Fevotte and Godsill, 2006 ; Dobigeon et al., 2009). The key feature in their approach is the introduction of auxiliary "indicator" variables to embed all models as special case of fixed-dimensional "big model". As a result, plain Gibbs or Metropolis-within-Gibbs sampling is sufficient to explore the augmented posterior distributions.

1.4.1 Mixture of proposal kernels

Metropolis-Hastings-Green ratio for mixture of proposal kernels

To generalize the MH sampler of Section 1.2.2 to the trans-dimensional case, it is often convenient to consider a proposal kernel Q built as a mixture of simpler transition kernels Q_m, with m in some finite or countable index set $\mathbb{M}$. In this case we have

$$Q\left(\boldsymbol{x},\, \mathrm{d}\boldsymbol{x}'\right) \;=\; \sum_{m \in \mathbb{M}} j\left(\boldsymbol{x},\, m\right)\, Q_m\left(\boldsymbol{x},\, \mathrm{d}\boldsymbol{x}'\right), \tag{1.22}$$

where $j\left(\boldsymbol{x},\, m\right)$ is the probability of choosing the move type m given that the current state is $\boldsymbol{x}$. Note that the actual value of $Q_m(\boldsymbol{x},\, \cdot\,)$ is irrelevant when $j(\boldsymbol{x}, m) = 0$.

It turns out that, under some assumptions, the MHG ratio for a mixture kernel Q can be conveniently deduced from the elementary ratios computed for each individual kernel Q_m using the formula

$$r\left(\boldsymbol{x}, \boldsymbol{x}'\right) \;=\; \frac{j\left(\boldsymbol{x}',\, m'\right)}{j\left(\boldsymbol{x},\, m\right)} \;\cdot\; \frac{\pi\left(\mathrm{d}\boldsymbol{x}'\right) Q_{m'}\left(\boldsymbol{x}', \mathrm{d}\boldsymbol{x}\right)}{\pi\left(\mathrm{d}\boldsymbol{x}\right) Q_m\left(\boldsymbol{x}, \mathrm{d}\boldsymbol{x}'\right)}. \tag{1.23}$$

where $m \in \mathbb{M}$ denotes the specific move that has been used to propose $\boldsymbol{x}'$, and $m' \in \mathbb{M}$ is the corresponding "reverse move". Equation (1.23) is routinely used in applications of the RJ-MCMC algorithm, and is alluded to in Green's paper (Green, 1995, p. 717) in the sentence : "*If [other] discrete variables are generated in making proposals, the* *probability functions of their realised values are multiplied into the move probabilities*". Note that, however, the acceptance ratio (1.23) is not true in the general case when a single MH kernel is used with mixture of proposal distributions (see, e.g., Tierney, 1998, Section 4). In the general case, to compute the MH acceptance ratio, evaluation of all transition kernels Q_m, $m \in \mathbb{M}$, is usually necessary, which can be computationally expensive. Sufficient conditions for Equation (1.23) to hold are provided by the following result:

Proposition 1.10. *Let* $R_m(\mathrm{d}\boldsymbol{x}, \mathrm{d}\boldsymbol{x}') = j(\boldsymbol{x}, m)\, \pi(\mathrm{d}\boldsymbol{x})\, Q_m(\boldsymbol{x}, \mathrm{d}\boldsymbol{x}')$. *Assume that there exists a family of disjoint sets* $\mathbb{W}_m \in \mathcal{B} \otimes \mathcal{B}$, *indexed by* $\mathbb{M}$, *such that :*

i) For each $m \in \mathbb{M}$, R_m *is supported by* $\mathbb{W}_m$, *which means* $R_m\left(\mathbb{X}^2 \setminus \mathbb{W}_m\right) = 0$.

ii) Each move $m \in \mathbb{M}$ *has a unique "reverse move"* $\varphi(m) \in \mathbb{M}$ *in the sense that* $\mathbb{W}_{\varphi(m)} = \mathbb{W}_m^{\mathrm{T}}$, *where* $\mathbb{W}_m^{\mathrm{T}} = \{(\boldsymbol{x}', \boldsymbol{x}) : (\boldsymbol{x}, \boldsymbol{x}') \in \mathbb{W}_m\}$.

Then, then MHG ratio (1.7) *is given by Equation* (1.23) *with* $m' = \varphi(m)$.

Proof. For $\pi(\mathrm{d}\boldsymbol{x})\, Q(\boldsymbol{x}, \mathrm{d}\boldsymbol{x}')$-almost everywhere on $\mathbb{X}^2$, there is a unique $m = m_{\boldsymbol{x},\boldsymbol{x}'} \in \mathbb{M}$ such that $(\boldsymbol{x}, \boldsymbol{x}') \in \mathbb{W}_m$. Indeed, the sets $\mathbb{W}_m$, $m \in \mathbb{M}$, are disjoint and

$$\begin{aligned}\iint_{\mathbb{X}^2 \setminus \cup \mathbb{W}_m} \pi(\mathrm{d}\boldsymbol{x})\, Q(\boldsymbol{x}, \mathrm{d}\boldsymbol{x}') &= \sum_{m \in \mathbb{M}} R_m(\mathbb{X}^2 \setminus \cup \mathbb{W}_m) \\ &\leq \sum_{m \in \mathbb{M}} R_m(\mathbb{X}^2 \setminus \mathbb{W}_m) \;=\; 0.\end{aligned}$$

Equation (1.23) can be rewritten as:

$$r\left(\boldsymbol{x}, \boldsymbol{x}'\right) \;=\; \frac{R_{\varphi(m_{\boldsymbol{x},\boldsymbol{x}'})}(\mathrm{d}\boldsymbol{x}', \mathrm{d}\boldsymbol{x})}{R_{m_{\boldsymbol{x},\boldsymbol{x}'}}(\mathrm{d}\boldsymbol{x}, \mathrm{d}\boldsymbol{x}')}.$$

Then, for all $A \in \mathcal{B} \otimes \mathcal{B}$,

$$
\begin{aligned}
&\iint_A r(\boldsymbol{x},\boldsymbol{x}')\, R(\mathrm{d}\boldsymbol{x}, \mathrm{d}\boldsymbol{x}') \\
&\quad = \iint_A \frac{R_{\varphi(m_{\boldsymbol{x},\boldsymbol{x}'})}(\mathrm{d}\boldsymbol{x}', \mathrm{d}\boldsymbol{x})}{R_{m_{\boldsymbol{x},\boldsymbol{x}'}}(\mathrm{d}\boldsymbol{x}, \mathrm{d}\boldsymbol{x}')} \cdot \sum_{m_0 \in \mathbb{M}} R_{m_0}(\mathrm{d}\boldsymbol{x}, \mathrm{d}\boldsymbol{x}') \\
&\quad = \sum_{m_0 \in \mathbb{M}} \iint_{A \cap \mathbb{W}_{m_0}} \frac{R_{\varphi(m_0)}(\mathrm{d}\boldsymbol{x}', \mathrm{d}\boldsymbol{x})}{R_{m_0}(\mathrm{d}\boldsymbol{x}, \mathrm{d}\boldsymbol{x}')} R_{m_0}(\mathrm{d}\boldsymbol{x}, \mathrm{d}\boldsymbol{x}') \\
&\quad = \sum_{m_0 \in \mathbb{M}} \iint_{A \cap \mathbb{W}_{m_0}} R_{\varphi(m_0)}(\mathrm{d}\boldsymbol{x}', \mathrm{d}\boldsymbol{x}) \\
&\quad = \sum_{m_0 \in \mathbb{M}} \iint_{A^{\mathrm{T}} \cap \mathbb{W}_{m_0}^{\mathrm{T}}} R_{\varphi(m_0)}(\mathrm{d}\boldsymbol{x}, \mathrm{d}\boldsymbol{x}') \\
&\quad = \iint_{A^{\mathrm{T}}} R(\mathrm{d}\boldsymbol{x}, \mathrm{d}\boldsymbol{x}') \qquad \text{because } \mathbb{W}_{m_0}^{\mathrm{T}} = \mathbb{W}_{\varphi(m_0)} \\
&\quad = \iint_A R(\mathrm{d}\boldsymbol{x}', \mathrm{d}\boldsymbol{x})\,.
\end{aligned}
$$

□

Mixture representation of trans-dimensional kernels

Consider the case of a variable-dimensional space , that can be written as $\mathbb{X} = \cup_{k \in \mathcal{K}} \{k\} \times \Theta_k$. A point $\boldsymbol{x} \in \mathbb{X}$ is a pair $(k, \boldsymbol{\theta}_k)$ with $k \in \mathcal{K}$ and $\boldsymbol{\theta}_k \in \Theta_k$.

Set $\mathbb{X}_k = \{k\} \times \Theta_k$. Any kernel Q on $\mathbb{X}$ admits a natural representation as a mixture of fixed-dimensional and trans-dimensional kernels :

$$
Q\left(\boldsymbol{x}, \mathrm{d}\boldsymbol{x}'\right) \;=\; \sum_{(k,l) \in \mathcal{K}^2} p_{k,l}(\boldsymbol{x})\, Q_{k,l}\left(\boldsymbol{x}, \mathrm{d}\boldsymbol{x}'\right)\,, \tag{1.24}
$$

where

$$
\begin{aligned}
p_{k,l}(\boldsymbol{x}) \;&=\; \mathbb{1}_{\mathbb{X}_k}(\boldsymbol{x})\, Q(\boldsymbol{x}, \mathbb{X}_l)\,, \\
Q_{k,l}(\boldsymbol{x}, \cdot\,) \;&=\; \frac{1}{p_{k,l}(\boldsymbol{x})}\, Q\left(\boldsymbol{x}, \cdot \cap \mathbb{X}_l\right)\,.
\end{aligned}
$$

(An arbitrary value can be chosen for $Q_{k,l}(\boldsymbol{x}, \cdot\,)$ when $p_{k,l}(\boldsymbol{x}) = 0$ to make it a completely defined transition kernel.) The kernels $Q_{k,k}$, $k \in \mathcal{K}$, correspond to the “fixed-dimensional” part of the transition kernel Q; while the kernels $Q_{k,l}$, $(k,l) \in \mathcal{K}^2$, $k \neq l$, correspond to the “trans-dimensional” part.

The mixture representation (1.24) satisfy the assumptions of Proposition 1.10 with $\mathbb{M} = \mathcal{K}^2$, $\mathbb{W}_{k,l} = \mathbb{X}_k \times \mathbb{X}_l$ for all $(k,l) \in \mathbb{M}$ and $\varphi(k,l) = (l,k)$. Therefore, if the current state x is in $\mathbb{X}_k$ and the proposed state x' in $\mathbb{X}_l$, the MHG ratio (1.23) reads

$$
r(\boldsymbol{x}, \boldsymbol{x}') \;=\; \frac{p_{l,k}(\boldsymbol{x}')}{p_{k,l}(\boldsymbol{x})} \cdot \frac{\pi(\mathrm{d}\boldsymbol{x}')\, Q_{l,k}(\boldsymbol{x}', \mathrm{d}\boldsymbol{x})}{\pi(\mathrm{d}\boldsymbol{x})\, Q_{k,l}(\boldsymbol{x}, \mathrm{d}\boldsymbol{x}')}\,. \tag{1.25}
$$

In most "tutorial" papers about the RJ-MCMC method, this expression is directly written in the special case where Green's dimension matching argument can be applied (see, e.g., Green (2003), Sections 2.2 and 2.3). Unfortunately, the dimension matching argument does not apply directly to the commonly used Birth-or-Death kernels (see next section) if the mixture representation (1.24), which leads to (1.25), is used.

1.4.2 Birth-or-Death kernels

Birth-or-Death kernels on (unsorted) vectors

Let us consider the situation where a point $\boldsymbol{x} \in \mathbb{X}$ describes a set of k objects $s_1, \dots, s_k \in \mathbb{S}$, with $(\mathbb{S}, \nu)$ an atomless[3] measure space and $k \in \mathbb{N}$. One possible—and commonly used—way of representing this is to consider pairs $(k, \boldsymbol{s})$, where the objects s_i, $1 \le i \le k$, have been arranged in a vector $\boldsymbol{s} = (s_1, \dots, s_k) \in \mathbb{S}^k$. The corresponding space is $\mathbb{X} = \cup_{k \ge 0} \mathbb{X}_k$, $\mathbb{X}_k = \{k\} \times \boldsymbol{\Theta}_k$, with $\boldsymbol{\Theta}_k = \mathbb{S}^k$ and using the convention that $\mathbb{S}^0 = \{\varnothing\}$.

Birth-or-death kernels are the most natural kind of trans-dimensional moves in such spaces. Given $k \in \mathbb{N}$, $\boldsymbol{s} = (s_1, \dots, s_k) \in \mathbb{S}^k$ and $s^* \in \mathbb{S}$, we introduce the notations

$$\begin{aligned} \boldsymbol{s}_{-i} &= (s_1, \dots, s_{i-1}, s_{i+1}, \dots, s_k) \in \mathbb{S}^{k-1}, \\ \boldsymbol{s} \oplus_i s^* &= (s_1, \dots, s_{i-1}, s^*, s_i, s_{i+1}, \dots, s_k) \in \mathbb{S}^{k+1}, \end{aligned}$$

where $1 \le i \le k$ in the first case and $1 \le i \le k+1$ in the second case. Starting from $\boldsymbol{x} = (k, \boldsymbol{s})$, a birth move inserts a new component $s^* \in \mathbb{S}$, generated according to some proposal distribution $q(s)\,\nu(\mathrm{d}s)$, at a randomly selected location:

$$Q_{\mathrm{b}}\left(\boldsymbol{x}, \,\cdot\,\right) \;=\; \frac{1}{k+1} \sum_{i=1}^{k+1} \int_{\mathbb{S}} \delta_{(k+1, \boldsymbol{s} \oplus_i s^*)}\, q(s^*)\, \nu(\mathrm{d}s^*)\,. \tag{1.26}$$

A death move, on the contrary, removes a randomly selected component form the current state:

$$Q_{\mathrm{d}}\left(\boldsymbol{x}, \,\cdot\,\right) \;=\; \frac{1}{k} \sum_{i=1}^{k} \delta_{(k-1, \boldsymbol{s}_{-i})}\,. \tag{1.27}$$

Finally, the birth-or-death kernel is a mixture of the two:

$$Q(\boldsymbol{x}, \,\cdot\,) \;=\; p_{\mathrm{b}}(\boldsymbol{x})\, Q_{\mathrm{b}}(\boldsymbol{x}, \,\cdot\,) \;+\; p_{\mathrm{d}}(\boldsymbol{x})\, Q_{\mathrm{d}}(\boldsymbol{x}, \,\cdot\,)\,, \tag{1.28}$$

with $p_{\mathrm{b}}(\boldsymbol{x})$, $p_{\mathrm{d}}(\boldsymbol{x}) \ge 0$, $p_{\mathrm{b}}(\boldsymbol{x}) + p_{\mathrm{d}}(\boldsymbol{x}) = 1$, and $p_{\mathrm{d}}\left((0, \varnothing)\right) = 0$. Moreover, if $\mathcal{K}$ has an upper bound $k_{\max}$ then $p_{\mathrm{b}}\left((k_{\max}, \boldsymbol{s})\right) = 0$.

[3] See, e.g., Fremlin (2001). As a concrete example, think of $\mathbb{S} = \mathbb{R}^d$ endowed with its usual Borel σ-algebra and ν equal to Lebesgue's measure. We will use the following property in the proof of Proposition 1.11: if $(\mathbb{S}, \nu)$ is atomless, then the diagonal $\Delta = \{(s,s) : s \in \mathbb{S}\}$ is $\nu \otimes \nu$-negligible in $\mathbb{S} \times \mathbb{S}$.

Expression of the MHG ratio

The following proposition provides the expression of the MHG ratio for the Birth-or-Death kernel.

Proposition 1.11. *Assume that, for all $k \geq 1$, the target measure π restricted to $\mathbb{X}_k$ admits a probability density function f_k with respect to $\nu^{\otimes k}$. Then the MHG ratio is*

$$r(\boldsymbol{x}, \boldsymbol{x}') \;=\; \frac{f_{k+1}(\boldsymbol{x}')}{f_k(\boldsymbol{x})} \cdot \frac{p_{\mathrm{d}}(\boldsymbol{x}')}{p_{\mathrm{b}}(\boldsymbol{x})} \cdot \frac{1}{q(s^*)} \tag{1.29}$$

for a birth move from $\boldsymbol{x} = (k, \boldsymbol{s})$ to $\boldsymbol{x}' = (k+1, \boldsymbol{s} \oplus_i s^*)$.

Proof. Although a direct computation of the MHG ratio would be possible based on Equations (1.26)–(1.28), we find it much more illuminating to deduce the result from Proposition 1.10 using kernels which are simpler than Q_{b} and Q_{d}. To do so, let us consider the family of elementary kernels Q_m, with m in the index set

$$\mathbb{M} = \Big\{ (\alpha, k, i) \in \{0, 1\} \times \mathbb{N}^2 : 1 \leq i \leq k + \alpha \Big\}$$

where $Q_{1,k,i}$ is the kernel from $\mathbb{X}_k$ to $\mathbb{X}_{k+1}$ that inserts a new component $s^* \sim q(s)\nu(ds)$ in position i, and $Q_{0,k,i}$ is the kernel from $\mathbb{X}_k$ to $\mathbb{X}_{k-1}$ that removes the i^{th} component. Then we can write

$$Q(\boldsymbol{x}, \,\cdot\,) \;=\; \sum_{m \in \mathbb{M}} j(\boldsymbol{x}, m)\, Q_m(\boldsymbol{x}, \,\cdot\,), \tag{1.30}$$

with $j(\boldsymbol{x}, m)$ defined for all $\boldsymbol{x} = (k, \boldsymbol{s}) \in \mathbb{X}$ as

$$j(\boldsymbol{x}, m) \;=\; \begin{cases} p_{\mathrm{b}}(\boldsymbol{x})/(k+1) & \text{if } m = (1, k, i), 1 \leq i \leq k+1, \\ p_{\mathrm{d}}(\boldsymbol{x})/k & \text{if } m = (0, k, i), 1 \leq i \leq k, \\ 0 & \text{otherwise.} \end{cases}$$

Denote by $\widetilde{\mathbb{X}}_k$ the set of all $\boldsymbol{x} \in \mathbb{X}_k$ in which no two components are equal. For all k, $\pi(\mathbb{X}_k \setminus \widetilde{\mathbb{X}}_k) = 0$, since $\pi_{|\mathbb{X}_k}$ admits a density with respect to the product measure $\nu^{\otimes k}$. The mixture representation (1.30) thus satisfies the assumptions of Proposition 1.10 with

$$\begin{aligned} \mathbb{W}_{(1,k,i)} \;=\; \Big\{ & (\boldsymbol{x}, \boldsymbol{x}') \in \widetilde{\mathbb{X}}_k \times \widetilde{\mathbb{X}}_{k+1} : \; \exists \boldsymbol{s} \in \mathbb{S}^k, \, \exists s^* \in \mathbb{S}, \\ & \boldsymbol{x} = (k, \boldsymbol{s}), \, \boldsymbol{x}' = (k+1, \boldsymbol{s} \oplus_i s^*) \Big\}, \end{aligned}$$

$\mathbb{W}_{(0,k,i)} = \mathbb{W}_{(1,k-1,i)}^{\mathrm{T}}$, $\varphi(1, k, i) = (0, k{+}1, i)$ and $\varphi(0, k, i) = (1, k{-}1, i)$. As a consequence, the MHG ratio for a birth move $m = (1, k, i)$ is

$$r(\boldsymbol{x}, \boldsymbol{x}') \;=\; \frac{p_{\mathrm{d}}(\boldsymbol{x}')}{p_{\mathrm{b}}(\boldsymbol{x})} \cdot \frac{\pi(\mathrm{d}\boldsymbol{x}')\, Q_{0,k+1,i}(\boldsymbol{x}', \mathrm{d}\boldsymbol{x})}{\pi(\mathrm{d}\boldsymbol{x})\, Q_{1,k,i}(\boldsymbol{x}, \mathrm{d}\boldsymbol{x}')}.$$

Observe that the $1/(k+1)$ terms, in the move selection probabilities, cancel each other. To complete the proof, it remains to show that

$$\frac{\pi(\mathrm{d}\boldsymbol{x}')\, Q_{0,k+1,i}(\boldsymbol{x}', \mathrm{d}\boldsymbol{x})}{\pi(\mathrm{d}\boldsymbol{x})\, Q_{1,k,i}(\boldsymbol{x}, \mathrm{d}\boldsymbol{x}')} \;=\; \frac{f_{k+1}(\boldsymbol{x}')}{f_k(\boldsymbol{x})} \cdot \frac{1}{q(s^*)}. \tag{1.31}$$

This can be obtained, in the general case, by a direct computation of the densities with respect to the symmetric measure

$$\begin{aligned}\xi\left(\mathrm{d}(k, \boldsymbol{s}), \mathrm{d}\boldsymbol{x}'\right) \;=\; \nu^{\otimes k}(\mathrm{d}\boldsymbol{s}) \Big[\; & \delta_{(k-1, \boldsymbol{s}_{-i})}(\mathrm{d}\boldsymbol{x}') \\ & + \int_{\mathbb{S}} \delta_{(k+1, \boldsymbol{s} \oplus_i s^*)}\, \nu(\mathrm{d}s^*) \Big].\end{aligned}$$

In the important special case where $\mathbb{S} \subset \mathbb{R}^d$ and ν is (the restriction of) the d-dimensional Lebesgue measure, (1.31) can be seen as the result of Green's dimension matching argument (Green, 1995, Section 3.3), in a very simple case where the Jacobian is equal to one. □

Remark 1.9. We emphasize that (1.30) is *not* the usual mixture representation of trans-dimensional kernels introduced in Section 1.4.1. Indeed, starting, e.g., from $\mathbb{X}_k$, there are several elementary kernels that can propose a point in $\mathbb{X}_{k+1}$.

Birth-or-Death kernels on sorted vectors

Let us assume now that the objects are "sorted", in some sense, before being arranged in the vector $\boldsymbol{s} = (s_1, \ldots, s_k) \in \mathbb{S}^k$. This happens, in practice, either when there is a natural ordering on the set of objects (e.g., the jump times in signal segmentation or multiple change-point problems Green (1995) ; Punskaya et al. (2002)) or when artificial constraints are introduced to restore identifiability in the case of exchangeable components (see Richardson and Green (1997, 1998) ; Stephens (2000) ; Cappé et al. (2003) ; Jasra et al. (2005) for the case of mixture models).

To formalize this, let us consider the same space $\mathbb{X}$ as in Section 1.4.2. Assume that $\mathbb{S}$ is endowed with a total order and that the corresponding "sort function" $\psi : \mathbb{X} \to \mathbb{X}$ is measurable. What we are assuming now is that the target measure, denoted by $\widetilde{\pi}$ in this section, is supported by $\psi(\mathbb{X})$—in other words, the components of $\boldsymbol{x} \in \mathbb{X}$ are $\widetilde{\pi}$-almost surely sorted.

In such a setting, the definition of the Birth-or-Death kernel has to be slightly modified in order to accommodate the sort constraint: the death kernel is unchanged, but new components are inserted *deterministically* at the only location that makes the resulting vector sorted (instead of being added at a random location). Mathematically, for $\boldsymbol{x} = (k, \boldsymbol{s}) \in \mathbb{X}_k$, we now have:

$$\begin{aligned} \widetilde{Q}_{\mathrm{b}}\left(\boldsymbol{x},\,\cdot\,\right) &= \int_{\mathbb{S}} \delta_{\psi(k+1,\boldsymbol{s}\oplus_1 s^*)}\, q(s^*)\, \nu(\mathrm{d}s^*)\,, \\ \widetilde{Q}_{\mathrm{d}}\left(\boldsymbol{x},\,\cdot\,\right) &= \frac{1}{k}\sum_{i=1}^{k} \delta_{(k-1,\boldsymbol{s}_{-i})} \;=\; Q_{\mathrm{d}}\left(\boldsymbol{x},\,\cdot\,\right)\,. \end{aligned}$$

Proceeding as in the proof of Proposition 1.11, it can be proved that the MHG ratio for a birth move from $\boldsymbol{x} = (k, \boldsymbol{s})$ to $\boldsymbol{x}' = (k+1, \boldsymbol{s} \oplus_i s^*)$ is

$$r(\boldsymbol{x}, \boldsymbol{x}') \;=\; \frac{\widetilde{f}_{k+1}(\boldsymbol{x}')}{\widetilde{f}_k(\boldsymbol{x})} \cdot \frac{p_{\mathrm{d}}(\boldsymbol{x}')/(k+1)}{p_{\mathrm{b}}(\boldsymbol{x})\,\eta_i(\boldsymbol{x})} \cdot \frac{1}{q(s^*)/\eta_i(\boldsymbol{x})}\,, \tag{1.32}$$

where $\widetilde{f}_k$ denotes the pdf of $\widetilde{\pi}$ on $\mathbb{X}_k$ and $\eta_i(\boldsymbol{x})$ the probability that $s^* \sim q(s)\,\nu(\mathrm{d}s)$ is inserted at location i in $\boldsymbol{x}$. (Note that $p_{\mathrm{b}}(\boldsymbol{x})\,\eta_i(\boldsymbol{x})$ is the probability of performing a birth move at location i, and $p_{\mathrm{d}}(\boldsymbol{x}')/(k+1)$ the probability of the reverse death move; this is the appropriate way of decomposing this kernel as mixture in order to use Proposition 1.10.)

Let us now consider the case where, in the setting of Section 1.4.2, the target probability measure π is invariant under permutations of the components indices (in other words, the corresponding random variables are *exchangeable* (Bernardo and Smith, 2000, Chapter 4)). Sorting the components (as an identifiability device) is equivalent to looking at the image measure $\widetilde{\pi} = \pi^{\psi}$, which has the pdf $\widetilde{f}_k \;=\; k!\, f_k\, \mathbb{1}_{\psi(\mathbb{X})}$ on $\mathbb{X}_k$. As a consequence, the MHG ratios (1.29) and (1.32) are equal.

Remark 1.10. Another option, when the components of the vector $(s_1, \ldots, s_k)$ are exchangeable, is to forget about the indices and consider the set $\{s_1, \ldots, s_k\}$ instead. The object of interest is then a (random) finite set of points in $\mathbb{S}$—in other words, a point process on $\mathbb{S}$. The expression of the MHG ratio for Birth-or-Death moves in the point process framework, with the Poisson point process as a reference measure, has been given in Geyer and Møller (1994) (one year before the publication of Green (1995)). Point processes have been widely used, since then, in image processing and object identification (see, e.g., Rue and Hurn (1999) ; Descombes et al. (2004) ; Stoica et al. (2004) ; Lacoste et al. (2005)).

1.5 Summary

In this chapter, we reviewed different Monte Carlo sampling methods with a focus on Markov Chain Monte Carlo techniques. We presented two well-known fixed-dimensional MCMC samplers, namely the Metropolis-Hastings (Metropolis et al., 1953 ; Hastings,

1970) and Gibbs samplers (Geman and Geman, 1984) in Section 1.2.2. Through a toy example in Section 1.2.3, we emphasize the influence of proposal distributions in the performance of the MH sampler. Algorithms for tuning parameters of proposal distributions are clearly important, but were beyond the scope of this review. We also discussed importance sampling based methods, esp. sequential Monte Carlo samplers (see, e.g., Doucet et al., 2001 ; Liu, 2001 ; Chopin, 2002 ; Del Moral et al., 2006) in Section 1.3. We will use the SMC sampler in Section 3.2.5 for analyzing the sensitivity of posterior distributions to the parameters of prior distributions.

Due to the widespread existence of wrong acceptance ratios in the signal processing literature, in Section 1.4, we explicitly presented the process of between-models moves, particularly the birth-or-death moves, in trans-dimensional MCMC samplers to clarify in what manner these moves can be *correctly* employed. More precisely, in Section 1.4.2, we established results asserting under what conditions the trans-dimensional MCMC sampler admits the target distribution π as a stationary distribution. Furthermore, in Section 1.4.2, we extended the results to the case of sorted vectors and stated that the MHG ratio would be similar. We finish the discussion by an interesting quotation from Jannink and Fernando (2004): *"The fact that this error has remained in the literature for over 5 years underscores the view that while Bayesian analysis using Markov chain Monte Carlo is incredibly flexible and therefore powerful, the devil is in the details. Furthermore, incorrect analyses can give results that seem quite reasonable."* Surprisingly, this mistake last more than 12 years in this community. In Section 3.2.4, to illuminate the arguments concerning trans-dimensional samplers stated in Section 1.4, we will study the effect of using the erroneous acceptance ratio in the trans-dimensional example of joint Bayesian detection and estimation of sinusoids in white Gaussian noise (Andrieu and Doucet, 1999).

Chapter 2

Summarizing Variable-dimensional Posterior Distributions

2.1 Introduction

In Chapter 1, we described Monte Carlo sampling methods that enable us to draw samples from posterior distribution of interest. Nevertheless, practical challenges remain at the inference level to extract, from the (possibly very large number of) 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, respectively) are typically provided in addition to a graphical summary of the distribution (e.g., a histogram or a kernel density estimate); see the toy example of Section 1.2.3 of Chapter 1. In the case of multimodal distributions summarization becomes more difficult but can be carried out using, for instance, the approximation of the posteriors by Gaussian Mixture Models (GMMs); see, for example, West (1993) ; McLachlan and Peel (2000) ; Frühwirth-Schnatter (2006) ; Mengersen et al. (2011). Summarizing or approximating posterior distributions has also been used in designing proposal distributions of MH samplers in an adaptive MCMC framework; see, e.g., Bai et al. (2011).

In this chapter, we address the issue of summarizing variable-dimensional posterior distributions defined over a union of subspaces with differing dimension. These distributions are encountered in the "trans-dimensional problems" in which the observed signal, or, its distribution, is assumed to be made of an unknown number of individual components. In these problems, there is a countable set of models, $\mathcal{M} = \{\mathcal{M}_1, \mathcal{M}_2, \cdots\}$, indexed by $k \in \mathcal{K} \subset \mathbb{N}$, to describe the observed data $\mathbf{y}$. The model $\mathcal{M}_k$ assumes that $\mathbf{y}$ is made up of k components and a noise parameter. Thus, $\mathcal{M}_k$ has an associated vector of component-specific parameters $\boldsymbol{\theta}_k \in \boldsymbol{\Theta}^k$, where $\boldsymbol{\Theta}$ is the space over which the component-specific parameters are defined. Then, the posterior distribution of interest operates over a measurable space $(\mathbb{X}, \mathcal{B})$ where $\mathbb{X} = \cup_{k \geq 0} \mathbb{X}_k$, $\mathbb{X}_k = \{k\} \times \boldsymbol{\Theta}^k$, with the convention that $\boldsymbol{\Theta}^0 = \{\varnothing\}$. The problems of signal decomposition and Gaussian mixture modeling when the number of components is unknown are two important examples of such problems.

Remark 2.1. Observe that, for example, the problem of estimating the coefficients of the Autoregressive (AR) models when the AR order is unknown cannot be considered in the above framework.

One of the most challenging issues when summarizing posterior distributions, that even occurs in fixed-dimensional situation, is the "label-switching" phenomenon (see, e.g., Jasra et al., 2005 ; Stephens, 2000 ; Celeux et al., 2000 ; Mengersen et al., 2011), which is a consequence of the lack of identifiability of component labels. Hence, to summarize variable-dimensional posteriors, the proposed method should be able to "undo" switching of labels.

For this purpose, in Section 2.2.1, we will explain the label-switching phenomenon and briefly review the proposed relabeling algorithms in the case of fixed-dimensional posterior distributions (indeed, all the proposed methods are limited to the fixed-dimensional posteriors). Then, we will show that in the variable-dimensional posteriors, label-switching shows up in a more complicated form due to the fact that the trans-dimensional sampler jumps from one model to another one leading to "birth" or "death" of component labels—in addition to the usual fixed-dimensional label-switching.

Next, in Section 2.2.2, classical Bayesian approaches for summarizing posteriors defined over the union of subspaces of varying-dimensions, namely Bayesian model selection and Bayesian model averaging approaches, are described and their limitations, which encouraged us for designing a novel method, are pointed out. In the rest of the Chapter, the method we propose to summarize variable-dimensional posterior distributions is introduced. In the following chapters, we will use the proposed method for summarizing posterior distributions of varying-dimensions in three different applications.

2.2 State-of-the-art and outline of the proposed approach

2.2.1 The label-switching problem and its extension to variable-dimensional posteriors

The so-called label-switching issue arises when attempting to make inference from a posterior distribution that is invariant to the permutation of the components' labels, inducing a problem of *identifiability*. More precisely, in many fixed-or variable-dimensional problems, such as sinusoid detection in white Gaussian noise (described in Chapter 3) or parameter estimation in mixture models (see, for example, Diebolt and Robert, 1994 ; Richardson and Green, 1997), the likelihood is invariant under relabeling of the components. Then, in a Bayesian context, if the assigned prior distribution does not provide enough information to distinguish the components, e.g., when *exchangeable* prior distributions are used, the resulting posterior distribution will also be invariant under permutation of the components' labels.

A fixed-dimensional posterior distribution that is invariant under permutation of the labels, assuming that the model has k components, has $k!$ symmetric modes. As a result, during Monte Carlo simulation, e.g., using MCMC methods, the interpretation of the components corresponding to a given label switches from one iteration to another one; thus, leading to the marginal posteriors of the component-specific parameters being highly multimodal which consequently makes the process of drawing inference more difficult (see, e.g., Celeux et al., 2000 ; Stephens, 2000 ; Jasra et al., 2005).

Figure 2.1 illustrates the marginal posterior density estimates of the radial frequencies conditional to the number k of components from the output of the RJ-MCMC sampler for a sinusoid detection experiment that will be defined later (see Table 3.1, second experiment with SNR = 7dB). It can be observed that all marginal posterior distributions depicted in one row, that is, conditional to one value of k, are "nearly identical".

Identifiability Constraints (ICs), such as sorting the sinusoidal components based on radial frequencies, are one of the first—and simplest—remedies to deal with the label-switching problem used in mixture model analysis literature (see, for example, Diebolt and Robert, 1994 ; Richardson and Green, 1997). It breaks the symmetry of the posterior distribution by imposing a constraint, that is only satisfied by one relabeling $\mathcal{C}(\boldsymbol{x})$, for each $\boldsymbol{x} \in \mathbb{X}$. Assuming that the posterior π restricted to $\mathbb{X}_k$ admits a probability density f_k, the constrained posterior density $\tilde{f}_k$ becomes

$$\tilde{f}_k\left(\boldsymbol{x}\right) = k!\, f_k\left(\boldsymbol{x}\right) \mathbb{1}_{\mathcal{C}(\mathbb{X}_k)}(\boldsymbol{x}).$$

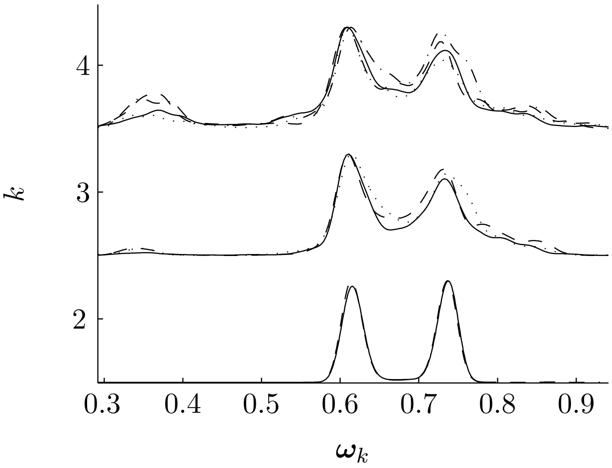


Figure 2.1 – *Marginal posterior density estimates of the unsorted radial frequencies $\boldsymbol{\omega}_k$ given k from the output of the RJ-MCMC sampler for the second sinusoid detection experiment defined in Table 3.1 with SNR $= 7$dB. Each row is dedicated to one value of k, for $2 \leq k \leq 4$. Note that in this figure, the components are not sorted to highlight the label-switching phenomenon.*

Remark 2.2. Imposing identifiability constraints indeed amounts to modifying the prior distribution by restricting the space $\mathbb{X}$.

For example, a possible IC for the problem of sinusoid detection is to sort the samples based on the radial frequencies. However, comparing the unsorted marginal densities illustrated in Figure 2.1 with the sorted ones shown in Figure 2.2, it can be seen that ICs cannot always be fruitful (see the rows related to $k = 3$ and $k = 4$ in Figure 2.2). Moreover, selecting an appropriate IC is not possible when there is no prior information to elicit one, particularly in multivariate problems, and inappropriate ICs can lead to results which are at odd with anticipation (see, e.g., the arguments in Celeux et al., 2000 ; Jasra et al., 2005). In the following, we will review briefly the relabeling algorithms that have been proposed so far for the fixed-dimensional case (see Jasra (2005) ; Celeux et al. (2000) ; Sperrin et al. (2010) ; Yao (2011) ; Papastamoulis and Iliopoulos (2010) for more details).

In relabeling algorithms, the goal is to permute each sample point of MCMC sampler so that the marginal posteriors become as unimodal and normally distributed as possible. Apart from ICs, in general, the proposed relabeling algorithms can be divided into two main classes; *deterministic* and *probabilistic* algorithms. The former category includes many of the proposed algorithms such as the ones similar to k-means clustering algorithm proposed by Stephens (1997a) and Celeux (1998) and decision theoretic approaches that select a relabeling for each sample point by optimizing the posterior expectation of some loss function; see, the relabeling algorithms of Stephens (2000) and Celeux et al. (2000) which aim at finding an appropriate IC for each sample point, as examples of the decision theoretic algorithms. Furthermore, the allocation-based algorithm of Papastamoulis and Iliopoulos (2010) and pivotal reordering algorithm of Marin et al. (2005) can also be classified into this category. One of the main drawbacks of both the ICs and the deterministic methods is that they are assuming that there is a "single" (or even "true") relabeling and the objective is to find it. This leads to neglecting the uncertainty of permutations.

On the other hand, in the more recently developed *probabilistic* algorithms, the permutation of the labels are assumed to be random variables to account for their uncertainty. This idea has been first developed in Jasra (2005, Chapter 4.5) by approximating the posterior distribution of interest and then, using it to derive conditional posterior distributions for permutations. Later, Sperrin et al. (2010) and Yao (2011) continued this idea and proposed EM-type algorithms for fitting an approximate model to the fixed-dimensional posterior distribution. Nevertheless, both methods proposed in Sperrin et al. (2010) and Yao (2011) study the uncertainty of all $k!$ possible permutations which is practically restrictive when the number k of components takes a moderate value.

Turning to the specific type of variable-dimensional posterior distributions introduced in Section 2.1, it is evident that the lack of identifiability of components together with uncertainty concerning their "presence" in the model leads to a more complicated situation for making inference about their labels. More explicitly, the trans-dimensional sampler jumps between models, with different number of components, from one iteration to another, in addition to switching the labels. Consequently, as described in Section 1.4, in a between models birth move, say, from $\mathcal{M}_k$ to $\mathcal{M}_{k+1}$, a component is added to the model; thus, we have a "birth" in the set of labels too. In contrary, when moving from $\mathcal{M}_{k+1}$ to $\mathcal{M}_k$, a component is removed from the model and, consequently, we have a "death" in the set of labels. We call this phenomenon "birth, death, and switching of labels".

This notion is illustrated in the Figure 2.2 (same data as Figure 2.1). It can be observed that, when going from $\mathcal{M}_2$ to $\mathcal{M}_3$, an additional sinusoidal component appears between the two others. In other words, in this specific example, *the second component in $\mathcal{M}_2$ is not the second one in $\mathcal{M}_3$, it is indeed the third one!* Following colors associated to the labeled components in the Figure 2.2 makes this issue more tangible. This indicates that a method designed for summarizing variable-dimensional posteriors should be capable of dealing with this change of dimensions in both the parameter and label spaces.

Therefore, in variable-dimensional posterior distributions, there is an extra uncertainty about the “presence” of components, in addition to their location. This challenging problem has hindered previous attempts to “undo” label-switching in the variable-dimensional scenario, where, according to Robert (1997) “*the meaning of individual components is vacuous*”.

2.2.2 Variable-dimensional summarization: classical Bayesian approaches

In a Bayesian setting, model uncertainty is studied through the posterior model probabilities

$$p(k \mid \mathbf{y}) = \frac{p(\mathbf{y} \mid k)p(k)}{\sum_{k'} p(\mathbf{y} \mid k')p(k')}, \tag{2.1}$$

where

$$p(\mathbf{y} \mid k) = \int p(\mathbf{y} \mid \boldsymbol{\theta}_k, k)p(\boldsymbol{\theta}_k \mid k)\mathrm{d}\boldsymbol{\theta}_k$$

is the marginal likelihood of $\mathcal{M}_k$. Then, the posterior (2.1) can be used to analyze and compare models (or even select one “best” model). Roughly speaking, two classical Bayesian approaches co-exist in the literature for such situations: Bayesian Model Selection (BMS) and Bayesian Model Averaging (BMA).

The BMS approach ranks models according to their posterior probabilities $p(k|\mathbf{y})$, selects one model (with the highest posterior support), say, $\mathcal{M}_{k_{\mathrm{MAP}}}$, where MAP stands for Maximum A Posteriori, and then summarizes the posterior of parameters under the (fixed-dimensional) selected model, i.e., $p(\boldsymbol{\theta}_{k_{MAP}} \mid \mathbf{y}, k_{\mathrm{MAP}})$. Due to the simplicity of the BMS approach, it has been used extensively in the literature, particularly when a new method is developed for a trans-dimensional problem and a comparison with the previous ones seems to be necessary to justify the efficiency of it (see Andrieu and Doucet (1999) ; Larocque and Reilly (2002) ; Davy et al. (2006) ; Punskaya et al. (2002) ; Andrieu et al. (1998) ; Hong et al. (2010) for signal processing, and George and Foster (2000) ; Chipman et al. (2001), and references therein, for Bayesian variable selection examples). Nevertheless, this is at the price of losing valuable information provided by the other (discarded) models.

To highlight the pros and cons of both the BMS and BMA approaches, we use again the example of sinusoid detection in white Gaussian noise. Figure 2.2 illustrates the posterior of the number k of components along with the posterior of sorted radial frequencies obtained from the output of the RJ-MCMC on the sinusoid detection problem discussed in Chapter 3. In this experiment, there are three sinusoidal components with the parameters $\boldsymbol{\omega}_k = (0.63, 0.68, 0.73)^t$ (see Table 3.1 for more details). The SNR is set to the moderate value of 7 dB. Following the BMS approach, by inspecting the posterior of k shown in the left panel of the Figure 2.2, the model with two sinusoidal components with

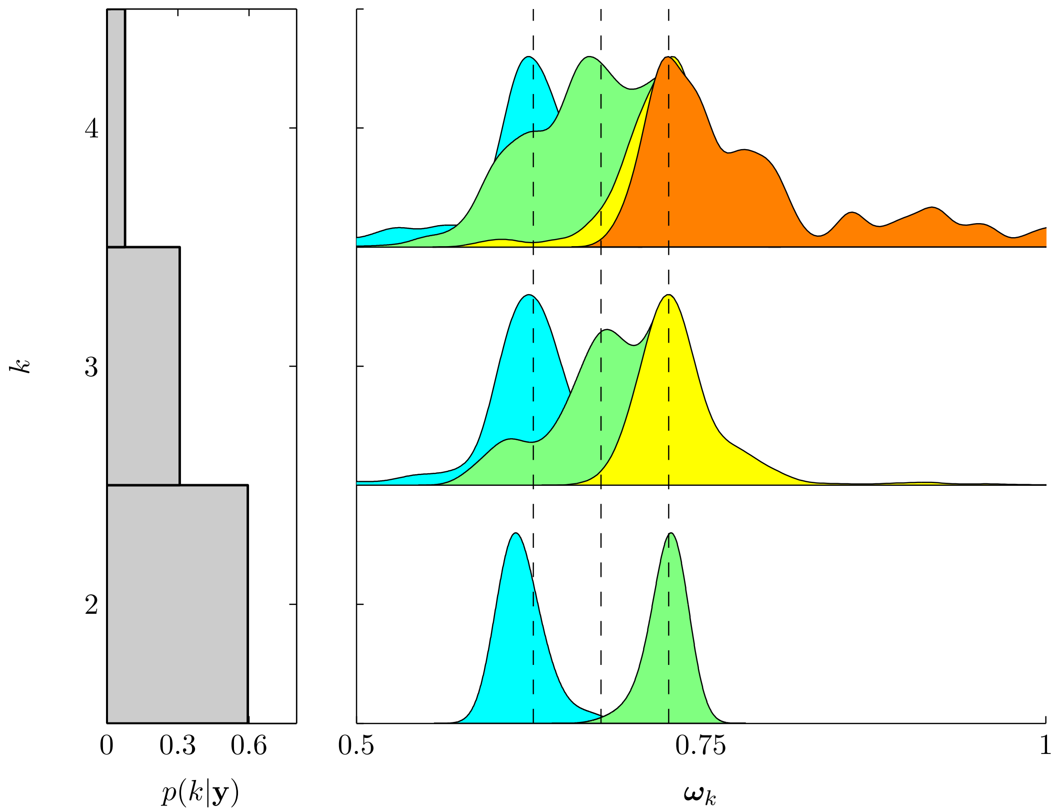


Figure 2.2 – *Posteriors of k (left) and sorted radial frequencies given k (right) from the output of the RJ-MCMC sampler for the second sinusoid detection experiment defined in Table 3.1 with SNR = 7dB. The true number of components is three. The vertical dashed lines in the right figure locate the true radial frequencies.*

$p(k = 2|\mathbf{y}) = 59.5\%$ would be selected. As a result, all information about the small—and therefore harder to detect—middle component would be lost, while it is clearly present in the posterior of sorted radial frequencies given $\mathcal{M}_3$ and $\mathcal{M}_4$, despite less posterior support ($p(k > 2|\mathbf{y}) = 40.5\%$). Therefore, in certain situations, selecting just one model and discarding all the others might not only be restricting but also be undesirable.

An alternative Bayesian approach to the BMS approach is the BMA in which the uncertainties of different models are incorporated—rather than selecting one "best" model—by reporting the results that are averaged over all possible models (see, e.g., Clyde and George, 2004 ; Hoeting et al., 1999 ; Kass and Raftery, 1995, and references therein). Suppose Δ is some quantity of interest, then, its posterior given the observed data $\mathbf{y}$ is

$$p(\Delta \,|\, \mathbf{y}) \;=\; \sum_{k \in \mathcal{K}} p(\Delta \,|\, \mathbf{y}, k) p(k \,|\, \mathbf{y}). \tag{2.2}$$

For example, the quantity Δ can be a future observation or the noiseless signal. Note that, however, it cannot be a component-specific parameter, the number of which changes in each model. Thus, we conclude here that the BMA approach is not appropriate for the kind of variable-dimensional posterior summarization we are interested in, while rather well-suited for predictive purposes.

Nonetheless, the BMA approach can still be used to produce informative summaries; for example, in the sinusoid detection problem, one may opt for dividing the interval $(0, \pi)$ into T bins denoted by Δ_t, for $t = 1, \dots, T$, and report the expected number of sinusoidal components in the bin Δ_t as

$$\mathbb{E}(N(\Delta_t) \,|\, \mathbf{y}) \;=\; \sum_{k=1}^{k_{\max}} \mathbb{E}(N(\Delta_t) \,|\, k, \mathbf{y})\, p(k \,|\, \mathbf{y}). \tag{2.3}$$

As an illustration, using the BMA approach to compute expected number of components from the RJ-MCMC output samples shown in Figure 2.2, as explained above, we obtain a histogram estimator of the intensity of radial frequencies illustrated in Figure 2.3. It can be seen from the histogram that the RJ-MCMC samples are concentrated around two peaks.

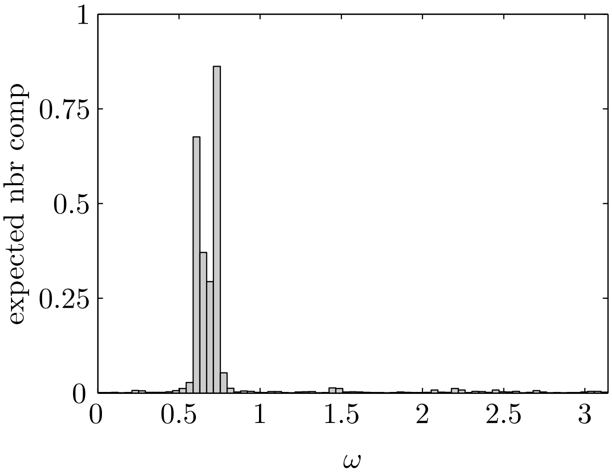


Figure 2.3 – *Histogram of the expected number of components for the second sinusoid detection experiment defined in Table 3.1 with SNR = 7dB obtained using the BMA approach* (2.3) *on the output samples of the RJ-MCMC sampler; see Figure 2.2 for the posterior distributions of the number k of sinusoidal components and sorted radial frequencies given k.*

To the best of our knowledge, no generic method is currently available, that would allow to summarize the information that is easily read on Figure 2.2 for this simple example: 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.*

Remark 2.3. Note that there is also one more recent Bayesian approach named "Median

Probability Model" (MPM) proposed by Barbieri and Berger (2004) for Bayesian variable selection problems. Since it depends on the posterior probability of presence of individual components—the property that is not yet available in the problems we are addressing such as signal decomposition and mixture modeling—we will not discuss it here. Later in Chapter 3, we will however return to this idea, as the technique we propose in this chapter assigns to each component a probability of presence.

2.2.3 Variable-dimensional summarization: the proposed approach

In this chapter, we will propose a novel approach to summarize the posterior distributions over variable-dimensional subspaces that typically arise in signal decomposition and mixture modeling problems with an unknown number of components. In a nutshell, it consists in approximating the complex posterior distribution with a parametric model of varying-dimensionality, by minimization of a divergence measure between the two distributions. We use two divergence measures, namely the Kullback-Leibler (KL) (Kullback and Leibler, 1951) and the more robust α-divergence measure proposed by Basu et al. (1998)—called hereafter BHHJ α-divergence. Then, a Stochastic EM (SEM)-type algorithm (Broniatowski et al., 1983 ; Celeux and Diebolt, 1985), driven by the output of an RJ-MCMC sampler, is used to estimate the parameters of the approximate model.

Our approach shares some similarities with the relabeling algorithms proposed in Stephens (2000) ; Sperrin et al. (2010) ; Yao (2011) to solve the label switching issue, and also with the EM algorithm used in Bai et al. (2011) in the context of adaptive MCMC algorithms (all in a *fixed*-dimensional setting). The main contribution of the proposed algorithm 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 proposed can be seen as a realization of the idea that M. Stephens had in mind when he stated (Stephens, 1997b, 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 ks, and grouping together components which are "similar", in that they have similar predictive density estimates. However, attempts to do this have failed to produce an easily interpretable results.*"

2.3 Variable-dimensional parametric model

In this section, we describe the original parametric model used for approximating the posterior π of interest defined on $\mathbb{X} = \bigcup_{k \geq 0} \{k\} \times \boldsymbol{\Theta}^k$. One point in $\mathbb{X}$ is a pair $\boldsymbol{x} = (k, \boldsymbol{\theta}_k)$ with $k \in \mathcal{K}$ and $\boldsymbol{\theta}_k = (\boldsymbol{\theta}_{1,k}, \ldots, \boldsymbol{\theta}_{k,k}) \in \boldsymbol{\Theta}^k$. The special point $(0, ())$ will be denoted by $\varnothing$. Let ρ denote the natural *reference measure* on $\mathbb{X}$, which is defined by $\rho(A) = \delta_{\varnothing}(A) + \sum_{k \geq 1} \int_{\boldsymbol{\Theta}^k} \mathbb{1}_A(k, \boldsymbol{\theta}_k) \, \mathrm{d}\boldsymbol{\theta}_k$ for all measurable $A \subset \mathbb{X}$. The integral of a measurable function $\varphi : \mathbb{X} \to \mathbb{R}$ with respect to ρ is given by $\int \varphi \, \mathrm{d}\rho = \varphi(\varnothing) + \sum_{k \geq 1} \int_{\boldsymbol{\Theta}^k} \varphi(k, \boldsymbol{\theta}_k) \, \mathrm{d}\boldsymbol{\theta}_k$. We assume that π admits a pdf f, with respect to ρ. The proposed parametric model will also be defined on the variable-dimensional space $\mathbb{X}$ (i.e., it is not a fixed-dimensional approximation as in the BMS approach).

We introduce the proposed parametric model in two steps; first, a "simple" version consisting of only Gaussian components is introduced in Section 2.3.1. Then, we argue the sensitivity of the "simple" parametric model to the observed samples that can be considered as "outliers" with respect to the majority of the observed samples. A "robustified" parametric model equipped with a Poisson point process component to account for the outliers is proposed in Section 2.3.2.

2.3.1 "Simple" parametric model

The proposed parametric model is established on two arguments. First, we have seen in Section 2.1 that summarizing fixed-dimensional posterior distributions is, often, implicitly or explicitly carried out by fitting Gaussian distributions or Gaussian mixture models. Hence, as in a traditional GMM, we assume that there is a certain number L of "Gaussian components" in the (approximate) posterior, i.e., parametric model, each generating a d-variate Gaussian vector with mean $\boldsymbol{\mu}_l$ and covariance matrix $\boldsymbol{\Sigma}_l$, $1 \leq l \leq L$. Second, it has been mentioned that in order for a summarizing method to be capable of dealing with "birth, death, and switching" of components labels, the parametric model should be able to generate variable-dimensional samples (see Section 2.2.1). We thus introduce in the parametric model binary indicator variables $\xi_l \in \{0, 1\}$ corresponding to each Gaussian component, for $l = 1, \ldots, L$, where $\xi_l = 1$ indicates that Gaussian component l is *present*; otherwise it is *absent*. These binary variables are assumed to be independently Bernoulli distributed, and we denote by $\pi_l \in (0; 1]$ the "probability of presence" of the l^{th} Gaussian component. Therefore, the probability distribution of the binary indicator vector $\boldsymbol{\xi}$ reads

$$p\,(\boldsymbol{\xi} \,|\, \boldsymbol{\pi}) \;=\; \prod_{l=1}^{L} \mathcal{B}er(\xi_l | \pi_l) \;=\; \prod_{l=1}^{L} \pi_l^{\xi_l} (1 - \pi_l)^{(1-\xi_l)}, \tag{2.4}$$

where, $\mathcal{B}er(\cdot|a)$ denotes a Bernoulli distribution with probability a and $\boldsymbol{\pi} = (\pi_1, \dots, \pi_L)$.

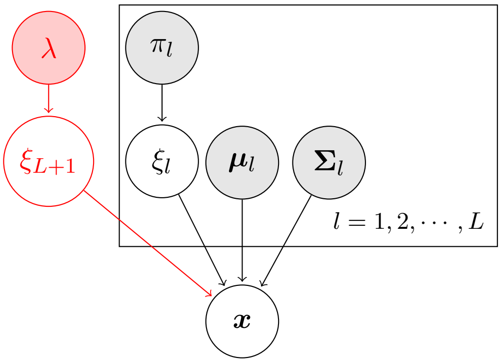


Figure 2.4 – *The proposed variable-dimensional parametric model in a generative viewpoint. It is assumed that there are L Gaussian components in the model with individual parameters $\boldsymbol{\mu}_l$ and $\boldsymbol{\Sigma}_l$, $0 \leq l \leq L$. Each component can be either present or absent according to a binary indicator variable $\xi_l \in \{0, 1\}$, where ξ_l is Bernoulli distributed with the probability π_l. The red part shows the intensity parameter λ and the indicator variable ξ_{L+1} of the Poisson point process component added to account for diffuse observed samples in Section 2.3.2.*

Let us describe the proposed parametric model from a generative point of view; An $\mathbb{X}$-valued random variable $\boldsymbol{x} = (k, \boldsymbol{\theta}_k)$, with $0 \leq k \leq L$, is generated as follows. First, each of the L Gaussian components can be either present or absent according to the binary indicator variable $\xi_l \in \{0, 1\}$ drawn from $\mathcal{B}er(\cdot|\pi_l)$. Second, given the indicator variables $\boldsymbol{\xi}$, $k = \sum_{l=1}^{L} \xi_l$ Gaussian vectors are generated by the Gaussian components that are present (that is, $\xi_l = 1$) and randomly arranged in a vector $\boldsymbol{\theta}_k = (\boldsymbol{\theta}_{k,1}, \dots, \boldsymbol{\theta}_{k,k})$. Figure 2.4 illustrates the corresponding DAG (parameter λ and indicator variable ξ_{L+1} are related to the Poisson point process component, which will be explained in Section 2.3.2).

Remark 2.4. Observe that the proposed variable-dimensional parametric model *is not* a GMM. In GMMs, only one component is present at a time (i.e., $k = 1$ in our notations), while in the proposed model at each realization up to L components can be present (see Example 2.1). As a consequence, in contrast with GMMs, there is no constraint here on the sum of the probabilities of presence. That is, $\sum_{l=1}^{L} \pi_l \neq 1$ in general.

Example 2.1. This simple example illustrates what the random samples generated from such a variable-dimensional parametric model look like. We assume that there are $L = 3$ univariate Gaussian components in the model, the individual parameters of which (means μ_l, variances s_l^2, and probabilities of presence π_l, with $1 \leq l \leq L$) presented in Table 2.1. Figure 2.5 depicts the pdf's of the three Gaussian components along with six random samples generated from this parametric model. Moreover, the kernel density estimates of 10 000 random samples generated from the "simple" parametric model of Table 2.1 are depicted in Figure 2.6 (a). It can be seen from both figures that the dimension of the generated samples varies from $k = 0$ to $k = L = 3$.

Observe also that the distributions of the sorted generated samples from the parametric models shown in Figure 2.6 (a) do not follow completely a Gaussian pattern. Note the tails of the distributions when $k = 2$ caused due to both the presence of samples generated from the middle Gaussian component, with less frequency, and sorting the random generated samples. An interesting point to mention is that there are bimodal distributions in the figure under $\mathcal{M}_1$.

l	μ	s^2	π
1	0.63	0.01	0.8
2	0.68	0.02	0.3
3	0.73	0.01	0.8

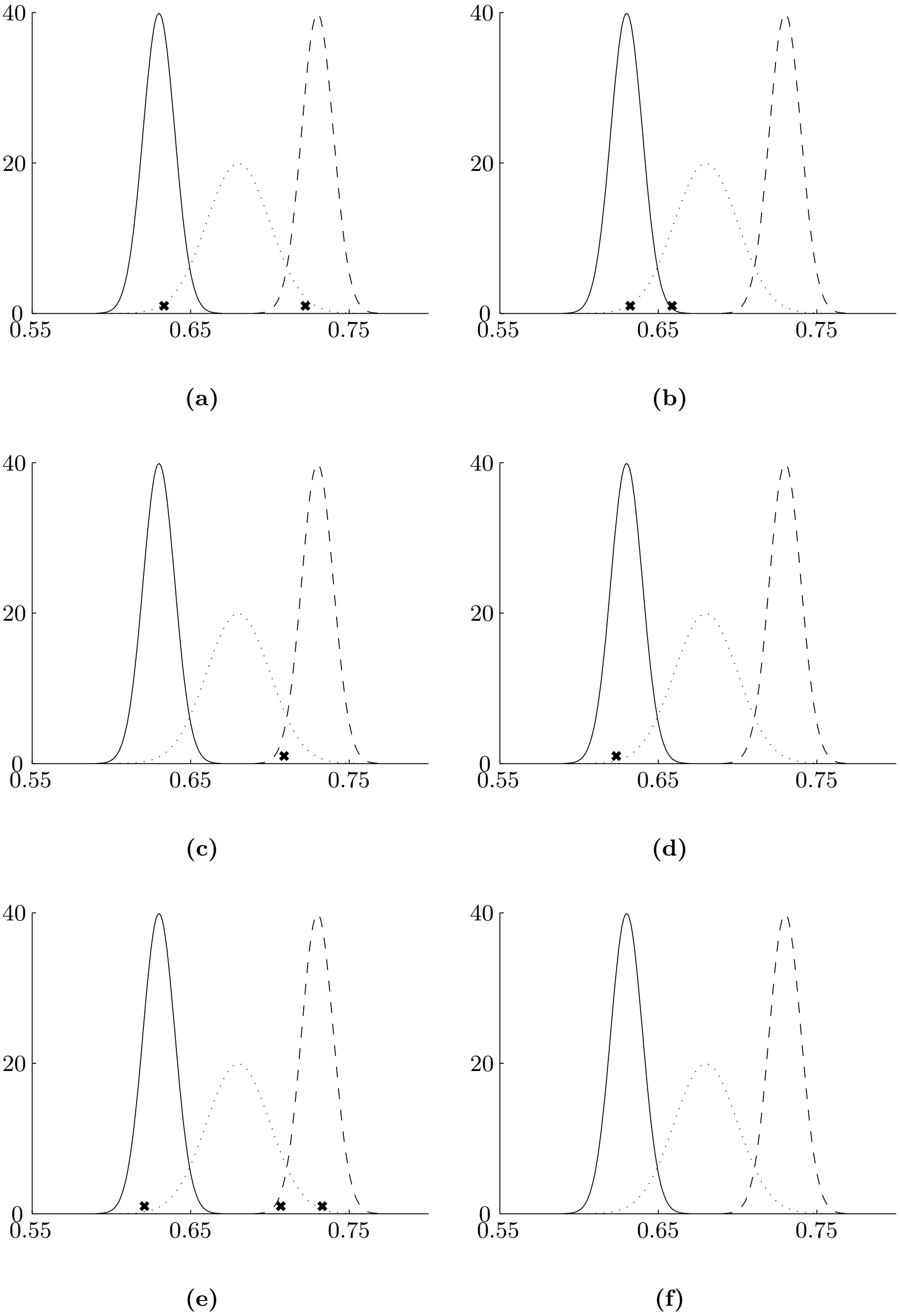
Table 2.1 – *Parameters of the model used in the Example 2.1.*

Contemplating the posterior distributions of the sorted radial frequencies depicted in the right panel of Figure 2.2, particularly the plots related to the models with three and four sinusoidal components, it can be observed that there are "diffuse parts" in the RJ-MCMC output samples resulting in the heavy asymmetric tails of some components. It is evident that a model constructed by only Gaussian components is not capable of describing these diffuse samples, at least not in a parsimonious way. These *abnormal* observations, with respect to the bulk of the observed data, or, simply *outliers*, can adversely influence the process of fitting the approximate posterior to the true posterior distribution of interest and consequently lead to meaningless parameter estimates.

We will propose two solutions to cope with this critical robustness issue: one in the modeling and another in the parameter estimation steps. In the next section, a modification in the parametric model towards a "robustified" parametric model will be proposed, while other solutions in the estimation procedure will be described in Sections 2.4 and 2.5.

2.3.2 "Robustified" parametric model

At this step, to robustify the "simple" parametric model of the previous section, we propose to include a Poisson point process (see, e.g., Karr, 1991 ; Van Lieshout, 2000, for



(a)

(b)

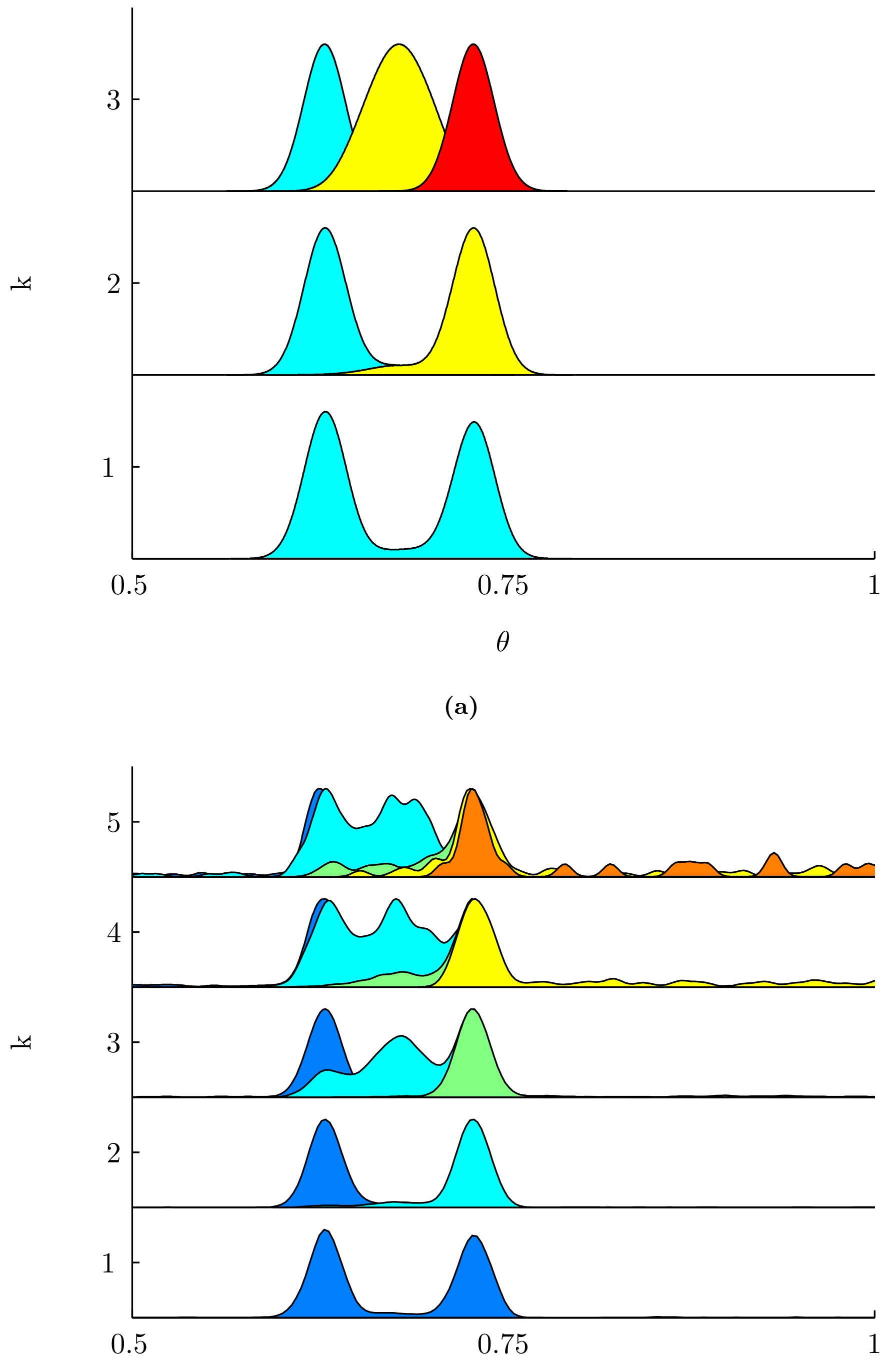
(c)

(d)

(e)

(f)

Figure 2.5 – *Generated samples from an example of the proposed "simple" variable-dimensional parametric model. There are $L = 3$ Gaussian components in the model with the parameters presented in Table 2.1. The × signs indicate the location of the random generated 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 $\boldsymbol{\theta}_0 = ()$.*



(a)

(b)

Figure 2.6 – *(a) Estimated kernel densities of 10 000 sorted random samples generated from the "simple" model of Example 2.1. (b) Estimated kernel densities of 10 000 sorted random samples generated from the parametric model of Example 2.1 equipped with a Poisson point process component with* $\lambda = 0.5$ *and uniform intensity on* $(0, \pi)$.

more information) component to account for the outliers in the observed samples. More precisely, the samples generated from the point process component, the number of which follows a Poisson distribution of mean $\lambda > 0$, are assumed to be uniformly distributed on the space $\boldsymbol{\Theta}$ of component-specific parameters. Therefore, they can present non-Gaussian patterns.

To be consistent with our previous notations, we denote by $\xi_{L+1} \in \mathbb{N}$ the number of points generated from the Poisson point process. Hence, it is distributed according to

$$p(\xi_{L+1}|\lambda) \;=\; \frac{e^{-\lambda} \cdot \lambda^{\xi_{L+1}}}{\xi_{L+1}!}. \tag{2.5}$$

Note that other elements of $\boldsymbol{\xi}$, i.e., $\xi_1, \ldots, \xi_L$, still take their values in $\{0, 1\}$. Then, from (2.4) and (2.5), we obtain the following distribution for the vector $\boldsymbol{\xi}$ of length $L+1$

$$p\,(\boldsymbol{\xi}\,|\,\boldsymbol{\pi}, \lambda) \;=\; \frac{e^{-\lambda} \cdot \lambda^{\xi_{L+1}}}{\xi_{L+1}!} \prod_{l=1}^{L} \pi_l^{\xi_l}(1-\pi_l)^{(1-\xi_l)}. \tag{2.6}$$

Finally, given $\boldsymbol{\xi}$, ξ_{L+1} random samples are generated uniformly on $\boldsymbol{\Theta}$ and randomly inserted among the samples drawn from the present Gaussian components.

Figure 2.4 demonstrates the DAG of the "robustified" parametric model. Setting $\boldsymbol{\Theta}$ to the interval $(0, \pi)$ and $\lambda = 0.5$, Figure 2.6(b) shows the intensities of generated samples from the parametric model of Example 2.1 equipped with the Poisson point process component. It can be observed that the robustified model 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 2.2). Another interesting point that can be seen in Figure 2.6(b) is that the model with the Poisson point process component is able to generate samples with dimensions greater than the number L of Gaussian components. This latter point allows to deal with the vector of observed samples with dimension greater than L (We will clarify it in the process of parameter estimation in following sections).

Henceforth, we only use the robustified variable-dimensional parametric model shown in Figure 2.4. Therefore, we define the parameters $\boldsymbol{\eta}$ of the model as $\boldsymbol{\eta} = (\boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_L, \lambda)$, where $\boldsymbol{\eta}_l = (\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l, \pi_l)$ is the vector of parameters of the l^{th} Gaussian component, $1 \leq l \leq L$. Thus, the space of parameters for the l^{th} Gaussian component, in the uni-variate case, is $\mathrm{N}_l = \mathbb{R} \times \mathbb{R}_+ \times (0, 1]$ and $\mathrm{N} = \prod_{l=1}^{L} \mathrm{N}_l \times \mathbb{R}_+$.

Remark 2.5. For some parameters, distributions other than Gaussians could be used in the parametric model; for example, one could use a log-normal or an inverse gamma distribution for a variance parameter, as in Stephens (1997a). In fact this is a heuristic assumption, which is quite common in label-switching literature (see, for example, Yao, 2011 ; Stephens, 2000 ; Celeux, 1998 ; Celeux et al., 2000), that, if there is no genuine multimodality (see, e.g., Grün and Leisch, 2009), suitably relabeled MCMC samples should be approximately unimodal and normally distributed.

2.3.3 Distribution of the labeled samples

Allocation vectors and space of labeled samples

To estimate the parameters $\boldsymbol{\eta} \in \mathrm{N}$, we first introduce a latent variable interpretation of the parametric model shown in Figure 2.7 by defining latent allocation vectors $\mathbf{z}^{(i)} = (z_1^{(i)}, \dots, z_{k^{(i)}}^{(i)})$, with the same length as $\boldsymbol{x}^{(i)}$, corresponding to the observed sample $\boldsymbol{x}^{(i)}$, for $i = 1, \dots, M$. The element $z_j^{(i)} = l$ indicates that $\boldsymbol{x}_j^{(i)}$ comes from the l^{th} Gaussian component, if $1 \leq l \leq L$; otherwise, if $l = L + 1$, $\boldsymbol{x}_j^{(i)}$ is assumed to have arisen from the Poisson point process component. Note that we had already the vector of latent indicator variables $\boldsymbol{\xi}$. However, for estimation purposes it is more convenient to work with allocation vectors. In order to better present the problem, we introduce the following notations.

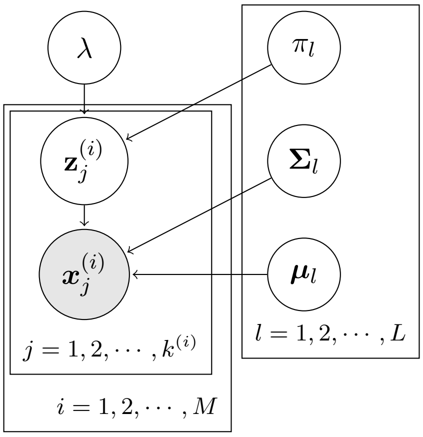


Figure 2.7 – *Latent variable presentation of the proposed parametric model. There are L Gaussian components with the mean $\boldsymbol{\mu}_l$, the covariance matrix $\boldsymbol{\Sigma}_l$, and the probability of presence π_l in the model, $1 \leq l \leq L$. The proposed model also includes a Poisson point process component with the intensity parameter λ to account for outliers. For $i = 1, \dots, M$, $\boldsymbol{x}^{(i)}$ of length $k^{(i)}$ denotes the observed samples (e.g., output of RJ-MCMC) while $\mathbf{z}^{(i)}$ is the corresponding allocation vector. The element $z_j^{(i)} = l$ indicates that $\boldsymbol{x}_j^{(i)}$ is allocated to the l^{th} Gaussian component, if $1 \leq l \leq L$; otherwise, if $l = L+1$, $\boldsymbol{x}_j^{(i)}$ is assumed to have arisen from the Poisson point process component.*

Set of labeled samples. Let $\mathbb{X}_{\mathcal{L}} = \cup_{k \geq 0} \{k\} \times (\boldsymbol{\Theta} \times \mathcal{L})^k$ denote the set of *labeled samples*, where $\mathcal{L} = \{1, 2, \dots, L+1\}$ for some $L \in \mathbb{N}$. This is the set where the "completed" or "augmented" samples live. One point in $\mathbb{X}_{\mathcal{L}}$ can also be considered as a triplet

$(k, \boldsymbol{\theta}_k, \mathbf{z}) = (\boldsymbol{x}, \mathbf{z})$ with $(k, \boldsymbol{\theta}_k) \in \mathbb{X}$ and $\mathbf{z} = (z_1, \ldots, z_k) \in \mathcal{L}^k$. Such a $\mathbf{z}$ will be called an *allocation vector*. The special point $(0, (), ())$ will be denoted by $\varnothing_{\mathcal{L}}$.

Allocation vectors. An allocation vector $\mathbf{z} = (z_1, \ldots, z_k)$, which allocates the elements of the vector of the observed sample $\boldsymbol{x}$ to the components in the parametric model, is a point in the set $\mathcal{Z} = \cup_{k \geq 0} \mathcal{L}^k$. To each $\mathbf{z} \in \mathcal{Z}$ of length k we associate a "counting vector" $\mathbf{n}(\mathbf{z}) = (n_1(\mathbf{z}), \ldots, n_{L+1}(\mathbf{z}))$, where

$$n_l(\mathbf{z}) = \sum_{j=1}^{k} \mathbb{1}_{z_j = l}.$$

Note that $\mathbf{n}(\mathbf{z})$ corresponds to the vector $\boldsymbol{\xi}$ in the generative model point of view introduced in Section 2.3. We define by $\mathcal{Z}_0$ the set of all $\mathbf{z} \in \mathcal{Z}$ such that $n_l(\mathbf{z}) \leq 1$ for all $l \leq L$. In other words, for $\mathbf{z} \in \mathcal{Z}_0$, for each vector of the observed samples $\boldsymbol{x}$, the allocation vector $\mathbf{z} \in \mathcal{Z}_0$ is imposed to not allocate more than one observed element $\boldsymbol{x}_j$, $1 \leq j \leq k$, to one individual Gaussian component. On the other hand, several observed elements can be allocated to the Poisson point process component $(L + 1)$.

Reference measure on $\mathbb{X}_{\mathcal{L}}$. Let $\rho_{\mathcal{L}}$ denote the natural *reference measure* on $\mathbb{X}_{\mathcal{L}}$, which is defined by $\rho_{\mathcal{L}}(A) = \delta_{\varnothing_{\mathcal{L}}}(A) + \sum_{k \geq 1} \sum_{\mathbf{z} \in \mathcal{L}^k} \int_{\boldsymbol{\Theta}^k} \mathbb{1}_A(k, \boldsymbol{\theta}_k, \mathbf{z}) \, \mathrm{d}\boldsymbol{\theta}_k$ for all measurable $A \subset \mathbb{X}_{\mathcal{L}}$. The integral of a measurable function $\varphi : \mathbb{X}_{\mathcal{L}} \to \mathbb{R}$ with respect to $\rho_{\mathcal{L}}$ is given by $\int \varphi \, \mathrm{d}\rho_{\mathcal{L}} = \varphi(\varnothing_{\mathcal{L}}) + \sum_{k \geq 1} \sum_{\mathbf{z} \in \mathcal{Z}} \int_{\boldsymbol{\Theta}^k} \varphi(k, \boldsymbol{\theta}_k, \mathbf{z}) \, \mathrm{d}\boldsymbol{\theta}_k$.

Derivation of the distribution of the labeled samples

Let π be a (variable-dimensional) probability distribution on the set of *unlabeled* samples $\mathbb{X}$. Let $\left\{ Q^{\mathcal{L}}_{\boldsymbol{\eta}}, \, \boldsymbol{\eta} \in \mathrm{N} \right\}$ be a parametric family of probability distributions on the set of *labeled* samples $\mathbb{X}_{\mathcal{L}}$. Each $Q^{\mathcal{L}}_{\boldsymbol{\eta}}$ induces a probability measure $Q_{\boldsymbol{\eta}}$ on $\mathbb{X}$ through the mapping $(\boldsymbol{x}, \mathbf{z}) \mapsto \boldsymbol{x}$. The measure $Q_{\boldsymbol{\eta}}$ is the probability distribution of the unlabeled sample $\boldsymbol{x}$ when $(\boldsymbol{x}, \mathbf{z}) \sim Q^{\mathcal{L}}_{\boldsymbol{\eta}}$.

Observing that the vector $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_{L+1})$ is a deterministic function of $\mathbf{z} : \boldsymbol{\xi} = n(\mathbf{z})$, with $n_l = \sum_{j=1}^{k} \mathbb{1}_{z_j = l}$, for $1 \leq l \leq L + 1$, we can write

$$q_{\boldsymbol{\eta}}(\mathbf{z}) \; = \; q_{\boldsymbol{\eta}}(\mathbf{z} \,|\, \boldsymbol{\xi}) \, q_{\boldsymbol{\eta}}(\boldsymbol{\xi}). \tag{2.7}$$

To compute the first term, remember that the points generated by both the Gaussian components and the Poisson component are *randomly* arranged in $\boldsymbol{\theta}_k$. Therefore, for all $\boldsymbol{\xi} \in \{0, 1\}^L \times \mathbb{N}$ such that $\sum_{l=1}^{L+1} \xi_l = k$,

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

since two arrangements that differ only by the position of the points corresponding to the point process component give rise to the same allocation vector. The second term in (2.7) is given by (2.6):

$$p\,(\boldsymbol{\xi}) \;=\; \frac{e^{-\lambda}\cdot\lambda^{\xi_{L+1}}}{\xi_{L+1}!}\prod_{l=1}^{L}\pi_l^{\xi_l}(1-\pi_l)^{(1-\xi_l)}\,\mathbb{1}_{\{0,1\}^L\times\mathbb{N}}(\boldsymbol{\xi})$$

and therefore,

$$q_{\boldsymbol{\eta}}(\mathbf{z}) \;=\; \frac{(L+\xi_{L+1}-k)!}{L!}\times\lambda^{n_{L+1}}e^{-\lambda}\prod_{l=1}^{L}\pi_l^{n_l}\,(1-\pi_l)^{1-n_l}\,\mathbb{1}_{\mathcal{Z}_0}(\mathbf{z})\,, \tag{2.8}$$

since $\mathcal{Z}_0 = n^{-1}(\{0,1\}^L\times\mathbb{N})$.

The other density needed to be defined is the conditional likelihood of the parametric model, i.e., $q_{\boldsymbol{\eta}}(\boldsymbol{x}|\mathbf{z})$. Recall that the generated points from the point process component are distributed uniformly over $\boldsymbol{\Theta}$ and they are independent given their number ξ_{L+1} (see Section 2.3.2). Thus for each element of the vector of the observed samples $\boldsymbol{x}$, we have

$$q_{\boldsymbol{\eta}}(\boldsymbol{x}_j\,|\,z_j) \;=\; \begin{cases}\mathcal{N}\left(\boldsymbol{x}_j\,|\,\boldsymbol{\mu}_{z_j},\,\boldsymbol{\Sigma}_{z_j}\right) & \text{if } z_j \;\le\; L,\\ \frac{1}{|\boldsymbol{\Theta}|} & \text{if } z_j \;=\; L+1.\end{cases} \tag{2.9}$$

As a result, we obtain

$$q_{\boldsymbol{\eta}}(\boldsymbol{x}\,|\,\mathbf{z}) \;=\; |\boldsymbol{\Theta}|^{-\xi_{L+1}}\prod_{\substack{1\le j\le k\\ z_j\ne L+1}}\mathcal{N}\left(\boldsymbol{x}_j\,|\,\boldsymbol{\mu}_{z_j},\boldsymbol{\Sigma}_{z_j}\right), \tag{2.10}$$

and the joint pdf of $(\boldsymbol{x},\mathbf{z})$ under $Q_{\boldsymbol{\eta}}^{\mathcal{L}}$ reads

$$q_{\boldsymbol{\eta}}(\boldsymbol{x},\mathbf{z}) \;=\; q_{\boldsymbol{\eta}}(\boldsymbol{x}\,|\,\mathbf{z})\,q_{\boldsymbol{\eta}}(\mathbf{z}). \tag{2.11}$$

2.4 Estimating the model parameters: Algorithm I

In this section, we propose a first algorithm to estimate the parameters $\boldsymbol{\eta}\in\mathrm{N}$ of the variable-dimensional parametric model $q_{\boldsymbol{\eta}}$ shown in Figure 2.4. This algorithm, loosely speaking, consists in fitting $q_{\boldsymbol{\eta}}$ to the true variable-dimensional posterior density f through minimizing the KL divergence from f to $q_{\boldsymbol{\eta}}$, denoted by $\mathcal{D}_{KL}\left(f\,\|\,q_{\boldsymbol{\eta}}\right)$. Remember that both densities are defined on $\mathbb{X}=\bigcup_{k\ge 0}\{k\}\times\boldsymbol{\Theta}^k$, $k\in\mathcal{K}$. In what follows, we assume that M variable-dimensional samples $\boldsymbol{x}^{(i)}$ of length $k^{(i)}$, $i=1,\ldots,M$, have been generated from the true posterior π, with density f, by a trans-dimensional MCMC sampler (such as the RJ-MCMC sampler explained in Section 1.4).

Figure 2.8 illustrates the block diagram of the proposed summarization approach. Note that there are two models in the problem we are dealing with here; one is the hierarchical model of the observations $\mathbf{y}$ from which the true posterior f is defined (see Figure 3.1 for an example such a hierarchical model in the problem of sinusoid detection). The other one is the variable-dimensional parametric model $q_{\boldsymbol{\eta}}$ we proposed as an approximate posterior to summarize f.

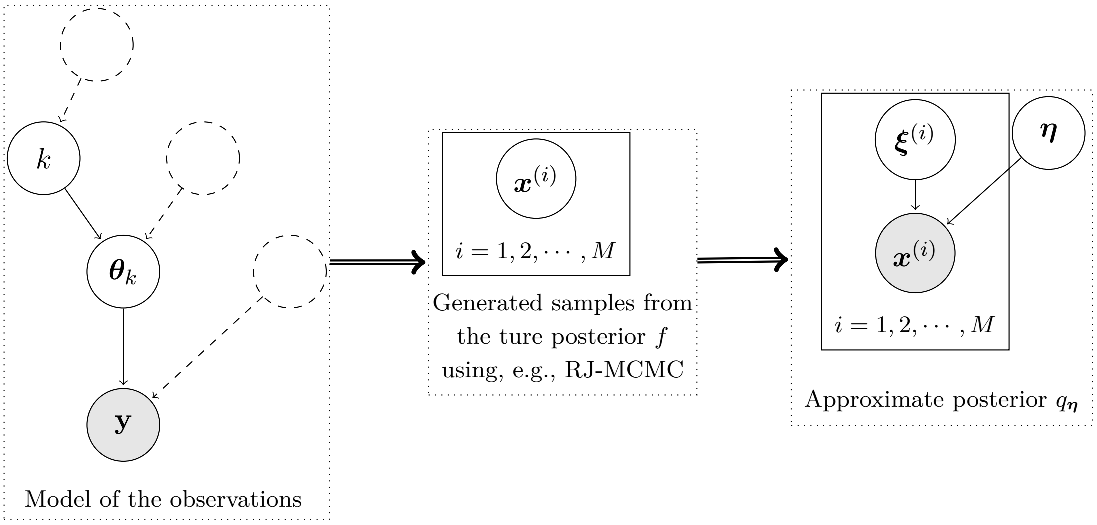


Figure 2.8 – *Block diagram showing the structure of the proposed summarization approach that consists in fitting an approximate model $q_{\boldsymbol{\eta}}$ to the posterior f of interest by minimizing a divergence measure of f from $q_{\boldsymbol{\eta}}$ using samples $\boldsymbol{x}^{(i)}$ generated from f. The left block illustrates the DAG of the observations $\mathbf{y}$ of length N for the trans-dimensional problem under study where the unknown parameters are the number k of components and the vector of component-specific parameters. The empty dashed nodes present other unknown parameters, which are not to be summarized (hyperparameters, for instance). See Figure 3.1 for an example such a hierarchical model in the problem of sinusoid detection. The middle block represents the process of drawing samples, $\boldsymbol{x}^{(i)}$ of length $k^{(i)}$, for $i = 1, \ldots, M$, from the posterior f of interest using, e.g., RJ-MCMC. The right block demonstrates the graphical model of the approximate posterior (the parametric model) $q_{\boldsymbol{\eta}}$ introduced in Section 2.3. Note that the generated samples $\boldsymbol{x}^{(i)}$ are considered as the observed data for the parametric model $q_{\boldsymbol{\eta}}$.*

We derive a criterion based on the KL divergence from f to $q_{\boldsymbol{\eta}}$ using the observed samples $\boldsymbol{x}^{(i)}$, $i = 1, \ldots, M$, in Section 2.4.1. Next, Section 2.4.2 describes a SEM-type algorithm proposed to estimate the parameters of the model.

2.4.1 Divergence measures and randomized allocation procedures

We propose to fit the parametric distribution $q_{\boldsymbol{\eta}}$ to the posterior f of interest by minimizing a divergence measure from f to $q_{\boldsymbol{\eta}}$. We use the KL divergence as a divergence measure in this Section and a more robust α-divergence measure proposed by Basu et al. (1998) (BHHJ α-divergence) in Section 2.5. Minimizing the KL divergence is used in the derivation of Maximum Likelihood Estimator (MLE). Other examples of parameter estimation methods based on minimizing density-based divergences can be found in the recent work of Broniatowski and Keziou (2009) and Basu et al. (1998).

Let P and Q be two probability measures on $(\mathbb{X}, \mathcal{B})$ such that $P \ll Q$, i.e., P is absolutely continuous with respect to Q. We first introduce the family of "ϕ-divergence" measures which have the KL divergence as a special case (see, e.g., Parclo, 2005 ; Csiszár, 1967):

Definition 2.1. *The ϕ-divergence measure from the probability distributions P to Q is*

$$\mathcal{D}_{\phi}\left(P \,\|\, Q\right) \;=\; \int_{\mathbb{X}} \phi\left(\frac{\mathrm{d}P}{\mathrm{d}Q}\right) \mathrm{d}Q, \tag{2.12}$$

where ϕ is a convex functions, such that, $\phi(1) = 0$.

Then, the KL divergence is a special case of the above family of ϕ-divergence measures obtained by setting $\phi(\boldsymbol{x}) = \boldsymbol{x}\log(\boldsymbol{x})$ in (2.12) which reads

$$\mathcal{D}_{KL}\left(P \,\|\, Q\right) \;=\; \int_{\mathbb{X}} \log\left(\frac{\mathrm{d}P}{\mathrm{d}Q}\right) \mathrm{d}P. \tag{2.13}$$

Now, in order to provide a meaningful summary of the (variable-dimensional) probability distribution π on the set $\mathbb{X}$, we want to approximate it by a member of the family $\{Q_{\boldsymbol{\eta}},\, \boldsymbol{\eta} \in \mathrm{N}\}$. Working with the distributions Q_{η} directly is not convenient, however, since they are defined as the marginal distribution of $\boldsymbol{x}$ in a sample $(\boldsymbol{x}, \mathbf{z}) \sim Q_{\boldsymbol{\eta}}^{\mathcal{L}}$ and, therefore, involve summations over the set of all possible allocation vectors.

Proposition 2.2. *For the family of ϕ-divergence measures defined in Definition 2.1, minimizing the divergence measure from π to $Q_{\boldsymbol{\eta}}$ is equivalent to minimizing the ones from the augmented distribution $\pi_{\boldsymbol{\eta}}^{\mathcal{L}}(\mathrm{d}\boldsymbol{x}\mathrm{d}\mathbf{z}) = \pi(\mathrm{d}\boldsymbol{x})\, Q_{\boldsymbol{\eta}}^{\mathcal{L}}(\mathrm{d}\mathbf{z}|\boldsymbol{x})$ on $\mathbb{X}_{\mathcal{L}}$ to $Q_{\boldsymbol{\eta}}^{\mathcal{L}}$.*

Proof.

$$\begin{aligned}\mathcal{D}_\phi\left(\pi \,\|\, Q_{\boldsymbol{\eta}}\right) &= \int_{\mathbb{X}} \phi\left(\frac{\mathrm{d}\pi}{\mathrm{d}Q_{\boldsymbol{\eta}}}\right) \mathrm{d}Q_{\boldsymbol{\eta}} \\ &= \int_{\mathbb{X}} \phi\left(\frac{\pi(\mathrm{d}\boldsymbol{x})Q^{\mathcal{L}}_{\boldsymbol{\eta}}(\mathrm{d}\mathbf{z}|\boldsymbol{x})}{Q^{\mathcal{L}}_{\boldsymbol{\eta}}(\mathrm{d}\boldsymbol{x}\mathrm{d}\mathbf{z})}\right) Q_{\boldsymbol{\eta}}(\mathrm{d}\boldsymbol{x}) \\ &= \int_{\mathbb{X}_{\mathcal{L}}} \phi\left(\frac{\mathrm{d}\pi^{\mathcal{L}}_{\boldsymbol{\eta}}}{\mathrm{d}Q^{\mathcal{L}}_{\boldsymbol{\eta}}}\right) \mathrm{d}Q^{\mathcal{L}}_{\boldsymbol{\eta}} \;=\; \mathcal{D}_\phi\left(\pi^{\mathcal{L}}_{\boldsymbol{\eta}} \,\|\, Q^{\mathcal{L}}_{\boldsymbol{\eta}}\right).\end{aligned}$$

□

Proposition 2.2 allows us to calculate the ϕ-divergence measure from π to $Q_{\boldsymbol{\eta}}$ by augmenting the unlabeled observed samples $\boldsymbol{x}$ using the allocation vectors $\mathbf{z}$ which are distributed according to the conditional posterior distribution $Q^{\mathcal{L}}_{\boldsymbol{\eta}}(\,\cdot\,|\boldsymbol{x})$. The conditional distribution $Q^{\mathcal{L}}_{\boldsymbol{\eta}}(\mathrm{d}\mathbf{z}|\boldsymbol{x})$ can be thought of as a *randomized allocation procedure* that allows to draw an allocation vector $\mathbf{z} \sim Q^{\mathcal{L}}_{\boldsymbol{\eta}}(\,\cdot\,|\boldsymbol{x})$ given an unlabeled sample $\boldsymbol{x}$.

Remark 2.6. The general relabeling algorithm proposed in (Stephens, 2000, Section 4) can be understood in this framework as trying to minimize a posterior expected loss, e.g., the KL divergence measure with respect to the labeled posterior $\tilde{\pi}^{\mathcal{L}}_{\boldsymbol{\eta}}(\mathrm{d}\boldsymbol{x}\mathrm{d}\mathbf{z}) = \pi(\mathrm{d}\boldsymbol{x})\, a_{\boldsymbol{\eta}}(\mathrm{d}\mathbf{z}|\boldsymbol{x})$ on $\mathbb{X}_{\mathcal{L}}$, where $a_{\boldsymbol{\eta}}$ is a deterministic allocation procedure related to the chosen model.

Now, setting $f = \frac{\mathrm{d}\pi}{\mathrm{d}\rho}$ and $q_{\boldsymbol{\eta}} = \frac{\mathrm{d}Q_{\boldsymbol{\eta}}}{\mathrm{d}\rho}$, we can define the criterion to be minimized based on the KL divergence (2.13) as

$$\mathcal{J}(\boldsymbol{\eta}) \;=\; \mathcal{D}_{KL}\left(\pi \,\|\, Q_{\boldsymbol{\eta}}\right) \;=\; \int_{\mathbb{X}} f \log\left(\frac{f}{q_{\boldsymbol{\eta}}}\right) \mathrm{d}\rho.$$

Furthermore, using available (variable-dimensional) samples $\boldsymbol{x}^{(i)}$, for $i = 1, \dots, M$, generated according to the posterior f, the above criterion can be approximated by

$$\hat{\mathcal{J}}_M(\boldsymbol{\eta}) \;=\; \mathcal{D}_{KL}\left(\pi \,\|\, Q_{\boldsymbol{\eta}}\right) \;\simeq\; -\frac{1}{M}\sum_{i=1}^{M} \log\left(q_{\boldsymbol{\eta}}(\boldsymbol{x}^{(i)})\right) + C, \tag{2.14}$$

where C is a constant that does not depend on the parameters $\boldsymbol{\eta}$. One should note that minimizing (2.14) amounts to estimating $\boldsymbol{\eta}$ by

$$\hat{\boldsymbol{\eta}} \;=\; \mathrm{argmax}_{\boldsymbol{\eta}} \sum_{i=1}^{M} \log\left(q_{\boldsymbol{\eta}}(\boldsymbol{x}^{(i)})\right), \tag{2.15}$$

which is formally the MLE of $\boldsymbol{\eta}$ for an iid samples distributed according to $q_{\boldsymbol{\eta}}$.

In the following section, we propose an SEM-type algorithm to compute this estimator.

2.4.2 SEM-type algorithm

To estimate the model parameters $\boldsymbol{\eta} \in \mathrm{N}$, one of the extensively used algorithms for Maximum Likelihood (ML) parameter estimation in latent variable models is the EM algorithm proposed by Dempster et al. (1977) (see, e.g., McLachlan and Krishnan, 2008, for more information). It consists of two steps; the Expectation (E)-step and the Maximization (M)-step. The E-step, at iteration $(r+1)$, consists in computing the expectation of the completed-data log-likelihood, i.e.,

$$\sum_{i=1}^{M} \log\left(q_{\boldsymbol{\eta}}(\boldsymbol{x}^{(i)}, \mathbf{z}^{(i)})\right),$$

where $q_{\boldsymbol{\eta}}(\boldsymbol{x}, \mathbf{z})$ is defined in (2.8)–(2.11), with respect to the conditional posterior of the latent variables given the estimated parameters in the previous step, $\hat{\boldsymbol{\eta}}^{(r)}$, that is,

$$q_{\hat{\boldsymbol{\eta}}^{(r)}}(\mathbf{z}^{(i)} \mid \boldsymbol{x}^{(i)}) \;=\; \frac{q_{\hat{\boldsymbol{\eta}}^{(r)}}(\boldsymbol{x}^{(i)},\, \mathbf{z}^{(i)})}{q_{\hat{\boldsymbol{\eta}}^{(r)}}(\boldsymbol{x}^{(i)})}. \tag{2.16}$$

It turns out, however, that the EM-type algorithms, which have been used in similar works (Stephens, 2000 ; Sperrin et al., 2010 ; Bai et al., 2011), but only when k is small, are not appropriate for solving this problem, as computing the expectation in the E-step is intricate. More explicitly, in the problem we are dealing with, the computational burden of the summation in the E-step over the set of all possible allocation vectors $\mathbf{z}$ increases very rapidly with 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 $\frac{L!}{(L-k)!} \approx 1.1 \times 10^{10}$ terms, assuming $\xi_{L+1} = 0$.

In the literature, there are a few methods proposed for overcoming this limitation based on approximating the E-step by Monte Carlo simulation, such as the SEM algorithm developed by Broniatowski et al. (1983) ; Celeux and Diebolt (1985, 1992) and the Monte Carlo EM (MCEM) algorithm proposed by Wei and Tanner (1990). It is also possible, in a Bayesian setting, to assign prior distributions over the unknown parameters and study their posterior distributions, for example, using MCMC methods, in the spirit of the "data augmentation" algorithm proposed by Tanner and Wong (1987). Here, we opt for the SEM algorithm to estimate the unknown parameters. We describe briefly the SEM algorithm and derive a SEM-type algorithm to estimate the model parameters $\boldsymbol{\eta}$ in the following sections.

Stochastic EM

In the SEM algorithm (Broniatowski et al., 1983 ; Celeux and Diebolt, 1985, 1992) (see also Gilks et al. (1996, Chapter 15) for a more recent review), the E-step of the EM algorithm is substituted with stochastic simulation of the latent variables or missing data from their conditional posterior distributions given the previous estimates of the unknown parameters, i.e., $q_{\hat{\boldsymbol{\eta}}^{(r)}}(\mathbf{z} \,|\, \boldsymbol{x})$ in our notations. This step is called Stochastic (S)-step. Then, these random samples are used to construct the so-called pseudo-completed log-likelihood, that is,

$$\sum_{i=1}^{M} \log \left(q_{\hat{\boldsymbol{\eta}}^{(r)}}(\boldsymbol{x}^{(i)}, \mathbf{z}^{(i)}) \right).$$

Next, in the M-step, the sum over all pseudo-completed log-likelihoods is maximized. The "proposed SEM-type" algorithm for the problem we are dealing with is described in Algorithm 2.1.

Algorithm 2.1. *At the* $(r+1)^{th}$ *iteration of the SEM algorithm,*

S-step: *For* $i = 1, \ldots, M$,

- *draw allocation vectors* $\mathbf{z}^{(i)} \sim q_{\hat{\boldsymbol{\eta}}^{(r)}}(\,\cdot \,|\, \boldsymbol{x}^{(i)})$ *defined in* (2.16).

E-step: *construct the pseudo-completed log-likelihood*

$$\hat{\eth}_M(\boldsymbol{\eta}) \;=\; -\sum_{i=1}^{M} \log \left(q_{\boldsymbol{\eta}}(\boldsymbol{x}^{(i)}, \mathbf{z}^{(i)}) \right).$$

M-step: *Estimate* $\hat{\boldsymbol{\eta}}^{(r+1)}$ *such that*

$$\hat{\boldsymbol{\eta}}^{(r+1)} \;=\; \operatorname{argmin}_{\boldsymbol{\eta}} \hat{\eth}_M(\boldsymbol{\eta}). \tag{2.17}$$

Stochastic step

Here, it is assumed that we are at iteration $(r+1)$ of the SEM-type algorithm. As explained in Algorithm 2.1, the S-step of the SEM-type algorithm consists in generating the allocation vectors $\mathbf{z}^{(i)}$ from the conditional posterior distribution $q_{\hat{\boldsymbol{\eta}}^{(r)}}(\,\cdot \,|\, \boldsymbol{x}^{(i)})$ expressed in (2.16), for $i = 1, \ldots, M$. Unfortunately, there seem to be no "easy" way to sample directly from $q_{\hat{\boldsymbol{\eta}}^{(r)}}(\,\cdot \,|\, \boldsymbol{x}^{(i)})$, which is a complex discrete distribution on $\mathcal{Z}_0$. Moreover, we have to deal with the unnormalized conditional posterior distribution

$$\hat{q}_{\hat{\boldsymbol{\eta}}^{(r)}}(\mathbf{z} \,|\, \boldsymbol{x}) \;\propto\; q_{\hat{\boldsymbol{\eta}}^{(r)}}(\boldsymbol{x},\, \mathbf{z}), \tag{2.18}$$

since computing the normalizing constant involves summing over all possible permutations of the allocation vector $\mathbf{z}^{(i)}$ circumventing which has been the main reason for using the SEM algorithm (note that this summation is the one encountered in the E-step of the EM algorithm).

Therefore, we propose to carry out this step using MCMC methods. We develop an Independent-MH (I-MH) sampler (see Section 1.2.2) in which the proposition of the next state does not depend on the current state of the Markov chain. The elements of the allocation vector $\mathbf{z}^{(i)}$ are sampled using a “sampling without replacement” strategy, as one element of the vector of the observed samples cannot be allocated to more than one Gaussian component (see the conditions imposed on the allocation vector in Section 2.3.3).

Let $\mathbb{PC}$ be the set of possible components that the elements of the vector of the observed samples can be allocated to. When sampling each allocation vector, at the very beginning $\mathbb{PC} = \{1, \ldots, L+1\}$. But, as we proceed allocating samples to the (Gaussian) components, $\mathbb{PC}$ will be modified to respect the one-to-one allocation condition of Gaussian components. Moreover, let

$$\tilde{g}_{\boldsymbol{\eta}}(\boldsymbol{x}_j,\, l) \;=\; \begin{cases} \mathcal{N}\left(\boldsymbol{x}_j \,|\, \boldsymbol{\mu}_l,\, \boldsymbol{\Sigma}_l\right) \frac{\pi_l}{1-\pi_l} & \text{if } 1 \,\leq\, l \,\leq\, L \\ \frac{\lambda}{|\boldsymbol{\Theta}|} & \text{if } l \,=\, L+1. \end{cases} \tag{2.19}$$

Then, we can rewrite $q_{\boldsymbol{\eta}}(\boldsymbol{x},\, \mathbf{z})$ as

$$q_{\boldsymbol{\eta}}(\boldsymbol{x},\, \mathbf{z}) \;\propto\; \prod_{\substack{1\leq j \leq k \\ l \,=\, z_j}} \tilde{g}_{\boldsymbol{\eta}}(\boldsymbol{x}_j,\, l) \;=\; \frac{\lambda}{|\boldsymbol{\Theta}|}^{n_{L+1}} \prod_{\substack{1\leq j \leq k \\ z_j \neq L+1}} \mathcal{N}\left(\boldsymbol{x}_j \,|\, \boldsymbol{\mu}_{z_j}, \boldsymbol{\Sigma}_{z_j}\right) \frac{\pi_{z_j}}{1-\pi_{z_j}}, \tag{2.20}$$

which serves as the target distribution for the proposed I-MH sampler. The mechanism to propose an allocation vector $\mathbf{z}^\star = (z_1^\star, \ldots, z_k^\star)$ given the vector of the observed samples $\boldsymbol{x}$ and the previous estimated parameters $\hat{\boldsymbol{\eta}}^{(r)}$ in the proposed I-MH sampler is described in Algorithm 2.2.

Next, the proposed vector of allocations $\mathbf{z}^\star$ is accepted and replaced the old one denoted by $\mathbf{z}^o$ with the MH acceptance probability of

$$\alpha\left(\mathbf{z}^o,\, \mathbf{z}^\star\right) \;=\; \left\{ 1,\, \underbrace{\frac{q_{\hat{\boldsymbol{\eta}}^{(r)}}(\boldsymbol{x},\, \mathbf{z}^\star)}{q_{\hat{\boldsymbol{\eta}}^{(r)}}(\boldsymbol{x},\, \mathbf{z}^o)}}_{\text{target ratio}} \times \underbrace{\frac{\prod_{\substack{1\leq j \leq k \\ l \,=\, z_j^o}} g_{\hat{\boldsymbol{\eta}}^{(r)}}(\boldsymbol{x}_j,\, l)}{\prod_{\substack{1\leq j \leq k \\ l \,=\, z_j^\star}} g_{\hat{\boldsymbol{\eta}}^{(r)}}(\boldsymbol{x}_j,\, l)}}_{\text{proposal ratio}} \right\}, \tag{2.21}$$

Algorithm 2.2. *Mechanism to propose an allocation vector* $\mathbf{z}^\star$*:*
Set $\mathbb{PC} = \{1, \dots, L+1\}$*, then, for* $j = 1, \dots, k$*, do*

i) Compute unnormalized probabilities $\tilde{g}_{\hat{\boldsymbol{\eta}}^{(r)}}(\boldsymbol{x}_j,\, l)$*, for* $l \in \mathbb{PC}$*, as expressed in* (2.19)*.*

ii) Normalize $\tilde{g}_{\hat{\boldsymbol{\eta}}^{(r)}}(\boldsymbol{x}_j,\, l)$ *to sum to one to obtain normalized probabilities* $g_{\hat{\boldsymbol{\eta}}^{(r)}}(\boldsymbol{x}_j,\, l)$*.*

iii) Draw the j^{th} *element of the allocation vector* $z_j^\star$ *using a multinomial random generator with* $\{g_{\hat{\boldsymbol{\eta}}^{(r)}}(\boldsymbol{x}_j,\, l)\}_{l \in \mathbb{PC}}$ *as probabilities.*

iv) If $z_j^\star \leq L$*, remove the selected label from* $\mathbb{PC}$ *(to respect the one-to-one allocation condition of the Gaussian components).*

with $g_{\hat{\boldsymbol{\eta}}^{(r)}}(\boldsymbol{x}_j,\, l)$ is defined in Algorithm 2.2.

In our experiments, we observed that the performance of the algorithm can be improved if the order in which the elements of the vector of observed sample are scanned is chosen randomly. In other words, in the proposition of $\mathbf{z}^\star$, instead of systematically scanning the elements of the observed sample $\boldsymbol{x}$ from 1 to k, they are scanned randomly with equal probabilities. In particular, this scanning strategy becomes beneficial when there are two (or more) elements in the vector of observed sample competing for the same Gaussian component, i.e., they both (all) have non negligible probabilities of being allocated to the Gaussian component l, say. Then, by random scan strategy, we grant both (all) rather "fair" situation[1]. We use the random scan strategy in the I-MH sampler hereafter.

Remark 2.7. Selecting the random scan order is not considered as part of the proposal distribution. In other words, we are using a mixture of I-MH moves rather than an I-MH move with a mixture of proposal distributions (which would make the computation of the MH ratio very expensive).

Remark 2.8. One might think that the self normalized importance sampling method introduced in Section 1.3.1 can be used as well to generate the allocation vectors from the unnormalized density (2.18), using Algorithm 2.2 as a instrumental distribution. This is not possible in an SEM algorithm, however, since the unknown normalizing constant depends on $\boldsymbol{x}^{(i)}$. A possible workaround would be to generate several allocation vectors $\mathbf{z}^{(i,j)}$

[1] Similar scanning strategies, but with different justifications, exist in the literature of the Gibbs sampler; see, for example, (Liu, 2001, pages 130–131).

for each $\boldsymbol{x}^{(i)}$, in the spirit of the MCEM algorithm, but the computational cost would be much more important.

Maximization step

Turning to the M-step of the SEM-type algorithm described in Algorithm 2.1, the parameters $\boldsymbol{\eta} = \{\boldsymbol{\eta}_l\}_{1\leq l\leq L}$, with $\boldsymbol{\eta}_l = \{\pi_l, \boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l\}$, of the model together with the intensity λ of the Poisson point process component are estimated by maximizing (2.17) through the following equations:

$$\begin{aligned}
\hat{\boldsymbol{\mu}}_l^{(r+1)} &= \frac{1}{M_l}\sum_{i=1}^{M_l} \boldsymbol{x}_{\to l}^{(i)}, \\
\hat{\boldsymbol{\Sigma}}_l^{(r+1)} &= \frac{1}{M_l}\sum_{i=1}^{M_l} \left(\boldsymbol{x}_{\to l}^{(i)} - \hat{\boldsymbol{\mu}}_l^{(r+1)}\right)\left(\boldsymbol{x}_{\to l}^{(i)} - \hat{\boldsymbol{\mu}}_l^{(r+1)}\right)^t, \\
\hat{\pi}_l^{(r+1)} &= \frac{M_l}{M}, \\
\hat{\lambda}^{(r+1)} &= \frac{\sum_{i=1}^{M} n_{L+1}^{(i)}}{M},
\end{aligned} \tag{2.22}$$

where M_l is the number of observed samples in which one element is allocated to the Gaussian component l and $\boldsymbol{x}_{\to l}^{(i)}$ is the element of the i^{th} observed samples where $\mathbf{z}^{(i)} = l$, with $1 \leq l \leq L$.

Convergence of the SEM algorithm

The following convergence results have been proved for the SEM algorithm in the general form by Nielsen (2000a,b) and in the particular example of mixture analysis problems by Diebolt and Celeux (1993). Assume that, for $i = 1, \ldots, M$, the observed data samples $\boldsymbol{x}^{(i)}$ are *i.i.d* and it is possible to sample the latent variables $\mathbf{z}^{(i)}$ independently. Then, under the assumptions given in Nielsen (2000a), for a fixed number M of observed samples:

i) For each i, the Markov chain $\{\mathbf{z}_{(r)}^{(i)}\}_{r\in\mathbb{N}}$ is irreducible and aperiodic.

ii) The random sequence $\{\hat{\boldsymbol{\eta}}^{(r)}\}_{r\in\mathbb{N}}$ generated by the SEM algorithm is a homogeneous Markov chain.

iii) The Markov chain $\{\hat{\boldsymbol{\eta}}^{(r)}\}_{r\in\mathbb{N}}$ is ergodic, in most cases, with Ψ_M as its stationary distribution.

Moreover, denoting by Ψ_M the stationary distribution of $\{\hat{\boldsymbol{\eta}}^{(r)}\}_{r\in\mathbb{N}}$ and by $\hat{\boldsymbol{\eta}}_M^{SEM} = \text{mean}(\Psi_M)$ the estimated value of the unknown parameters provided by the SEM algorithm, and letting M tends to infinity, the following asymptotic results are proved

i) $\hat{\boldsymbol{\eta}}_M^{SEM}$ is a consistent estimator of $\boldsymbol{\eta}$,

ii) $\sqrt{M}\,(\tilde{\boldsymbol{\eta}}_M - \boldsymbol{\eta}_M)$ is asymptotically normal distributed with zero mean and positive variance matrix, where $\tilde{\boldsymbol{\eta}}_M$ is a random variable drawn from the stationary distribution Ψ_M and $\boldsymbol{\eta}_M$ is the unique consistent solution of the likelihood.

Unfortunately, the assumptions in Diebolt and Celeux (1993) ; Nielsen (2000a,b) do not hold in the problem we are dealing with as, 1) the observed samples $\boldsymbol{x}^{(i)}$ are *correlated* owing to the fact that they are generated from the complex posterior distribution using some MCMC methods, e.g., the RJ-MCMC sampler; 2) an I-MH sampler is used in Algorithm 2.2 to draw $\mathbf{z}^{(i)}$ from the conditional posterior distribution (2.16). Empirical evidence of the "good" convergence properties of our SEM-type algorithm will be provided in the next two chapters.

2.5 Estimating the model parameters: Algorithms II & III

2.5.1 Robustness issue

Preliminary experiments with the SEM-type algorithm described in Section 2.4 were not satisfactory, because the sample mean and (co)variance estimates expressed in (2.22) obtained from minimizing the KL divergence from the posterior distribution f to the parametric model $q_{\boldsymbol{\eta}}$ still suffers from sensitivity to the outliers in the observed samples, even after including the Poisson point process component (see Section 2.3). In this section, we propose two robustified SEM-type algorithms which differ mainly from Algorithm 2.1 in the M-step.

First solution: using robust estimators in the M-step

The first solution is to use robust estimates of the mean and (co)variance parameters in the spirit of "robust statistics" literature (Huber and Ronchetti, 2009 ; Maronna et al., 2006). More explicitly, in the univariate case, i.e., when the dimension d of the observed sample $\boldsymbol{x}$ is one, we use the median and the Normalized InterQuartile Range (N-IQR) instead of the empirical mean and variance estimates (2.22) in the M-step. Denoting by $\boldsymbol{x}_{\to l}$ the samples allocated to the l^{th} Gaussian component, as in (2.22), N-IQR is the estimator of the standard deviation defined by

$$\text{N-IQR}(\boldsymbol{x}_{\to l}) \;=\; \frac{Q_3(\boldsymbol{x}_{\to l}) - Q_1(\boldsymbol{x}_{\to l})}{2\,\Phi^{-1}(0.75)},$$

where Φ is the CDF of standard normal distribution and Q_j is the j^{th} empirical quartile.

Similar robustness concerns are widespread in the clustering literature; see, e.g., Davé and Krishnapuram (1997) and the references therein. In the multivariate case (i.e., $d > 1$), if we assume that the covariance matrix is diagonal, then, we can still use the median and interquartile range of each coordinate separately as robust alternatives. Otherwise, more complicated (iterative) robust algorithms should be used; see, e.g., Maronna et al. (2006, Chapter 6), for more information.

To clarify the benefit of this modification, we ran 100 iterations of the SEM-type algorithm on the sinusoid detection example for the experiment shown in Figure 2.2, twice: once with the empirical mean and variance estimates (2.22) in the M-step (called Algorithm I) and once with the corresponding robust estimates (called Algorithm II). Note that we only focus here on the radial frequencies of sinusoidal components; see Chapter 3 for more information. Hence, $d = 1$ and $\boldsymbol{\Theta} = (0, \pi)$. The number L of Gaussian components was set to three (the posterior probability of $\{k \leq 3\}$ is approximately 90.3%) and the initial values for means μ_l and variances s_l^2, with $1 \leq l \leq L$, were estimated from the posterior distribution of sorted radial frequencies given $k = L$.

Figures 2.9 and 2.10 illustrate the histogram of the "labeled samples" $(\boldsymbol{x}^{(i)}, \mathbf{z}^{(i)}) \in \mathbb{X}_{\mathcal{L}}$, with $i = 1, \ldots, M$, i.e., the samples allocated to the Gaussian and Poisson point process components (see Section 2.3.3 for more information), along with the pdf's of the estimated Gaussian components. To obtain those histograms, we ran the randomized allocation procedure, i.e., the I-MH sampler described in Algorithm 2.2 developed for generating the allocation vectors $\mathbf{z}^{(i)}$, with $i = 1, \ldots, M$, given the estimated parameters $\hat{\boldsymbol{\eta}}$. To reduce the randomness effect of the allocation procedure, we generated 10 allocation vectors for each vector of observed sample. Furthermore, the final estimated parameters after running both algorithms for 100 iterations are presented in the corresponding panels. The true values of radial frequencies were $\boldsymbol{\omega}_k = (0.628, 0.677, 0.726)$. It can be easily read from both figures that the most notable difference between using the robust and simple estimates in the M-step is in summarizing the information pertaining to the middle component, particularly its dispersion (variance) parameter.

Comparing the histogram of the allocated samples to the middle (or second) Gaussian component and the corresponding pdf of fitted distribution obtained using simple estimates with the ones obtained using robust estimates shown in the top right panel of Figures 2.9 and 2.10, respectively, it can be observed that in the former case not only the resulting Gaussian distribution has a larger variance but also the distribution of its allocated samples is bimodal. This is indeed due to the fact that using simple estimates allows a Gaussian component to have a large variance by considering the samples located in the tails of the distribution. As a consequence, it makes the probability of catching samples located far away from its mean—allocating which expands the pdf by increasing again its variance—non-negligible. In this specific example, this phenomenon results in two modes in the histogram of the labeled samples to the second Gaussian component (see the top right panel of Figure 2.9) around the means of the other two adjacent components. On the other hand, using robust estimates in the M-step, the estimated means and variances are not significantly affected neither by the existing outliers nor by the samples of the adjacent components; as a result, the middle component does not catch samples located too far away from its mean that indeed should be allocated to the other components. .

Second solution: Modifying the divergence measure

In the following, we propose to deal with this robustness issue in a theoretically sounder way by replacing the KL divergence with a divergence measure that enjoys robustness properties. In the literature, there have been several attempts at estimating model parameters by minimizing robust divergence measures. However, in most of the proposed robust divergences, it is necessary to use some nonparametric smoothing method, e.g., kernel density estimation, of the true density from the observed data samples which consequently makes the algorithm sensitive to the parameters of the smoothing method; for more information see Basu et al. (1998) ; Jones et al. (2001) and references therein.

Remark 2.9. Another approach to avoid using kernel density estimates, called dual ϕ-divergence estimates, has been proposed by Broniatowski and Keziou (2009). However, in this work, we only use the divergence proposed by Basu et al. (1998) as a robust divergence.

We describe briefly the BHHJ α-divergence and its properties in Section 2.5.2. Then, we propose a new SEM-type algorithm to fit the parametric model $q_{\boldsymbol{\eta}}$ to the posterior f of interest by minimizing the BHHJ α-divergence. In fact, it turns out that by modifying the divergence measure only the M-step of the SEM-type algorithm is changed. In other words, the I-MH sampler described in Section 2.4.2 for stochastic simulation of the allocation vectors in the S-step remains unaltered. The optimization procedure developed for minimizing the obtained criterion based on BHHJ α-divergence is explained in Section 2.5.3.

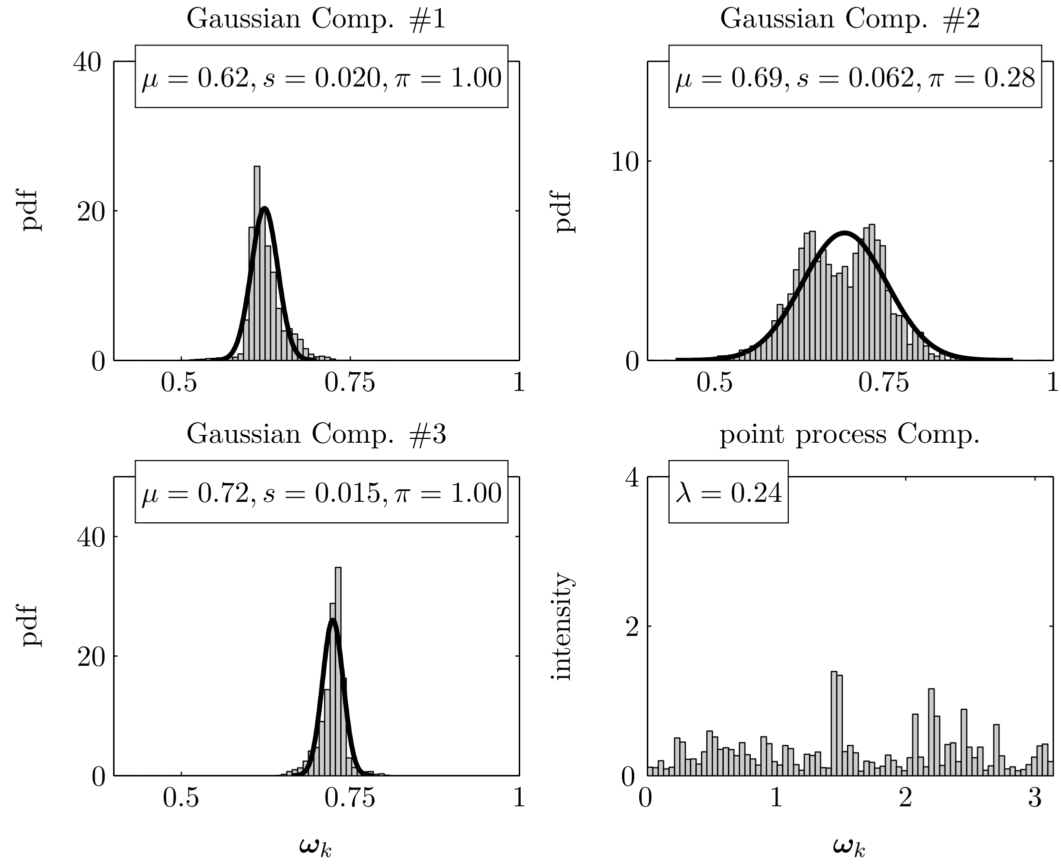


Figure 2.9 – *Histograms of the labeled samples* $(\boldsymbol{x}^{(i)}, \mathbf{z}^{(i)})$*, with* $i = 1, \ldots, M$*, that is, the samples allocated to the Gaussian and Poisson point process components, versus the pdf's of estimated Gaussian components in the model (black solid line) for the summarizing algorithm derived from minimizing the KL divergence from* $p(k, \boldsymbol{\omega}_k)$ *to* $q_{\boldsymbol{\eta}}$ *without using the robust estimators. Moreover, the estimated parameters of each component are presented in the corresponding panel. To generate these histograms the randomized allocation procedure was run 10 times.*

2.5.2 BHHJ α-divergence measures

Basu et al. (1998) have proposed a robust divergence measure indexed by a parameter $\alpha \geq 0$, called BHHJ α-divergence throughout the thesis, as a robust alternative to the KL divergence. For example, Fujisawa and Eguchi (2006) and Miyamura and Kano (2006) have derived robust estimators using the BHHJ α-divergence to estimate the parameters of Gaussian mixture models. This divergence has also been used in Mihoko and Eguchi (2002) to separate sources in a robust fashion.

To allow for an easier comparison with the family of ϕ-divergence measures defined in Definition 2.1, we describe the BHHJ α-divergence as follows. Let $(\mathbb{X}, \mathcal{B}, \rho)$ be a measure space and let P and Q be two probability measures on $(\mathbb{X}, \mathcal{B})$ such that $P \ll Q \ll \rho$

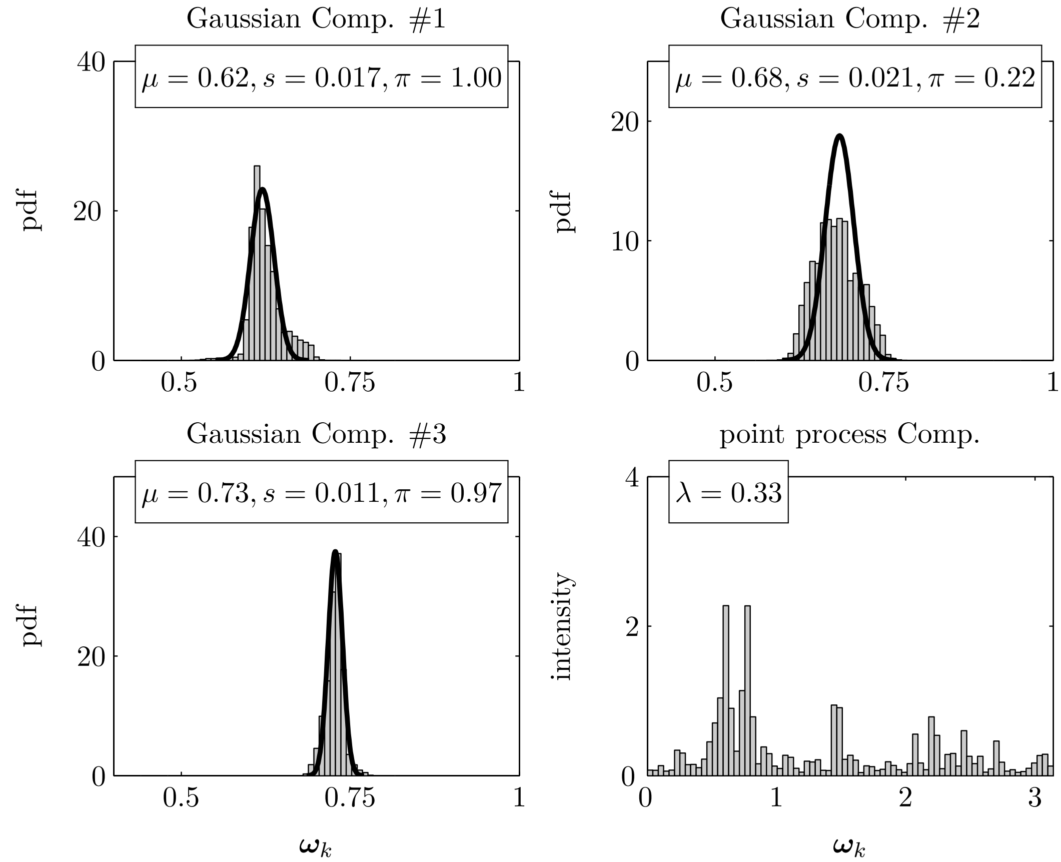


Figure 2.10 – *Histogram of the labeled samples* ($\boldsymbol{x}^{(i)}$, $\mathbf{z}^{(i)}$)*, with* $i = 1, \ldots, M$*, that is, the samples allocated to the Gaussian and Poisson point process components, versus the pdf's of estimated Gaussian components in the model (black solid line) for the summarizing algorithm derived from minimizing the KL divergence from* $p(k, \boldsymbol{\omega}_k)$ *to* $q_{\boldsymbol{\eta}}$ *with using the robust estimators. Moreover, the estimated parameters of each component are presented in the corresponding panel. To generate these histograms the randomized allocation procedure was run 10 times.*

(for instance both P and Q admit a strictly positive pdf with respect to ρ). As a natural generalization of the BHHJ α-divergence (Basu et al., 1998) in this setting, we define the *α-divergence from P to Q* as

$$\mathcal{D}_\alpha(P \,\|\, Q) \;=\; \int_{\mathbb{X}} \phi_\alpha\left(\frac{\mathrm{d}P}{\mathrm{d}Q}\right)\left(\frac{\mathrm{d}Q}{\mathrm{d}\rho}\right)^{1+\alpha} \mathrm{d}\rho \tag{2.23}$$

where

$$\phi_\alpha(u) = 1 - (1 + \alpha^{-1})u + \alpha^{-1}u^{1+\alpha}. \tag{2.24}$$

Observe that this is not a ϕ-divergence because of the exponent $(1 + \alpha)$ on $\frac{\mathrm{d}Q}{\mathrm{d}\rho}$. Observe also that the definition relies on the choice of a reference measure ρ on $(\mathbb{X}, \mathcal{B})$, which is not the case for the family of ϕ-divergences.

Proposition 2.3. *Let P and Q be probability measures on $(\mathbb{X}, \mathcal{B})$ such that $P \ll Q \ll \rho$. Set $p = \frac{\mathrm{d}P}{\mathrm{d}\rho}$ and $q = \frac{\mathrm{d}Q}{\mathrm{d}\rho}$. Then, for $\alpha \geq 0$,*

i) $\mathcal{D}_\alpha(P \,\|\, Q) \geq 0$.

ii) $\mathcal{D}_\alpha(P \,\|\, Q) = 0$ *iff (if and only if)* $Q = P$ *iff* $p = q$ *ρ-almost everywhere.*

iii) $\phi_\alpha(u) \to u \log u - u + 1$ *when* $\alpha \to 0$. *Or, equivalently,*

$$\lim_{\alpha \to 0} \mathcal{D}_\alpha(P \,\|\, Q) \;=\; \mathcal{D}_{KL}(P \,\|\, Q).$$

iv) $\mathcal{D}_{\alpha=1}(P \,\|\, Q) \;=\; \mathcal{D}_{MSE}(P \,\|\, Q)$, *where*

$$\mathcal{D}_{MSE}(P \,\|\, Q) = \int_{\mathbb{X}} (p - q)^2 \mathrm{d}\rho \tag{2.25}$$

is the mean squared error (MSE) divergence measure.

v) Plugging (2.24) *into* (2.23), *the expression of the BHHJ α-divergence from p to q becomes*

$$\mathcal{D}_\alpha(P \,\|\, Q) \;=\; \int_{\mathbb{X}} \left(q^{1+\alpha} - \left(1 + \frac{1}{\alpha}\right) p q^\alpha + \frac{1}{\alpha} p^{1+\alpha} \right) \mathrm{d}\rho. \tag{2.26}$$

In other words, $\mathcal{D}_\alpha$ is a divergence, in the sense that it is positive and vanishes iff $P = Q$. We recover the KL divergence in the limit $\alpha \to 0$, and the method is the maximum likelihood which is efficient but not robust; while when $\alpha = 1$, the method is the MSE estimator which is *robust but inefficient.* Therefore, the parameter α can be considered as a tuning parameter that controls the compromise between efficiency ($\alpha \to 0$) and robustness ($\alpha \to 1$). Thus, the obtained estimator is sensitive to the chosen value of α. However, it is indicated in Basu et al. (1998), that for $\alpha > 1$, the estimator suffers from a great loss of efficiency; thus, the region of interest is $0 < \alpha \leq 1$.

An interesting feature of this divergence measure is that, as in the KL divergence, to approximate the divergence there is no need to carry out any smoothing of the data. More precisely, given that $\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(M)}$ are samples from P, (2.26) becomes

$$\mathcal{D}_\alpha(P \,\|\, Q) \;\simeq\; \int_{\mathbb{X}} q^{1+\alpha} \mathrm{d}\rho - \left(1 + \frac{1}{\alpha}\right) \frac{1}{M} \sum_{i=1}^{M} q^\alpha(\boldsymbol{x}^{(i)}) + C. \tag{2.27}$$

2.5.3 Algorithm III: Using the BHHJ α-divergence in the SEM-type algorithm

According to Proposition 2.2, the M-step of Algorithm 2.1 can be seen as minimizing (approximately) the KL divergence from $\pi^{\mathcal{L}}_{\hat{\boldsymbol{\eta}}^{(r)}}$ to $Q^{\mathcal{L}}_{\boldsymbol{\eta}}$. We propose to replace the KL divergence by the BHHJ α-divergence between the labeled distributions. (Note that the equivalence result stated in Proposition 2.2 for ϕ-divergences does *not* hold for the BHHJ α-divergence). Setting $P = \pi^{\mathcal{L}}_{\boldsymbol{\eta}}$ and $Q = Q^{\mathcal{L}}_{\boldsymbol{\eta}}$ in (2.26), and assuming that both $\pi^{\mathcal{L}}_{\boldsymbol{\eta}}$ and $Q^{\mathcal{L}}_{\boldsymbol{\eta}}$ have pdf's with respect to $\rho_{\mathcal{L}}$, denoted respectively by $f_{\boldsymbol{\eta}}$ and $q_{\boldsymbol{\eta}}$, we have the following:

i) The BHHJ $\boldsymbol{\alpha}$-divergence from $\pi^{\mathcal{L}}_{\boldsymbol{\eta}}$ to $Q^{\mathcal{L}}_{\boldsymbol{\eta}}$ reads

$$\mathcal{D}_\alpha(\pi^{\mathcal{L}}_{\boldsymbol{\eta}} \,\|\, Q^{\mathcal{L}}_{\boldsymbol{\eta}}) \;=\; \int_{\mathbb{X}_{\mathcal{L}}} q^{1+\alpha}_{\boldsymbol{\eta}}\, \mathrm{d}\rho_{\mathcal{L}} \;-\; (1+\frac{1}{\alpha}) \int_{\mathbb{X}_{\mathcal{L}}} q^{\alpha}_{\boldsymbol{\eta}}\, \mathrm{d}\pi^{\mathcal{L}}_{\boldsymbol{\eta}} \;+\; \frac{1}{\alpha} \int_{\mathbb{X}_{\mathcal{L}}} f^{1+\alpha}_{\boldsymbol{\eta}} \mathrm{d}\rho_{\mathcal{L}}\,. \tag{2.28}$$

ii) Moreover, according to (2.27), given samples $\boldsymbol{x}^{(1)}, \dots, \boldsymbol{x}^{(M)}$ from π, possibly using some Monte Carlo method, and the corresponding allocation vectors, $\mathbf{z}^{(1)}, \dots, \mathbf{z}^{(M)}$, distributed according to $q_{\hat{\boldsymbol{\eta}}}(\,\cdot \mid \boldsymbol{x}^{(i)})$, the second integral can be approximated, and we derive the following criterion

$$\hat{\mathcal{J}}^{\alpha}_{M}(\boldsymbol{\eta}) \;=\; \int_{\mathbb{X}_{\mathcal{L}}} q^{1+\alpha}_{\boldsymbol{\eta}}\, \mathrm{d}\rho_{\mathcal{L}} \;-\; (1+\frac{1}{\alpha}) \cdot \frac{1}{M} \sum_{i=1}^{M} q^{\alpha}_{\boldsymbol{\eta}}(\boldsymbol{x}^{(i)},\, \mathbf{z}^{(i)}). \tag{2.29}$$

Note that in the process of the SEM-type algorithms, the third integral in (2.28) depends on $\hat{\boldsymbol{\eta}}$ and thus, becomes irrelevant when estimating $\boldsymbol{\eta}$. Now, we can introduce the third SEM-type algorithm to fit the parametric model $q_{\boldsymbol{\eta}}$ to the posterior f of interest by minimizing the BHHJ α-divergence as follows:

Algorithm 2.3. *At the $(r+1)^{th}$ iteration of the SEM-type algorithm based on the BHHJ α-divergence,*

S-step: *For $i = 1, \dots, M$,*

- *draw allocation vectors $\mathbf{z}^{(i)} \sim q_{\hat{\boldsymbol{\eta}}^{(r)}}(\,\cdot \mid \boldsymbol{x}^{(i)})$ defined in equations (2.8)–(2.16).*

E-step: *Construct the criterion*

$$\hat{\mathcal{J}}^{\alpha}_{M}(\boldsymbol{\eta}) \;=\; \int_{\mathbb{X}_{\mathcal{L}}} q^{1+\alpha}_{\boldsymbol{\eta}}\, \mathrm{d}\rho_{\mathcal{L}} \;-\; (1+\frac{1}{\alpha}) \cdot \frac{1}{M} \sum_{i=1}^{M} q^{\alpha}_{\boldsymbol{\eta}}(\boldsymbol{x}^{(i)},\, \mathbf{z}^{(i)}).$$

M-step: *Estimate $\hat{\boldsymbol{\eta}}^{(r+1)} = \operatorname{argmin}_{\boldsymbol{\eta} \in \mathrm{N}} \hat{\mathcal{J}}^{\alpha}_{M}(\boldsymbol{\eta})$.*

As mentioned before, the S-step is the same as the one used in Algorithm 2.1. However, the M-step, described in the next section, is more involved. For the M-step, we propose to use the BFGS optimization method which is the most popular quasi-Newton method (for more information, see Nocedal and Wright, 1999, Chapter 6). In this method, the Hessian matrix of the second-order partial derivatives of objective function is approximated using the first-order partial derivatives.

Remark 2.10. Asymptotic and robustness properties of estimators based on the BHHJ α-divergence can be found in Basu et al. (1998) ; Jones et al. (2001) ; Fujisawa and Eguchi (2006).

In the following, we derive the expressions to evaluate the objective function (2.29) and its first-order partial derivatives.

Computation of the criterion and its gradients

Recall the parametric model densities expressions derived in (2.8)–(2.11). To carry out the optimization of the objective function (2.29), first, the integral term should be dealt with. It can be written as

$$\int_{\mathbb{X}_\mathcal{L}} q_{\boldsymbol{\eta}}^{1+\alpha}\, \mathrm{d}\rho_\mathcal{L} \;=\; \sum_{k\geq 0} \sum_{\mathbf{z}\in\mathcal{Z}} q_{\boldsymbol{\eta}}^{1+\alpha}(\mathbf{z}) \prod_{j=1}^{k} \int_{\boldsymbol{\Theta}} q_{\boldsymbol{\eta}}^{1+\alpha}(k, \boldsymbol{\theta}_{j,k}|z_j)\, \mathrm{d}\boldsymbol{\theta}_{j,k}\,, \tag{2.30}$$

where $q_{\boldsymbol{\eta}}(\mathbf{z})$ is the density of allocation vectors and $q_{\boldsymbol{\eta}}(k, \boldsymbol{\theta}_{j,k}|z_j)$ is the conditional likelihood of the element j of the vector of the observed samples $\boldsymbol{x} = (k,\, \boldsymbol{\theta}_k)$ defined, respectively, in (2.8) and (2.9). The integral with respect to $\boldsymbol{\theta}_{j,k}$ on the right hand side of (2.30) have a closed-form expression

$$\int_{\boldsymbol{\Theta}} q_{\boldsymbol{\eta}}^{1+\alpha}(k, \boldsymbol{\theta}_{j,k}|z_j)\, \mathrm{d}\boldsymbol{\theta}_{j,k} \;\triangleq\; \tilde{q}_{\boldsymbol{\eta}}^{\alpha}(z_j) \;=\; \begin{cases} (1+\alpha)^{-1/2} \left|2\pi\boldsymbol{\Sigma}_{z_j}\right|^{-\alpha/2} & \text{if } z_j \leq L\,, \\ |\boldsymbol{\Theta}|^{-\alpha} & \text{otherwise.} \end{cases}$$

Hence, we can rewrite (2.30) as

$$\int_{\mathbb{X}_\mathcal{L}} q_{\boldsymbol{\eta}}^{1+\alpha}\, \mathrm{d}\rho_\mathcal{L} \;=\; \sum_{k\geq 0} \sum_{\mathbf{z}\in\mathcal{Z}} q_{\boldsymbol{\eta}}^{1+\alpha}(\mathbf{z}) \prod_{j=1}^{k} \tilde{q}_{\boldsymbol{\eta}}^{\alpha}(z_j)\,. \tag{2.31}$$

Note that the summation in (2.31) involves an infinite number of terms. We propose two approaches, one “exact” computation and another one based on Monte Carlo approximation, to evaluate (2.31) and compute its partial derivatives. The former one is appealing for moderate values of the number L of Gaussian components, while for the problems with large L the latter one is recommended. These two approaches are explained in Appendix A.1.

With some computations, we obtain the following first-order partial derivatives with respect to the model parameters $\boldsymbol{\eta} = \{\boldsymbol{\eta}_l\}_{1\le l\le L}$, with $\boldsymbol{\eta}_l = \{\pi_l, \boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l\}$, along with the mean λ of the Poisson point process component:

$$\begin{aligned}
\frac{\partial\,\hat{\eth}^{\alpha}_{M}(\boldsymbol{\eta})}{\partial\,\boldsymbol{\mu}_l} &= -\frac{\alpha+1}{M}\sum_{\substack{1\le i\le M\\ n_l^{(i)}=1}} (\boldsymbol{x}^{(i)}_{\to l} - \boldsymbol{\mu}_l)\, q^{\alpha}_{\boldsymbol{\eta}}(\boldsymbol{x}^{(i)},\, \mathbf{z}^{(i)}),\\
\frac{\partial\,\hat{\eth}^{\alpha}_{M}(\boldsymbol{\eta})}{\partial\,\boldsymbol{\Sigma}_l} &= \frac{\partial \int q^{1+\alpha}_{\boldsymbol{\eta}}\,\mathrm{d}\rho}{\partial\,\boldsymbol{\Sigma}_l}\\
&\quad - \frac{\alpha+1}{2M}\boldsymbol{\Sigma}_l^{-1}\sum_{\substack{1\le i\le M\\ n_l^{(i)}=1}} \Big((\boldsymbol{x}^{(i)}_{\to l} - \boldsymbol{\mu}_l)(\boldsymbol{x}^{(i)}_{\to l} - \boldsymbol{\mu}_l)^t\,\boldsymbol{\Sigma}_l^{-1} - 1\Big)\; q^{\alpha}_{\boldsymbol{\eta}}(\boldsymbol{x}^{(i)},\, \mathbf{z}^{(i)}),\\
\frac{\partial\,\hat{\eth}^{\alpha}_{M}(\boldsymbol{\eta})}{\partial\,\pi_l} &= \frac{\partial \int q^{1+\alpha}_{\boldsymbol{\eta}}\,\mathrm{d}\rho}{\partial\,\pi_l} - \frac{\alpha+1}{M\,\pi_l(1-\pi_l)}\sum_{i=1}^{M}(n_l^{(i)} - \pi_l)\, q^{\alpha}_{\boldsymbol{\eta}}(\boldsymbol{x}^{(i)},\, \mathbf{z}^{(i)}),\\
\frac{\partial\,\hat{\eth}^{\alpha}_{M}(\boldsymbol{\eta})}{\partial\lambda} &= \frac{\partial \int q^{1+\alpha}_{\boldsymbol{\eta}}\,\mathrm{d}\rho}{\partial\,\lambda} - \frac{\alpha+1}{M}\sum_{i=1}^{M}\left(\frac{n_{L+1}^{(i)}}{\lambda} - 1\right)\, q^{\alpha}_{\boldsymbol{\eta}}(\boldsymbol{x}^{(i)},\, \mathbf{z}^{(i)}),
\end{aligned}\tag{2.32}$$

where, as before, $\boldsymbol{x}^{(i)}_{\to l}$ is the element of the i^{th} vector of observed samples where $\mathbf{z}^{(i)} = l$ and $n_l^{(i)}$ shows the number of samples allocated to the component l. Recall that, for $1 \le l \le L$, i.e., the Gaussian components, $n_l^{(i)}$ is binary, while for the Poisson point process component, $n_{L+1}^{(i)} \in \mathbb{N}$. Moreover, note that from (2.31), it can be observed that the integral term in (2.29), i.e., $\int q^{1+\alpha}_{\boldsymbol{\eta}}\,\mathrm{d}\rho$, does not contribute in estimating the mean parameters, $\boldsymbol{\mu}_l$, $1 \le l \le L$. Its partial derivatives with respect to the other parameters is expressed in Appendix A.1.

Remark 2.11. At each iteration of the SEM-type algorithm, we use the robust estimates of the parameters as initial values for the BFGS algorithm (fminunc() in Matlab).

2.6 Summary

In this chapter, we have proposed a novel approach to summarize posterior distributions defined over union of subspaces of differing dimensionality that typically arise, in a Bayesian framework, when the number of components is unknown. We pointed out the limitations of the two well-known classical Bayesian approaches, i.e., the Bayesian model selection and Bayesian model averaging. Using the BMS approach leads to not only losing information from the discarded models but also ignoring the uncertainties concerning the presence of components. On the other hand, the BMA approach, which uses the information from all (plausible) models, is not appropriate to study the posterior of component-specific parameters, the number of which changes in each model.

An "ideal" summarization approach should be able to provide posterior summaries for component-specific parameters along with measures of uncertainties about presence of components using information from all (plausible) models. It should also be capable of dealing with label-switching problem in a variable-dimensional setting. Indeed, as discussed in Section 2.2.1, the lack of identifiability along with the uncertainty about the number of components results in a phenomenon that we called "birth, death, and switching of labels".

For this purpose, we proposed a novel approach which consists in fitting an original variable-dimensional parametric model to the true posterior distribution. The variable-dimensional parametric model $q_{\boldsymbol{\eta}}$, which serves as an approximate posterior, consists of a certain number L of Gaussian components, the presence of which controlled by binary indicator variables ξ_l Bernoulli distributed with probabilities π_l, with $1 \leq l \leq L$. Furthermore, due to robustness issues, a Poisson point process component of intensity λ was added to the model to account for the observed outliers and allow for a number L of Gaussian components smaller than the maximum observed $k^{(i)}$.

Turning to the estimation of the model parameters $\boldsymbol{\eta} \in \mathrm{N}$, we proposed three SEM-type algorithms to fit the approximate model $q_{\boldsymbol{\eta}}$ to the true posterior f by minimizing divergence measures from f to $q_{\boldsymbol{\eta}}$, using samples from the posterior f generated by a trans-dimensional Monte Carlo sampler, e.g., RJ-MCMC. We used the KL divergence and the BHHJ α-divergence for this purpose. We discussed that there is a serious robustness issue in the problem we are dealing with. In order to cope with the lack of robustness of maximum likelihood-type estimates resulting from minimizing the KL divergence, in addition to introducing the Poisson point process component to capture the outliers, modifications of the first SEM-type algorithm have been proposed. More specifically, first, as an intuitive solution, the empirical mean and (co)variance estimates in the M-step are substituted with the robust estimators (see Section 2.5.1). Second, in Section 2.5, we used a robust divergence measure, i.e., the BHHJ α-divergence proposed by Basu et al. (1998), instead of the KL divergence. Using the BHHJ α-divergence resulted in the third SEM-type algorithm with a difference in the M-step which was carried out using the BFGS optimization algorithm. In this case, we get a tangible criterion to be minimized, which might be used later, for example, to study the behavior of the proposed algorithm.

Following chapters investigate the performance of the proposed algorithm, both for summarizing and for relabeling variable-dimensional posterior distributions, on two problems:

i) detection and estimation of sinusoidal components in white Gaussian noise (Chapter 3).

ii) detection and estimation of astrophysical particles in the Auger project (Chapter 4).

Chapter 3

Bayesian Detection and Estimation of Sinusoids in White Gaussian Noise

3.1 Introduction

In this chapter, we present the problem of detection and estimation of sinusoids in white Gaussian noise and use it to illustrate the performance of the summarizing approach proposed in Chapter 2. Methods for detecting and estimating frequencies in a noisy signal have applications in various fields including communications, seismology, and radar—to name but a few.

A host of frequency estimation techniques have been proposed in the literature since Schuster's celebrated periodogram (Schuster, 1898), including for instance correlation-based methods, such as the Yule-Walker algorithm, and maximum likelihood methods (see, e.g., Stoica et al., 1989, and references therein for more information). It was shown much later (Jaynes, 1987 ; Bretthorst, 1988) that the periodogram is in fact a special case of a more general Bayesian estimator. However, the Bayes estimator of Bretthorst (1988) is based on crude approximations of the posterior distribution to avoid computing high-dimensional integrals, which do not hold in the case of small sample size or closely located radial frequencies (see, e.g., Dou and Hodgson, 1996). Dou and Hodgson (1995, 1996) derived an MCMC sampler to approximate the posterior distribution of the parameters.

Turning to the detection of sinusoidal components, Djurić (1996) provides a review of criterion-based methods along with proposing a new penalty term that can be seen as "corrected BIC" for this specific problem. Dou and Hodgson (1995, 1996) carried out the model selection part by comparing the Bayes factors using the MCMC samples generated from the posterior distribution of the parameters for each model separately. Later, Andrieu and Doucet (1999) proposed an original hierarchical model and RJ-MCMC sampler for the problem of joint Bayesian model selection and estimation of sinusoids in white Gaussian noise. Concerning this problem, in this thesis, we follow the model and RJ-MCMC sampler proposed in Andrieu and Doucet (1999) unless otherwise stated.

This chapter is organized as follows. In Section 3.2, we describe the problem and the ingredients of the Bayesian method, i.e., the hierarchical model and the RJ-MCMC sampler, developed by Andrieu and Doucet (1999). Then, we explain issues regarding both the computation of the birth-or-death ratio (Section 3.2.4), following the discussions in Section 1.4, and specification of the model hyperparameters (Section 3.2.5). In the rest of the chapter, we investigate the capability of the approach we proposed in Chapter 2 for summarizing variable-dimensional posterior distributions encountered in this problem. For this purpose, Section 3.3 provides illustrative results and discusses the performance of the proposed algorithms in detail whereas Section 3.4 studies the performance of the proposed approach "on average". Finally, Section 3.5 provides a summary of the chapter and discusses the obtained results.

3.2 Bayesian framework

In this section, we first present the problem of joint detection and estimation of sinusoidal components in white Gaussian noise in Section 3.2.1. Then, we briefly describe the hierarchical model (Section 3.2.2) and RJ-MCMC sampler (Section 3.2.3) proposed by Andrieu and Doucet (1999) for this problem. Then, following our result in Section 1.4 concerning the mistake committed in the computation of the birth-or-death ratio by Andrieu and Doucet (1999) and their followers, in Section 3.2.4, we provide an experiment and study its influence on the posterior of the number k of components. Finally, in Section 3.2.5, we briefly address the sensitivity of the posterior distributions to the model's hyperparameters.

3.2.1 Problem statement

Let $\mathbf{y} = (y_1, y_2, \ldots, y_N)^t$ be a vector of N samples of an observed signal. We consider a finite family of embedded models $\{\mathcal{M}_k,\ k \in \mathcal{K}\}$, with $\mathcal{K} = \{0, \ldots, k_{\max}\}$, where $\mathcal{M}_k$ assumes that $\mathbf{y}$ can be written as a linear combination of k sinusoids observed in white

Gaussian noise, as follows:

$$\begin{aligned} \mathcal{M}_0 : y(i) &= n(i), \\ \mathcal{M}_k : y(i) &= \sum_{j=1}^{k} \Big(a_{c_{j,k}} \cos(\omega_{j,k} i) + a_{s_{j,k}} \sin(\omega_{j,k} i) \Big) + n(i). \end{aligned}$$

Let $\boldsymbol{\omega}_k = (\omega_{1,k}, \dots, \omega_{k,k})$ and $\mathbf{a}_k = \Big(a_{c_{1,k}}, a_{s_{1,k}}, \dots, a_{c_{k,k}}, a_{s_{k,k}} \Big)$ be the vectors of radial frequencies and amplitudes under model $\mathcal{M}_k$, respectively; moreover, let $\mathbf{D}_k$ be the corresponding $N \times 2k$ design matrix defined by

$$\mathbf{D}_k(i+1, 2j-1) \triangleq \cos(\omega_{j,k} i), \quad \mathbf{D}_k(i+1, 2j) \triangleq \sin(\omega_{j,k} i) \tag{3.1}$$

for $i = 0, \dots, N-1$ and $j = 1, \dots, k$. Then, the observed signal $\mathbf{y}$ follows under $\mathcal{M}_k$ a normal linear regression model:

$$\mathbf{y} = \mathbf{y}_0 + \mathbf{n} = \mathbf{D}_k . \mathbf{a}_k + \mathbf{n} ,$$

where $\mathbf{y}_0$ is the noiseless signal and $\mathbf{n}$ is a white Gaussian noise of variance σ^2. The unknown parameters are, then, assumed to be the number k of components, the component-specific parameters $\boldsymbol{\theta}_k = (\mathbf{a}_k, \boldsymbol{\omega}_k)$ and the noise variance σ^2 which is common to all models. Hence, the space of unknown component-specific parameters is $\boldsymbol{\Theta}_k = \boldsymbol{\Theta}^k$, under $\mathcal{M}_k$, with $\boldsymbol{\Theta} = \mathbb{R}^2 \times (0, \pi)$ and the convention that $\boldsymbol{\Theta}_0 = \{\varnothing\}$, and the overall parameter space is $\mathbb{X} = \Big(\bigcup_{k=0}^{k_{\max}} \{k\} \times \boldsymbol{\Theta}_k \Big) \bigcup \mathbb{R}_+$.

3.2.2 Hierarchical model and prior distributions

Assuming that no (or little) information is available about the vector of amplitudes $\mathbf{a}_k$ and the noise variance σ^2, it is usually recommended to use Zellner's conditionally conjugate g-prior as a default prior in the Bayesian variable selection literature (Zellner, 1986 ; George and Foster, 2000 ; Fernández et al., 2001 ; Cui and George, 2008 ; Liang et al., 2008). Under this prior, the distribution of $\mathbf{a}_k$, conditionally to σ^2, k and $\boldsymbol{\omega}_k$, is a multivariate Gaussian distribution with $\sigma^2 \delta^2 \left(\mathbf{D}_k^t \mathbf{D}_k \right)^{-1}$ as its covariance matrix, where δ^2 is a possible hyperparameter. Moreover, the noise variance σ^2 is endowed with Jeffreys' improper prior, i.e. $p(\sigma^2) \propto 1/\sigma^2$. Note that δ^2 is the inverse of the conventional g parameter in the g-prior, i.e., $\delta^2 = 1/g$. Following Andrieu and Doucet (1999), a zero-mean g-prior for $\mathbf{a}_k$ and σ^2 will be used in this thesis. Conditional on k, the radial frequencies $\boldsymbol{\omega}_k$ are assumed independent and identically distributed with a uniform distribution on the interval $(0, \pi)$. The number of components k is given a Poisson distribution with mean Λ, truncated to $\{0, 1, \dots, k_{\max}\}$, where Λ and $k_{\max}$ are two additional hyperparameters.

Furthermore, the hyperparameters δ^2 and Λ are also treated as random variables. In fact, the parameter δ^2, called the Expected SNR (ESNR) by Andrieu and Doucet (1999), controls the expected size of the amplitudes. Owing to its influence on the performance of the algorithm, and assuming again that no (or little) information is available, the hyperparameter δ^2 is given in Andrieu and Doucet (1999) a conjugate inverse gamma prior with parameters α_{δ^2} and β_{δ^2}, that we denote hereafter by $\mathcal{IG}\left(\alpha_{\delta^2}, \beta_{\delta^2}\right)$. The hyperparameter Λ is endowed with a conjugate Gamma distribution with parameters α_Λ and β_Λ denoted by $\mathcal{G}\left(\alpha_{\delta^2}, \beta_{\delta^2}\right)$. Such a hierarchical Bayes approach is usually hoped to increase the robustness of the statistical analysis; see Robert (2007, Section 10.2) for more information. We will discuss more about prior specification arguments and the sensitivity of the posterior distribution to their parameters in Section 3.2.5.

Figure 3.1 shows the DAG of the complete hierarchical model designed for this problem using graphical model conventions: filled (solid) circles denote deterministic parameters, that are either observed or set to a fixed value, while unfilled circles are used for random variables. The full joint prior distribution of the unknown parameters has the following hierarchical structure:

$$p\left(k, \mathbf{a}_k, \boldsymbol{\omega}_k, \sigma^2, \delta^2, \Lambda\right) \;=\; p(\mathbf{a}_k \mid k, \boldsymbol{\omega}_k, \sigma^2, \delta^2)\; p(\boldsymbol{\omega}_k \mid k)\; p(k \mid \Lambda)\; p(\sigma^2)\; p(\delta^2)\; p(\Lambda). \tag{3.2}$$

In fact, due to using conditionally conjugate prior distributions, it is possible to analytically integrate $\mathbf{a}_k$ and σ^2 out. Therefore, doing so, the target distribution for the RJ-MCMC sampler becomes

$$p\left(k, \boldsymbol{\omega}_k, \delta^2, \Lambda \,|\, \mathbf{y}\right) \;\propto\; (\mathbf{y}^t \mathbf{P}_k \mathbf{y})^{-N/2} \frac{\Lambda^k \pi^{-k}}{k!\,(\delta^2+1)^k} \mathbb{1}_{(0,\pi)^k}(\boldsymbol{\omega}_k)\; p(\delta^2)\; p(\Lambda), \tag{3.3}$$

with

$$\mathbf{P}_k \;=\; \mathbf{I}_N \;-\; \frac{\delta^2}{1+\delta^2} \mathbf{D}_k \left(\mathbf{D}_k^t \mathbf{D}_k\right)^{-1} \mathbf{D}_k^t$$

when $k \geq 1$ and $\mathbf{P}_0 = \mathbf{I}_N$.

3.2.3 RJ-MCMC sampler

In the following, the RJ-MCMC sampler proposed by Andrieu and Doucet (1999) to generate samples from the target distribution (3.3) is briefly described. For more detailed expressions refer to Andrieu and Doucet (1999).

The MH-within-Gibbs sampler, that leaves the target density (3.3) invariant, consists of a MH move for updating the value of k and $\boldsymbol{\omega}_k$, followed by a sequence of Gibbs moves to update the hyperparameters δ^2 and Λ. The proposal kernel of the MH move designed

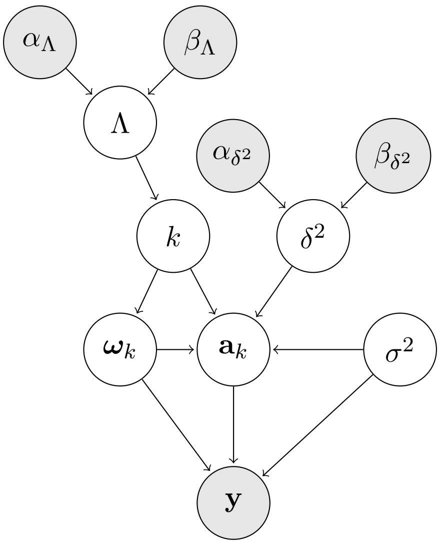


Figure 3.1 – *DAG showing the hierarchical model used for the problem of sinusoid detection in white Gaussian noise. Filled and unfilled circles indicate deterministic and random variables, respectively.*

for updating k and $\boldsymbol{\omega}_k$ is in fact a mixture of proposal kernels performing within-model moves (updating radial frequencies without changing k) and between-models moves (birth and death moves, which respectively add and remove components). More explicitly, the proposal kernel is

$$Q(\boldsymbol{x},\,\cdot\,) \;=\; p_{\mathrm{b}}(\boldsymbol{x})\,Q_{\mathrm{b}}(\boldsymbol{x},\,\cdot\,) \,+\, p_{\mathrm{d}}(\boldsymbol{x})\,Q_{\mathrm{d}}(\boldsymbol{x},\,\cdot\,) \,+\, p_{\mathrm{u}}(\boldsymbol{x})\,Q_{\mathrm{u}}(\boldsymbol{x},\,\cdot\,) \tag{3.4}$$

where the probabilities for choosing birth, death, and update moves are

$$\begin{aligned}
p_{\mathrm{b}}(k,\boldsymbol{\omega}_k) \;&=\; \begin{cases} c\,\cdot\,\min\left\{1,\,\frac{p(k+1)}{p(k)}\right\} & \text{if} \quad k \,<\, k_{\max}, \\ 0 & \text{otherwise}, \end{cases} \\
p_{\mathrm{d}}(k+1,\boldsymbol{\omega}_{k+1}) \;&=\; \begin{cases} c\,\cdot\,\min\left\{1,\,\frac{p(k)}{p(k+1)}\right\} & \text{if} \quad k \,>\, 0, \\ 0 & \text{otherwise}, \end{cases} \\
p_{\mathrm{u}}(\boldsymbol{x}) \;&=\; 1 - p_{\mathrm{b}}(\boldsymbol{x}) \,-\, p_{\mathrm{d}}(\boldsymbol{x}).
\end{aligned} \tag{3.5}$$

where c is set to 0.5. Note that $k_{\max} = N/2$ here to avoid occurrence of linearly dependent columns in $\mathbf{D}_k$. The birth and death kernels, i.e., $Q_{\mathrm{b}}(\boldsymbol{x},\,\cdot\,)$ and $Q_{\mathrm{d}}(\boldsymbol{x},\,\cdot\,)$, respectively, are as defined in expressions (1.26) and (1.27). The proposal distribution used to generate a new radial frequency ω^* in the birth kernel (1.26), denoted by $q(\omega)$, is a uniform distribution on the interval $(0,\,\pi)$. Then, following Proposition 1.11 and setting $\boldsymbol{x} = (k,\,\boldsymbol{\omega}_k)$ and

$\boldsymbol{x}' = (k+1,\, \boldsymbol{\omega}_k \oplus_i \omega^*)$, the birth ratio becomes

$$r(\boldsymbol{x}, \boldsymbol{x}') \;=\; \frac{p\left(k+1, \boldsymbol{\omega}_k \oplus_i \omega^*, \delta^2, \Lambda \,|\, \mathbf{y}\right)}{p\left(k, \boldsymbol{\omega}_k, \delta^2, \Lambda \,|\, \mathbf{y}\right)} \cdot \frac{p_{\mathrm{d}}(\boldsymbol{x}')}{p_{\mathrm{b}}(\boldsymbol{x})} \cdot \frac{1}{q\left(\omega^*\right)} \;=\; \left(\frac{\mathbf{y}^t \mathbf{P}_{k+1} \mathbf{y}}{\mathbf{y}^t \mathbf{P}_k \mathbf{y}}\right)^{-N/2} \frac{1}{1+\delta^2} \,. \tag{3.6}$$

In the within-model move for updating the radial frequencies assuming k is fixed, Andrieu and Doucet (1999) proposed to update each component's radial frequency using a mixture of MH moves, that is, a Fourier Transform (FT) based global move and a local normal random walk move. Then, the update move acceptance ratio follows from the simple MHG ratio (1.7).

Turning to the Gibbs sampler for updating the hyperparameter δ^2, one should note that direct sampling from the conditional posterior distribution

$$p(\delta^2 \,|\, \mathbf{y}, k, \boldsymbol{\omega}_k) \;\propto\; \frac{(\mathbf{y}^t \mathbf{P}_k \mathbf{y})^{-N/2}}{(\delta^2+1)^k}\, p(\delta^2)$$

is not feasible. On the other hand, the conditional posterior distribution of δ^2 given $\mathbf{y}$, k, $\mathbf{a}_k$, $\boldsymbol{\omega}_k$, σ^2 can be written as

$$p(\delta^2 \;|\; \mathbf{y},\, k,\, \mathbf{a}_k,\, \boldsymbol{\omega}_k,\, \sigma^2) \;\propto\; \mathcal{IG}\left(k + \alpha_{\delta^2},\, \frac{\mathbf{a}_k^t \mathbf{D}_k^t \mathbf{D}_k \mathbf{a}_k}{2\sigma^2} + \beta_{\delta^2}\right),$$

from which direct samples can be generated. Therefore, to be able to carry out the Gibbs move, Andrieu and Doucet (1999) proposed to demarginalize σ^2 and $\mathbf{a}_k$, in the spirit of data augmentation arguement. Finally, Λ is updated by a Gibbs move. Algorithm 3.1 presents the RJ-MCMC sampler used for generating samples from (3.3). Samplers having this structure are also known as *partially collapsed Gibbs samplers* (see Van Dyk and Park, 2008, for more discussion).

3.2.4 The effect of using the wrong birth-or-death ratio on the results

One should note that the birth ratio computed in Andrieu and Doucet (1999) differs from the one expressed in (3.6) by a $1/(k+1)$ factor. A similar mistake in computing RJ-MCMC ratios has been reported in the field of genetics (Jannink and Fernando, 2004 ; Sillanpaa et al., 2004). In fact, using the expression of the birth ratio with an additional factor of $1/(k+1)$, as in Andrieu and Doucet (1999), amounts to assigning a different prior distribution over k called "accelerated Poisson distribution" (Sillanpaa et al., 2004) which reads

$$p_2(k) \;\propto\; \frac{e^{-\Lambda}\Lambda^k}{(k!)^2}\, \mathbb{1}_{\mathbb{N}}(k). \tag{3.7}$$

Algorithm 3.1. *RJ-MCMC sampler for the problem of joint detection and estimation of sinusoids in white Gaussian noise.*

Trans-dimensional move: *the birth, death, and update moves*

- *Generate a random number* $u \sim \mathcal{U}(0, 1)$.
- *If* $p_{\mathrm{b}}(\boldsymbol{x}) \geq u$ *do a birth move;*
    - *Generate the insertion location* i *on* $\{1, \ldots, k+1\}$.
    - *Propose a new radial frequency* $\omega^* \sim \mathcal{U}(0, \pi)$.
    - *Accept the proposed move with the probability* $\alpha(\boldsymbol{x}, \boldsymbol{x}') = min\{1, r(\boldsymbol{x}, \boldsymbol{x}')\}$, *where* $r(\boldsymbol{x}, \boldsymbol{x}')$ *is the acceptance ratio expressed in* (3.6).
- *Else if* $p_{\mathrm{b}}(\boldsymbol{x}) + p_{\mathrm{d}}(\boldsymbol{x}) \geq u$ *do a death move;*
    - *Generate the index of the component to be removed* i *on* $\{1, \ldots, k\}$.
    - *Accept the proposed move with the probability* $\alpha(\boldsymbol{x}, \boldsymbol{x}') = min\{1, 1/r(\boldsymbol{x}, \boldsymbol{x}')\}$, *where* $r(\boldsymbol{x}, \boldsymbol{x}')$ *is the acceptance ratio expressed in* (3.6).
- *Otherwise, update the radial frequencies without altering the* k *number of components using the within-model move.*

Demarginalization: *updating the noise variance and amplitudes*

- $\sigma^2 \mid \mathbf{y}, k, \boldsymbol{\omega}_k \sim \mathcal{IG}\left(N/2, \frac{\mathbf{y}^t \mathbf{P}_k \mathbf{y}}{2}\right)$
- $\mathbf{a}_k \mid \mathbf{y}, k, \boldsymbol{\omega}_k, \sigma^2 \sim \mathcal{N}\left(\mathbf{m}_k, \sigma^2 \mathbf{M}_k\right)$ *with* $\mathbf{M}_k^{-1} = \left(1 + \delta^{-2}\right) \mathbf{D}_k^t \mathbf{D}_k$ *and* $\mathbf{m}_k = \mathbf{M}_k \mathbf{D}^t \mathbf{y}$

Hyperparameters: *update* δ^2 *and* Λ *using*

- $\delta^2 \mid \mathbf{y}, k, \mathbf{a}_k, \boldsymbol{\omega}_k, \sigma^2 \sim \mathcal{IG}\left(k + \alpha_{\delta^2}, \frac{\mathbf{a}_k^t \mathbf{D}_k^t \mathbf{D}_k \mathbf{a}_k}{2\sigma^2} + \beta_{\delta^2}\right)$
- $\Lambda \mid \mathbf{y}, k \sim \mathcal{G}\left(\alpha_\Lambda + k, 1 + \beta_\Lambda\right)$

Figure 3.2 illustrates the difference between both the accelerated (black) and the usual (gray) Poisson distributions when mean $\Lambda = 5$. It can be observed that the accelerated Poisson distribution (3.7) puts a stronger emphasis on "sparse" models, i.e., models with a small number of components.

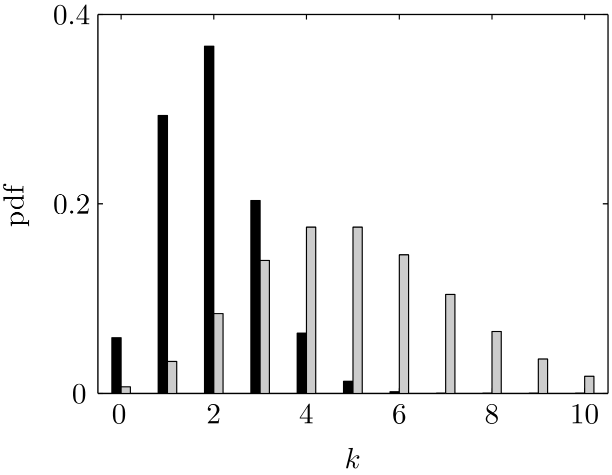


Figure 3.2 – *Probability distribution functions of the Poisson (gray) and the accelerated Poisson (black) distributions with mean* $\Lambda = 5$.

To highlight the influence of using an erroneous birth ratio on the posterior distribution of k, let us consider an observed signal $\mathbf{y}$ of length $N = 64$ from the first experiment defined in Table 3.1 consisting of $k = 3$ sinusoidal components with a moderate value SNR of 7dB. Samples from the posterior distribution of k are obtained using the RJ-MCMC sampler described in Algorithm 3.1, with an inverse Gamma prior $\mathcal{IG}(2, 100)$ on δ^2 and a Gamma prior $\mathcal{G}(1, 10^{-3})$ on Λ. For each observed signal in 100 replications of the experiment, the sampler was run twice: once with the correct expression of the ratio, given by (3.6), and once with the erroneous expression from Andrieu and Doucet (1999). Figure 3.3 shows the frequency of selection of each model using MAP under both the Poisson and the accelerated Poisson distribution as a prior for k. It appears that the (unintended) use of the accelerated Poisson distribution, induced by the erroneous expression of the MHG ratio, can result in a significant shift to the left of the posterior distribution of k.

Remark 3.1. Working with "sorted" vectors of frequencies would be quite natural in this problem, since the frequencies are exchangeable under the posterior (3.3). As explained in Section 1.4.2, the expression of the MHG ratio would be the same.

Remark 3.2. The reason why the ratio in Andrieu and Doucet (1999) is wrong can be understood from a subsequent paper (Andrieu et al., 2001b), where the same computation is explained in greater detail. There we can see that the authors, working with an "unsorted

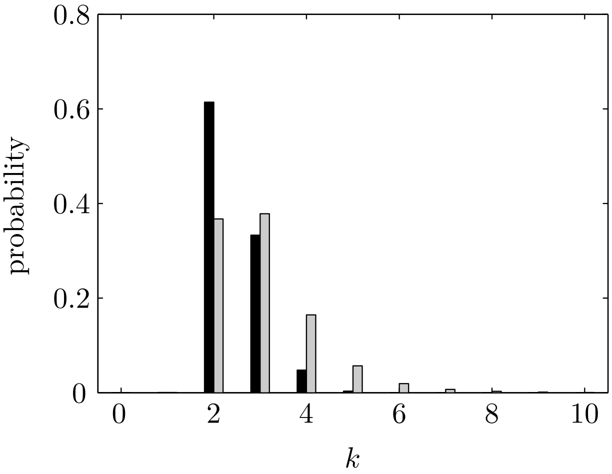


Figure 3.3 – *Frequency of selection for each model $\mathcal{M}_k$ using MAP for 100 replications of the experiment described in Section 3.2.4, using the expression of the ratio given in Andrieu and Doucet (1999, Equation (20)) (black) and the corrected ratio* (3.6) *(gray). There are $k = 3$ sinusoidal components in the observed signal $\mathbf{y}$ and the SNR $= 7$dB. 100k samples were generated using RJ-MCMC sampler and the first 20k were discarded as burn-in period.*

vector" representation, consider that the new component in a birth move is *inserted at the end*. The death move, however, is defined as in the present paper: a sinusoid to be removed is *selected randomly* among the existing components. Here is the mistake: if the new component is inserted at the end during a birth move, then any attempt at removing a component which is not the last one should be rejected during a death move. In other words, the acceptance probability should be zero when any component but the last one is picked to be removed during a death move.

3.2.5 Prior specification for signal-to-noise ratio hyperparameter and Bayesian sensitivity analysis

Every Bayesian method contains the delicate step of prior specification over the model's unknown parameters; refer to, e.g., Kass and Wasserman (1996) and Robert (2007, Chapter 2) for more discussion. Here, we discuss briefly the issues concerning the sensitivity of the posterior distributions to the values of the hyperparameter δ^2 and its scale β_{δ^2} in the hierarchical model defined in Section 3.2.2. Using the g-prior over the amplitudes, the task of prior specification boils down to the selection of the scalar parameter g. Nevertheless, it is well-known from the Bayesian variable selection literature that the g parameter—or, δ^2 in our notation—of the Zellner's g-prior, which controls the expected relative size of the amplitudes with respect to σ, plays an important role from the model selection viewpoint (Zellner, 1986 ; George and Foster, 2000 ; Fernández et al., 2001 ; Cui and George, 2008 ; Liang et al., 2008 ; Celeux et al., 2012). Indeed, fixing δ^2 is not recommended in the literature, as not only there is no default value for δ^2 (setting δ^2 to a large value, in an attempt of being non-informative, results in the Barlett or Lindley-Jeffreys paradoxes (see, e.g., Liang et al., 2008 ; Celeux et al., 2012)) but also it results in underestimating the uncertainties.

In a fully Bayesian solution, one assigns a prior distribution over δ^2. Often, assuming that no (or little) prior information is available, prior distributions are chosen to be as non-informative as possible to reduce their influence on the resulting posterior distributions. Usual Bayesian default non-informative prior distributions are the Jeffreys and "Reference" priors (see, e.g., Bernardo et al., 1992 ; Berger et al., 2009). Following Berger et al. (2001), both the Jeffreys and reference (improper) prior distributions for δ^2 are

$$p^{\mathrm{REF}}(\delta^2) \;\propto\; \frac{1}{1+\delta^2}.$$

Note that, however, the use of improper prior distributions over δ^2 is not allowed, because δ^2 is not included under $\mathcal{M}_0$, and, consequently, using improper priors results in indeterminate Bayes factors. Celeux et al. (2012) sidestepped this limitation by including the intercept parameter in the design matrix $\mathbf{D}$, at the price of loosing the location invariance. Other attempts at making the prior $p^{\mathrm{REF}}(\delta^2)$ proper can be found in Cui and George (2008) and Liang et al. (2008) by introducing a power factor b as follows

$$p^{\mathrm{REF}\star}(\delta^2) \;\propto\; \left(\frac{1}{1+\delta^2}\right)^{-b/2},$$

where they recommended to set $b = 3$ and $b = 4$. However, Berger et al. (2001) strongly advise against making improper priors proper by truncating or adding extra parameters, owing to the fact that the resulting posterior would be very sensitive to its parameters.

Therefore, neither can δ^2 be fixed to a default value, nor can an improper non-informative reference prior distribution can be assigned over it. Moreover, the proposed proper priors are not completely satisfactory. The other possibility, that we have decided to use in this chapter, is to use a weakly-informative conjugate $\mathcal{IG}(\alpha_{\delta^2}, \beta_{\delta^2})$ prior distribution as proposed by Andrieu and Doucet (1999). However, it is expected that the posterior distribution would be sensitive to β_{δ^2} but with lesser extent. Results of numerical experiments provided in Appendix B show that the posterior distribution is sensitive to the value of β_{δ^2} in moderate to low SNR situations. We have proposed to either estimate an appropriate value for β_{δ^2} from the observed data in the spirit of the EB approach using an IS based Monte Carlo EM (MCEM) algorithm (see, e.g., Quintana et al., 1999 ; Levine and Casella, 2001) or to integrate it out by assigning a conjugate Gamma prior over it. However, both approaches failed in low SNR situations, while in high SNR situations the sensitivity to β_{δ^2} is negligible (see Appendix B for numerical results).

Using the SMC sampler for Bayesian sensitivity analysis

Instead of fixing β_{δ^2} to an arbitrary value, one can opt for communicating the sensitivity of the posterior distribution to its variations; see, e.g., Berger (1990). For this purpose, a sequence of reasonable values of β_{δ^2}, say, $\{\beta^t_{\delta^2}\}_{t\in\mathbb{T}}$, $\mathbb{T} = \{1, 2, \ldots, T\}$, is considered. Then, we are interested in generating samples from the sequence of posterior distributions $\{\pi_t\}_{t\in\mathbb{T}}$, where $\pi_t = p\left(k, \boldsymbol{\omega}_k, \delta^2, \Lambda \,|\, \mathbf{y}, \beta^t_{\delta^2}\right)$. It is evident that, for large values of T, using the RJ-MCMC sampler to draw samples from every posterior distribution in the sequence would be computationally very expensive.

The SMC sampler described in Section 1.3.2 is well suited to generate samples efficiently from the sequence of posterior distributions $\{\pi_t\}_{t\in\mathbb{T}}$ in order to investigate the sensitivity of the posterior distribution to β_{δ^2} following the idea developed by Bornn et al. (2010). For this purpose, we use the RJ-MCMC sampler described in Algorithm 3.1 to generate samples from π_1 corresponding to $\beta^1_{\delta^2}$ which, after discarding the burn-in period, serve as particles for the SMC sampler. Moreover, choosing a large T, we can assume that $\pi_{t-1} \approx \pi_t$. Then, as in Bornn et al. (2010), to reduce the computational burden, we only resample and move the particles when ESS $= 1/\sum_{i=1}^{M}(W_t^{(i)})^2$ is lower than a certain threshold, say, $M/2$, where $W_t^{(i)}$, $i = 1, \ldots, M$, are the normalized weights. Otherwise, we simply copy the particles and update the corresponding weights.

To show the performance of the sensitivity analysis algorithm, we consider the first experiment of Table 3.1 with $k = 3$ sinusoidal components and SNR $= 5$ dB. Figure 3.4 illustrates the sensitivity of the posterior distribution of k to the variations in the scale parameter β_{δ^2}. It can be seen that, for example, if one is interested in selecting a model with the highest posterior probability, then the obtained result would be different by modifying the values of β_{δ^2}. Observe also the decreasing behavior of the posterior mean of k shown on Figure 3.4 (b). An interesting point to note is that, only four times the particles were resampled and moved, and in the rest they were simply copied. Similar graphics could be produced for any posterior quantity of interest.

As a concluding remark, we recommend, in practice, to use the described SMC sam-

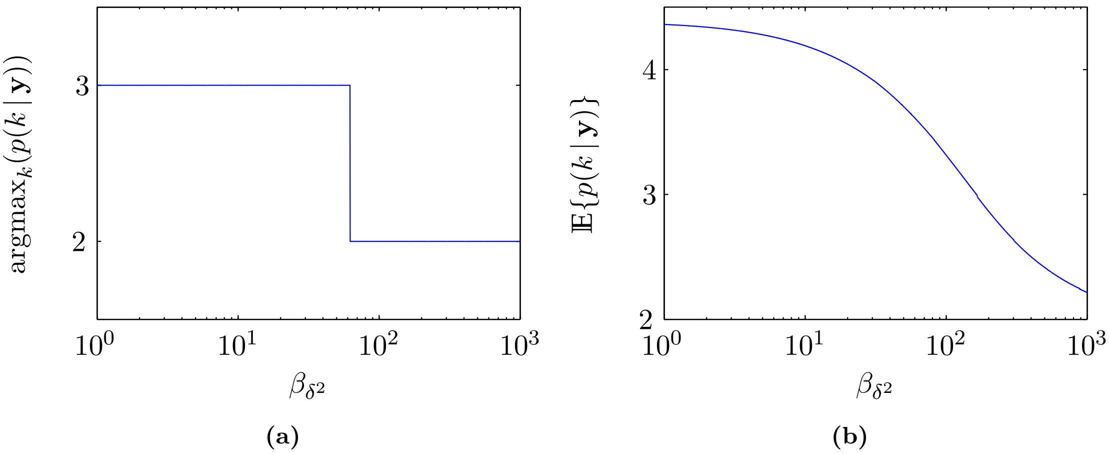


Figure 3.4 – *Sensitivity of the posterior distribution of k to the variations in the scale parameter β_{δ^2} for the first experiment of Table 3.1 with $SNR = 5$dB. We consider $T = 10\,000$ points in [1, 1 000] for β_{δ^2} and used $M = 40\,000$ particles.*

pler for analyzing the sensitivity of posterior distributions to the hyperparameters. Nevertheless, in the rest of the chapter, since our goal is the summarization of the posterior distributions, we set β_{δ^2} to a fixed value of 20.

3.3 Summarizing variable-dimensional posteriors: illustrative examples

3.3.1 Objectives

In this section, we investigate the capability of the algorithms proposed in Chapter 2 for summarizing variable-dimensional posterior distributions encountered in the problem of joint detection and estimation of sinusoids in white Gaussian noise. We emphasize again that the output of the trans-dimensional Monte Carlo sampler, i.e., the RJ-MCMC sampler described in Algorithm 3.1, is considered as the observed data for the proposed algorithms (see Figure 2.8). For the sake of simplicity, we concentrate here on summarizing the joint posterior distribution of the number k of components and the radial frequencies, i.e., $p(k,\, \boldsymbol{\omega}_k)$. As a result, the Gaussian components used in the parametric model, $q_{\boldsymbol{\eta}}$, shown in Figure 2.4 are considered to be univariate, and the space of component-specific parameters is $\boldsymbol{\Theta} = (0, \pi) \subset \mathbb{R}$. Therefore, in this section, each Gaussian component in $q_{\boldsymbol{\eta}}$ has a mean μ, a variance s^2, and a probability of presence π to be estimated.

We consider three summarizing algorithms in this chapter. Two of them are derived from minimizing the KL divergence from the true posterior $p(k,\, \boldsymbol{\omega}_k)$ to the approximate posterior $q_{\boldsymbol{\eta}}$, with a difference in the M-step: the first one uses the empirical mean and variance estimates, expressed in (2.22), and the second one uses the median and interquartile range as their robust alternatives (see Section 2.5.1). They will be denoted by TAP-KL1 and TAP-KL2, respectively, where TAP stands for "Trans-dimensional Approximate Model". The third summarizing algorithm, called TAP-BHHJ, is the one derived from minimizing the BHHJ α-divergence from $p(k, \boldsymbol{\omega}_k)$ to $q_{\boldsymbol{\eta}}$ described in Section 2.5. Moreover, we compare the obtained results with the ones obtained using the BMS and BMA approaches. In the BMS approach, the radial frequencies are estimated using the median of the posterior distributions of the sorted radial frequencies given the selected model.

In this section, we concentrate on both the summarization and the relabeling properties of the proposed algorithms. For this purpose, the performance of the proposed summarizing algorithms is illustrated on three specific examples. The objectives are:

i) to study the behavior of the proposed algorithms and the impact of the solutions proposed to deal with the robustness issue caused by the outliers (see Section 2.5.1),

ii) to assess the convergence properties of the proposed algorithms,

iii) to assess whether the proposed algorithms are able to solve the label-switching issue in the trans-dimensional problems,

iv) to assess how well the information contained in the true posterior distribution is captured by the approximate parametric model.

In all the experiments, the RJ-MCMC sampler explained in Algorithm 3.1 was used to generate 100 000 samples from the target distribution (3.3) and the first 20 000 iterations were discarded as the burn-in period. Next, to reduce both the correlation of the samples and the computational burden of the summarization algorithms, we "thinned" the generated samples to every fifth. Hence, the total number M of samples used as observation for the summarizing algorithms was 16 000. The hyperparameters were set as follows (see Section 3.2.5); the shape parameter of the prior distribution over δ^2, i.e., $p(\delta^2)$, was set to $\alpha_{\delta^2} = 2$, in order to have a heavy-tailed "weakly informative" prior (with infinite variance). We set its scale parameter, β_{δ^2}, to an arbitrary moderate value of 20, while acknowledging the fact that the sampler is sensitive to its value in low SNR situations. Furthermore, the parameters of the Gamma prior over Λ are set to $\alpha_\Lambda = 1$ and $\beta_\Lambda \approx 0$ to have a flat prior over the number k of components.

Following (Djurić, 1996 ; Andrieu and Doucet, 1999), two experiments are considered in this section to demonstrate both the performance of the summarizing algorithms and the usefulness of the obtained new summaries. The parameters of both experiments are given in Table 3.1. We set the number N of observations to 64. The parameter r in the second experiment defines the resolution of the problem.

First experiment ($k = 3$)			
j	E_j	ϕ_j	$\omega_{j,k}/2\pi$
1	20	0	0.2
2	6.3246	$\pi/4$	$0.2 + 1/N$
3	20	$\pi/3$	$0.2 + 2/N$

Second experiment ($k = 2$)			
j	E_j	ϕ_j	$\omega_{j,k}/2\pi$
1	20	0	0.2
2	20	$\pi/4$	$0.2+\frac{1}{r \cdot N}$

Table 3.1 – *Parameters of the experiments in the problem of detection and estimation of sinusoids in white Gaussian noise. We define the energy* $E^2 \triangleq a_c^2 + a_s^2$*, the phase* $\phi \triangleq -\arctan(a_s/a_c)$*, and SNR* $\triangleq \frac{\|\mathbf{D}_k \mathbf{a}_k\|^2}{N\sigma^2}$.

As illustrative examples, we show results on three specific observed signals. One from the first experiment defined in Table 3.1, where there is a hard-to-detect component. Next, an observed signal from the second experiment of Table 3.1, where there are two very closely located sinusoidal components. Finally, as third illustrative example, we study the performance of the proposed summarizing approach in situations where the number k of components is large.

3.3.2 First illustrative example

The first illustrative example is an instance of the first experiment defined in Table 3.1 with SNR = 7dB. The goal of this example is to detect a hard-to-detect sinusoidal component located in the middle of two other “stronger” ones. So, the true number k of components is three. Figure 3.5 shows the observed and noiseless signals, i.e., $\mathbf{y}$ and $\mathbf{y}_0$, along with the periodogram of $\mathbf{y}$. It can be observed from the periodogram that there are two significant peaks corresponding to the two strong sinusoidal components, whereas the middle harder-to-detect sinusoidal component is masked by them.

Figure 3.6 shows the posterior distributions of the number k of components and the sorted radial frequencies $\boldsymbol{\omega}_k$ obtained from the output of the RJ-MCMC sampler for this sinusoid detection example. We ran the algorithms proposed in Chapter 2 on the specific example shown in Figure 3.6, for 100 iterations, with $L = 3$ Gaussian components (the posterior probability of $\{k \leq 3\}$ is approximately 90.3%). To initialize the means and variances of the Gaussian components, we used the median and normalized interquartile range of the marginal posterior distributions of sorted radial frequencies given $k = L$

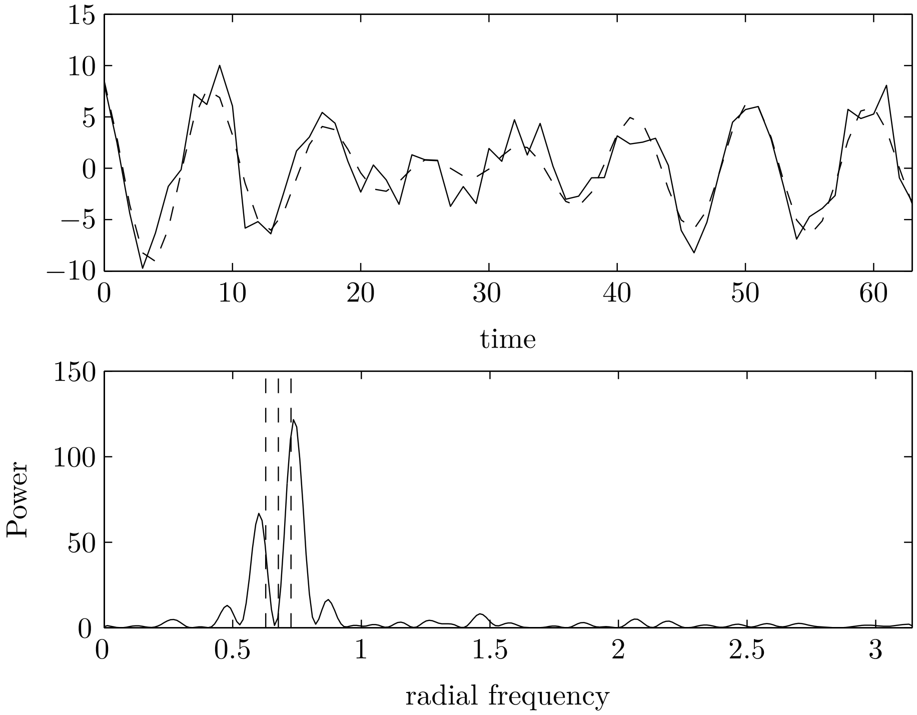


Figure 3.5 – *The top panel shows the observed (solid curve) and the noiseless (dashed curve) signals for the first experiment expressed in Table 3.1 with SNR = 7 dB. The bottom panel illustrates the periodogram of* $\mathbf{y}$. *The vertical dashed lines show the location of the true radial frequencies.*

(middle row of the right panel of Figure 3.6). We will call this approach of initialization the "naive" initialization procedure, hereafter.

Convergence assessment

Figures 3.7–3.9 illustrate the evolution of the model parameters, i.e., $\boldsymbol{\eta}_l = \{\mu_l,\, s_l^2,\, \pi_l\}$, with $1 \leq l \leq L$, and the mean parameter λ of the Poisson point process component together with the criteria to be minimized. Two substantial facts can be deduced from these figures:

i) the "generally" decreasing behavior of the criteria obtained from minimizing both the KL divergence and the BHHJ α-divergence, defined in Equations (2.14) and (2.29), respectively. They are almost constant after the 20^{th} iteration of the SEM-type algorithms.

ii) the convergence of the parameters of the parametric model, esp. the means μ and the probabilities of presence π, though using a naive initialization procedure. Indeed

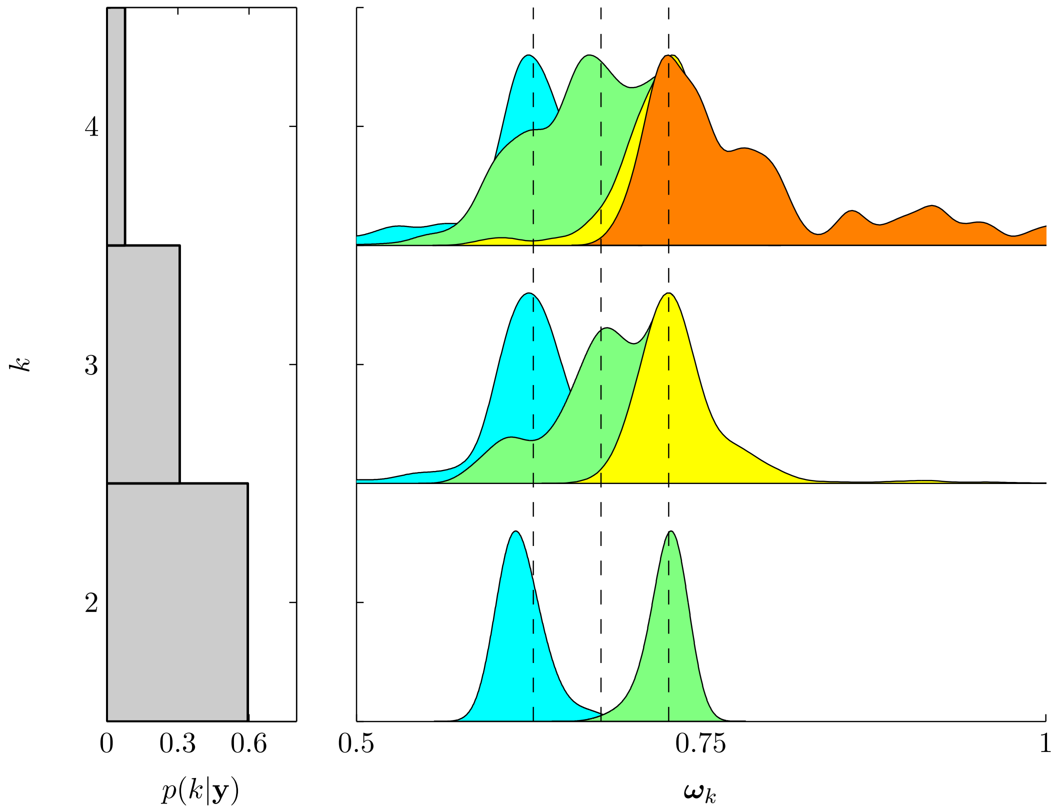


Figure 3.6 – *Posterior distributions of k (left) and sorted radial frequencies, $\boldsymbol{\omega}_k$, given k (right) from the output of the RJ-MCMC sampler for the second sinusoid detection experiment defined in Table 3.1 with SNR = 7dB (i.e., the first illustrative example). The true number of components is three. The vertical dashed lines in the right panel locate the true radial frequencies.*

after the 50^{th} iteration there is no significant move in the parameter estimates. Note that TAP-KL1 shows to be the fastest algorithm in the sense of convergence rate, but careful inspection in the obtained summary presented in Figure 3.10 reveals that it has converged to a solution that is not desirable due to the bimodality of the distribution of the samples allocated to the second Gaussian component (top right panel of Figure 3.10). The other two relatively robust algorithms, i.e., TAP-KL2 and TAP-BHHJ, have converged to similar solutions; see the estimated values presented in Figures 3.11 and 3.12. However, the latter one, converged in fewer iterations.

To inspect better the convergence of the algorithms, Table 3.2 presents the values of the KL and BHHJ criteria evaluated at the estimated model parameters using TAP-KL1, TAP-KL2, and TAP-BHHJ algorithms. Comparing the presented values of the KL criterion (2.14), it can be seen that, TAP-KL1 algorithm, which minimizes directly this

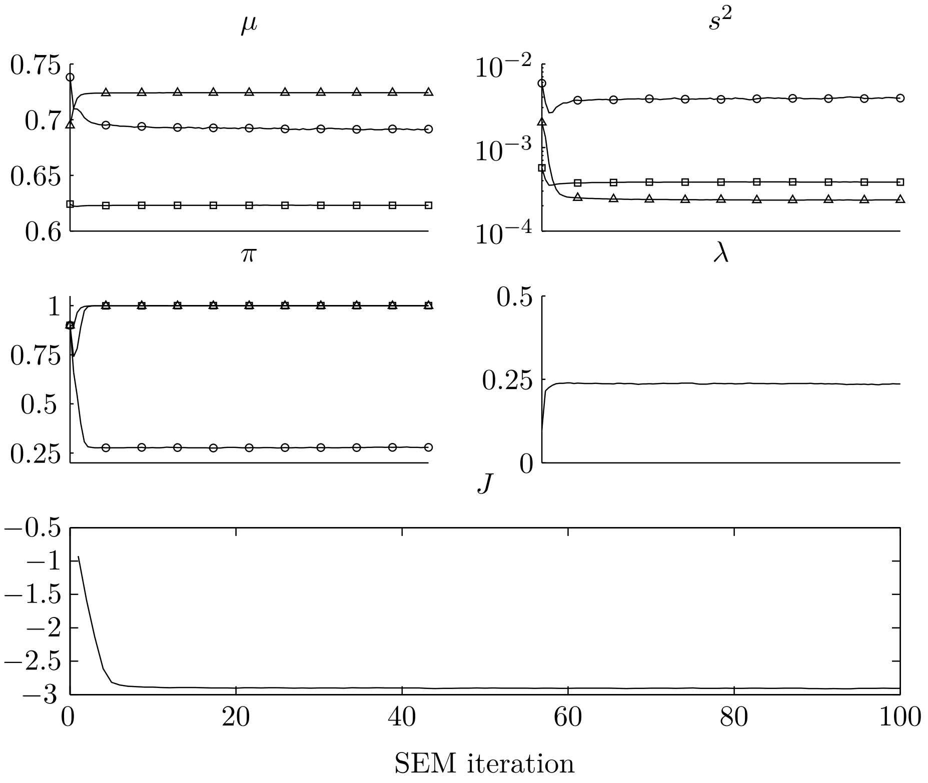


Figure 3.7 – *Evolution of the model parameters along with the criterion* $\hat{\mathcal{J}}$ *defined in* (2.14) *using TAP-KL1 with* $L = 3$ *on the first illustrative sinusoid detection example.*

criterion, has converged to a (local) minimum with the lowest value of the criterion. On the other hand, despite the summary obtained by TAP-KL2 presented in Figure 3.11 is preferable to the one of TAP-KL1, it has converged to a point in the parameter space with a greater value of the KL criterion. This might be due to the fact that, in TAP-KL2, the KL criterion is minimized indirectly by plugging the robust estimators of the mean and variance into the M-step. Tuning to the BHHJ criterion (2.29), one can see that both TAP-KL2 and TAP-BHHJ have converged to summaries with comparative values of the criterion; whereas, the evaluated value of the criterion at the solution of TAP-KL1 is quite higher than the others.

Relabeling properties

Figures 3.10–3.12 show the histograms of the labeled samples, i.e., $(\boldsymbol{x}^{(i)}, \mathbf{z}^{(i)})$, with $i = 1, \ldots, M$, along with the pdf's of the estimated Gaussian components (black solid line). Moreover, the summaries provided by the proposed algorithms for each component are presented in its corresponding panel. We used the average of the last 50 SEM iterations

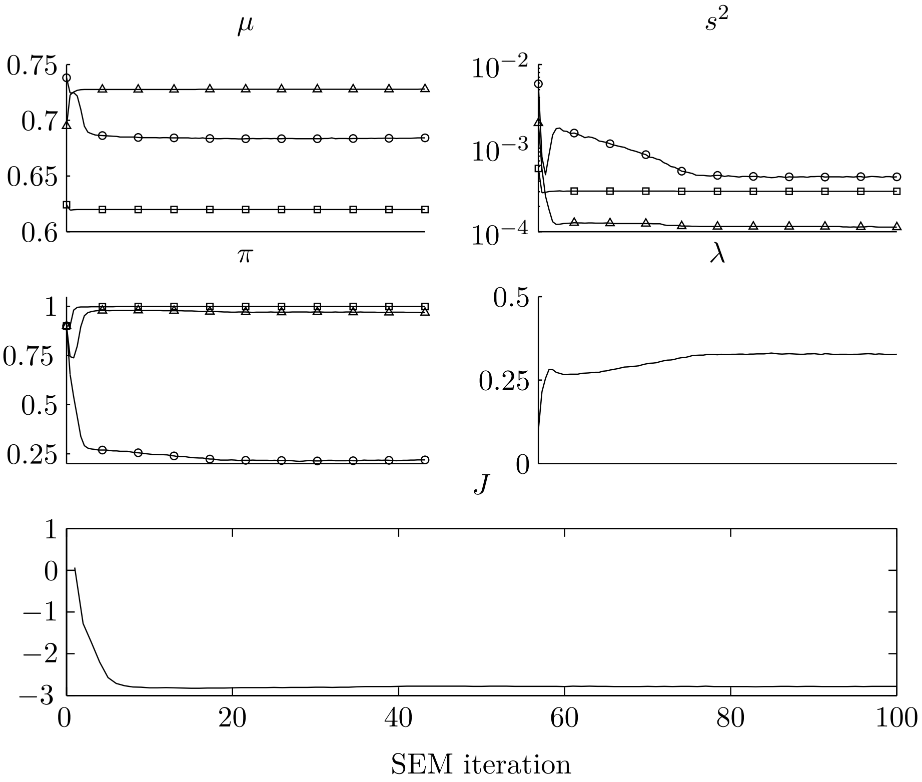


Figure 3.8 – *Evolution of the model parameters along with the criterion* $\hat{\mathfrak{J}}$ *defined in* (2.14) *using TAP-KL2 with* $L = 3$ *on the first illustrative sinusoid detection example.*

Criterion	KL	BHHJ
TAP-KL1	**-2.39**	-13.97
TAP-KL2	-2.20	-15.27
TAP-BHHJ	-2.17	**-15.34**

Table 3.2 – *KL and BHHJ criteria evaluated at the solutions obtained using different summarizing algorithms for the first illustrative sinusoid detection example. The smallest value for each criterion is highlighted in bold.*

as parameter estimates, as recommended in the SEM literature (see, for example, Celeux and Diebolt, 1992 ; Nielsen, 2000a). To reduce the variability of the histograms of the labeled samples due to randomness of allocation procedure, we labeled each sample 10 times using the S-step of the SEM-type algorithm given the estimated parameters of the model and then produce the histograms using all 10 labels.

The efficiency of the proposed algorithms for relabeling the variable-dimensional output samples of the RJ-MCMC sampler can be well observed in these figures. Comparing the

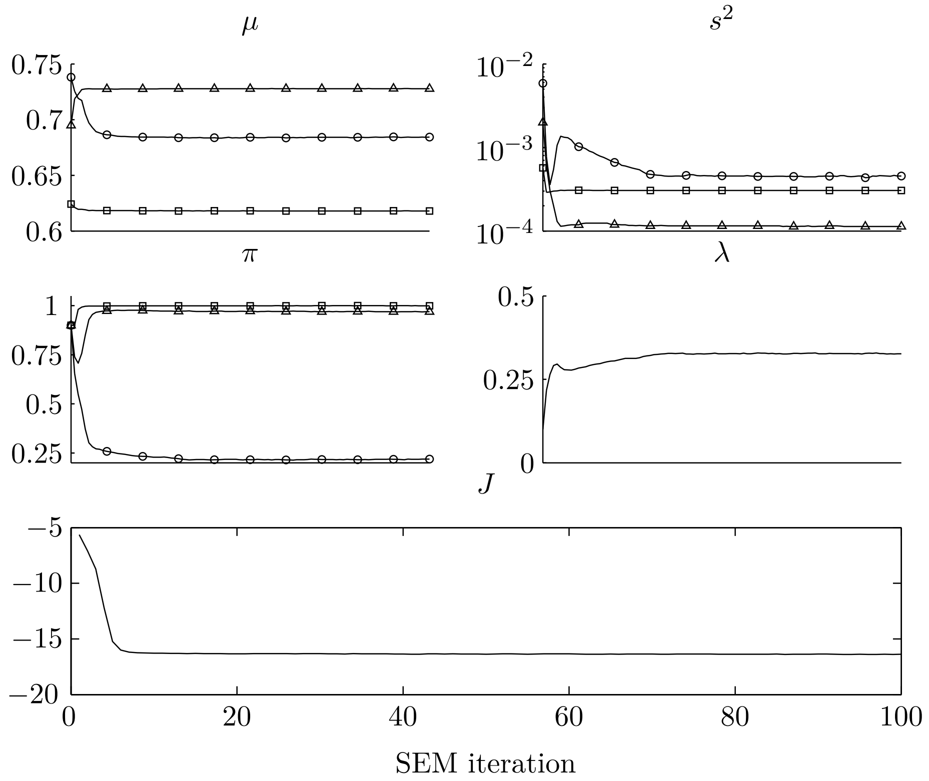


Figure 3.9 – *Evolution of the model parameters along with the criterion $\hat{\mathcal{J}}$ defined in* (2.29) *using TAP-BHHJ with $\alpha = 0.5$ and $L = 3$ on the first illustrative sinusoid detection example.*

distributions of the labeled samples with the ones of the posterior distributions of the sorted radial frequencies given $k = 3$ shown in Figure 3.13(a), which are highly multimodal, reveals the capability of the proposed summarizing algorithms to solve the label-switching in a variable-dimensional setting. Note, however, that the histogram of the allocated samples to the second Gaussian component, which corresponds to the middle harder-to-detect sinusoidal component of the example under study, is bimodal when using TAP-KL1 (see the top right panel of Figure 3.10). On the other hand, using the robust algorithms resulted in distributions of the samples labeled as the second Gaussian component to be nearly unimodal and enjoy compact dispersion; see the top right panel of Figures 3.11 and 3.12.

Looking at the bottom right panels, the role of the point process component in capturing the outliers in the observed samples that cannot be described by the Gaussian components becomes clearer. Note that, without the point process component, these outliers would be allocated to the Gaussian components which can, consequently, yield a significant deterioration of the parameter estimates.

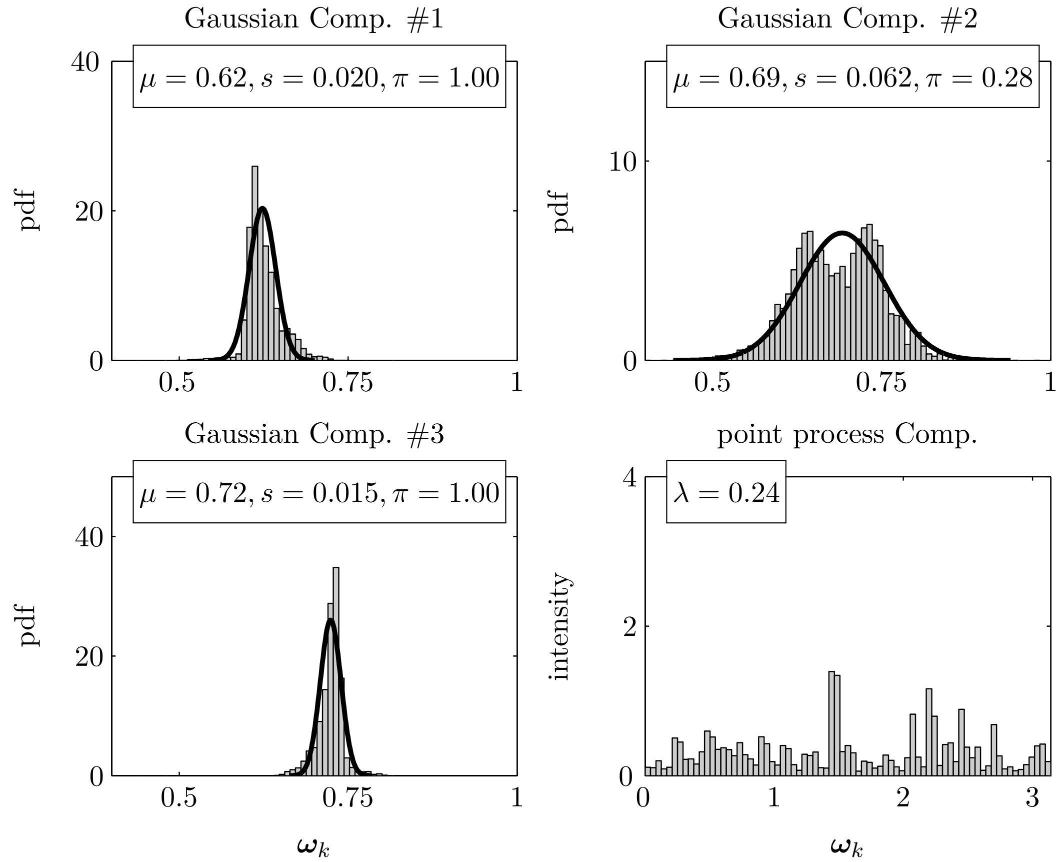


Figure 3.10 – *Histograms of the labeled samples, that is, the samples allocated to the Gaussian and Poisson point process components, versus the pdf's of estimated Gaussian components in the model (black solid line) using TAP-KL1 on the first illustrative sinusoid detection example. The estimated parameters of each component are presented in the corresponding panel. To generate these histograms the randomized allocation procedure was run 10 times.*

Turning to the comparison of the provided summaries using the proposed algorithms with the ones of the BMS approach, the first point to note is that using the BMS approach on this specific example results in loosing the middle sinusoidal component by selecting $\mathcal{M}_2$. Then the estimated summaries for the two detected components using the robust estimates of the posterior distributions of the sorted radial frequencies given $\mathcal{M}_2$ are: $\boldsymbol{\mu} = (0.62, 0.73)$ and $\boldsymbol{s} = (0.016, 0.012)$. Contrary to the BMS approach, the approach that we proposed enabled us to benefit from the information of all probable models to give summaries about the middle harder-to-detect component. It can be seen from the estimated summaries presented in Figures 3.10–3.12 that the estimated means are compatible with the true radial frequencies, i.e., (0.628, 0.677, 0.727). Furthermore, the estimated probabilities of presence are consistent with the uncertainties of the sinusoidal components

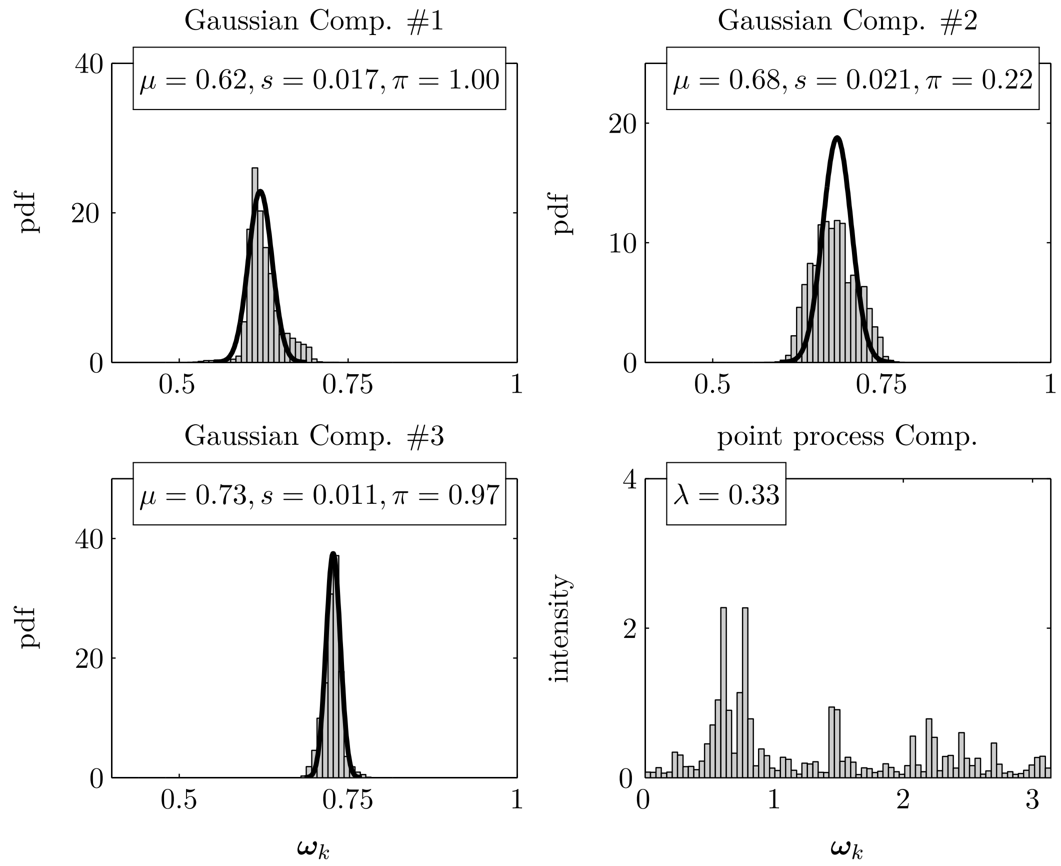


Figure 3.11 – *Histogram of the labeled samples, that is, the samples allocated to the Gaussian and Poisson point process components, versus the pdf's of estimated Gaussian components in the model (black solid line) using TAP-KL2 on the first illustrative sinusoid detection example. The estimated parameters of each component are presented in the corresponding panel. To generate these histograms the randomized allocation procedure was run 10 times.*

in the experiment; that is, there are two components with high "confidence" and one in the middle with less "confidence". One should also note that the obtained summaries using TAP-KL2 and TAP-BHHJ with $\alpha = 0.5$ are identical.

Validation of the fitted models

To observe better the "goodness-of-fit" of the estimated Gaussian components, Figure 3.13(b) depicts the normalized densities [1] of them underneath the posterior distributions of the

[1] 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 height of the normalized densities amounts to the corresponding estimated probability of presence.

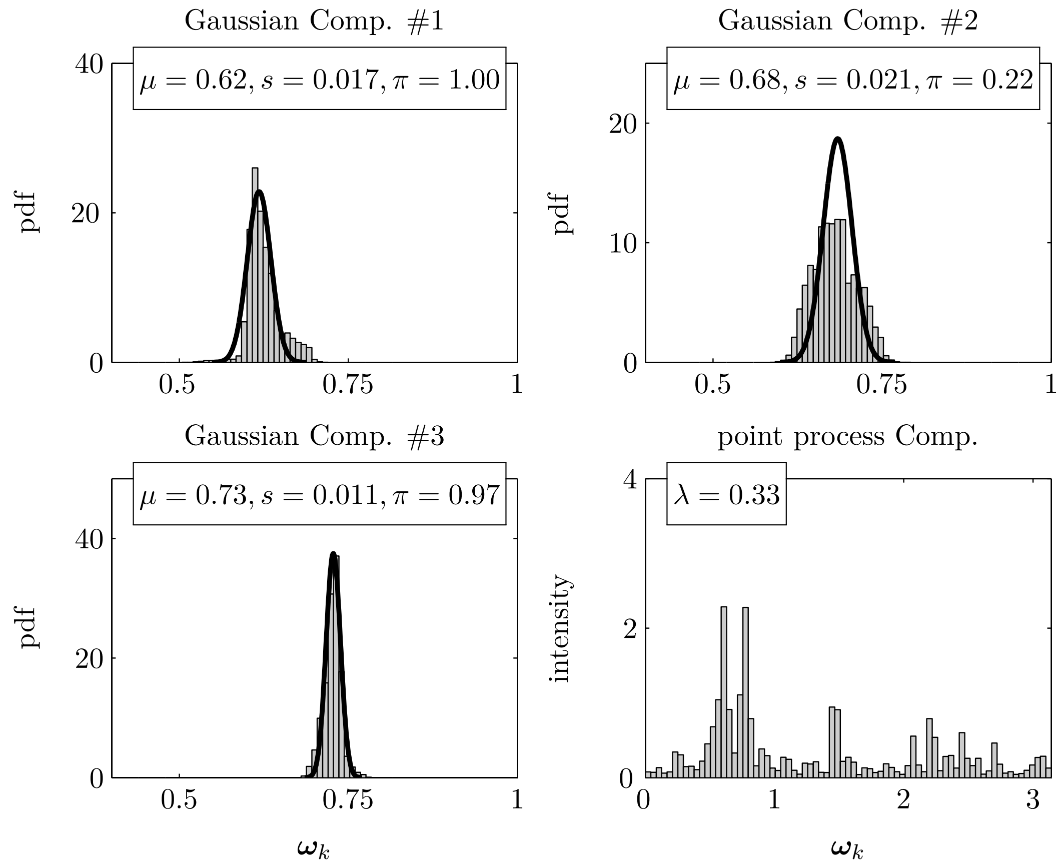


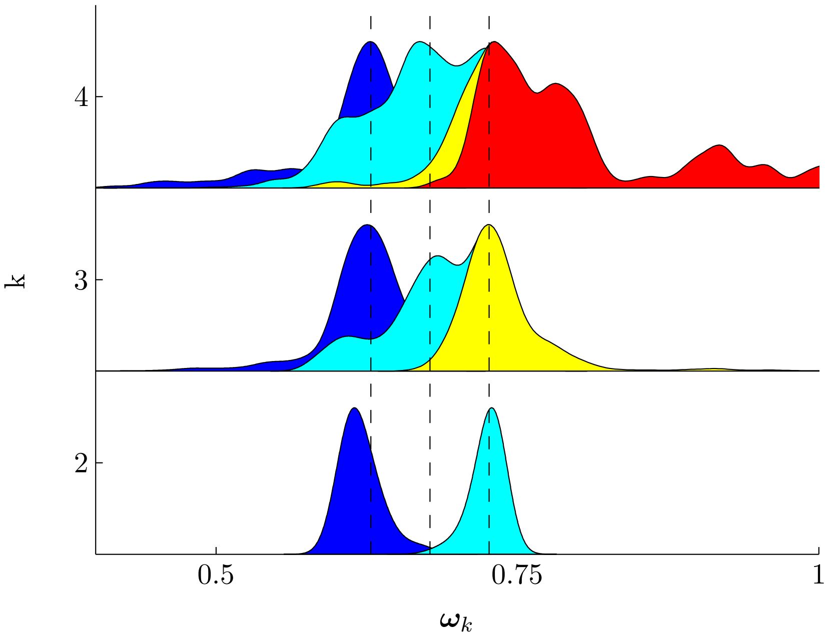
Figure 3.12 – *Histogram of the labeled samples, that is, the samples allocated to the Gaussian and Poisson point process components, versus the pdf's of estimated Gaussian components in the model (black solid line) using TAP-BHHJ with $\alpha = 0.5$ on the first illustrative sinusoid detection example. The estimated parameters of each component are presented in the corresponding panel. To generate these histograms the randomized allocation procedure was run 10 times.*

sorted radial frequencies given k illustrated in Figure 3.13(a). These figures 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. Note also the effect of using the robust algorithms on the estimated variance of the middle Gaussian component.

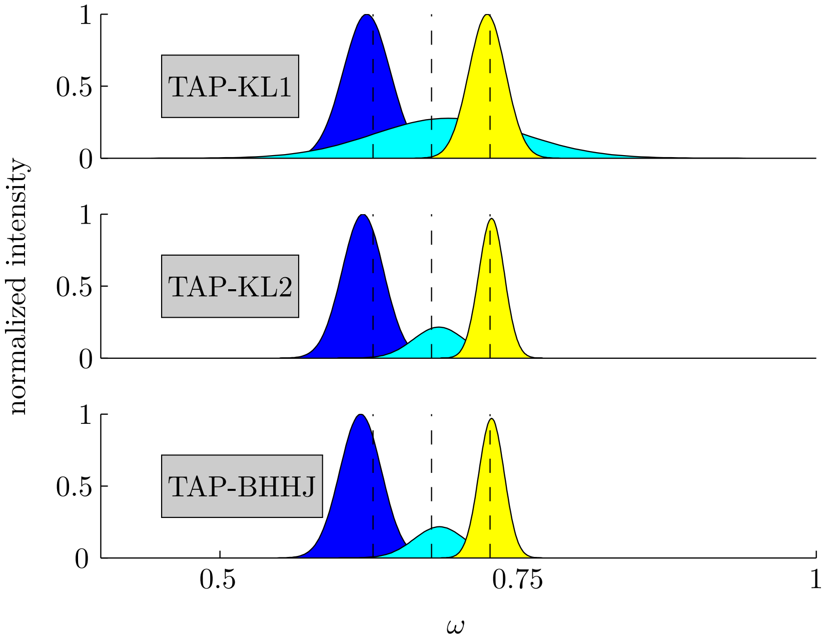
It is also useful for validating the estimated summaries to compare the intensity of the estimated parametric model $q_{\boldsymbol{\eta}}$ defined, in general, as

$$h(\boldsymbol{\eta}) \;=\; \sum_{l=1}^{L} \pi_l \,\cdot\, \mathcal{N}(\,\cdot\,|\,\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l), \tag{3.8}$$

where we ignore the point process component, with the histogram intensity of all radial



(a)



(b)

Figure 3.13 – *(a) The posterior distributions of the sorted radial frequencies given k. (b) corresponding normalized pdf's of fitted Gaussian components for three proposed summarization algorithms with $L = 3$ Gaussian components. The estimated parameters can be read in Figures 3.10–3.12. The dashed lines locate the true radial frequencies.*

frequencies obtained using the BMA approach explained in Section 2.2.2. Figure 3.14 shows such figures 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. It can be seen from the figures that the robust algorithms, i.e., TAP-KL2 and TAP-BHHJ, capture better the posterior information of the radial frequencies in comparison with TAP-KL1.

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 3.15 illustrates the posterior distribution of the number k of components together with its approximated versions using the proposed summarizing algorithms. It can be seen from the figure that the summarizing algorithms well captured the information provided in the true posterior of the number k of components.

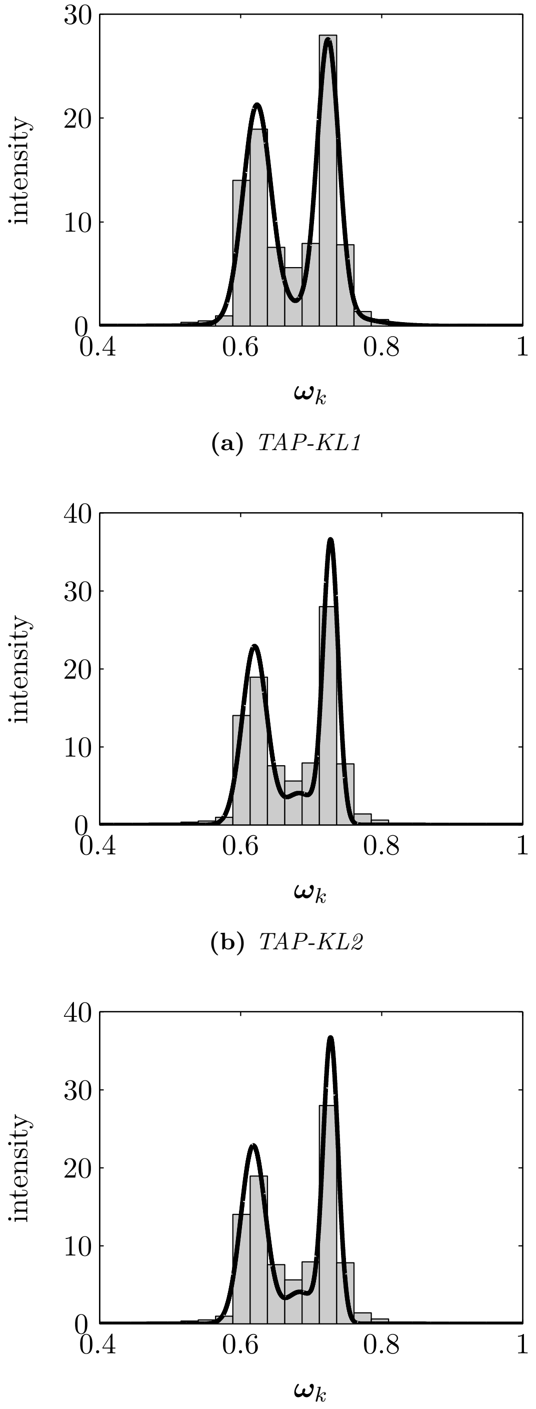
Remark 3.3. The expected number of components in the approximate posterior $q_{\boldsymbol{\eta}}$ is given by

$$\sum_{l=1}^{L} \pi_l + \lambda.$$

The posterior mean of $p(k \mid \mathbf{y})$ is 2.51, while the expected number of components in the approximate posteriors for all algorithms equal to 2.52 (see the estimated parameters presented in Figures 3.10–3.12).

Remark 3.4. Recall that the binary indicator variables ξ_l, $1 \leq l \leq L$, introduced in Chapter 2 to control the presence of the Gaussian components, are assumed to be *independently* Bernoulli distributed. Hence, the approximate posterior $q_{\boldsymbol{\eta}}$ by definition is not capable of reproducing the existing correlation in the presence of components in the true variable-dimensional posterior distribution. More precisely, for example, the presence of component, say, a, in the true posterior might preclude the presence of component, say, b. Although this characteristic cannot be preserved by the proposed approximate model, we can recover this information from the labeled samples $(\boldsymbol{x}^{(i)}, \mathbf{z}^{(i)})$, $i = 1, \dots, M$, given the estimated parameters $\hat{\boldsymbol{\eta}}$. Noting the fact that the vector of indicator vectors $\boldsymbol{\xi}$ can be obtained given the simulated allocation vector $\mathbf{z}$, we can easily compute their correlation. For the labeled samples shown in Figure 3.12, the matrix of correlation coefficients becomes

$$\begin{bmatrix} 1 & -0.01 & 0 \\ -0.01 & 1 & -0.37 \\ 0 & -0.37 & 1 \end{bmatrix}.$$



(a) *TAP-KL1*

(b) *TAP-KL2*

(c) *TAP-BHHJ with* $\alpha = 0.5$

Figure 3.14 – *Histogram intensity of all radial frequencies samples using BMA approach along with the intensity of the fitted parametric model obtained using the proposed summarizing algorithms.*

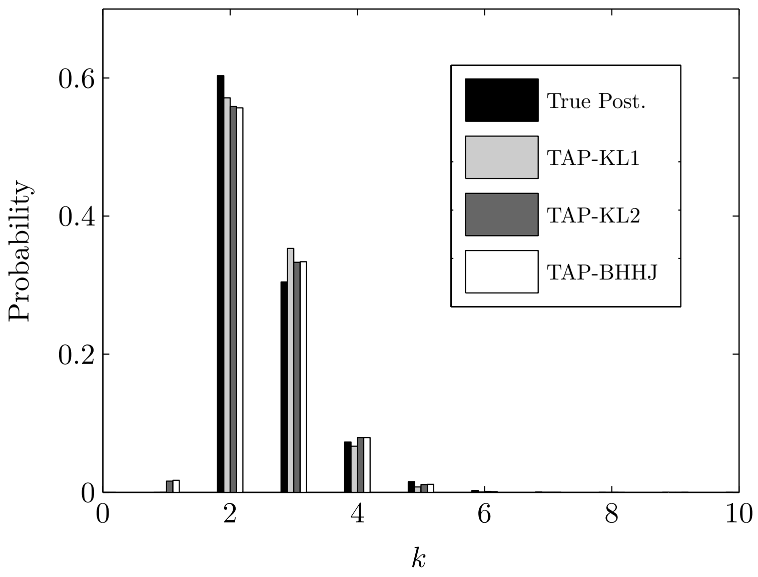
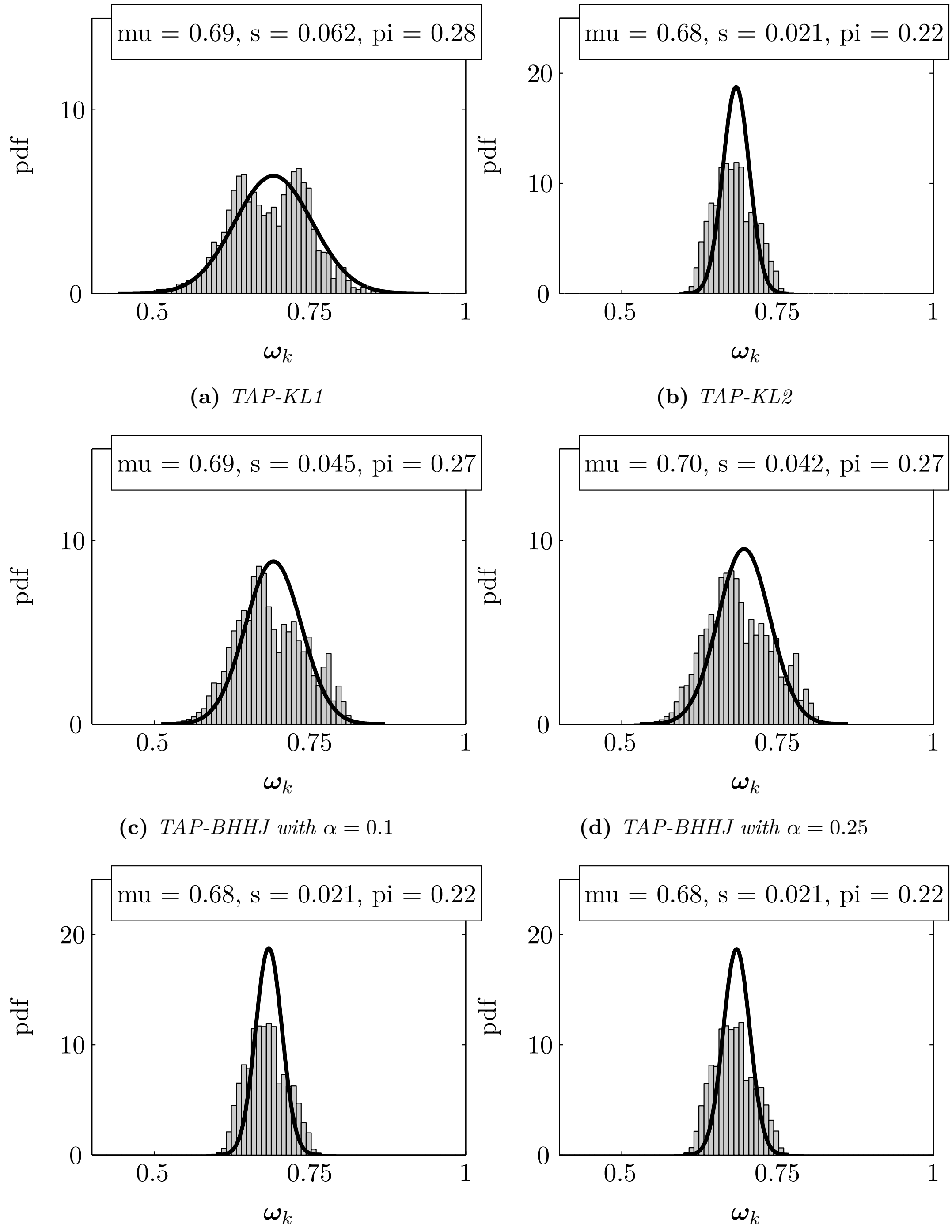


Figure 3.15 – *Posterior distribution of the number k of sinusoidal components along with its approximated ones using the proposed algorithms on the first illustrative sinusoid detection example.*

The above correlation coefficients can be justified as follows; the first component has the probability of presence equal to one, so it is always present no matter the presence of the others. The other two components are correlated, as, for example, when $k = 2$, presence of one of them forces absence of the other one.

Remark 3.5. As discussed in Section 2.5, the parameter α of BHHJ-α divergence can be considered as a tuning parameter that controls the compromise between efficiency ($\alpha \to 0$) and robustness ($\alpha \to \infty$) of the derived estimator. Thus, the obtained summary is sensitive to the chosen value of α. Basu et al. (1998) recommended to choose a value in the region $0 < \alpha \leq 1$.

Figure 3.16 shows the effect of the parameter α on both the second fitted Gaussian component and its labeled samples when using TAP-BHHJ on the first illustrative example along with the results obtained using TAP-KL1 and TAP-KL2. The parameters of the other two Gaussian components were almost similar for all cases (not shown here). It can be seen from the figure that for small values of α, i.e., $\alpha = 0.1$ and 0.25, the obtained summaries are close to the one of TAP-KL1. On the other hand, when α takes a larger value, i.e., $\alpha = 0.5$ and 1, the obtained summaries are completely identical—in this specific example—to the one of TAP-KL2.



(a) *TAP-KL1*

(b) *TAP-KL2*

(c) *TAP-BHHJ with* $\alpha = 0.1$

(d) *TAP-BHHJ with* $\alpha = 0.25$

(e) *TAP-BHHJ with* $\alpha = 0.5$

(f) *TAP-BHHJ with* $\alpha = 1$

Figure 3.16 – *Effect of* α *on the second fitted Gaussian component when applying TAP-BHHJ on the first illustrative sinusoid detection example.*

3.3.3 Second illustrative example

The second illustrative example highlights a situation in which the proposed summarizing approach might have difficulties. Figure 3.17 illustrates the variable-dimensional posterior distribution for this example, which is an instance of the second experiment given in Table 3.1 with SNR $=$ 7dB and $r = 2$. A remarkable feature of this example is that the location of the sinusoidal component under $\mathcal{M}_1$ is not "coherent" with the locations of the two sinusoidal components under $\mathcal{M}_2$. Moreover, both models are *a posteriori* nearly equiprobable ($p(\mathcal{M}_1 \,|\, \mathbf{y}) = 0.48$ and $p(\mathcal{M}_2 \,|\, \mathbf{y}) = 0.41$). Obviously, the RJ-MCMC output samples under $\mathcal{M}_1$ cannot be described by the two sinusoidal components under $\mathcal{M}_2$ properly. Therefore, this example can be considered as a challenging problem for the summarization approach we have developed.

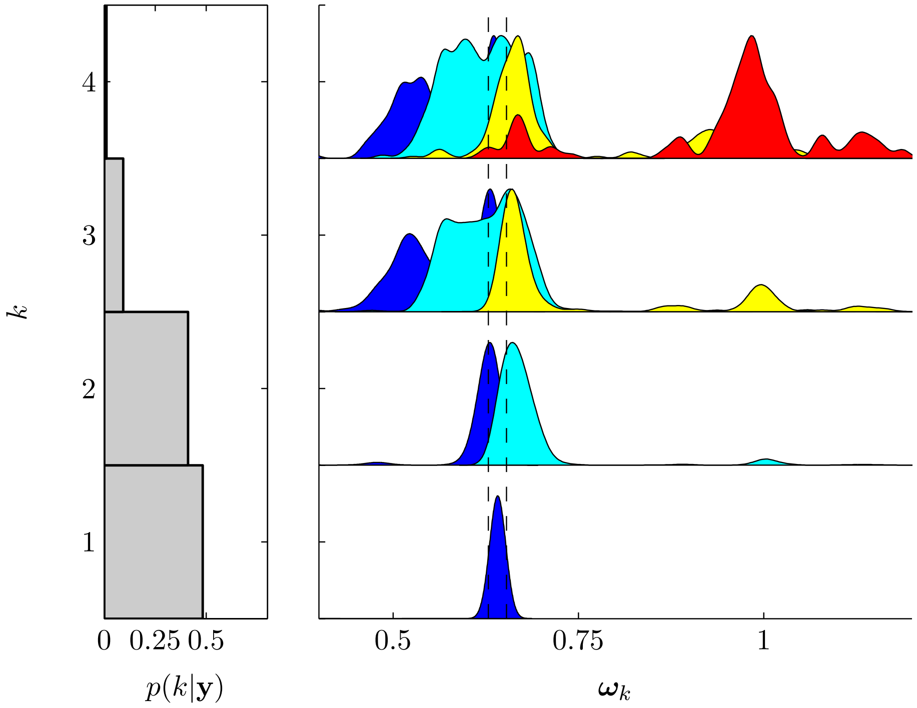


Figure 3.17 – *Posterior distributions of the number k of components (left) and the sorted radial frequencies, $\boldsymbol{\omega}_k$, given k (right) constructed using the 80 000 RJ-MCMC samples after discarding the first 20 000 samples as the burn-in period. The true number of components is two. It is indeed an example of the observed signal from the second experiment explained in Table 3.1 with SNR $= 7$dB and $r = 2$. The vertical dashed lines in the right figure locate the true radial frequencies, i.e.,* $(0.628,\, 0.653)$.

We ran the algorithms on the variable-dimensional samples generated using RJ-MCMC shown in Figure 3.17, for 100 iterations, with $L = 3$ Gaussian components (the posterior probability of $\{k \leq 3\}$ is approximately 98.7%). To initialize the parameters of the Gaussian components, we used the robust estimates of the mean and variance of the posterior distributions of the sorted radial frequencies given $k = L$, as in the previous example.

Figure 3.18 illustrate the evolution of the model parameters, i.e., $\boldsymbol{\eta}_l = \{\mu_l,\, s_l^2,\, \pi_l\}$, with $1 \leq l \leq L$, and the mean parameter λ of the Poisson point process component together with the criterion $\mathcal{J}$ when using TAP-BHHJ with $\alpha = 0.5$ to summarize the posterior samples shown in Figure 3.17. The decreasing behavior of $\mathcal{J}$ and convergence of the parameters can be observed from the figure.

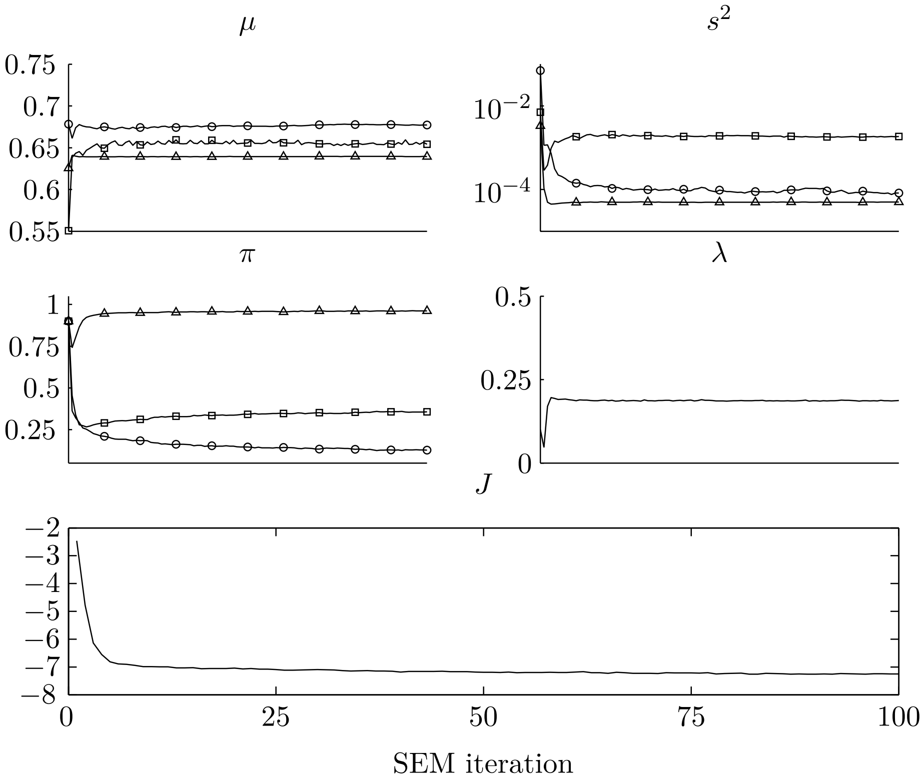


Figure 3.18 – *Evolution of the model parameters along with the criterion $\hat{\mathcal{J}}$ defined in* (2.29) *using TAP-BHHJ with $\alpha = 0.5$ and $L = 3$ on the second illustrative sinusoid detection example.*

The histograms of the labeled samples along with the pdf's of the estimated Gaussian components (black solid line) are shown in Figures 3.19 and 3.20, respectively, when TAP-KL2 and TAP-BHHJ with $\alpha = 0.5$ were used. Moreover, the summaries obtained by the proposed algorithms for each component are presented in its corresponding panel. We used the average of the last 50 SEM iterations as parameter estimates. As in the previous illustrative example, we ran the randomized allocation procedure 10 times to reduce variations in the histograms of the labeled samples.

The first point to note is that both summarizing algorithms have associated a Gaussian component to the RJ-MCMC output samples around the sinusoidal component under $\mathcal{M}_1$ concentrated around 0.64 (see Figure 3.17). Moreover, this Gaussian component is in the obtained summaries with a very high "confidence" (it has the probability of presence greater than 0.9 in both cases).

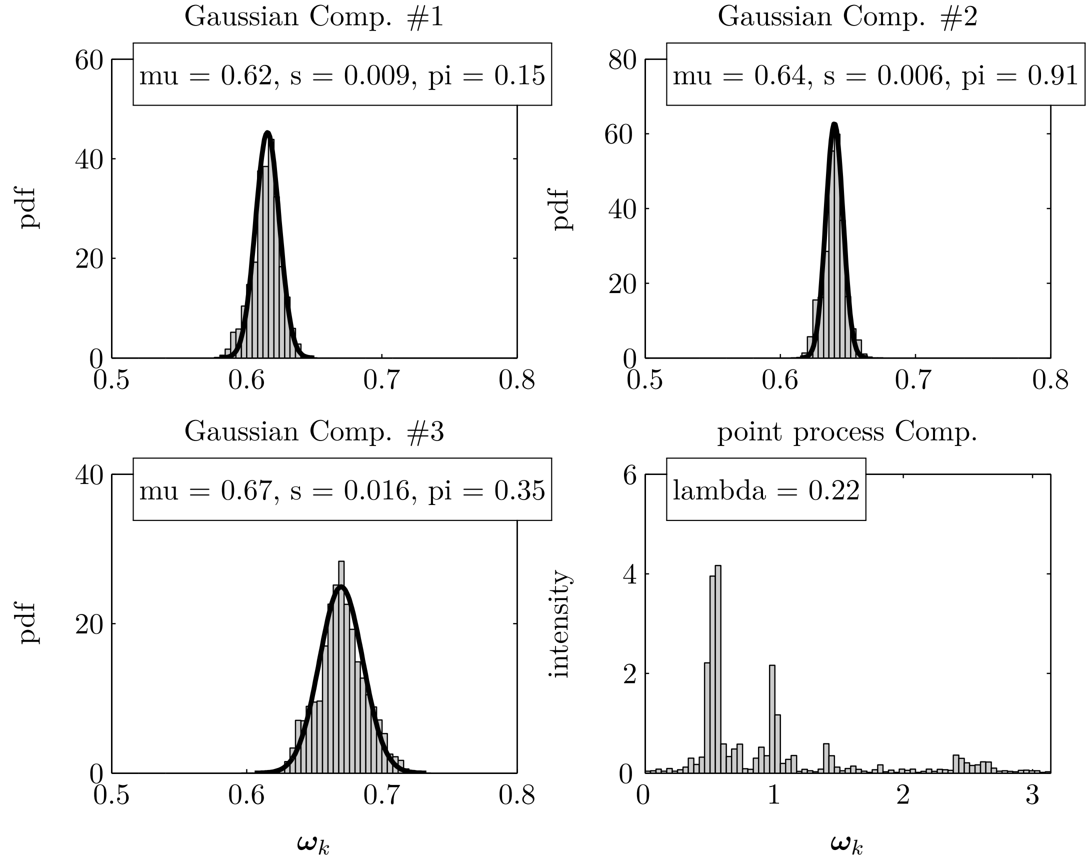


Figure 3.19 – *Histogram of the labeled samples versus the pdf's of estimated Gaussian components in the model (black solid line) using TAP-KL2 with $L = 3$ on the second illustrative sinusoid detection example. The estimated parameters of each component are presented in the corresponding panel. To generate these histograms the randomized allocation procedure was run 10 times.*

Comparing the two summaries shown in Figures 3.19 and 3.20, we observe that the one obtained using TAP-KL2 algorithm seems to be more appropriate as its Gaussian components enjoy smaller variances. The reason of the difference can be

i) TAP-BHHJ with $\alpha = 0.5$ was not robust enough;

ii) the model q_η with $L = 3$ Gaussian components was not a suitable approximate posterior to capture the information in the variable-dimensional posterior distribution shown in Figure 3.17;

iii) the naive initialization procedure used so far was not applicable in this example and the algorithm has been trapped in a local minimum.

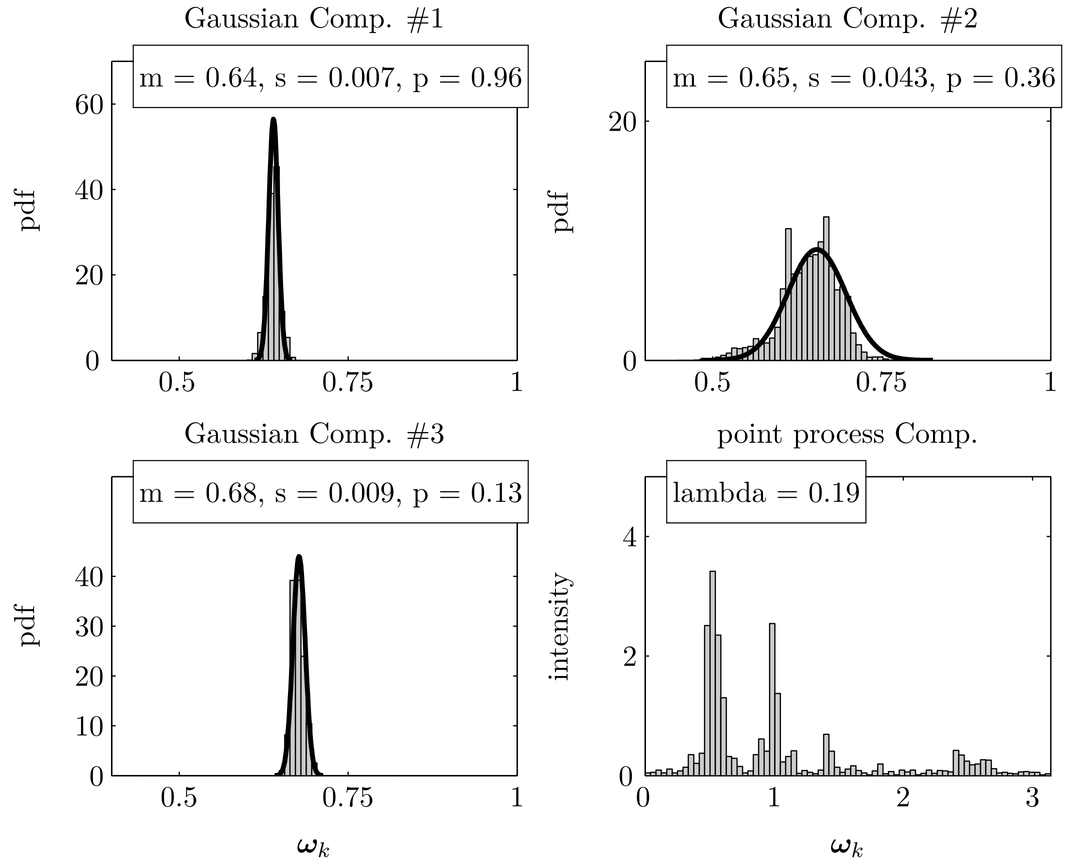


Figure 3.20 – *Histogram of the labeled samples versus the pdf's of estimated Gaussian components in the model (black solid line) using TAP-BHHJ with $\alpha = 0.5$ and $L = 3$ on the second illustrative sinusoid detection example. The estimated parameters of each component are presented in the corresponding panel. To generate these histograms the randomized allocation procedure was run 10 times.*

Indeed, as discussed in Section 2.5, TAP-BHHJ downweights the influence of the outliers and samples located at the tails of the distributions at a rate which depends on α, whereas TAP-KL2 "ignores" those samples using robust quantile-based estimators. Thus, one might attempt at increasing the value of α, say, set $\alpha = 1$, to obtain results close to the one of TAP-KL2. Although this approach improves the obtained summary on this specific example (results not shown here), in the following, we choose to investigate the effect of both the number L of Gaussian components and the initialization procedure.

Increasing the number L Gaussian components

Contemplating the top panels of Figure 3.17, i.e., the posterior distributions of the sorted radial frequencies given $k = 3$ and 4, it can be observed that there are non-negligible amount of samples concentrated around $\omega = 0.5$. Those samples were completely allocated to the point process component by TAP-KL2 algorithm, whereas TAP-BHHJ with $\alpha = 0.5$ allocated a portion of them to the Gaussian component with the largest estimated variance; see the histograms of the labeled samples shown in Figures 3.19 and 3.20, particularly the peak concentrated around 0.5 in the bottom right panels. Generally, having such large peaks in the histogram of the samples allocated to the point process component indicates that the chosen value of L was not sufficient. Therefore, we ran TAP-BHHJ with $\alpha = 0.5$ on the posterior shown in Figure 3.17 again, but this time using a parametric model $q_{\boldsymbol{\eta}}$ with $L = 4$ Gaussian components.

Remark 3.6. To initialize, however, using the naive initialization procedure is not reasonable here, as not only the amount of the posterior samples given $k = 4$ is not sufficient (note that $p(k = 4 \,|\, \mathbf{y}) = 0.01$) but also the posterior distributions under $\mathcal{M}_4$ exhibit a "severe" label-switching. Hence, we used an "advanced" initialization procedure that will be explained in Section 4.3. In a nutshell, it consists in allocating all the samples to the point process component and then, extracting Gaussian components from it progressively. After adding each Gaussian component, a few, say, five, iterations of TAP-KL2 is performed to estimate all parameters of the parametric model, including the probabilities of presence π_l, $1 \leq l \leq L$ and the mean λ (note that in the naive initialization, these parameters are set to arbitrary constants values).

Figure 3.21 shows the resulting summary when $L = 4$ and the advanced initialization was used. It can be seen that the first Gaussian component shown on the top left panel caught the samples concentrated around $\omega = 0.5$ which consequently resulted in the other Gaussian components to be of small dispersion. Figure 3.22 shows the posterior distributions of the sorted radial frequencies given k (on top) and the normalized pdf's of the fitted Gaussian components using the three summarizing algorithms (on bottom). In TAP-KL1 and TAP-KL2, L was set to three, whereas in TAP-BHHJ we set $L = \{3, 4, 5\}$. When $L = 3$, all algorithms were initialized using the naive initialization procedure; while, for $L > 3$, we used the advanced initialization procedure. These results allow us to study both the impact of the robust algorithms and the number L of Gaussian components.

It can be seen from Figure 3.22(b) that the summary obtained using TAP-KL1 contains a component with a very large variance. This large variance component exists in the summary obtained using TAP-BHHJ with $\alpha = 0.5$ and $L = 3$, but its variance is much lower than the one in summary of TAP-KL1. However, by increasing the number L of Gaussian components to four and five, it can be seen from the figure that the obtained summaries become fairly similar to the one of TAP-KL2, for the three Gaussian components with large probabilities of presence, but with additional components at $\omega = 0.54$ and $\omega = 1$, both with small probabilities of presence.

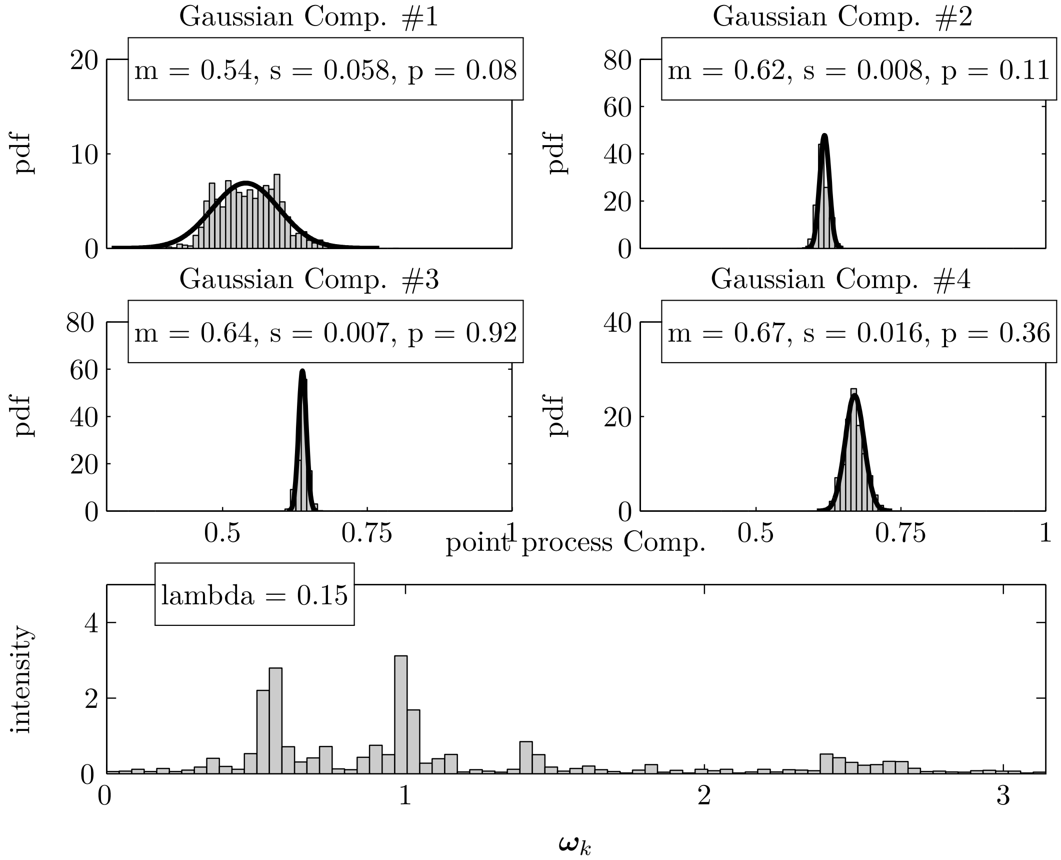
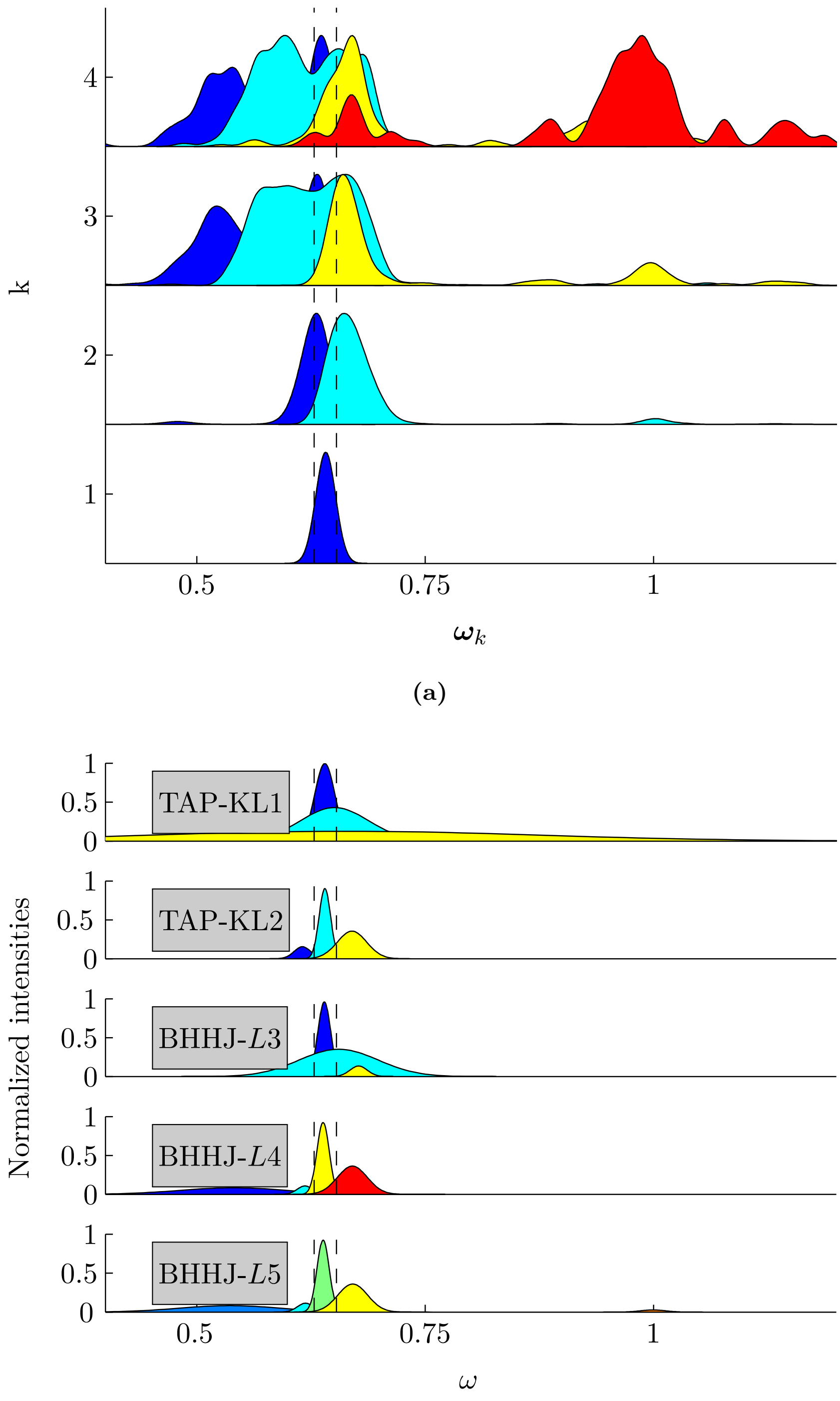


Figure 3.21 – *Histogram of the labeled samples versus the pdf's of estimated Gaussian components in the model (black solid line) using TAP-BHHJ with $\alpha = 0.5$ and $L = 4$ on the second illustrative sinusoid detection example. The estimated parameters of each component are presented in the corresponding panel. To generate these histograms the randomized allocation procedure was run 10 times.*

To compare the convergence of the algorithms, Table 3.3 presents the KL and BHHJ criteria evaluated at the estimated model parameters for different summarizing algorithms. For TAP-BHHJ, different values of L are considered. The first point to note from the table is that, in this specific example, the evaluated KL criterion for TAP-KL1 is greater than that of TAP-KL2. This suggests that TAP-KL1 might have been trapped in a local minimum. The second point is that by increasing the number L of components, when using TAP-BHHJ, the evaluated KL and BHHJ criteria decrease. Moreover, when there are $L = 5$ components in the model, both criteria have their lowest values.

Based on the presented results of the summarizing algorithms on the two sinusoid detection examples so far (and our exhaustive experiments not shown here), we can con-



(a)

(b)

Figure 3.22 – *(a) Posterior distributions of the sorted radial frequencies given k. (b) corresponding normalized pdf's of fitted Gaussian components for three proposed summarization algorithms applied on the second illustrative sinusoid detection example. In TAP-KL1 and TAP-KL2, L was set to three, whereas in TAP-BHHJ we set* $L = \{3, 4, 5\}$. *The dashed lines locate the true radial frequencies.*

Criterion	KL	BHHJ
TAP-KL1	-1.69	-6.72
TAP-KL2	-1.78	-6.93
TAP-BHHJ-$L3$	-1.63	-6.92
TAP-BHHJ-$L4$	-1.81	-7.20
TAP-BHHJ-$L5$	**-1.86**	**-7.26**

Table 3.3 – *KL and BHHJ criteria evaluated at the solutions obtained using different summarizing algorithms for the second illustrative sinusoid detection example. The smallest value for each criterion is highlighted in bold.*

clude that TAP-KL1 is less appropriate among the other two algorithms for this task. It often results in summaries containing large variance Gaussian components with the corresponding distribution of labeled samples being multimodal. On the other hand, both comparatively robust TAP-KL2 and TAP-BHHJ algorithms provide desirable summaries with compact components. Furthermore, since in TAP-BHHJ, the criterion is directly minimized, it allows for analyzing convergence of the algorithm.

Diagnosis of the lack-of-fit

As a concluding remark for this illustrative example, which has been intended to show situations where using the proposed summarizing approach has difficulties, we recall again the kind of trans-dimensional problems that using the proposed approach is meaningful. The notion of "birth, death, and switching of labels" introduced in Section 2.2.1 supposes that there is a certain relation between the locations of components when moving across models. More specifically, when moving from $\mathcal{M}_k$ to $\mathcal{M}_{k+1}$, it is assumed that the locations of the k components are aligned with the corresponding ones under $\mathcal{M}_k$ and only a new component with a new label is born. This is the phenomenon that is well illustrated in, for example, Figure 3.13. But, in the example presented in Figure 3.22, the component under $\mathcal{M}_1$ is not aligned with the two components under $\mathcal{M}_2$. As a results, the samples under $\mathcal{M}_1$ cannot easily be described by the two components under $\mathcal{M}_2$.

To detect such challenging situations, one should inspect both the posterior distributions of the number k of components and the sorted radial frequencies given k. Existence of discrepancies in the location of components across models together with nearly equiprobable models can be an indication of such situations. Note that, in the second illustrative example, if one the models had a lower posterior probability, then, it would have less affected the obtained summary. Finally, the correlation between the indicator variables can reveal also useful information about this kind of discrepancy. For TAP-BHHJ with $\alpha = 0.5$ and $L = 3$ (see Figure 3.20 for the obtained summary), the matrix of correlation coefficients is

$$\begin{bmatrix} 1 & -0.27 & -0.41 \\ -0.27 & 1 & -0.06 \\ -0.41 & -0.06 & 1 \end{bmatrix}.$$

Whereas, for TAP-BHHJ with $\alpha = 0.5$ and $L = 4$ (see Figure 3.21 for the obtained summary), the matrix of correlation coefficients is

$$\begin{bmatrix} 1 & 0.05 & -0.35 & 0.07 \\ 0.05 & 1 & -0.48 & 0.05 \\ -0.35 & -0.48 & 1 & -0.37 \\ 0.07 & 0.05 & -0.37 & 1 \end{bmatrix}.$$

From both correlation matrices, it can be seen that the component at $\hat{\mu} = 0.64$ (first row for $L = 3$ and third row for $L = 4$) has a significant correlation with the others.

Did the proposed approach completely fail?

In this specific example, the summaries, particularly, the estimated probabilities of presence, obtained using the summarizing algorithms are not in accordance with the variable-dimensional posterior distribution. For example, all the obtained summaries assert that there is a component with high probability of presence at $\omega = 0.64$,; whereas this is not the case in the true posterior distribution shown in Figure 3.17. Hence, one might argue that the obtained summary can result in misinterpretation.

Nonetheless, some features of the posterior distribution such as the location and dispersion of the sinusoidal components are well estimated. Moreover, the posterior distribution of the number k of components is preserved by the approximate parametric model (see Figure 3.23). Furthermore, we provided diagnoses of the lack-of-fit that warn us about this kind of situations. For example, existence of Gaussian components presence of which are highly correlated with the others in the matrix of correlation can be a sign of such issues.

3.3.4 Third illustrative example: Many components

In this section, we investigate the capability of the proposed SEM-type algorithms in dealing with challenging situations where the number k of components is large, say, $k > 10$. In

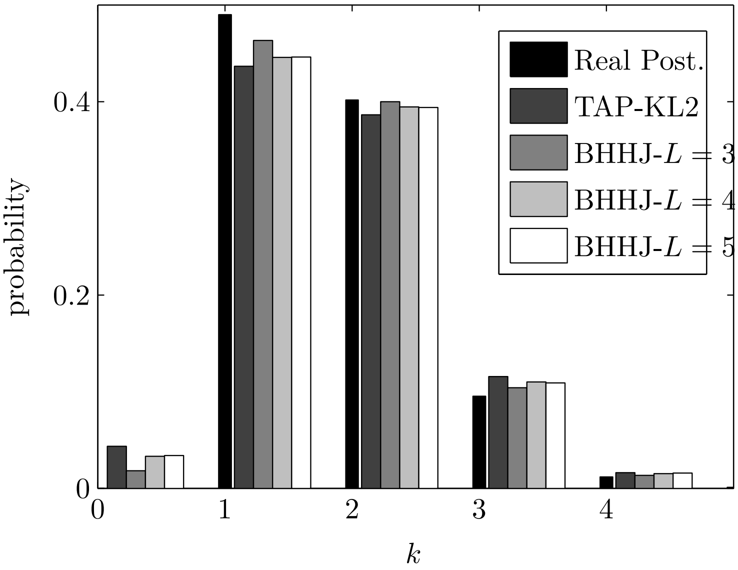


Figure 3.23 – *Posterior distribution of k, i.e., $p(k|\mathbf{y})$, along with its approximated versions for different scenarios for the second illustrative sinusoid detection example..*

these situations, the EM-type relabeling algorithms, such as the ones developed in Jasra et al. (2005) ; Sperrin et al. (2010) ; Papastamoulis and Iliopoulos (2010) ; Yao (2011) for the fixed-dimensional problems, cannot be used, as the summation over all possible permutation of the allocation vector $\mathbf{z}$ in the E-step is of cardinally $\frac{L!}{(L-k)!}$, assuming $\xi_{L+1} = 0$, which is computationally prohibitive; see discussions in Section 2.4.2 for more information.

For this purpose, we use here an experiment in which the observed signal $\mathbf{y}$ of length $N = 1024$ consists of $k = 30$ sinusoidal components, observed in white Gaussian noise with SNR $=$ 10dB. To locate the sinusoidal components, 10 blocks, each containing three components distributed according to the ones of the first experiment expressed in Table 3.1, situated 0.3 radian apart from each other, are considered. More precisely, the distance between the three sinusoidal components inside each block is π/N and they have the same amplitudes and phases as described in Table 3.1. Therefore, there are a total number of 10 hard-to-detect components, one in each block.

We generated 500 000 samples using the RJ-MCMC sampler described in Algorithm 3.1, initialized at the null model, and discarded the first 100 000 samples as the burn-in period. Next, we thinned the RJ-MCMC samples to one every 20^{th}. The posterior distribution of the number k of components is shown in Figure 3.24 (black bars). It can be seen from Figure 3.24 that k_{MAP} is 28. Thus, using the BMS approach, the model with $k = 28$ components will be selected. However, the resulting summary, not only contains components with "intra-block" label-switching but also a few ones with "inter-block" label-switching (see Figure 3.25). As a consequence, their estimated locations are meaningless and their estimated variances are large.

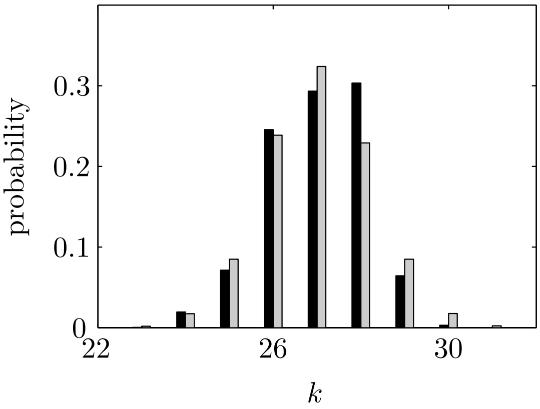
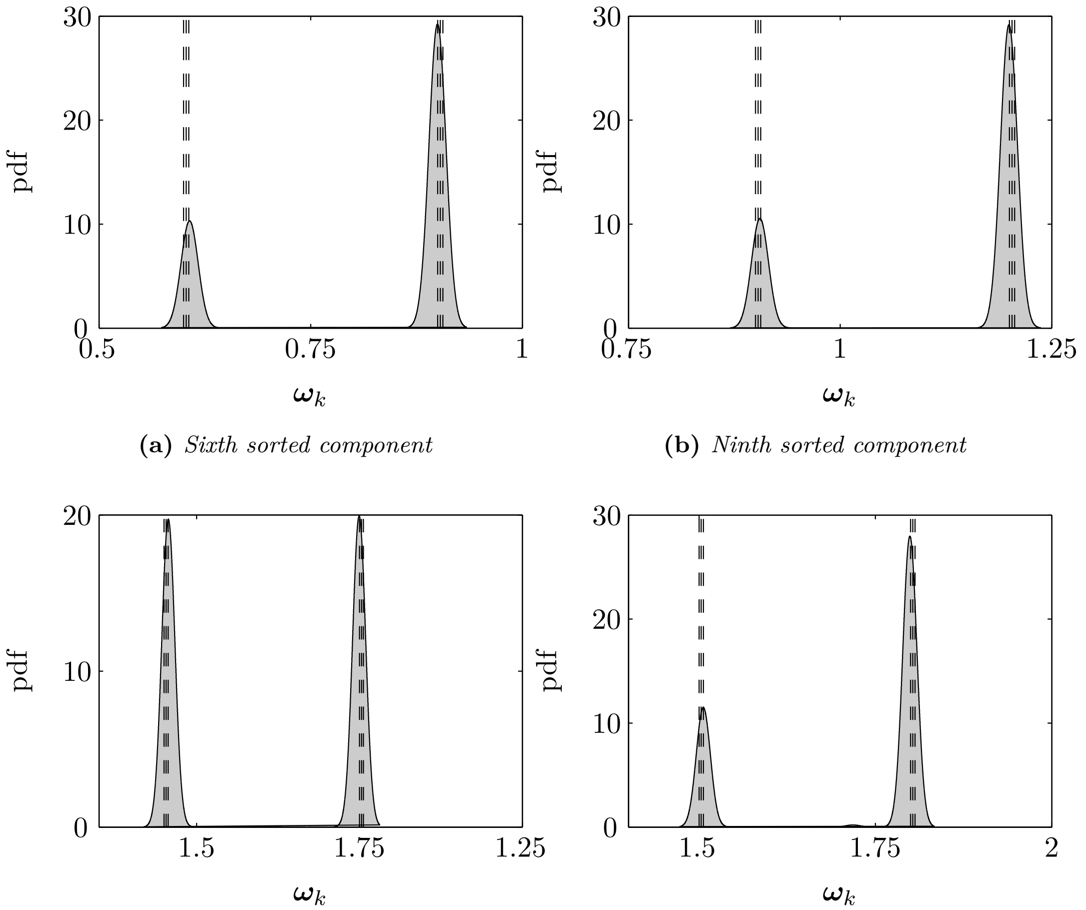


Figure 3.24 – *Posterior distribution of the number k of sinusoidal components, i.e., $p(k|\mathbf{y})$, (black) along with its approximated one obtained using 500 iterations of TAP-KL2 algorithm (gray) for the third illustrative sinusoid detection example. The mean of both posterior distributions equal 27.*

We ran 500 iterations of the TAP-KL2 algorithm with the number $L = 30$ of Gaussian components ($p(k \leq 30 \,|\, \mathbf{y}) \simeq 1$). To initialize, we used the robust estimators of the mean and variance of the posterior distributions of the sorted radial frequencies given $k = 30$. Moreover, we set $\pi_l = 0.5$, with $1 \leq l \leq L$, and $\lambda = 0.1$. Figure 3.26 illustrate the evolution of the model parameters together with the criterion $\mathfrak{J}$. It can be seen from the figure that, besides the means of the Gaussian components, other model parameters are evolving with SEM-iterations.

Remark 3.7. There is a remarkable move in the parameter space around the iteration 210 of the stochastic algorithm in which the variance of a component was substantially decreased. This also resulted in a spike in the evolution curve of the criterion $\mathfrak{J}$. This behavior is a result of not initializing appropriately the parameters of the model. To improve this issue, a new initialization procedure will be proposed in Section 4.3.

Figure 3.27 illustrates the obtained summaries for six out of ten blocks. Each panel corresponds to a block of three components. For each block, the histogram intensity of the RJ-MCMC samples obtained using the BMA approach along with the intensity of the fitted parametric model are show on the top, while the normalized pdf's of the fitted Gaussian components are shown in the bottom. It can be seen from the depicted summaries that the proposed approach provided a good fit to the true posterior distributions by solving both intra-block and inter-block label-switching issues (for comparison see the components with



(a) *Sixth sorted component*

(b) *Ninth sorted component*

(c) *Eleventh sorted component*

(d) *Fourteenth sorted component*

Figure 3.25 – *Posterior distributions of sorted radial frequencies given $k = k_{MAP} = 28$ (BMS) for some components with inter-block label-switching. The vertical dashed lines indicate the location of true radial frequencies.*

inter-block label-switching shown in Figure 3.25). Note that the probabilities of presence of the middle components in each block varies from $\hat{\pi}_{11} = 0.1$ (see Figure 3.27(c)) to $\hat{\pi}_{26} = 1$ (see Figure 3.27(f)).

The estimated model parameters, i.e., $\hat{\boldsymbol{\eta}}_l = \{\hat{\pi}_l, \hat{s}_l, \hat{\mu}_l\}$, $1 \leq l \leq L$, are presented in Figure 3.28. The left panel shows the estimated probabilities of presence versus the estimated means for the $L = 30$ Gaussian components in the model. The vertical dotted lines indicate the locations of each block of three components (so, there are three crosses around each vertical line). It can be seen that there are four Gaussian components with $\hat{\pi}_l < 0.5$, for $l \in \{5, 11, 14, 17\}$. The right panel shows the estimated probabilities of presence versus the estimated standard deviations. It reveals that the estimated variances are small.

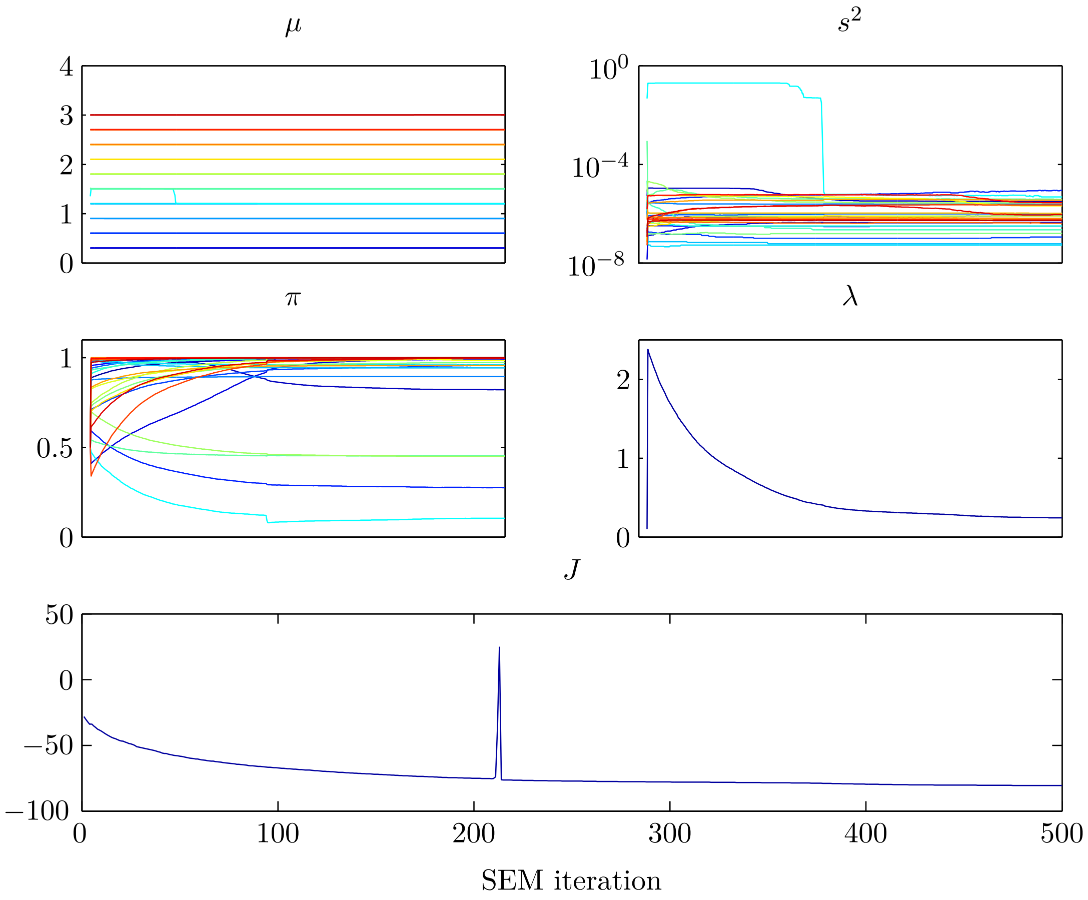
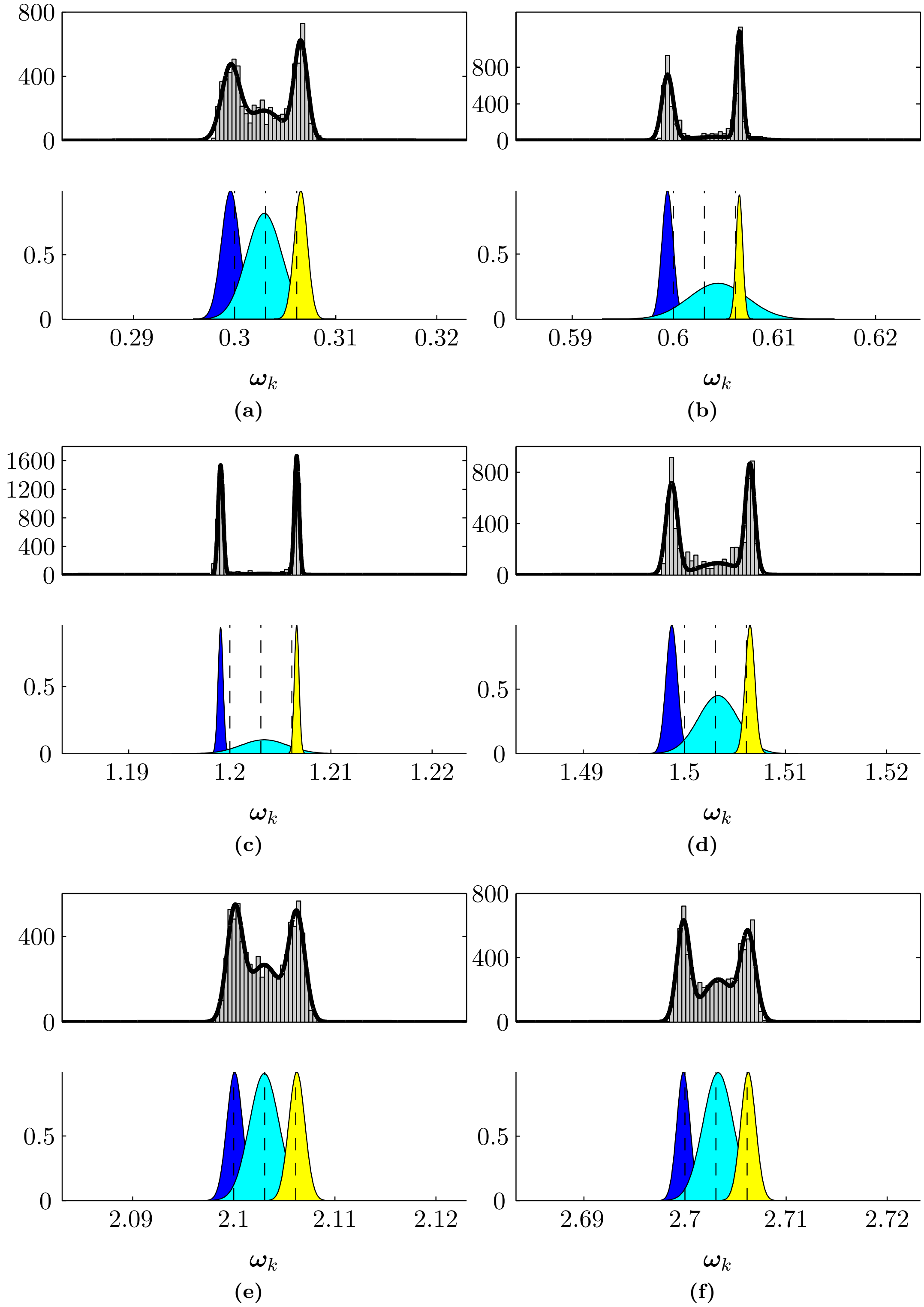


Figure 3.26 – *Evolution of the model parameters along with the criterion $\mathcal{J}$ using TAP-KL2 with $L = 30$ on the third illustrative sinusoid detection example.*

Indeed, the largest variance is $\hat{s}_5^2 = 8.24 \times 10^{-6}$.

One important point is that 500 iterations of TAP-KL2 implemented in MATLAB took 7862 seconds on a laptop with Intel Core i5 M540 running at 2.53 GHz and 4 GB of RAM. Given that the computation time of the SEM iterations are almost similar, each iteration of the algorithm took around 15.7 seconds which is justifiable considering the number $L = 30$ of components and number $M = 20\,000$ of observed samples.



(a) **(b)** **(c)** **(d)** **(e)** **(f)**

Figure 3.27 – *Results for six out of ten blocks; histogram intensity of the total observed samples along with the intensity of the fitted model, i.e., $q_{\hat{\boldsymbol{\eta}}}$, (top). Normalized pdf's of the estimated Gaussian components in each block (bottom). The vertical dashed lines indicate the locations of true radial frequencies.*

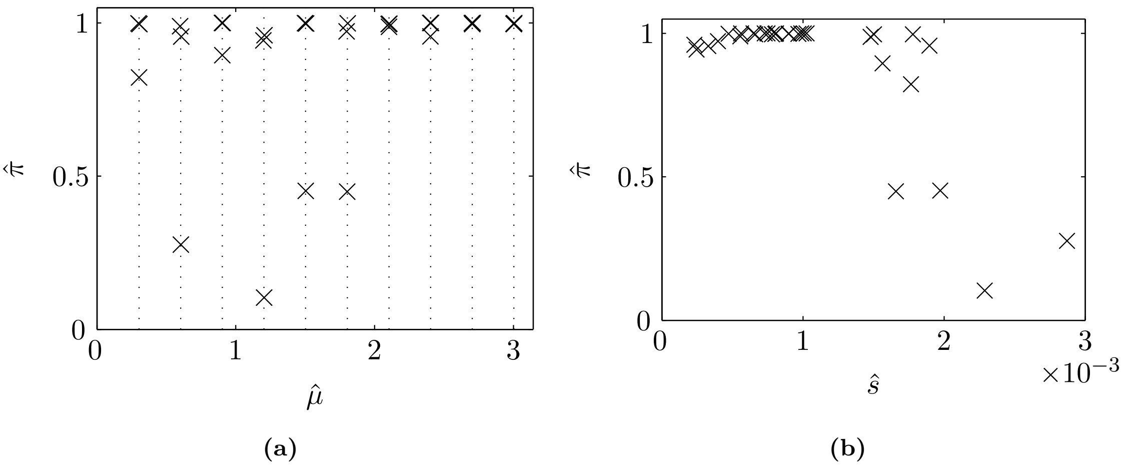


Figure 3.28 – *Estimated model parameters $\hat{\boldsymbol{\eta}}_l = \{\hat{\pi}_l, \hat{s}_l, \hat{\mu}_l\}$, $1 \leq l \leq L$, using 500 iterations of TAP-KL2 with $L = 30$ on the third illustrative example. (a) Estimated probabilities of presence versus estimated means. The vertical dotted lines indicate the locations of of each block of three components. (b) Estimated probabilities of presence versus estimated standard deviations.*

3.4 Summarizing variable-dimensional posteriors: average performance

3.4.1 Objectives

In the illustrative results shown in the previous section, we showed results confirming the capability of the proposed approach to summarize variable-dimensional posterior distributions encountered in the problem of joint detection and estimation of sinusoids in white Gaussian noise. In this section, we investigate using Monte Carlo simulations the frequentist properties of the proposed approach for summarizing variable-dimensional posterior distributions. The main question that we address in this section is: *can the summary provided by the proposed approach be considered as a faithful substitute for the "large" set of RJ-MCMC samples?* To answer this question, the following points need to be verified:

i) **Goodness-of-fit of the approximate posterior (Section 3.4.2):** Studying how faithfully the approximate posterior distribution represents the true posterior distribution. To do so, we will look at various features of the posterior distributions (posterior probabilities, reconstruction errors, . . .).

ii) **Comparison of the estimated vector of radial frequencies by the proposed approach with the ones obtained using the BMS approach (Section 3.4.3):** Considering the first sinusoid detection experiment defined in Table 3.1 with a middle hard-to-detect, we study the frequentist properties of the estimated vector of radial frequencies using both the proposed and the BMS approaches.

To begin, in what follows, we describe the experimental setup, the estimators of vector of radial frequencies, and the procedures to reconstruct the noiseless signal.

Experimental setup

We will consider the first sinusoid detection experiment defined in Table 3.1 with SNR = 5, 7, 10, 12 dB. Remember that the goal of this experiment is to detect the middle hard-to-detect sinusoidal component. To study the performance of the approaches when the middle component does not exist, we simply remove the middle component. We name these experiments H_1 and H_0, respectively. We ran the RJ-MCMC sampler explained in Algorithm 3.1 on 100 realizations of both experiments. 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. The parameters of the hierarchical model shown in Figure 3.1 were set as in the previous section.

Turning to the summarizing algorithms, first, we need to initialize the parametric model $q_{\boldsymbol{\eta}}$ in a systematic fashion. It is natural to deduce the number L of Gaussian components from the posterior distribution of k. In this section, 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 results in "richer" estimated parametric models in the sense that $L \geq k_{MAP}$, where $k_{MAP} = \underset{k}{\operatorname{argmax}}\, p(k|\mathbf{y})$—the selected model using the BMS approach. (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.) To initialize the Gaussian components' parameters, i.e., the mean μ and the variance s^2, we used the robust estimates of the posterior of the sorted radial frequencies given $k = L$.

Point estimates of the vector of radial frequencies

To estimate the vector of radial frequencies from the fitted parametric model $q_{\hat{\boldsymbol{\eta}}}$, in addition to Gaussian components' means $\hat{\mu}_l$, we have to take into account their probabilities of presence $\hat{\pi}_l$, with $1 \leq l \leq L$. We propose an estimator consisting in discarding the Gaussian components with the probabilities of presence smaller than a certain threshold, denoted by t_π, with $0 \leq t_\pi \leq 1$. Then, the means of the remaining components are used as the estimated frequency vectors.

This estimator can be seen as a post-processing procedure in which one can have a range of possible summaries by changing the value of the threshold t_π. For two extreme values of $t_\pi = 0$ and $t_\pi = 1$, there are, respectively, $L(0) = L$ and $L(1) = 0$ components in the parametric model, where $L(t_\pi)$ denotes the number components kept in the model after discarding the ones smaller than the threshold t_π. Furthermore, setting the threshold t_π to the special value of 0.5, the selection procedure becomes similar to the Median Probability Model approach introduced in Barbieri and Berger (2004) for Bayesian variable selection problems; see Section 2.2.2 for more discussions.

When using the BMS approach, a model with the highest posterior probability, i.e., k_{MAP}, is selected and, then, the estimated vector of radial frequencies is set to the median of the posterior distributions of the sorted radial frequencies given $k = k_{MAP}$.

Reconstructing the noiseless signal

To compare the performance of the approaches in reconstructing the noiseless signal $\mathbf{y}_0$, we first need to explain the reconstruction procedure for different approaches. To reconstruct the noiseless signal $\mathbf{y}_0$ from the summary provided by the proposed summarizing approach, we do as follows; for $r = 1, \dots, R$,

i) generate independent Bernoulli indicator variables $\boldsymbol{\xi}^{(r)} = (\xi_1^{(r)}, \dots, \xi_L^{(r)})$ given the estimated probabilities of presence $\hat{\boldsymbol{\pi}}$.

ii) set the number of components $\hat{k}^{(r)} = \sum_{l=1}^{L} \xi_l^{(r)}$.

iii) generate random variables denoted by $\hat{\omega}_{j,\hat{k}^{(r)}}^{(r)}$, with $j = 1, \dots, \hat{k}^{(r)}$, from the estimated Gaussian components that are present, i.e., their corresponding indicator variable $\xi_l^{(r)} = 1$. Then, arrange them in a vector $\hat{\boldsymbol{\omega}}_{\hat{k}^{(r)}}^{(r)} = (\hat{\omega}_{1,\hat{k}^{(r)}}^{(r)}, \dots, \hat{\omega}_{\hat{k}^{(r)},\hat{k}^{(r)}}^{(r)})$.

iv) estimate the corresponding amplitudes from their posterior mean, that is

$$\hat{\boldsymbol{a}}_{\hat{k}^{(r)}}^{(r)} = \frac{\hat{\delta}^2}{1+\hat{\delta}^2} \cdot ((\hat{\mathbf{D}}^{(r)})^t \hat{\mathbf{D}}^{(r)})^{-1} (\hat{\mathbf{D}}^{(r)})^t \mathbf{y},$$

where $\hat{\mathbf{D}}^{(r)}$ is the design matrix of the vector $\hat{\boldsymbol{\omega}}_{\hat{k}^{(r)}}^{(r)}$ as expressed in (3.1) and $\hat{\delta}^2$ is the estimated value of the amplitudes hyperparameter obtained from the median of the samples generated from $p(\delta^2|\mathbf{y})$.

Then, the reconstructed signal using the parameters of the fitted parametric model is

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

For comparison, we use both the BMS and BMA approaches. In the BMS approach, first, the vector of radial frequencies is estimated as explained before. Then, the amplitudes are estimated using their posterior mean. Finally, the reconstructed signal using the BMS approach is

$$\hat{\mathbf{y}}_0^{BMS} = \hat{\mathbf{D}}^{BMS}\, \hat{\boldsymbol{a}}_k^{BMS}.$$

In the BMA approach, we have

$$\hat{\mathbf{y}}_0^{BMA} = \mathbb{E}(\mathbf{y}_0 \,|\, \mathbf{y}) = \sum_{k=1}^{k_{\max}} \mathbb{E}(\mathbf{y}_0 \,|\, k, \mathbf{y}) \cdot p(k \,|\, \mathbf{y}) = \frac{1}{M} \sum_{i=1}^{M} \mathbf{D}^{(i)} \hat{\boldsymbol{a}}_k^{(i)}, \tag{3.9}$$

where $\mathbf{D}^{(i)}$ is the design matrix of the i^{th} vector of the sampled radial frequencies $\boldsymbol{\omega}_k^{(i)}$ given in (3.1) and $\hat{\boldsymbol{a}}_k^{(i)}$ is the posterior mean of the amplitudes given $\boldsymbol{\omega}_k^{(i)}$ and $\delta^{2(i)}$.

3.4.2 Verification of the goodness-of-fit of the approximate posterior distribution

In this section, we investigate how faithfully the approximate posterior distribution preserves the information of the true posterior distribution. In Section 3.3, such property has been studied on a few examples by comparing the posterior of k with its approximated version (see, e.g., Figure 3.15), and the histogram intensity of the observed samples with the intensity of the fitted parametric model (see, e.g., Figure 3.14).

Here, we compare the two posterior distributions using indicators that are explained in follows. Figures 3.29 and 3.30 show the comparisons of the fitted approximate posterior distribution $q_{\hat{\eta}}$ obtained using 100 iterations of TAP-BHHJ with $\alpha = 0.5$ with the true variable-dimensional posterior distribution for the first sinusoid detection experiment with SNR $=$ 5 and 7dB, respectively. The results of the other algorithms and other values of SNR were similar and thus are not shown here.

The scatter plots shown in panels (a), (b), and (c) of both figures compare the posterior distribution of the number k of components, i.e., $p(k|\mathbf{y})$, with its approximated version, i.e., $\hat{p}(k|\mathbf{y})$, in 100 runs. We only show the probabilities of $k = 2$ and $k = 3$ in this comparison as the other probabilities were 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 $k_{TAP} = \underset{k}{\operatorname{argmax}}\ \hat{p}(k|\mathbf{y})$. It can be seen from these three panels that the information in $p(k|\mathbf{y})$ was well preserved by the approximated posterior distributions.

Panel (d) of the figures compares the normalized reconstruction errors using TAP-BHHJ with the ones of the BMA approach in dB, defined as

$$10\ \log_{10}\left(\frac{\|\hat{\mathbf{y}}_0 - \mathbf{y}_0\|^2}{\|\mathbf{y}_0\|^2}\right), \tag{3.10}$$

where $\|\cdot\|$ is the L$_2$-norm and we set $\hat{\mathbf{y}}_0 = \hat{\mathbf{y}}_0^{BMA}$ and $\hat{\mathbf{y}}_0 = \hat{\mathbf{y}}_0^{TAP}$, when using the BMA approach and the proposed approach, respectively. It can be seen from the figures that the error of the reconstructed noiseless signals using the compact summary obtained by the proposed approach are quite comparable with the ones obtained using the BMA approach. Moreover, the normalized errors in both approaches are concentrated between -10 and -20 dB indicating a good reconstruction performance considering the values of SNR.

Finally, the scatter plots in the last two panels compare the expected number of components, i.e., $\mathbb{E}(N(\cdot))$, in the intervals $(0, \pi/4)$ and $(\pi/4, \pi/2)$ using the proposed approach with the ones obtained using the BMA approach as expressed in (2.3). For the proposed approach, the expected number of components in the interval Δ_j, for $j = 1, \ldots, J$, is given by

$$\mathbb{E}_{\hat{\boldsymbol{\eta}}}\left(N(\Delta_j)\right) = \sum_{l=1}^{L} \hat{\pi}_l(\mathrm{CDF}_{\hat{\boldsymbol{\eta}}_l}(\omega_{j+1}) - \mathrm{CDF}_{\hat{\boldsymbol{\eta}}_l}(\omega_j)) + \frac{\hat{\lambda}}{|\boldsymbol{\Theta}|}|\Delta_j|, \tag{3.11}$$

where $\mathrm{CDF}_{\hat{\boldsymbol{\eta}}_l}(\omega)$ is the CDF of the Gaussian component l with the estimated parameters $\hat{\boldsymbol{\eta}}_l$ at the point ω. 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 the proposed summarizing approach preserves faithfully the information lied in true posterior distribution. Moreover, the proposed approach has similar reconstruction performance to the BMA and BMS approaches.

3.4.3 Frequentist comparison of the estimated vector of radial frequencies

In this section, we address the frequentist performance of the proposed summarizing approach in detecting the middle hard-to-detect sinusoidal component. For comparison, we use the BMS approach.

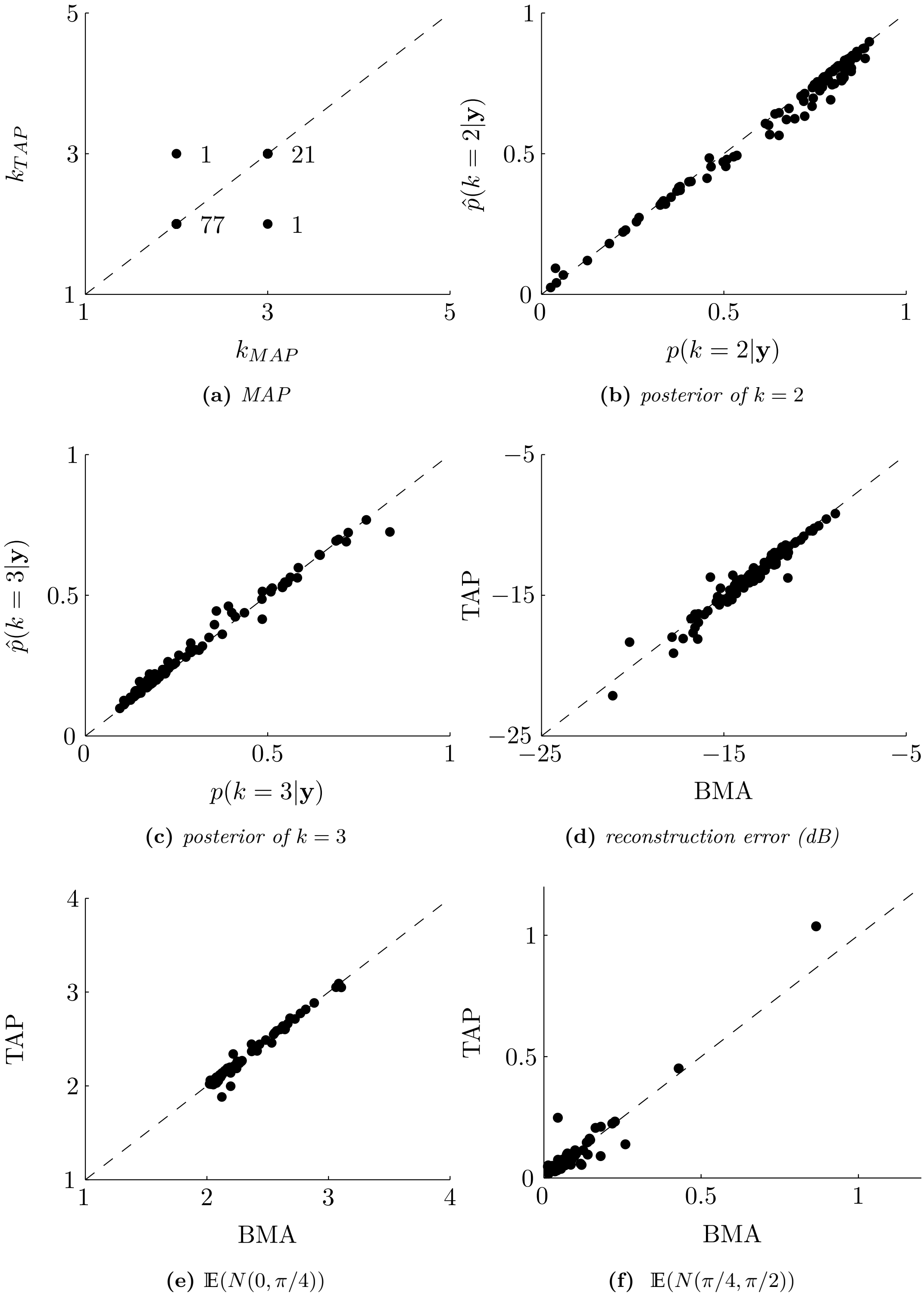
In fact, there is no natural distance between the estimated vectors of radial frequencies of possibly different dimensions obtained using the approaches aforementioned in a systematic way. In the following, we define such a distance and study the frequentist properties of the proposed approach by measuring two possible kinds of detection errors one can make; namely, we count the number of detected False Positives (FP's) and the number of omissions, denoted by N_{FP} and N_{O}, respectively.

Let us partition the parameter space $\boldsymbol{\Theta}$ into J subsets. Then, we define

$$\begin{aligned} N_{\mathrm{FP}} &\triangleq \sum_{j=1}^{J} \max\{0,\, \hat{N}^j - N^j_{\mathrm{True}}\}, \\ N_{\mathrm{O}} &\triangleq \sum_{j=1}^{J} \max\{0,\, N^j_{\mathrm{True}} - \hat{N}^j\}, \end{aligned} \tag{3.12}$$

where $\hat{N}^j$ and N^j_{True} are the estimated and the true number of components in the partition j, for $j = 1, \ldots, J$. Recall that when estimating the vector of radial frequencies from the fitted parametric model $q_{\hat{\boldsymbol{\eta}}}$, the estimated number of components depends on the value of the threshold t_π on the probabilities of presence.

When the number of partitions J is set to one, we have a general view of the parameter space, though, we can use the characteristics of the experiment under study to introduce more partitions and, consequently, present refined results. For example, in this experiment,



(a) *MAP*

(b) *posterior of* $k = 2$

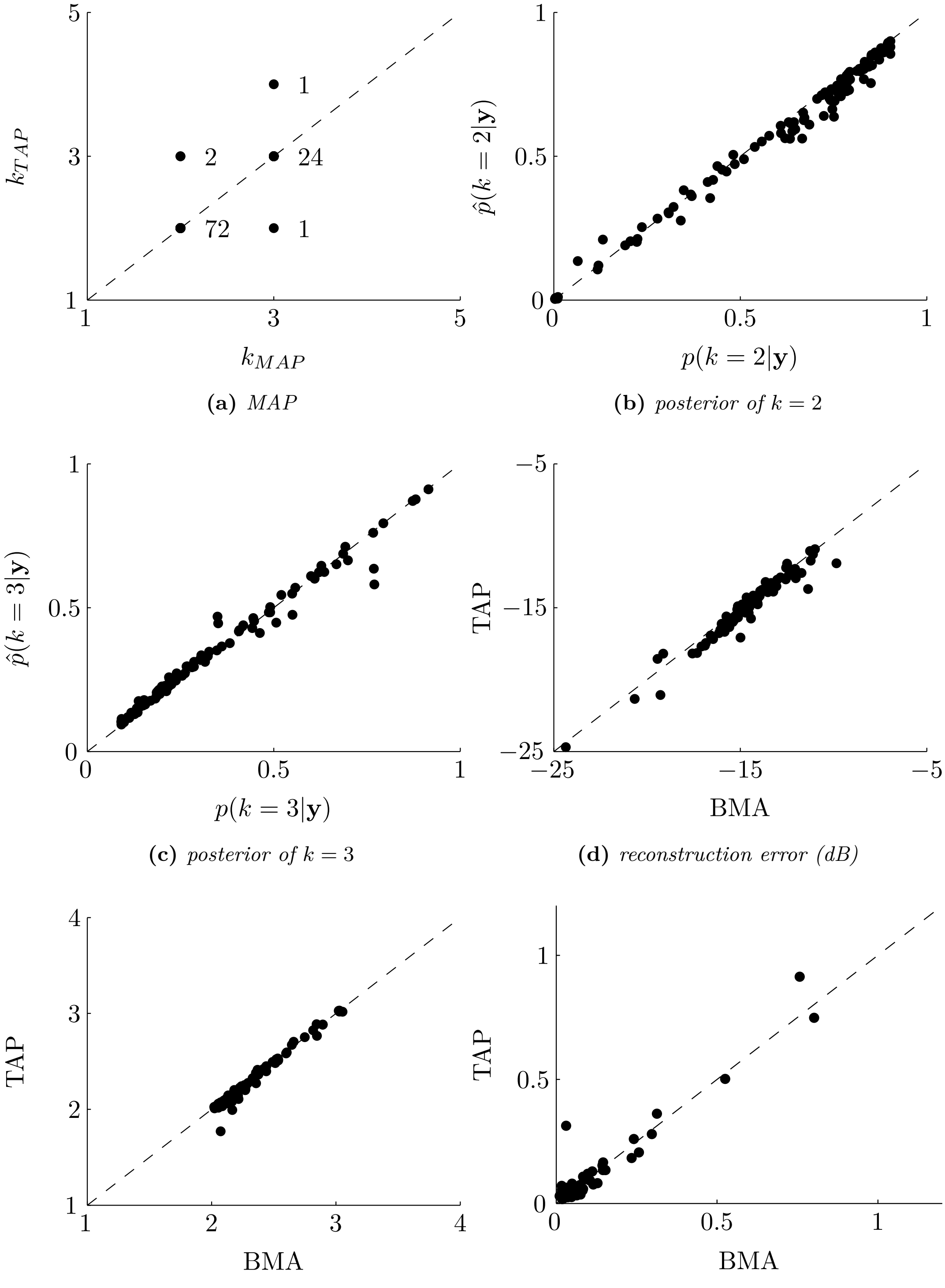
(c) *posterior of* $k = 3$

(d) *reconstruction error (dB)*

(e) $\mathbb{E}(N(0, \pi/4))$

(f) $\mathbb{E}(N(\pi/4, \pi/2))$

Figure 3.29 – *Comparison of the true posterior distribution with its approximated version when using TAP-BHHJ with* $\alpha = 0.5$ *on the experiment with SNR* $= 5dB$.



(a) *MAP*

(b) *posterior of* $k = 2$

(c) *posterior of* $k = 3$

(d) *reconstruction error (dB)*

(e) $\mathbb{E}(N(0, \pi/4))$

(f) $\mathbb{E}(N(\pi/4, \pi/2))$

Figure 3.30 – *Comparison of the true posterior distribution with its approximated version when using TAP-BHHJ with* $\alpha = 0.5$ *on the experiment with SNR* $= 7dB$.

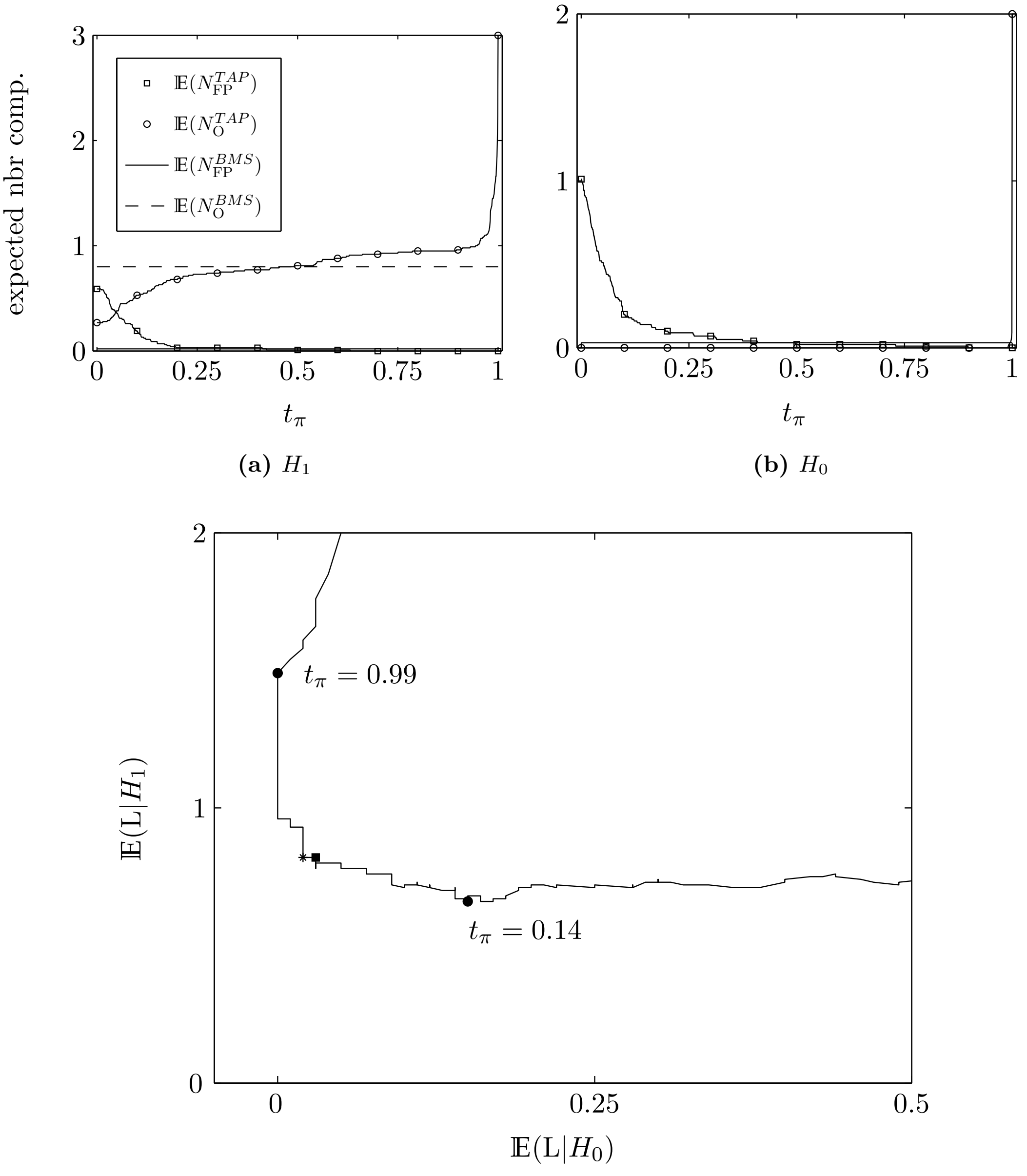
we define a Region Of Interest (ROI) $\subset \boldsymbol{\Theta}$, out of which any detected component will be considered as a FP. We set the ROI = (0.55, 0.8) and count the number of Gaussian components with the estimated means inside and outside this region when the summarizing approach is used. We denote the regions inside and outside ROI by $j = 1$ and $j = 2$, respectively. Hence, under H_1, we have $N^1_{\text{True}} = 3$. On the other hand, under H_0, there are two components in the ROI, i.e., $N^1_{\text{True}} = 2$ and zero outside.

Panels (a) and (b) of Figures 3.31–3.32 present the expectations of the number of detected FP's, i.e, N_{FP}, and the number of omissions, i.e., N_{O}, when using TAP-BHHJ with $\alpha = 0.5$ versus the threshold t_π in 100 runs for both experiments H_1 and H_0 with SNR = 5 and 7 dB. Moreover, the expectations of N_{FP} and N_{O} using the BMS approach are shown by the horizontal lines. The results of the other summarizing algorithms and for the other values of SNR were qualitatively similar and, thus, are not shown here.

Since both numbers N_{FP} and N_{O} measure the errors committed by the estimators, we want both of them to be close to zero. For the proposed summarizing approach, the expectations of both N_{FP} and N_{O} will change by altering the threshold of probabilities of presence t_π, due to the fact that some Gaussian components are removed. It can be seen from the figures that when $t_\pi = 0$, $\mathbb{E}(N^{TAP}_{\text{FP}})$ has its maximum value. As we increase the value of the threshold t_π, the value of $\mathbb{E}(N^{TAP}_{\text{FP}})$ decreases until a point around $t_\pi = 0.5$ that it becomes equal to $\mathbb{E}(N^{BMS}_{\text{FP}})$. On the other hand, $\mathbb{E}(N^{BMS}_{\text{FP}})$ has its minimum at $t_\pi = 0$ and it increases by increasing the threshold t_π. Again, at a certain point around $t_\pi = 0.5$, it coincides with the corresponding line of the BMS approach. Note that when $t_\pi = 1$, the number of Gaussian components $L(1)$ becomes zero. This is why, in the figures, when $t_\pi = 1$, we have $\mathbb{E}(N_{\text{O}}) = N_{\text{True}}$ and $\mathbb{E}(N_{\text{FP}}) = 0$. It other words, when $t_\pi = 1$, we lose even the two easy-to-detect sinusoidal components.

A remarkable point concerning the results illustrated in Figures 3.31 and 3.32 is that, the results of the proposed approach when $t_\pi = 0.5$ are comparable with the ones of the BMS approach. Thus, we conjecture that deriving an estimator in the spirit of the median probability model of Barbieri and Berger (2004), that is, discarding the Gaussian components with the probabilities of presence $\pi_l < 0.5$, the proposed summarizing approach would perform almost as well as the BMS approach in detecting the middle hard-to-detect component of this specific experiment. Note also that in the Bayesian variable selection problem, the MPM and BMS approaches coincide in some cases (see Barbieri and Berger, 2004, for more discussion).

In order to derive an effective comparison of the performance of the approaches, we

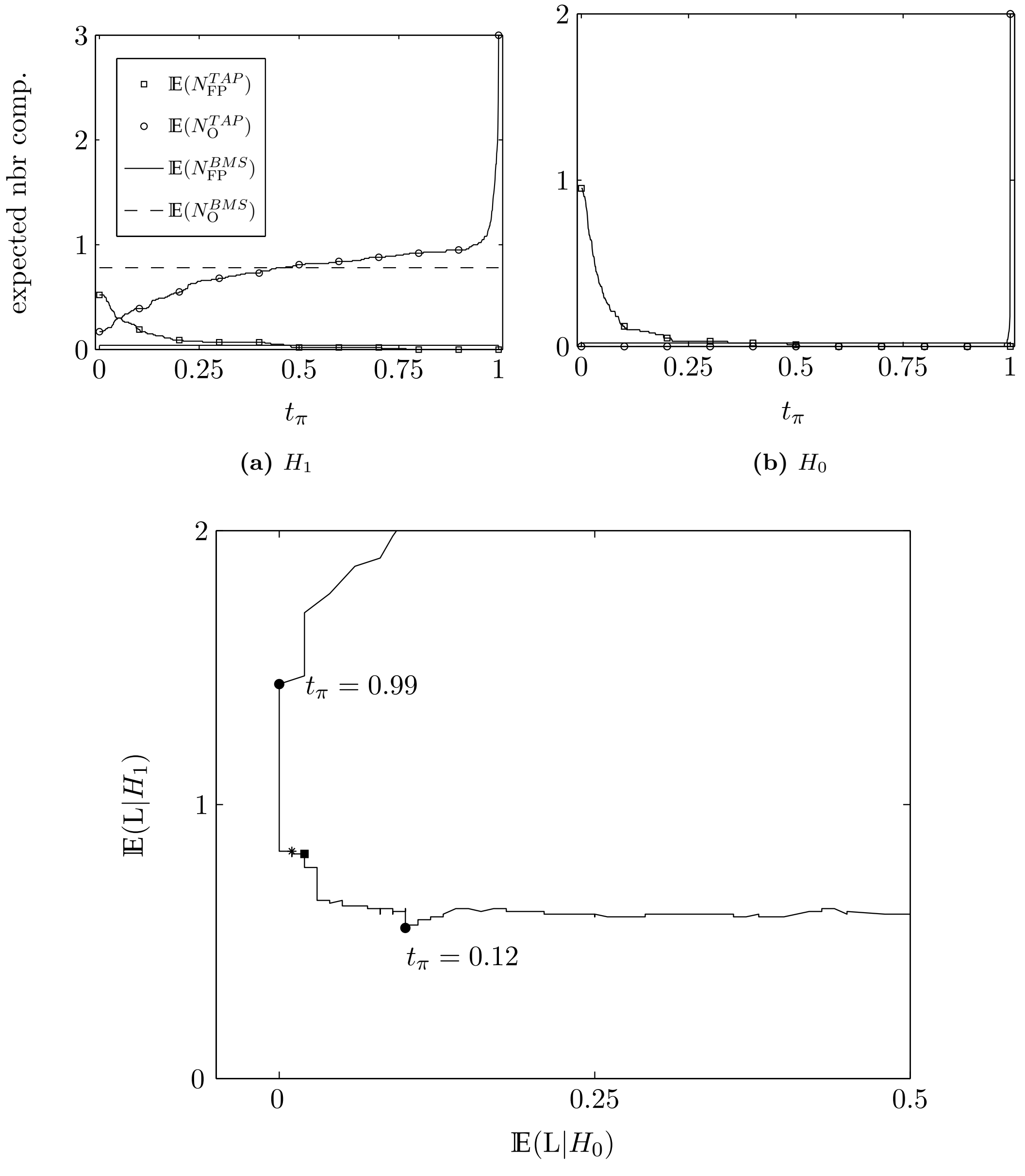


(a) H_1

(b) H_0

(c)

Figure 3.31 – *Comparison of the frequentist performances of TAP-BHHJ with* $\alpha = 0.5$ *and the BMS approach when SNR* $= 5dB$. *(a) and (b) show the expected number of detected FP's and omissions under* H_1 *and* H_0, *respectively. (c) illustrates the frequentist risks of the estimators derived from the compact summary under* H_1 *versus* H_0 *for different values of* t_π. *The filled square and asterisk, respectively, correspond to the risks of the BMS approach and TAP-BHHJ when* $t_\pi = 0.5$.



(a) H_1

(b) H_0

(c)

Figure 3.32 – *Comparison of the frequentist performances of TAP-BHHJ with* $\alpha = 0.5$ *and the BMS approach when SNR* $= 7dB$. *(a) and (b) show the expected number of detected FP's and omissions under* H_1 *and* H_0, *respectively. (c) illustrates the frequentist risks of the estimators derived from the compact summary under* H_1 *versus* H_0 *for different values of* t_π. *The filled square and asterisk, respectively, correspond to the risk of the BMS approach and TAP-BHHJ when* $t_\pi = 0.5$.

define a loss function in the spirit of decision theory literature (see, e.g., Robert, 2007, Chapter 2, for more discussion). A possible loss function can be defined based on the two error measures N_{FP} and N_{O} as

$$\mathrm{L} \triangleq N_{\mathrm{FP}} + N_{\mathrm{O}}. \tag{3.13}$$

Note that the proposed loss function penalizes both type of errors equally. Panel (c) of Figures 3.31 and 3.32 illustrate the *frequentist risk* (or average loss) under H_1, i.e., $\mathbb{E}(\mathrm{L} \mid H_1)$, versus the one under H_0, i.e., $\mathbb{E}(\mathrm{L} \mid H_0)$, when using TAP-BHHJ with $\alpha = 0.5$ for different values of the threshold t_π. Moreover, the corresponding risk of the BMS approach is shown by a filled square. It can be seen from the figures that by changing the threshold in the range $[0.12, 0.99]$ (or $[0.14, 0.99]$ in the case of $SNR = 7$dB), indicated by the filled circles, a family of *admissible* estimators, with respect to the class containing the BMS estimator and the whole family for $t_\pi \in [0, 1]$, is obtained that contains the solution attained using the BMS approach. Observe also that moving on the curve for $t_\pi > 0.99$, both risks increase, and for $t_\pi < 0.12$ (or $t_\pi < 0.14$ in Figure 3.32), $\mathbb{E}(\mathrm{L} \mid H_0)$ increases.

3.5 Summary and discussion

In this chapter, we addressed the important signal decomposition problem of joint detection and estimation of sinusoidal components in white Gaussian noise. We used the hierarchical model and the RJ-MCMC sampler developed by Andrieu and Doucet (1999), with a modification in the expression of the birth-or-death acceptance ratio (3.6), to generate samples from the target posterior distribution (3.3). We also discussed the issues concerning prior specification for the hyperparameter δ^2 (or, the g parameter in Zellner's g-prior). Moreover, assuming assigning a weakly-informative conjugate $\mathcal{IG}(\alpha_{\delta^2} = 2,\, \beta_{\delta^2})$ prior distribution over δ^2, as in Andrieu and Doucet (1999) we studied the sensitivity of the posterior distribution to the variations of the scale parameter β_{δ^2} using the SMC sampler developed by Bornn et al. (2010).

In the rest of the chapter, we used the problem of joint detection and estimation of sinusoids in white Gaussian noise as an example to investigate the capability of the summarizing algorithms we proposed in Chapter 2 for summarizing variable-dimensional posterior distributions. To this end, we presented two kinds of results, namely, the illustrative and average results.

3.5.1 Discussion of the illustrative results

Several properties of the proposed summarizing approach have been studied in Section 3.3 using three specific sinusoid detection examples. In the first example, we saw how using the proposed approach resulted in extracting useful information concerning the middle hard-to-detect sinusoidal component, while it was completely lost in the summary of the BMS approach.

The second illustrative example was chosen to highlight a situation where the proposed algorithms have difficulties. More precisely, posterior distributions in which the location of the component-specific parameters are not coherent across models (see Figure 3.17 for an example of such a posterior) can cause problem for the proposed summarizing approach. The results showed that the components' estimated probabilities of presence are not in accordance with their uncertainties in the true posterior distribution. But, their location and dispersion parameters were well estimated. Moreover, we provided diagnoses to detect this kind of situation; for example, existence of significant correlation between the presence of fitted components indicates the lack-of-fit of the fitted model. As a future work, one way to improve the performance of the proposed algorithms in this situation is to introduce correlation between the presence of components in the parametric model.

The third illustrative example investigated the usefulness of the proposed SEM-type algorithms in dealing with challenging situations where the number k of components is large, say, $k > 10$. In these situations, the EM-type algorithms cannot be used as the E-step is computationally prohibitive. For this purpose, in Section 3.3.4, we designed an experiment in which there were $k = 30$ sinusoids. From the posterior distributions of the sorted radial frequencies given $k = k_{MAP} = 28$, we saw that there are some components exhibiting severe label-switching. The obtained summary using the proposed approach confirmed that it is capable of not only dealing with challenging situations where k is large but also removing severe label-switching issues.

We particularly studied the following properties:

Convergence of the algorithms: To study the convergence of the proposed algorithms, we showed the evolution of the model parameters versus the SEM iterations. It was shown that both the KL and BHHJ criteria have "generally" decreasing behavior and the evolution of the models' estimated parameters become almost constant after a reasonable number of iterations.

Relabeling properties: In all the three example, comparing the distributions of the samples allocated to the fitted Gaussian components with the ones of the sorted radial frequencies given k, which are highly multimodal, revealed the capability of the proposed algorithms to provide the first solution, in the literature, to the "trans-dimensional label switching" problem.

Validation of the fitted model: Comparing the intensities of the fitted approximate models with the histogram intensities of the RJ-MCMC output samples confirmed the goodness-of-fit of the fitted models. Moreover, we saw that the information in the true posterior distribution of k was well preserved by its approximated versions.

Effect of the number L of Gaussian components: In the second example, we also showed the effect of the number L of Gaussian components on the obtained summaries. In fact, in general, to choose an appropriate value for L, in addition to the posterior of the number k of components, one should inspect the posterior distributions of the sorted component-specific parameters. Moreover, significant peaks in the distribution of the samples allocated to the point process component, i.e., the residuals of the fitted model, is another indication that the chosen value of L was small. We will discuss this issue in more detail in the next chapter.

3.5.2 Discussion of the average results

In the second kind of results presented in Section 3.4, we studied the average performance of the proposed approach. The results presented in Section 3.4.2 confirmed that the approximate posterior distribution $q_{\hat{\eta}}$ obtained using the proposed approach faithfully preserves the information of the true variable-dimensional posterior distribution. As an example, the obtained summaries have comparable signal reconstruction performance with respect to the BMA approach.

Section 3.4.3 was devoted to compare the frequentist performance of the proposed summarizing approach with the one of the BMS approach in 100 runs. To estimate the vector of radial frequencies from the fitted parametric model, we introduced the threshold $0 \leq t_\pi \leq 1$ on the probabilities of presence. Then, the Gaussian components with $\hat{\pi}_l \geq t_\pi$ were kept in the model and their estimated means were used as frequency estimates. The presented results showed that for a wide range of t_π, a family of admissible estimators is attained from the obtained summary containing the one of the BMS approach. Most notably, setting t_π to 0.5, provided estimators with comparable frequentist performance as the ones of the BMS approach.

Therefore, from the presented results, it can be inferred that the information provided by the approximate posterior distribution, esp. the probabilities of presence, is meaningful and can be used to construct estimates with good frequentist properties.

3.5.3 Choice of summarizing algorithms

Concerning the choice of summarizing algorithms, we do not recommend the use of the TAP-KL1 algorithm, as the summaries obtained using it, often, contained large variance components with corresponding distribution of labeled samples being multimodal. While on the contrary, both TAP-KL2 and TAP-BHHJ algorithms are capable of providing desirable summaries in which the effect of label-switching has been completely removed. An advantage of TAP-KL2 is that, in contrary to TAP-BHHJ, it does not depend on a tuning parameter. Moreover, it is less computationally involved and easier to implement in comparison with TAP-BHHJ.

On the other hand, TAP-BHHJ often converges faster than TAP-BHHJ. Moreover, the fact that it is a criterion driven algorithm (note that TAP-KL2 does not minimize directly the KL criterion) make it possible for further analysis of the convergence of the SEM-type summarizing algorithm.

In the next chapter, we will apply the proposed summarizing algorithms to another variable-dimensional problem encountered when analyzing cosmic rays signal in the Auger project.

Chapter 4

Bayesian Detection and Estimation of Astrophysical Particles in the Auger Project

4.1 Introduction

Auger project (see, e.g., Auger Collaboration, 1997, 2004) is an international project involving collaborators from more than 19 countries and 46 institutions. This 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 the Auger project is to answer the following questions:

1. What is the nature of the ultra-high energetic particles (Proton, Iron, etc.)?
2. Where is their origin in the universe?

When these cosmic ray particles collide the earth's atmosphere, air showers covering a vast surface are generated. However, such rays are quite rare, hitting an area of the size of a football field once every 10 000 years. Hence, an enormous observatory is needed to detect these ultra-high energy particles. To detect them, the Pierre Auger Cosmic Ray Observatory was built in Argentina. The observatory consists of two independent detectors; an array of Surface Detectors (SD) and a number of Fluorescence Detectors (FD). Each SD is a tank filled with around 11 000 liters of pure water and sit about 1.5 km away from the next tank. This array covers a surface of about 3000 km^2 which is about ten times the size of Paris. The second detection system, i.e., the FD's, consists of 24 fluorescence telescopes located on hills that on dark nights capture a faint light or fluorescence caused by the shower particles colliding with the atmosphere. Figure 4.1 illustrates a conceptual view of both detectors and an incoming shower.

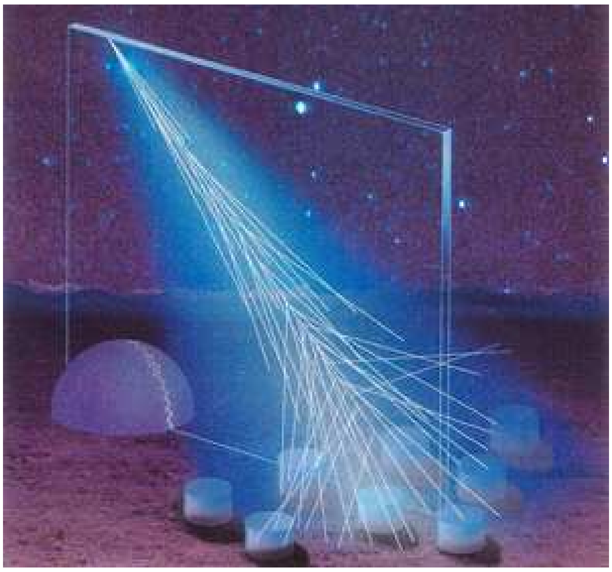


Figure 4.1 – *The conceptual picture of a fluorescence detector and surface detectors (tanks) with an incoming cosmic ray shower (source: http://www.auger.org/).*

In this work, we concentrate on analyzing the SD signal captured in the surface of the earth. The goal of this study is to count the number of generated muons, that is, the particles produced from the collision of cosmic rays and the earth's atmosphere, and estimate their individual parameters. In fact, determining the number of muons and their arrival times can be used as indications of the chemical composition of the original particles; for example, iron showers generate, in general, about 40% more muons than proton ones. Moreover, the proton showers are usually deeper which can be identified by the estimated arrival times.

A Bayesian algorithm for the trans-dimensional problem of joint detection and estimation of muons has been developed in Kégl (2008) ; Kégl and Veberic (2009) ; Bardenet et al. (2010, 2012). In this chapter, we address the problem of summarizing the variable-dimensional posterior distribution occurred in this problem. To show results, we use the data provided by Prof. Balázs Kégl from the Laboratoire de l'Accélérateur Linéaire (LAL), Université Paris Sud 11, containing several observed signals along with the samples generated from the corresponding variable-dimensional posterior distributions.

This chapter is outlined as follows. Section 4.2 is devoted to describing the hierarchical model and the RJ-MCMC sampler developed for this problem in Kégl (2008) ; Kégl and Veberic (2009) ; Bardenet et al. (2010, 2012). It turns out that the "naive" initialization procedure for the summarizing algorithms used in Chapter 3 causes convergence issues in some experiments. Section 4.3 shows the convergence issues caused by the "naive" initialization using an illustrative example and, then, describes a new initialization procedure for the proposed summarizing algorithms. Section 4.4 presents the results of the approach we proposed in Chapter 2 for summarizing the variable-dimensional posterior distributions encountered in this problem. Finally, Section 4.5 discusses the results presented in the chapter.

4.2 Hierarchical Model and RJ-MCMC sampler

When a muon crosses a SD tank, it generates "Cherenkov photons", the rate of which depends on the muon's energy, along its track. These photoelectrons (PE's) are then captured by detectors and create an analog signal which is consequently discretized using an analog-to-digital converter (ADC). The real signal from Pierre Auger observatory contains, in addition to the muonic part, a noise-like background part created by electromagnetic components (mostly gamma photons). In this work, we concentrate only on analyzing the muonic part of the observed signal.

Here, as observed signals, we have only been provided with the vector $\mathbf{n} = (n_1, \ldots, n_N) \in \mathbb{N}^N$ of the number of PE's in each bin simulated from the generative model for the muonic part of the signal (as in Bardenet et al. (2012)). 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).

In the following, we briefly describe the model and RJ-MCMC sampler developed for the problem of Bayesian detection and estimation of muons in the Auger project; see Kégl (2008) ; Kégl and Veberic (2009) ; Bardenet et al. (2010, 2012) for more information.

4.2.1 Hierarchical model

Each muon has two component-specific parameters; the arrival time t_μ and the signal amplitude a_μ. 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 numbers of PE's in each bin are independent, the likelihood is written as

$$p(\mathbf{n} \,|\, k, \boldsymbol{t}_\mu,\, \boldsymbol{a}_\mu) \;=\; \prod_{i=1}^{N} p(n_i \,|\, \bar{n}_i(k, \boldsymbol{a}_\mu,\, \boldsymbol{t}_\mu)), \tag{4.1}$$

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

$$\bar{n}_i(\boldsymbol{a}_\mu,\, \boldsymbol{t}_\mu) \;=\; \sum_{j=1}^{k} \bar{n}_i(a_{\mu,j},\, t_{\mu,j}). \tag{4.2}$$

To define $\bar{n}_i(a_{\mu,j},\, t_{\mu,j})$, one needs to model the absorption procedure of a photon in the detector. For this purpose, Bardenet et al. (2010, Section 2.2) modeled PE arrivals by a non-homogeneous Poisson point process with intensity

$$h(t \,|\, a_\mu, t_\mu) \;=\; a_\mu \, p_{\tau,t_d}(t - t_\mu), \tag{4.3}$$

where $p_{\tau,t_d}(t)$ is the time response distribution given by

$$p_{\tau,t_d}(t) \;=\; \frac{1}{t_d} \cdot \begin{cases} 0 & \text{if } t < 0, \\ 1 - \exp(-\frac{t}{\tau}) & \text{if } 0 \leq t < t_d, \\ \exp(-\frac{t-t_d}{\tau}) - \exp(-\frac{t}{\tau}) & \text{if } t_d \leq t, \end{cases} \tag{4.4}$$

where t_d is the risetime and τ is the exponential decay (both measured in ns). Figure 4.2(a) shows a typical time response distribution with parameters $t_d = 10$ ns and $\tau = 60$ ns. Then, the expected number of PE's in the bin i is obtained by integrating the intensity (4.3) in the corresponding bin, as follows

$$\bar{n}_i(a_\mu,\, t_\mu) \;=\; a_\mu \int_{t_{i-1}}^{t_i} p_{\tau,t_d}(t - t_\mu)\mathrm{d}t. \tag{4.5}$$

The muon's amplitude a_μ is defined by $a_\mu = s_\mu \phi_\mu \nu$, where s_μ is the tracklength of the muon, ϕ_μ is a factor that captures the energy dependence of the signal amplitude (see Bardenet et al., 2010, Section 2.3), and ν is the average number of PE's generated by a muon with kinetic energy of 1 GeV on a tracklength of 1 m. The tracklength s_μ depends on the zenith angle θ, i.e., the angle in which the muons are arrived, and the dimensions of the tank with the radius of 1.8 m and the height of 1.2 m. While in general the zenith angle θ is treated as a random variable (see Kégl (2008) ; Kégl and Veberic (2009)), here as in Bardenet et al. (2012), it is assumed to be fixed and set to $\theta = 45$. Moreover, in the generative model that our data was simulated from a simplified prior has been considered by Prof. Kégl and his colleagues over the amplitudes in which a uniform prior distribution was assigned over the tracklength, i.e., $p(s_\mu) = \mathcal{U}(0,\, 1.7)$, the energy factor ϕ_μ was set to one, and the average number of PE's was set to $\nu = 55$.

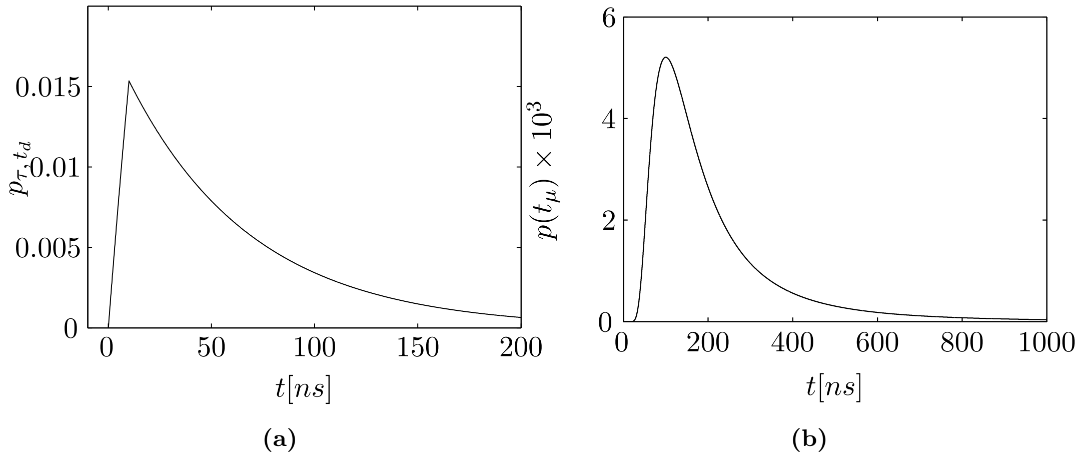


Figure 4.2 – *(a) Time response distribution* (4.4) *with parameters* $t_d = 10$ *ns and* $\tau = 60$ *ns. (b) Inverse Gamma prior distribution* $\mathcal{IG}(2.5,\, 350)$ *over the muon's arrival time* $p(t_\mu)$.

Turning to the muon's arrival time t_μ, an Inverse Gamma distribution is used as its prior distribution, i.e., $p(t_\mu) = \mathcal{IG}(a, b)$. To specify its parameters, the energy and geometry of the shower, and the distance of the tank from the shower core should be considered. The parameters of the Inverse Gamma prior are elicited by the disintegration of the observed showers and set to $a = 2.5$ and $b = 350$. Figure 4.2(b) shows the prior distribution assigned over the muon's time of arrival $p(t_\mu) = \mathcal{IG}(2.5,\, 350)$. Finally, from the expression of the likelihood (4.1) and the prior distributions mentioned above, the posterior distribution of the unknown parameters becomes

$$p(k, \boldsymbol{a}_\mu,\, \boldsymbol{t}_\mu \,|\, \mathbf{n}) \;\propto\; \prod_{i=1}^{N} p(n_i \,|\, \bar{n}_i(k, \boldsymbol{a}_\mu,\, \boldsymbol{t}_\mu))\, p(\boldsymbol{a}_\mu,\, \boldsymbol{t}_\mu \,|\, k)\, p(k), \tag{4.6}$$

where owing to the independence of $\boldsymbol{a}_\mu$ and $\boldsymbol{t}_\mu$ given k, we have

$$p(\boldsymbol{a}_\mu,\, \boldsymbol{t}_\mu \,|\, k) \;=\; p(\boldsymbol{a}_\mu \,|\, k)\, p(\boldsymbol{t}_\mu \,|\, k) \;=\; \prod_{j=1}^{k} p(a_{\mu,j})\, p(t_{\mu,j}),$$

and $p(k)$ is a Poisson distribution.

4.2.2 RJ-MCMC sampler

This section describes the RJ-MCMC sampler used to draw samples from the variable-dimensional posterior distribution (4.6). Similar to the RJ-MCMC sampler used in the previous chapter for the sinusoid detection problem, the sampler developed in Kégl (2008) ;

Bardenet et al. (2012) has both within-model moves, to update the model parameters without changing the number k of muons, and between-models moves, to change the dimension by adding or removing a muon.

For a given number k of muons, Bardenet et al. (2012) used an adaptive MCMC sampler that simultaneously learn appropriate identifiability constraints (following the approach of Celeux (1998)) and the covariance matrix of the normal random walk (similarly to the work of Haario et al. (2001) ; Roberts and Rosenthal (2001)).

Turning to the between-models moves, in the birth move a new muon is proposed by drawing its parameters, i.e., the amplitude a_μ and arrival time t_μ, from their corresponding prior distributions. In the death move, a muon is selected randomly and removed. Then, when a between-models move is accepted, the new vector of component-specific parameters is permuted to satisfy the learned identifiability constraint under the new model. The computation of the MHG acceptance ratio follows Proposition 1.11.

4.2.3 Related work

The problem of detection and estimation of the parameters, e.g., the locations and amplitudes, of filtered impulse (spike) trains (also known as deconvolution of filtered point processes) has applications in many fields including communication (see, e.g., Hero III, 1991), spectrometry (see, e.g., Andrieu et al., 2002 ; Barat and Dautremer, 2006), seismology (see, e.g., Rosec et al., 2003), and neural electrical activity (see, e.g., Mishchencko et al., 2011), to name a few; see also Cappé et al. (1999) and references therein.

4.3 A New Initialization Procedure for the Proposed Summarizing Algorithms

Before presenting the performance results of the proposed algorithms on the problem of joint detection and estimation of muons in Auger project, an important point to note here is that during our experiments, we observed that using the "naive" initialization procedure often yielded summaries in which the SEM-type algorithms got stuck in local minimums with inappropriate posterior summaries. In this section, we first show the convergence issues encountered when using the "naive" initialization procedure. Then, we introduce an "advanced" initialization procedure in which all the parameters of the model q_η are set be adding the Gaussian components progressively. Finally, in Section 4.3.3, we discuss ideas for selecting an appropriate value for the number L of Gaussian components.

4.3.1 Convergence issue with the naive initialization procedure

Recall that, in the "naive" initialization, after selecting the number L of the Gaussian components, for example, using the information provided in the posterior distribution $p(k \mid \mathbf{y})$ (see Section 4.3.3 for more discussion), the parameters of the Gaussian components are initialized using the robust estimates of the mean and variance of the posterior distributions of the sorted component-specific parameters given $k = L$. Observe that in the naive initialization procedure, the point process component and the probabilities of presence of the Gaussian components are neglected. The former one is of great importance, as it is supposed to capture the outliers. Therefore, in some experiments, the algorithm might be initialized near a local minimum in which some Gaussian components are of large variance and their corresponding labeled samples being multimodal.

To show the convergence issue, we use the following experiment, called hereafter the first illustrative example, in which there are $k = 3$ muons in the observed signal. Figure 4.3 shows the observed signal $\mathbf{n}$ in the top panel and the intensity of the model (4.3) in the bottom panel. The true arrival times are indicated by the vertical dashed lines. We were provided with 60 000 after burn-in variable-dimensional samples generated from the posterior distribution (4.6) using the RJ-MCMC sampler described in Section 4.2.2. Figure 4.4 demonstrates the posterior distribution of the number k of components together with the posterior distributions of the sorted arrival times given k.

Here, as the component-specific parameters, we concentrate only on the arrival times $\boldsymbol{t}_\mu$. Thus, the state space $\mathbb{X} = \bigcup_{k \in \mathcal{K}} \{k\} \times \boldsymbol{\Theta}^k$, with $\boldsymbol{\Theta} = \mathbb{R}_+$ being the space of arrival times. Moreover, we denote the samples on $\mathbb{X}$ by $\boldsymbol{x} = (k, \boldsymbol{\theta}_k)$, where the vector $\boldsymbol{\theta}_k$ of component-specific parameters only contains $\boldsymbol{t}_\mu$. Now, to initialize the summarizing algorithms, we choose, for example, $L = 3$ (note that $p(k \leq 3) = 0.82$). Note that we deliberately set $L = 3$ to highlight the convergence issue, while $L = 4$ is a more reasonable choice (see the next section for the results with $L = 4$). Then, if we follow the naive initialization procedure, we would use the robust estimates of the mean and variance of the posterior distributions of the sorted arrival times given $k = 3$ as the initial estimates[1]. Next, we set all the three probabilities of presence π_l to 0.5 and the mean λ of the Poisson point process component to 0.1 to avoid the point process component capturing too many samples in a few starting iterations of the SEM-type algorithm.

As it can be seen from the second row of Figure 4.4, related to $\mathcal{M}_3$, the middle compo-

[1] Note that these initial values would be the summary obtained if the BMS approach had been used to summarize the variable-dimensional posterior distribution shown in Figure 4.4.

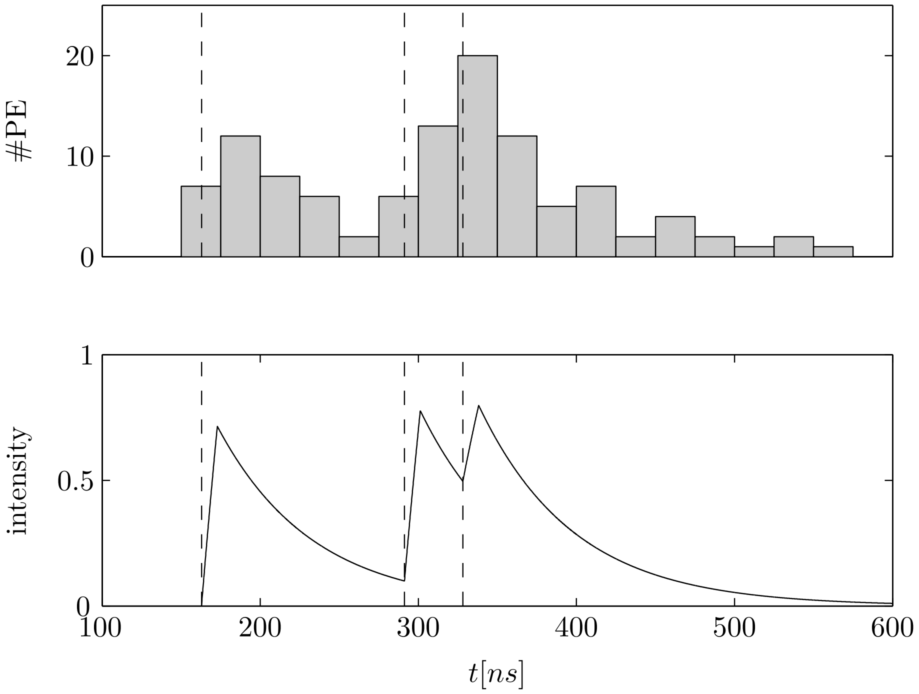


Figure 4.3 – *(top) Observed signal* $\mathbf{n}$ *of the first illustrative example. (bottom) Intensity of the model* $h(t \,|\, \boldsymbol{a}_\mu, \boldsymbol{t}_\mu)$ *defined in* (4.3). *There are* $k = 3$ *muons in the signal with the true arrival times, i.e.,* $\boldsymbol{t}_\mu = (163, 291, 328)$, *indicated by the vertical dashed lines.*

nent (colored in light blue) is highly bimodal with a very large variance ($s_2^2 = 7045.9$). We ran 100 iterations of the TAP-BHHJ with $\alpha = 0.1$ summarizing algorithm twice; once with the naive initialization procedure and once with the advanced one that will be described in Algorithm 4.1. Table 4.1 presents the initial and final estimated values of the model parameters for this example obtained using TAP-BHHJ with the naive initialization procedure. From the table, it can be observed that, except for the probabilities of presence, the final estimated values are very close to their corresponding initialized values. This evokes the question that the algorithm might have been trapped in a local minimum.

Figure 4.5 illustrates the normalized pdf's of the fitted Gaussian components for both scenarios along with the posterior distributions of the sorted arrival times given k. It can be seen that the main difference of the two summaries are in the second fitted Gaussian component. Therefore, for a better comparison, the histograms of the samples allocated to both the second Gaussian (the one with large variance in Table 4.1) and the point process components are illustrated in Figure 4.6.

It can be observed by comparing the two attained summaries presented in the figures that despite using the same SEM-type algorithms, clearly the summary obtained when the

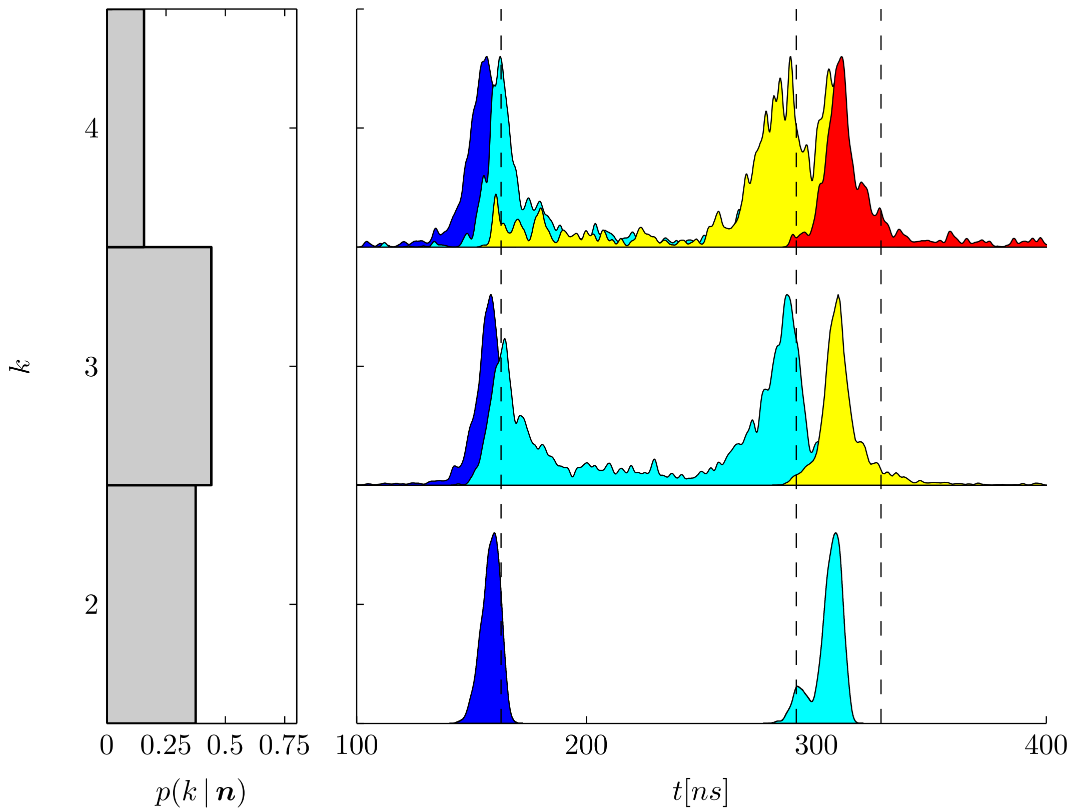


Figure 4.4 – *Posterior distributions of the number k of muons (left) and sorted arrival times, $\boldsymbol{t}_\mu$, given k (right) constructed using 60 000 RJ-MCMC output samples after discarding the burn-in period for the first illustrative example. The true number of components is three. The vertical dashed lines in the right figure locate the arrival times, i.e., $\boldsymbol{t}_\mu = (163, 291, 328)$.*

k	$\hat{\mu}$	$\hat{s}$	$\hat{\pi}$
Comp. 1	158.20 (156.84)	4.56 (5.30)	1 (0.5)
Comp. 2	274.92 (270.61)	67.49 (83.94)	0.57 (0.5)
Comp. 3	307.73 (309.89)	5.95 (7.13)	1 (0.5)

Table 4.1 – *The initial (in parentheses) and final estimates of the proposed summarizing algorithms for the first illustrative example (see Figure 4.4) using TAP-BHHJ with $\alpha = 0.1$ and the naive initialization procedures. The initialized and final estimated values of the mean parameter λ of the Poisson point process component was 0.1 and 0.32, respectively.*

advanced initialization was used is more desirable in the sense that the estimated middle component has a reasonable variance and the distribution of its corresponding labeled samples is unimodal. In fact, the Gaussian component with large variance shown on the

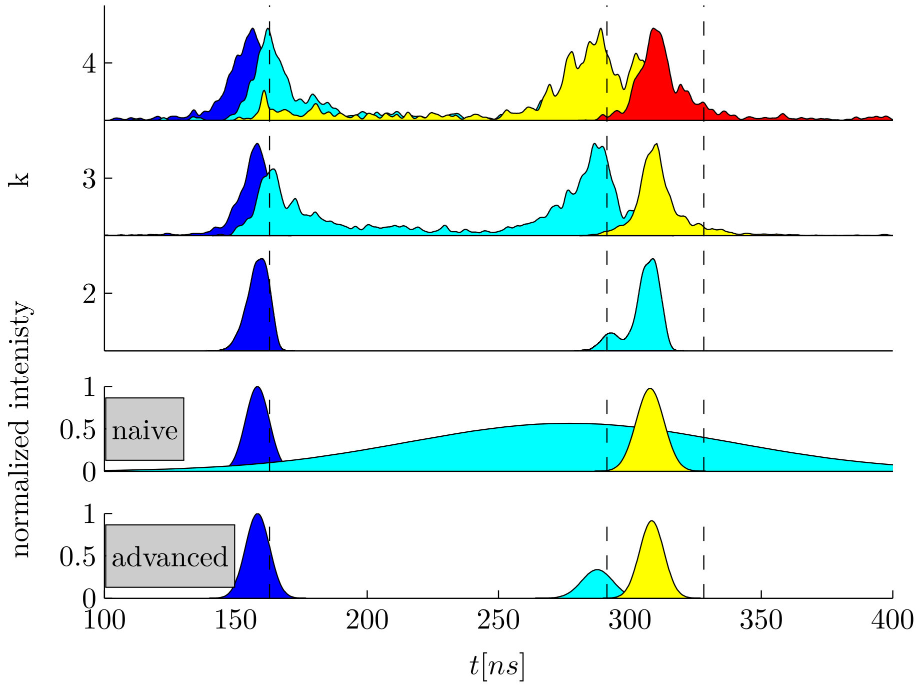


Figure 4.5 – *Normalized pdf's of the $L = 3$ fitted Gaussian components using 100 iterations of TAP-BHHJ summarizing algorithm with $\alpha = 0.1$ using once the naive and once the advanced initialization procedures. The top panel shows the posterior distributions of the sorted arrival times given k. The vertical dashed lines locate the true arrival times.*

top left panel of Figure 4.6 captured samples in a wide range, and thus, does not have a compact pattern. This also affected the estimated mean parameter $\hat{\lambda}$ of the Poisson point process component.

To emphasize again why we prefer the summary obtained using the advanced initialization shown in bottom row of Figure 4.5, Figure 4.7 compares the intensities of the fitted parametric model q_{η} defined in (3.8) with the histogram intensity of all arrival time samples obtained using the BMA approach. It can be seen that the fitted model when the advanced initialization procedure was used well captured the histogram intensity whereas the one of the naive procedure did not fit very well.

Remark 4.1. The approximated values of the BHHJ-α divergence criterion (2.29) in both cases are almost equal. This indicates that, in both cases, the SEM-type algorithm might have converged to two different minima with approximately equal values of the BHHJ criterion. Note that the results would have been different if we had introduced penalization for large variance components.

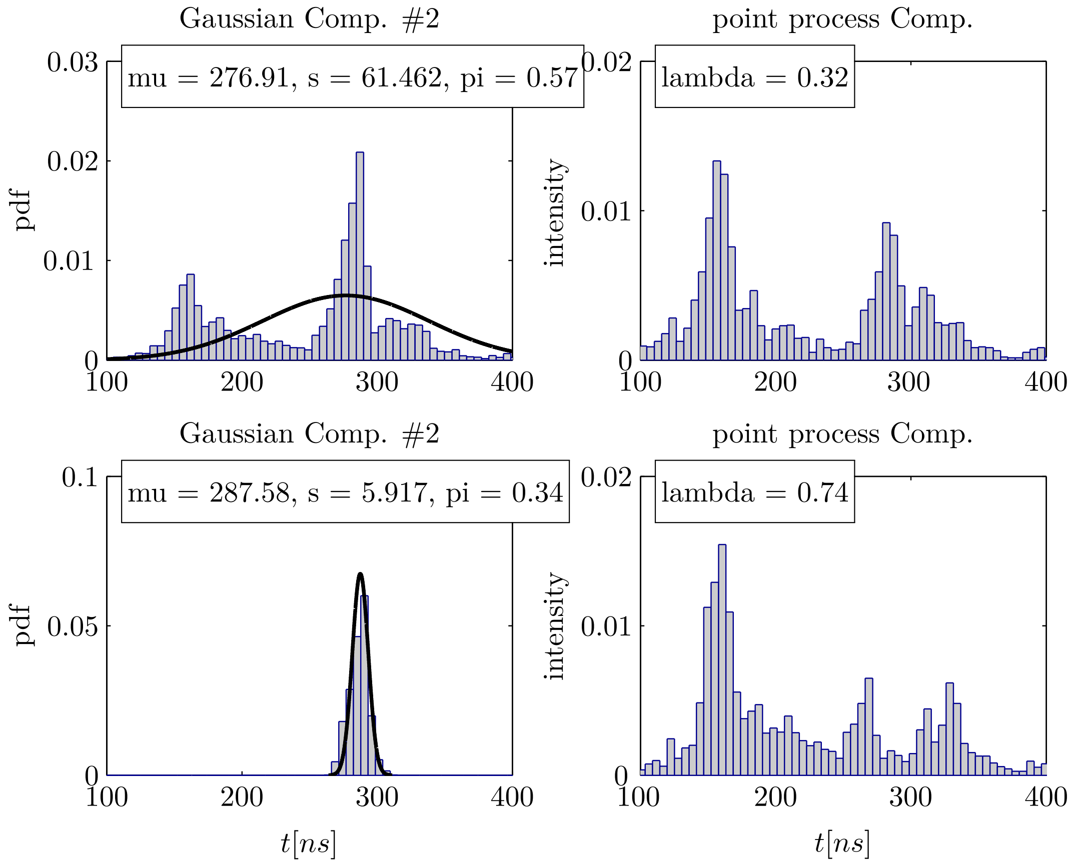


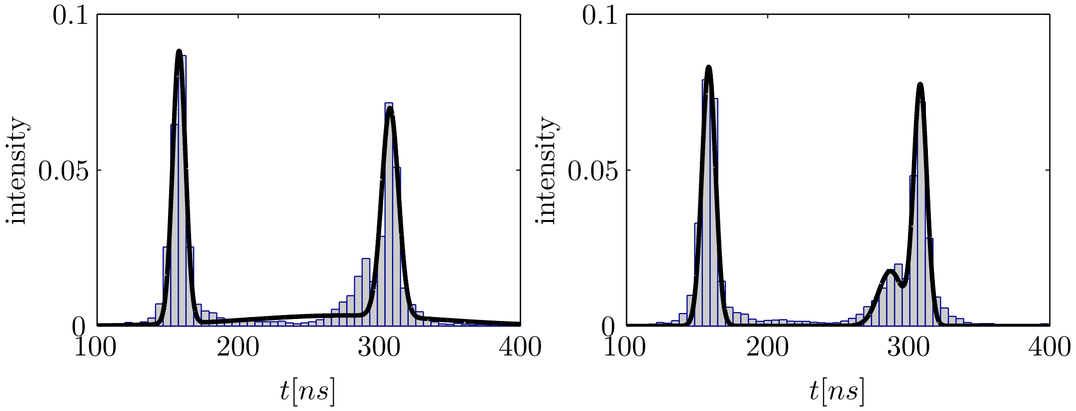
Figure 4.6 – *Histogram of the samples allocated to both the second Gaussian and the point process components, using once the naive initialization (top row) and once the advanced initialization procedure described in Algorithm 4.1 (bottom row).*

4.3.2 Advanced initialization procedure

In the advanced initialization procedure, as in the naive one, first, the maximum number of Gaussian components, denoted by L_{max} here, should be selected, for example, by inspecting the posterior distribution of the number k of components. Then, the parameters of both the Gaussian components and the Poisson point process component are initialized in a step-by-step fashion as described in Algorithm 4.1.

At the very beginning, it is assumed that there is no Gaussian component in the parametric model, i.e., $L = 0$, and all the observed samples are allocated to the point process component. Let us denote the matrix containing the bulk of samples allocated to the point process component by $\boldsymbol{X}_{L+1}$ of size $(d \times M_{L+1})$, where d is the dimension of $\boldsymbol{\theta}_{j,k}$, with $1 \leq j \leq k$, and

$$M_{L+1} = \sum_{i=1}^{M} \sum_{j=1}^{k^{(i)}} \mathbb{1}_{z_j^{(i)} = L+1}.$$



(a) *naive initialization*

(b) *advanced initialization*

Figure 4.7 – *Histogram intensity of all arrival time samples using BMA approach along with the intensity of the fitted parametric model obtained using TAP-BHHJ with* $\alpha = 0.1$.

Algorithm 4.1. *The advanced initialization procedure.*

- *Set* $L = 0$ *and allocate all the observed samples to the Poisson point process component.*
- *While* $L \leq L_{max}$ *do,*
  - *i) Set* $L = L + 1$*;*
  - *ii) Extract a Gaussian component from the matrix* $\boldsymbol{X}_{L+1}$ *containing the bulk of samples allocated to the point process component;*
  - *iii) Estimate the parameters of the model by doing a few iterations of the SEM-type algorithm (use robust estimators in the M-step). Update the matrix* $\boldsymbol{X}_{L+1}$.

We start by adding the first Gaussian component. To this end, a sample from the bulk of samples in the matrix $\boldsymbol{X}_{L+1}$ is selected such that a dense population of samples is concentrated around it. More precisely, we compute a distance matrix $\mathbf{D}$ of size $(M_{L+1} \times M_{L+1})$ for the columns of the matrix $\boldsymbol{X}_{L+1}$, each column $\boldsymbol{x}$ being a vector of observed sample allocated to the point process component, as

$$\mathbf{D}(i, j) \;=\; \|\boldsymbol{x}^{(i)} \,-\, \boldsymbol{x}^{(j)}\|.$$

Then, each row of the distance matrix $\mathbf{D}$ is sorted in an increasing order. Let us denote by $\widetilde{\mathbf{D}}$ the sorted distance matrix. Next, the sample index m with the minimum sum of the distances from its, say, r, nearest neighbors is chosen as the center of the cluster (or population). That is,

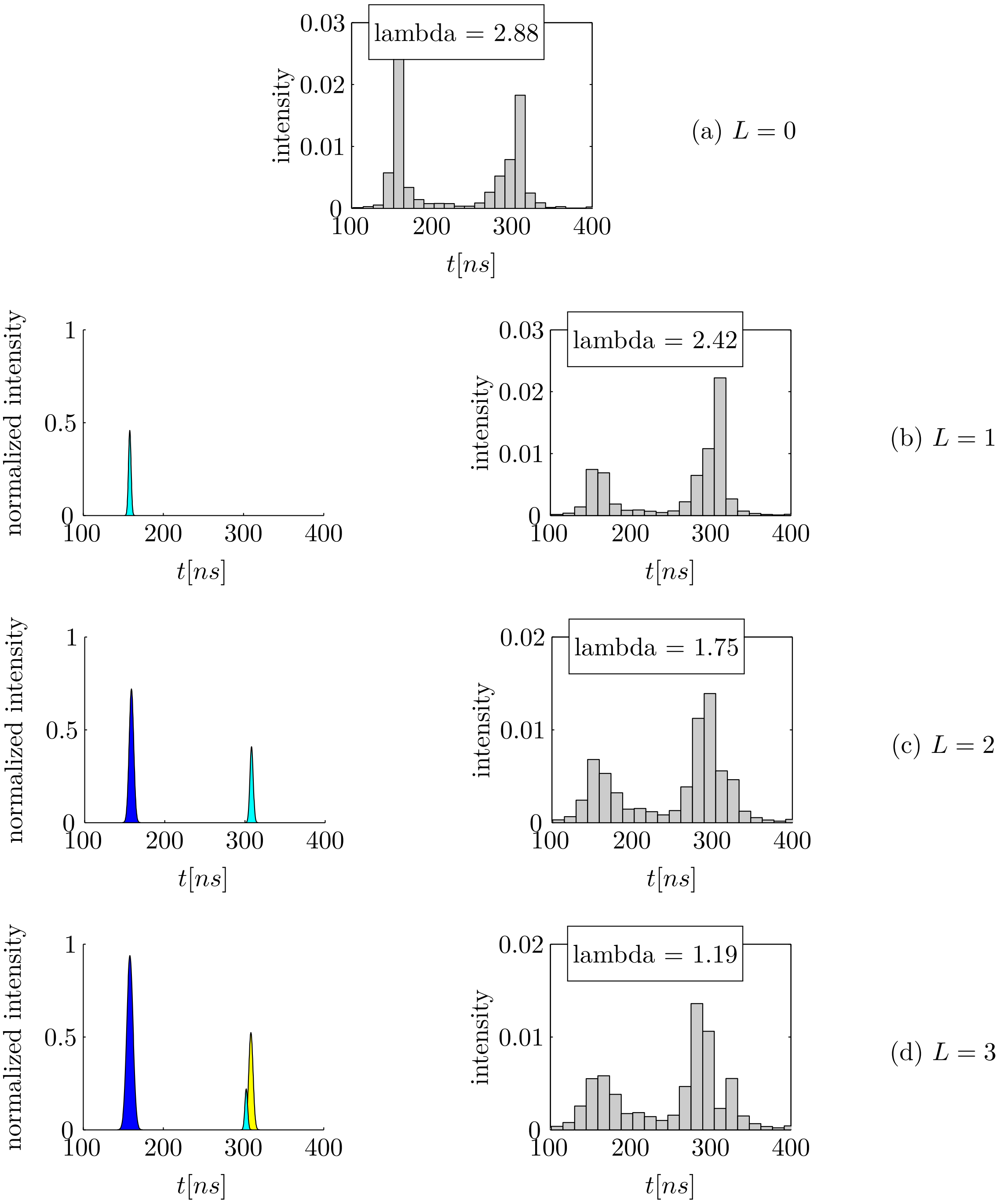
$$m \;=\; \underset{i}{\operatorname{argmin}} \sum_{j=1}^{r} \widetilde{\mathbf{D}}(i, j).$$

Finally, we estimate the parameters of the new Gaussian component from the selected $\boldsymbol{x}^{(m)}$ and its r nearest neighbors. After adding each Gaussian component, a few iterations, say, three, of the SEM-type algorithm are carried out to estimate all the parameters of the model $q_{\boldsymbol{\eta}}$. This procedure is repeated until there are $L = L_{max}$ Gaussian components in the model.

To clarify the advanced initialization procedure, Figure 4.8 illustrates the normalized pdf's of the estimated Gaussian components together with the histogram of the samples allocated to the point process component in different steps of the proposed initialization procedure. We set L_{max} to three and we used TAP-KL2 to estimate the parameters at each step of the advanced initialization procedure.

Figure 4.8(a) shows the histogram of the samples allocated to the Poisson point process component, i.e., $\boldsymbol{X}_{L+1}$, when $L = 0$. It can be seen that there are two regions of dense population (or modes); one located around 150 and another one around 300. In the next step, a Gaussian component was extracted from $\boldsymbol{X}_{L+1}$ and located at $\hat{\mu}_1 = 158$ using the distance matrix D as described before. Figure 4.8(b) shows the normalized pdf of the new Gaussian component and updated histogram of the point process components after estimating the parameters using three iterations of TAP-KL2 algorithm. As a result, the samples distributed around the first mode of Figure 4.8(a) were captured by the added Gaussian component and the mean λ was decreased from 2.88 to 2.42. Next, another Gaussian component located close to the second mode of Figure 4.8(a) was added to the model. Figure 4.8(c) shows the normalized pdf of both Gaussian components. Finally, a third Gaussian component was added and the final estimated parameters are presented in Figure 4.8(d).

Remark 4.2. The advanced initialization procedure can also be regarded as a complicated summarizing algorithm itself that starts from scratch (with no Gaussian component in the model), and, then, adds components progressively.



(a) $L = 0$

(b) $L = 1$

(c) $L = 2$

(d) $L = 3$

Figure 4.8 – *Normalized intensity of the estimated Gaussian components (left column) and histograms of the samples allocated to the point process component (right column) for different steps of the advanced initialization procedure described in Algorithm 4.1 used for the first illustrative example.*

Remark 4.3. We found that an approach which has some common features with the advanced initialization procedure has been developed independently by Melnykov and Melnykov (2011) for initializing the EM algorithm in Gaussian mixture models.

4.3.3 Remarks on how to choose L

Choosing an appropriate value for the number L of Gaussian components, which is indeed a model selection problem, is obviously a vital step in the approach that we proposed for summarizing variable-dimensional posterior distributions; see Section 3.3 for similar discussions. So far, we inspected visually both the posterior distributions of k and the sorted component-specific parameters to select a value for L.

Nevertheless, we saw in the second illustrative example of the previous chapter (see Section 3.3.3) that existence of the significant peaks in the distribution of samples allocated to the point process component is an indication of insufficiency of the chosen value for L (see the peak located around $t_\mu = 150$ in the distribution of the samples allocated to the point process component shown in Figure 4.6 for a similar issue). In fact, the samples allocated to the point process component 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. In the literature, it is always recommended to scrutinize residuals after fitting a model to an observed data (see, e.g., Draper and Smith, 1981, Chapter 3). More precisely, in our problem, existence of such peaks indicates our uniform assumption of the distribution of the residuals have been violated. Therefore, increasing L is an attempt to capture those non-uniform patterns (see Section 4.4 for more discussion). This approach can be seen as the *forward selection procedure* of the variable selection literature (see, e.g., Draper and Smith (1981, Chapter 6) and Miller (2002)).

On the other hand, in repeated experiments discussed in Section 3.4, in a systematic way, we set the value of L to the largest value of k such that under that model there is at least 5% of the total number of observed samples [2]. Then, during the process of the summarizing algorithms, Gaussian components with estimated probabilities of presence close to zero were removed. In other words, having Gaussian components with negligible probabilities of presence indicates that L was overestimated. This approach is similar to the *backward elimination procedure* in variable selection literature (see, e.g., Draper and

[2] Note that, however, when using the advanced initialization procedure, the chosen value for L can be even larger than the maximum value of k visited by the Markov chain, as in this approach we are not using the posterior distributions of component-specific parameters given $k = L$ (see Section 4.4 for more discussion).

Smith (1981, Chapter 6) and Miller (2002)). Note also that one can also remove the components with the estimated probabilities of presence smaller than a certain threshold in a post-processing step (see Section 3.4 for more discussion).

In the following section, we use a combination of the both aforementioned approaches to select an appropriate value for L. We run the summarizing algorithms with a "guess" value for L, obtained by inspecting visually the posterior distributions of k and component-specific parameters (for example, $L = 3$ was a guess value in the example analyzed in this section). Then, we inspect the residuals and study the goodness-of-fit of the model, e.g., through analyzing the intensity of the fitted model (see Figure 4.7), to see whether additional Gaussian components are needed. If so, we increase L and run the algorithms again. This procedure of adding Gaussian components might be repeated several times. In the meanwhile, Gaussian components with estimated probabilities of presence close to zero will be removed while running the algorithms.

4.4 Results

In this section, we investigate the performance of the proposed summarizing approach on three examples (including the one discussed in the previous section). For each example, we were provided with 60 000 post burn-in RJ-MCMC output samples from the posterior distribution (4.6). We further thinned the samples to every fifth. Hence, the number $M = 12\,000$ of RJ-MCMC samples used as observations for the summarizing algorithms. In all examples, we use the advanced initialization procedure proposed in Section 4.3 to initialize the algorithms.

4.4.1 First example

In the previous section, we saw the performance of the proposed summarizing approach on the variable-dimensional posterior distribution of the first illustrative example shown in Figure 4.4 when using a model with $L = 3$ Gaussian components. Nevertheless, inspecting the distribution of samples allocated to the point process component (bottom right panel of Figure 4.6), it can be seen that there is a dense population region of samples around $t_\mu = 160$, though a Gaussian component with $\hat{\mu}_1 = 158.31$ and $\hat{\pi}_1 = 1$ resides there. This observation suggests that an additional Gaussian component should be added to the parametric model for a better approximation of the true posterior distribution.

Hence, as described in Section 4.3.3, we increase L by one and study the solution obtained when there are four Gaussian components in the model. Figure 4.9 shows the histogram of the labeled samples along with the pdf's of the fitted Gaussian components using 100 iterations of TAP-BHHJ with $\alpha = 0.1$ and $L = 4$. Moreover, the estimated parameters of components are presented in the corresponding panels. Comparing the summary shown in the figure with the one shown in Figure 4.6, it can be seen that the additional Gaussian component with $\hat{\pi}_2 = 0.22$ (see top right panel of Figure 4.9) captured the extra samples around $t_\mu = 160$. As a result, the corresponding peak in the histogram of residuals is removed and the estimated mean $\hat{\lambda}$ goes from 0.55 to 0.31.

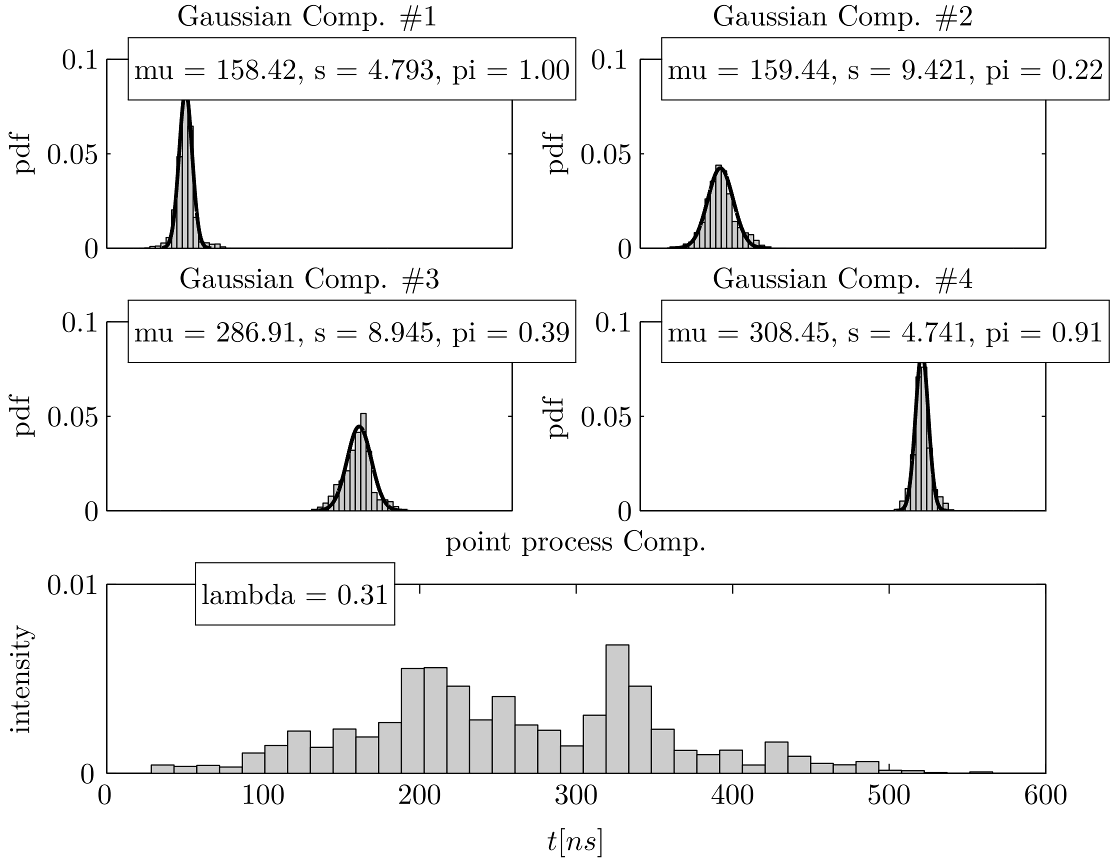


Figure 4.9 – *Histogram of the labeled samples along with the pdf's of estimated Gaussian components in the model (black solid line) using TAP-BHHJ with $\alpha = 0.1$ and $L = 4$ to analyze the first example of Auger project. The estimated parameters of each component are presented in the corresponding panel. To generate these histograms the randomized allocation procedure was run 10 times.*

It can be seen from the obtained summary presented in Figure 4.9 that there are two muons with high probabilities of presence at $\hat{\mu}_1 = 158.42$ and $\hat{\mu}_4 = 308.45$ in the variable-dimensional posterior distribution shown in Figure 4.4. There is also a third muon at $\hat{\mu}_3 = 286.91$ with probability of presence $\hat{\pi}_3 = 0.39$. Note that these are the same components as in the case when L was set to three. Finally, the posterior distribution contains a fourth muon at $\hat{\mu}_2 = 159.44$ with low probability of presence. Furthermore, inspecting the distributions of the samples allocated to the Gaussian components shown in Figure 4.9, it can be seen that the effects of label-switching were successfully removed by the proposed algorithm (observe that all distributions are unimodal).

However, still a peak around $t_\mu = 320$ns is apparent in the distribution of the residuals shown on the bottom panel of Figure 4.9. According to our discussion in Section 4.3.3, an extra component is needed to catch those samples in the residuals. Thus, we ran 100 iterations of TAP-BHHJ with $\alpha = 0.1$, this time with $L = 5$ Gaussian components in the model. Figure 4.10 illustrates the normalized pdf's of the fitted Gaussian components for both cases of $L = 4$ and $L = 5$. For comparison, the posterior distributions of sorted arrival times given k are also depicted. Comparing the two obtained summaries, it can be seen that when $L = 5$ a Gaussian component, with $\hat{\pi}_5 = 0.09$, shown in orange is added capturing samples around $\hat{\mu}_5 = 326.73$. As a result, the estimated value of λ becomes 0.22. The estimated parameters of the other four Gaussian components were almost identical in both cases.

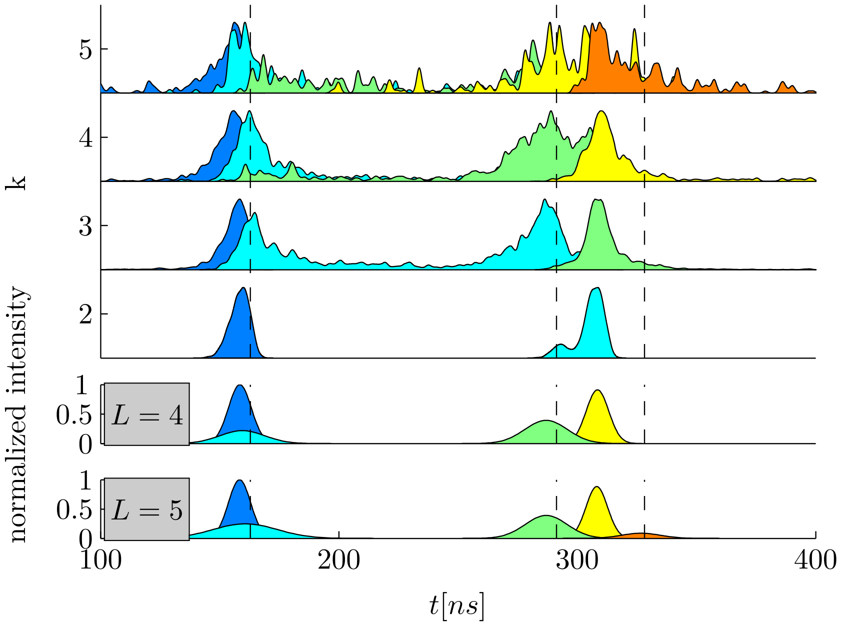
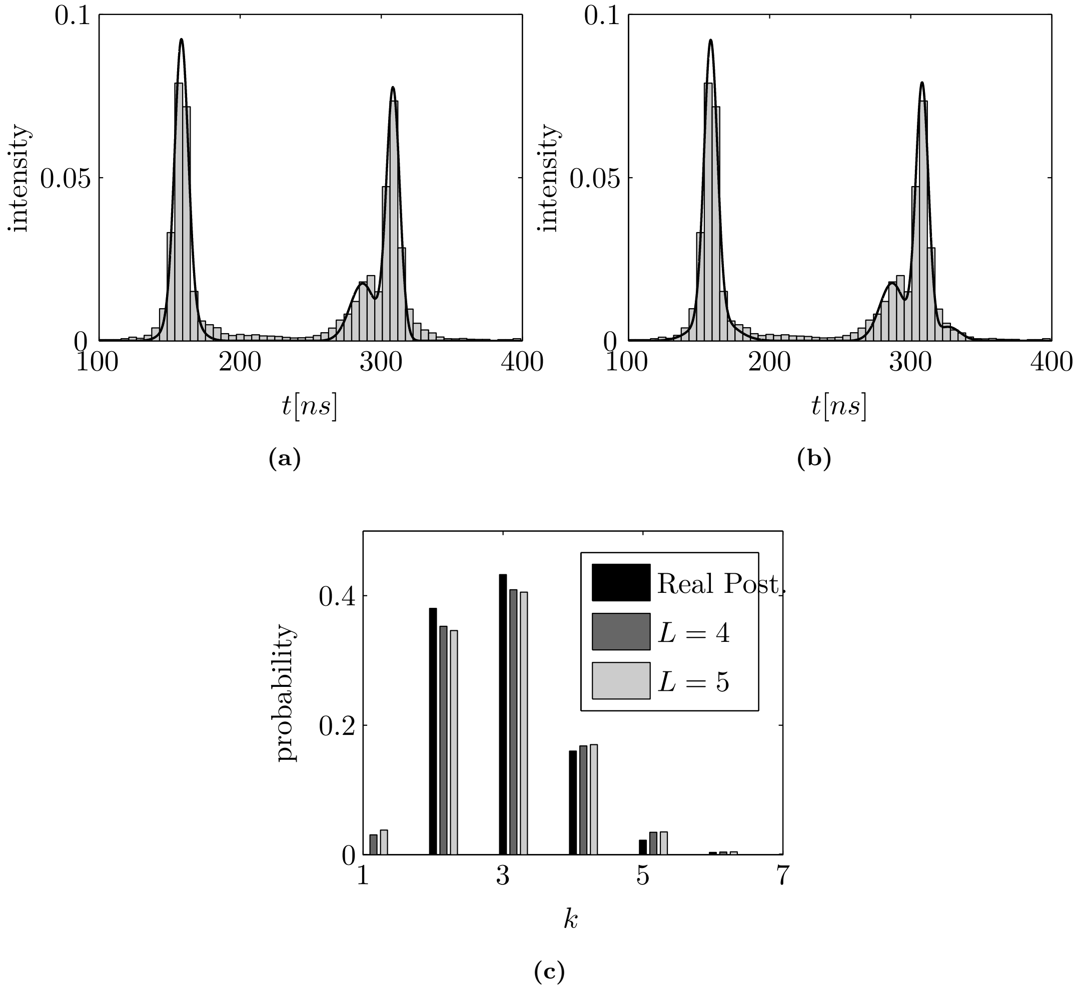


Figure 4.10 – *Normalized pdf's of the fitted Gaussian components using 100 iterations of TAP-BHHJ summarizing algorithm with $\alpha = 0.1$ using once with $L = 4$ and once with $L = 5$ on the first example. The top panel shows the posterior distributions of the sorted arrival times given k. The vertical dashed lines locate the true arrival times.*

To verify the goodness-of-fit of the approximate posterior distribution, i.e., $q_{\hat{\eta}}$, Figures 4.11(a) and (b) depict the histogram intensities of the observed samples obtained using the BMA approach along with the intensities of the fitted Gaussian components when there were four and five components in the model, respectively. It can be seen from the figures that when $L = 5$, the fitted model better captured the samples around $t_\mu = 320$ns. Moreover, Figure 4.11(c) compares the true posterior distribution of k with its approximated versions. The means of all posterior distributions are equal to 2.83. All three figures confirm that the approximate posterior distributions well captured the information in the true one.



(a) **(b)**

(c)

Figure 4.11 – *Goodness-of-fit of the approximate posterior distribution for the first example: (a, b) histogram intensities of the observed samples along with the intensities of the fitted Gaussian components for $L = 4$ and $L = 5$, respectively. (c) Posterior distribution of the number k of muons (black) versus its approximated versions (gray).*

4.4.2 Second example

In the second example considered in this section, the observed signal consists of eight muons located at $\boldsymbol{t}_\mu = (44,\, 46,\, 141,\, 266,\, 269,\, 346,\, 498,\, 644)$. Figure 4.12 shows both the observed signal and the intensity of the model (4.3) for this example. Note that there are two pairs of closely located components, i.e., one around $t_\mu = 45$ns and another one around $t_\mu = 267$ns. The posterior distributions of the number k of muons and the sorted arrival times are shown in Figure 4.13. Using the BMS approach, the model with $k = 7$ components would be selected ($p(k = 7 \,|\, \mathbf{n}) = 0.35$). However, as can be inspected from the figure, the second and third posterior distributions of sorted arrival times given $k = 7$ have large variances ($\hat{s}_2^2 = 2306$ and $\hat{s}_3^2 = 80$, using the normalized interquartile range explained in Section 2.5.1 as variance estimates). The former is highly multimodal, whereas the latter has a heavy tail on the right.

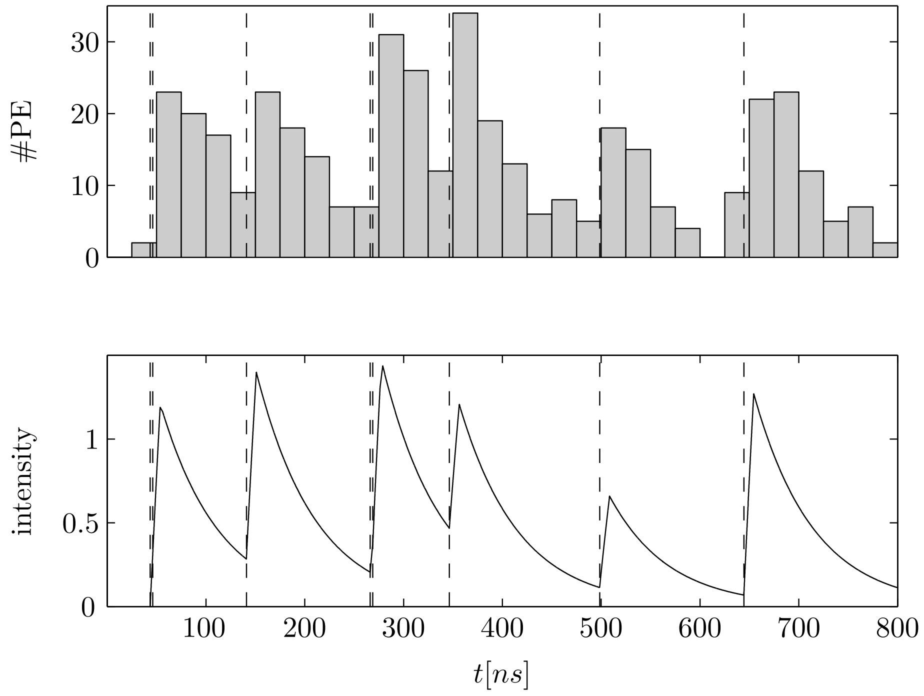


Figure 4.12 – *(top) Observed signal* $\mathbf{n}$ *of the second illustrative example. (bottom) Intensity of the model* $h(t \,|\, \boldsymbol{a}_\mu,\, \boldsymbol{t}_\mu)$ *defined in* (4.3). *There are* $k = 8$ *muons in the signal with the true arrival times, i.e.,* $\boldsymbol{t}_\mu = (44,\, 46,\, 141,\, 266,\, 269,\, 346,\, 498,\, 644)$, *indicated by vertical dashed lines.*

Considering the posterior probabilities given in $p(k \,|\, \mathbf{n})$ and the posterior distributions of the sorted arrival times shown in Figure 4.13, reasonable choices for L would be $L = \{8,\, 9\}$ (note that $p(k \leq 8 \,|\, \mathbf{n}) = 0.88$ and $p(k \leq 9 \,|\, \mathbf{n}) = 0.97$). Therefore, at a first

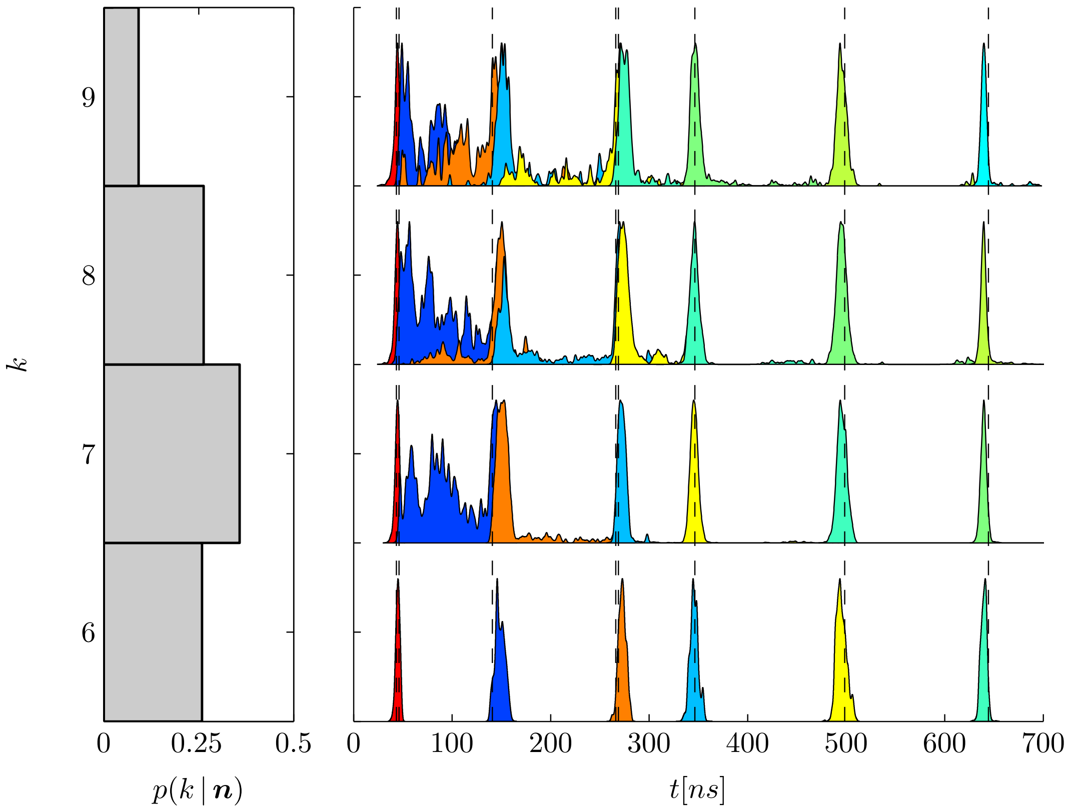


Figure 4.13 – *Posterior distributions of the number k of muons (left) and the sorted arrival times, $\boldsymbol{t}_\mu$, given k (right) constructed using 60 000 RJ-MCMC output samples after discarding the burn-in period for the second example. The true number of components is eight. The vertical dashed lines in the right figure locate the arrival times, i.e., $\boldsymbol{t}_\mu = (44, 46, 141, 266, 269, 346, 498, 644)$.*

attempt, we ran 100 iterations of TAP-BHHJ with $\alpha = 0.1$ with different values of the number $L = \{8, 9\}$ of Gaussian components on the variable-dimensional posterior shown in Figure 4.13.

When $L = 8$, Figure 4.14 shows the histogram of the labeled samples together with the pdf's of fitted Gaussian components. Moreover, the components' estimated parameters are presented in the corresponding panels. It can be seen from the figure that there are six components in the fitted model with probabilities of presence equal to one that correspond to the components shown in the last row of Figure 4.13. There is also the second Gaussian component with large estimated variance of $\hat{s}_2^2 = 1225$—almost half of the variance estimated using the BMS approach—that captured samples distributed in the range of [40, 180] (see Figure 4.15(c) for a zoomed view). Finally, a Gaussian component is associated with the samples around $\hat{\mu}_4 = 150.83$ but with low probability of presence.

Figure 4.15 shows the histogram of samples allocated to the Gaussian component with

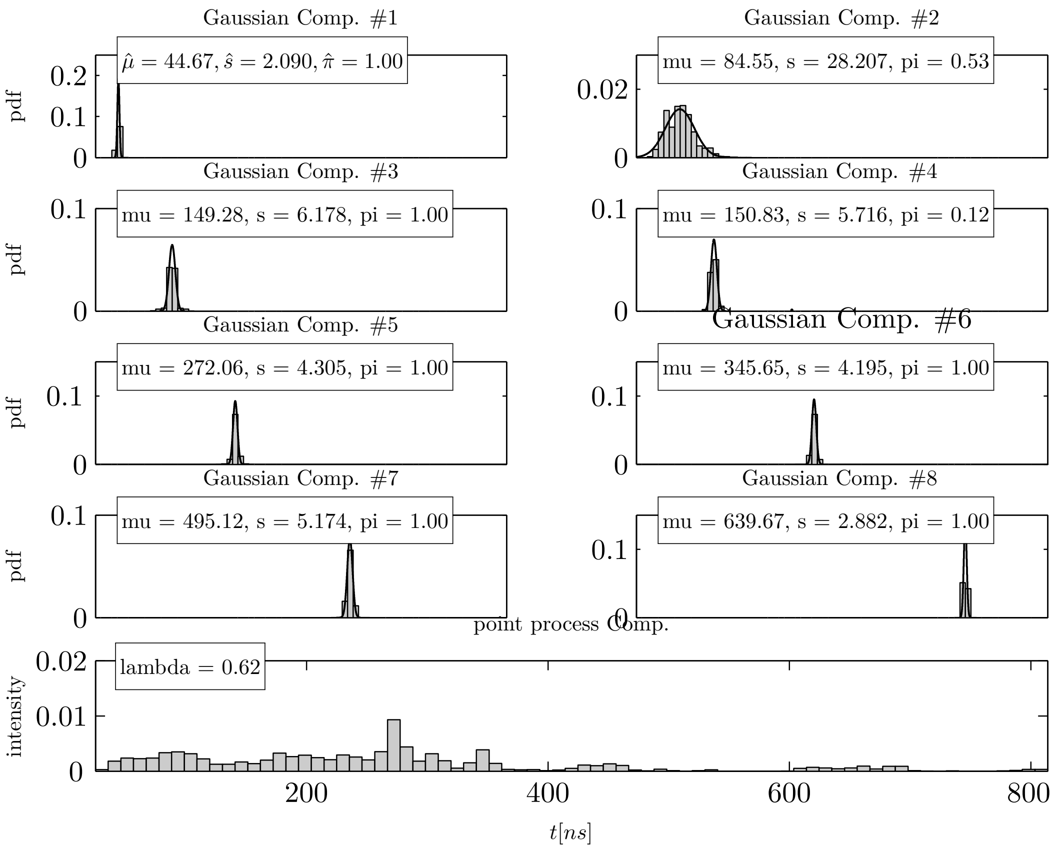
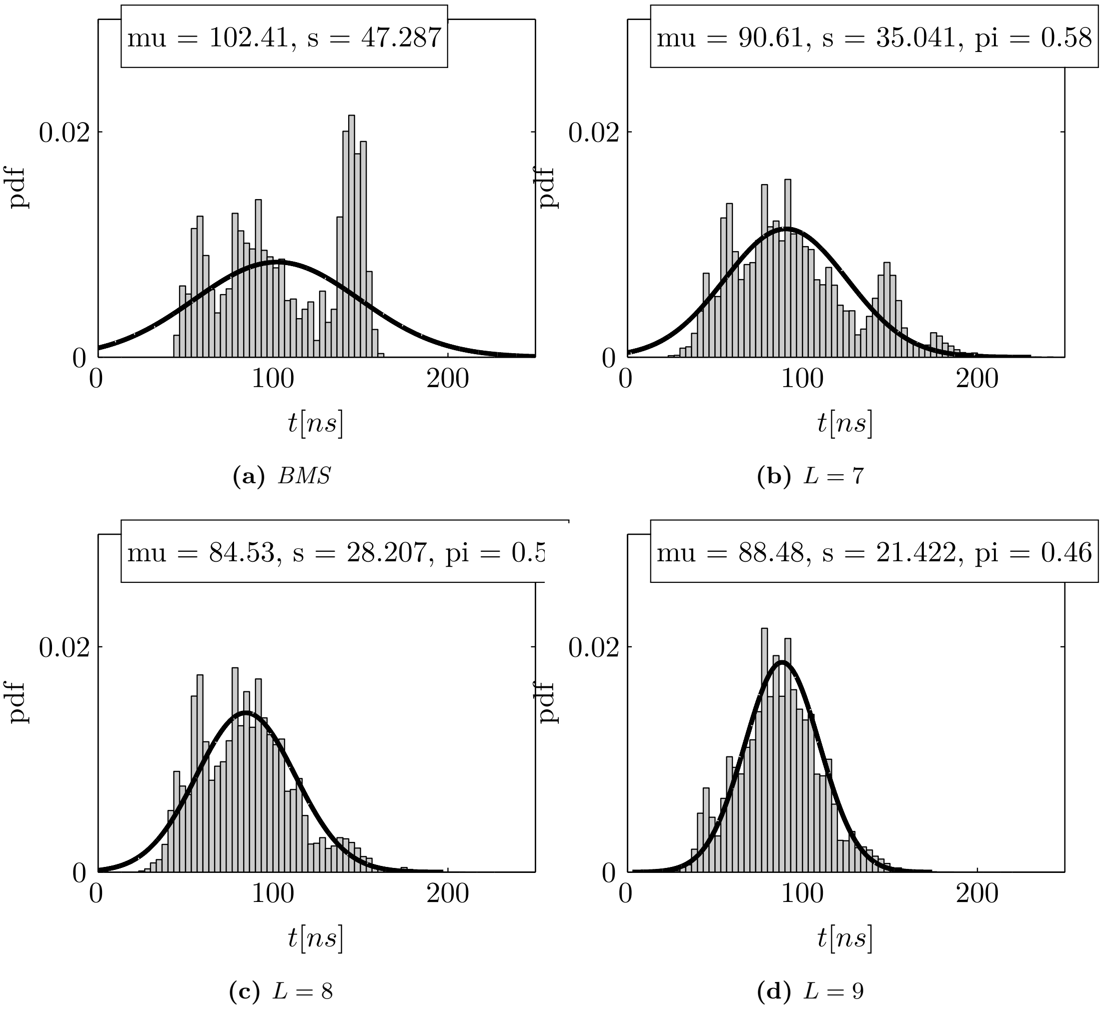


Figure 4.14 – *Histogram of the labeled samples along with the pdf's of estimated Gaussian components in the model (black solid line) using TAP-BHHJ with* $\alpha = 0.1$ *and* $L = 8$ *for the second example. The estimated parameters of each component are presented in the corresponding panel. To generate these histograms the randomized allocation procedure was run 10 times.*

large estimated variance along with its fitted pdf's for different values of $L = \{7, 8, 9\}$. Moreover, the corresponding summary when the BMS approach was used is also illustrated for comparison. It can be seen from Figure 4.15(c), corresponding to the case when $L = 8$, despite a Gaussian component being located at $\hat{\mu}_1 = 44.67$ with $\hat{\pi}_1 = 1$, some samples in this region were allocated to the Gaussian component with the large estimated variance; see the previous example for a similar discussion. Therefore, it is anticipated that by increasing L, additional components might catch the extra samples in that region, and thus, the variance of that component would be decreased.

It can be seen from the figures that going from $L = 7$ to $L = 8$, the mode on the right of Figure 4.15(b) (around $t_\mu = 150$ ns) disappears from the histogram of the labeled samples. Next, the mode on the left of Figure 4.15(c) was suppressed when $L = 9$, as the corresponding samples were captured by the additional component at $\hat{\mu}_2 = 55.31$

(with corresponding probability of presence of $\hat{\pi}_2 = 0.09$). So, in each case, the estimated variances were decreased. Furthermore, observe that in all cases the proposed approach provided a better summary for this component in comparison with the BMS approach.



(a) *BMS*

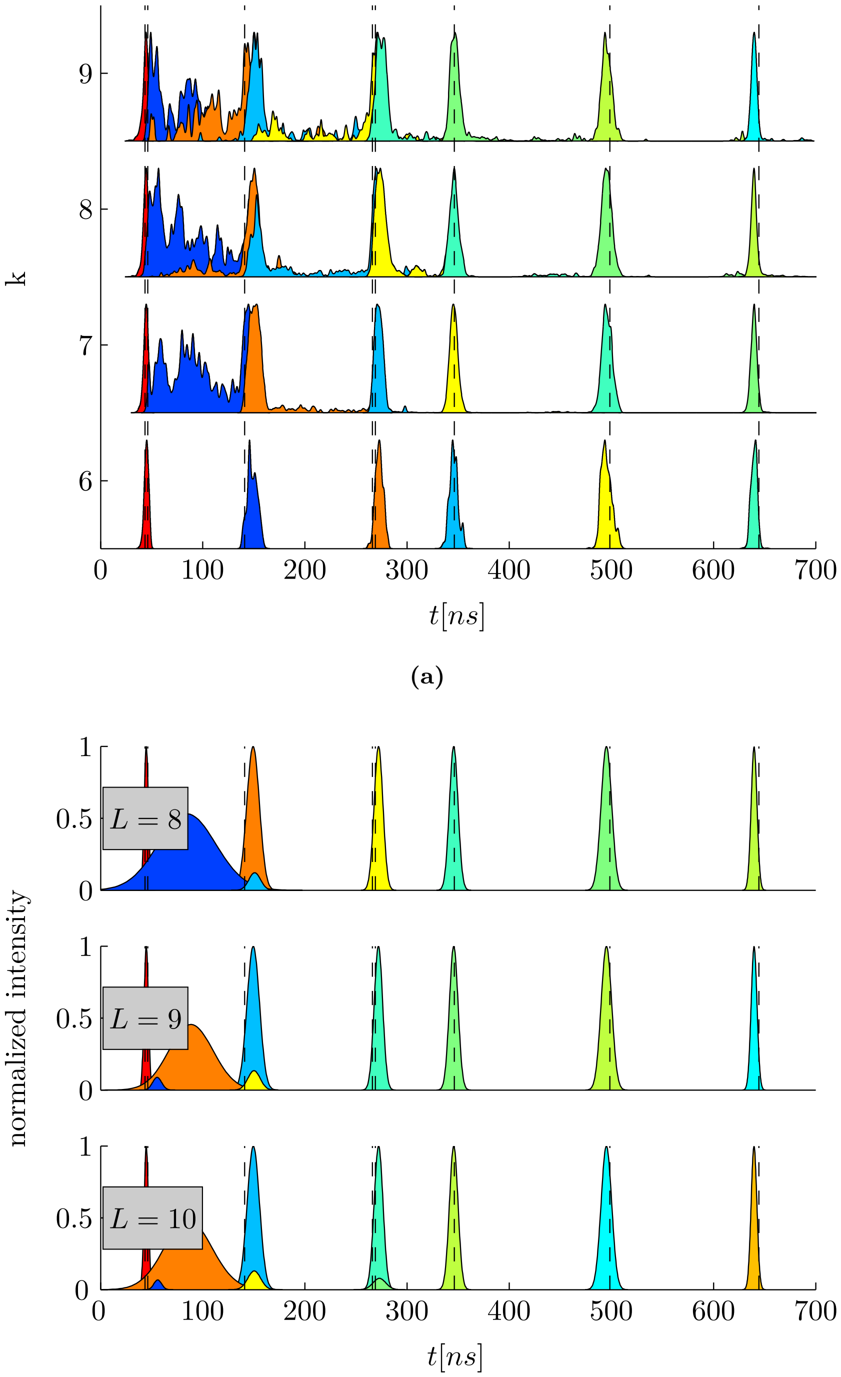
(b) $L = 7$

(c) $L = 8$

(d) $L = 9$

Figure 4.15 – *Histograms of the allocated samples to the Gaussian component with large estimated variance and its fitted pdf's using the BMS approach and TAP-BHHJ with $\alpha = 0.1$ and different values of $L = \{7, 8, 9\}$ for the second example.*

Inspecting the distribution of the samples allocated to the point process component shown in the bottom row of Figure 4.14, one observes a mode around $t_\mu = 270$ ns. This mode remained in the residual even for $L = 9$ (figure not shown here). In fact, computing the probabilities of the number of muons in the interval $t_\mu \in (260, 300)$ reveals that there is a non-negligible probability of having two components in this interval. That is, $p(k = 1 \,|\, t_\mu \in (260, 300), \mathbf{n}) = 0.89$ and $p(k = 2 \,|\, t_\mu \in (260, 300), \mathbf{n}) = 0.10$. These facts suggest that a value of $L = 10$ can lead to a better approximation of the true posterior



(a)

(b)

Figure 4.16 – *(a) Posterior distributions of the sorted arrival times given k for the second example. (b) Corresponding normalized pdf's of fitted Gaussian components using TAP-BHHJ with $\alpha = 0.1$ and different values of $L = \{8, 9, 10\}$.*

distribution. Therefore, we ran 100 iterations of TAP-BHHJ with $\alpha = 0.1$ and $L = 10$ Gaussian components.

Figure 4.16 illustrates the normalized pdf's of the estimated components for different values of $L = \{8,\ 9,\ 10\}$ along with the posterior distributions of sorted arrival times. It can be seen from the figure that, setting L to ten, an additional component is located at $\hat{\mu}_7 = 273$ with $\hat{\pi}_7 = 0.08$. Remarkably, in the true observed signal $\mathbf{n}$, there were two muons in this region (see Figure 4.12). Another point to note in the figure is that in all the summaries corresponding to different values of L, there are six components with probabilities of presence equal to one which are aligned with the ones in the bottom row of Figure 4.16(a).

Finally, to verify the goodness-of-fit of the fitted approximate posterior distribution $q_{\hat{\boldsymbol{\eta}}}$, Figure 4.17(a) shows the histogram intensity of the observed samples obtained using the BMA approach along with the intensity of the fitted Gaussian components when there were $L = 10$ components in the model. Moreover, Figure 4.17(b) compares the true posterior distribution of the number k of muons, i.e., $p(k \mid \mathbf{n})$, with its approximated versions obtained using TAP-BHHJ with different values of L. The true posterior mean is 7.28, while the means of the approximated posteriors for $L = 8$ to 10 are 7.27, 7.26, and 7.28, respectively. Both figures confirm that the approximate posterior well captured the information in the true posterior distribution.

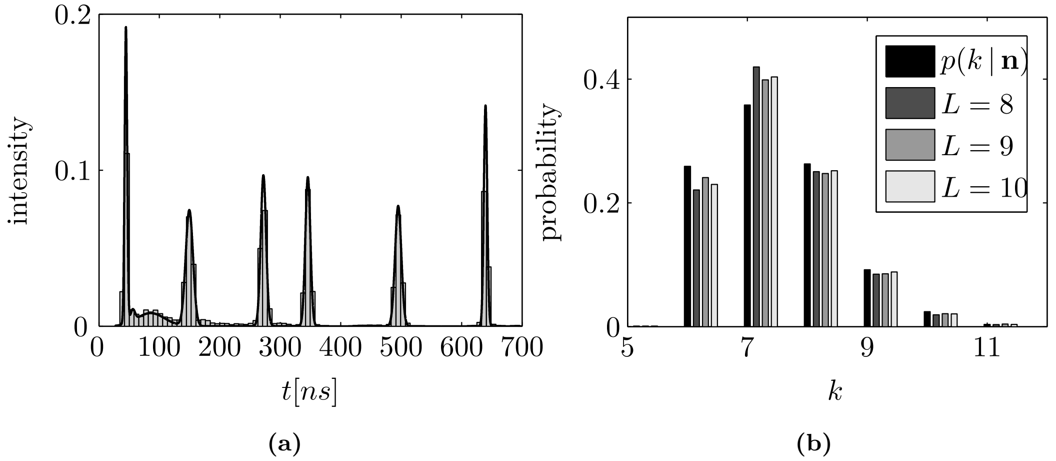


Figure 4.17 – *Goodness-of-fit of the approximate posterior distribution for the second example: (a) histogram intensity of the observed samples along with the intensity of the fitted Gaussian components when $L = 10$. (b) Posterior distribution of the number k of muons (black) versus its approximated versions.*

Remark 4.4. Observe that in this specific example, there are a number of RJ-MCMC samples non-concentrated distributed in the region $[40, 180]$; see Figure 4.13. Associating a Gaussian component to describe those samples by the algorithm resulted in its estimated variance to be large (or equivalently it has an uncertain location).

4.4.3 Third example

The observed signal of the third illustrative example, shown in Figure 4.18, consists of five muons located at $\boldsymbol{t}_\mu = (105, 169, 267, 268, 498)$. The posterior distributions of the number k of muons and the sorted arrival times are shown in Figure 4.19. Note that, in this example, there are two muons with almost equal arrival times, i.e., the third and fourth muons.

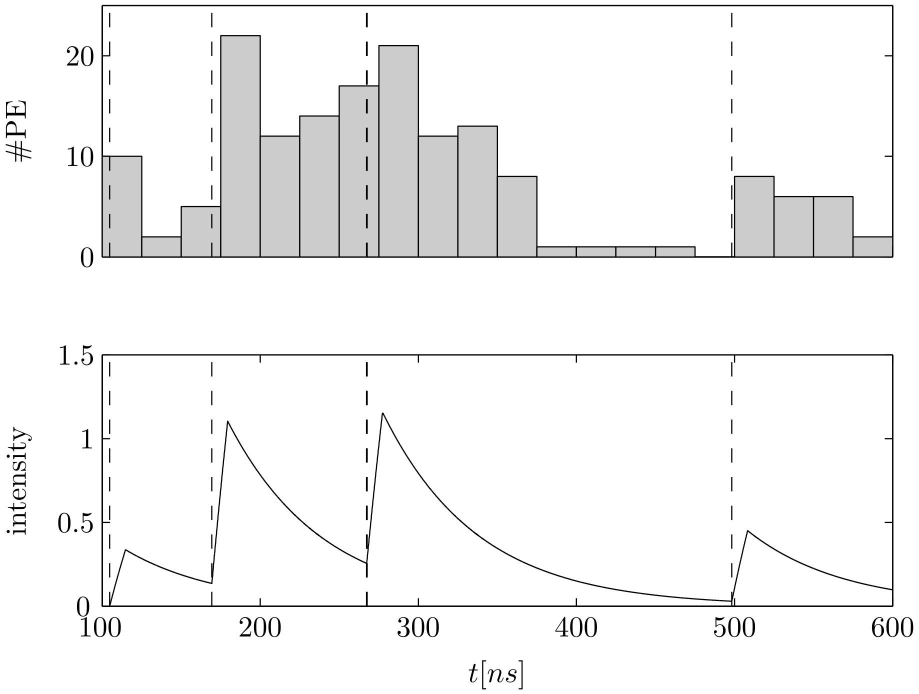


Figure 4.18 – *(top) Observed signal* $\mathbf{n}$ *of the third illustrative example. (bottom) Intensity of the model* $h(t \,|\, \boldsymbol{a}_\mu, \boldsymbol{t}_\mu)$ *defined in* (4.3). *There are* $k = 5$ *muons in the signal with the true arrival times, i.e.,* $\boldsymbol{t}_\mu = (105, 169, 267, 268, 498)$*, indicated by vertical dashed lines.*

Using the BMS approach, the model with four muons would be selected ($p(k = 4 \,|\, \mathbf{n}) = 0.4$), tough $\mathcal{M}_5$ has almost similar posterior probability of 0.38. However, the sorted posterior distributions of the third component, shown in orange color, under both models, particularly the one under $\mathcal{M}_5$, are bimodal. We ran TAP-BHHJ summarizing algorithm with $L = \{6, 7\}$ Gaussian components using the advanced initialization procedure on the RJ-MCMC output samples shown in Figure 4.19 (note that $p(k \leq 6 \,|\, \mathbf{n}) = 0.94$ and $p(k \leq 7 \,|\, \mathbf{n}) = 0.97$).

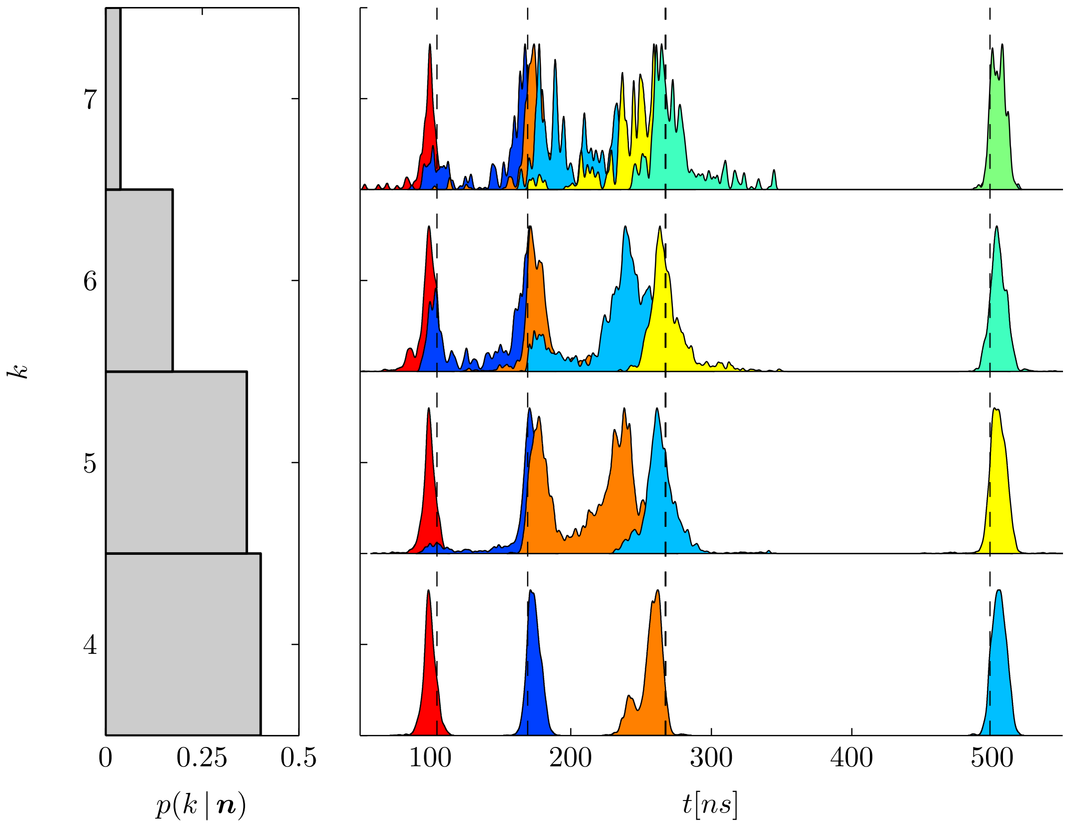


Figure 4.19 – *Posterior distributions of the number k of muons (left) and the sorted arrival times, $\boldsymbol{t}_\mu$, given k (right) constructed using 60 000 RJ-MCMC output samples after discarding the burn-in period for the third example. The true number of components is five. The vertical dashed lines in the right figure locate the arrival times, i.e., $\boldsymbol{t}_\mu = (105, 169, 267, 268, 498)$.*

When $L = 6$, Figure 4.20 shows the histogram of the labeled samples and the estimated parameters of the components. From the figure, it can be seen that the "severe" label-switching exhibited in Figure 4.19 is removed completely and the estimated Gaussian components enjoy reasonable variances. In the presented summary, there are four components with high probabilities of presence corresponding to the ones shown in the bottom row of Figure 4.19. There are also two other components with comparatively low probabilities of presence.

To discuss better the choice of L, Figure 4.21 illustrates the histograms of the residuals of the fitted model for different values of $L = \{6, 7, 8, 9\}$. It can be seen from the top left panel of Figure 4.21 that the distribution of the residuals corresponding to the case where $L = 6$ contains a few peaks. The peaks are gradually removed by adding Gaussian components. When $L = 7$, a component is added at $\hat{\mu} = 99.47$ with $\hat{\pi} = 0.06$ that captured

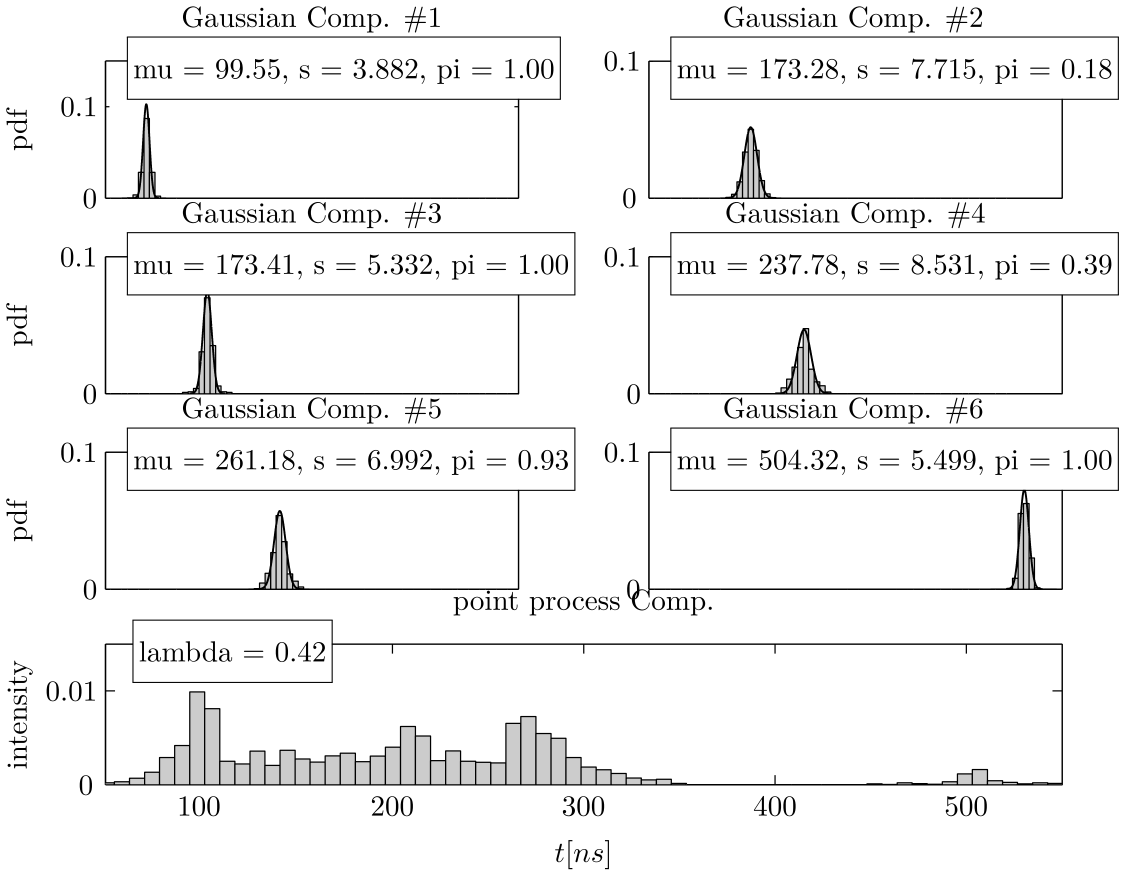


Figure 4.20 – *Histogram of the labeled samples along with the pdf's of estimated Gaussian components in the model (black solid line) using TAP-BHHJ with $\alpha = 0.1$ and $L = 6$ for the third example. The estimated parameters of each component are presented in the corresponding panel. To generate these histograms the randomized allocation procedure was run 10 times.*

samples distributed at the left peak of the top left panel of Figure 4.21. This is also coherent with the probabilities of having one or two components in the interval $t_\mu \in (80, 120)$ ($p(k = 1 \,|\, t_\mu \in (80, 120), \mathbf{n}) = 0.91$ and $p(k = 2 \,|\, t_\mu \in (80, 120), \mathbf{n}) = 0.08$.). Increasing the number L of components to eight, the samples corresponding to the peak around $t_\mu = 270$ were captured by the additional component at $\hat{\mu} = 268.4$ with $\hat{\pi} = 0.12$ (see the bottom left panel of Figure 4.21). Furthermore, Figure 4.22 compares the estimated Gaussian components for different values of L.

Figure 4.23 shows the histogram intensities of the observed samples along with the intensities of the fitted Gaussian components for $L = \{8, 9\}$ (note that the right column panels are zoomed versions of left column ones). It can be seen from the left column panels that, in both cases, the fitted model exhibits an acceptable fit to the distribution of the

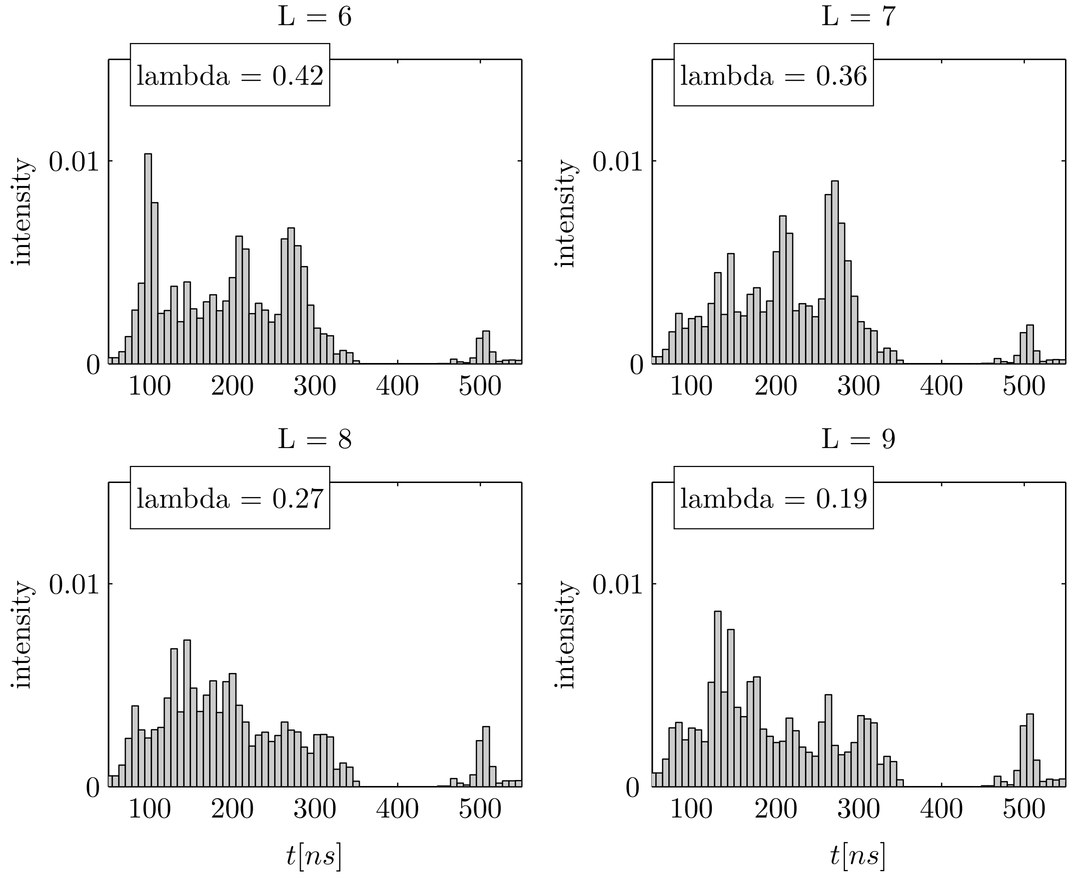


Figure 4.21 – *Histograms of the residuals of the fitted model using TAP-BHHJ with different values of L for the third example. To generate these histograms the randomized allocation procedure was run 10 times.*

observed samples. Nevertheless, inspecting the top right panel, one can see that the model with $L = 8$ components was not able to capture well the samples in the interval $t_\mu \in (200, 300)$. Moreover, computing the probabilities of the number k of muons in that interval reveals that there is a non-negligible probability of having four components in that region ($p(k = 4 \,|\, t_\mu \in (200, 300), \mathbf{n}) = 0.01$). These two facts suggest using a model with $L = 9$ components for this example. The bottom row of Figure 4.22 shows the normalized pdf's of the estimated components. See the bottom right panel of Figure 4.23 for improvements in the goodness-of-fit of the approximate model when nine Gaussian components were used.

Remark 4.5. Note that $p(k = 9 \,|\, \mathbf{n})$ is almost equal to zero. Hence, the naive initialization procedure cannot be used in this case.

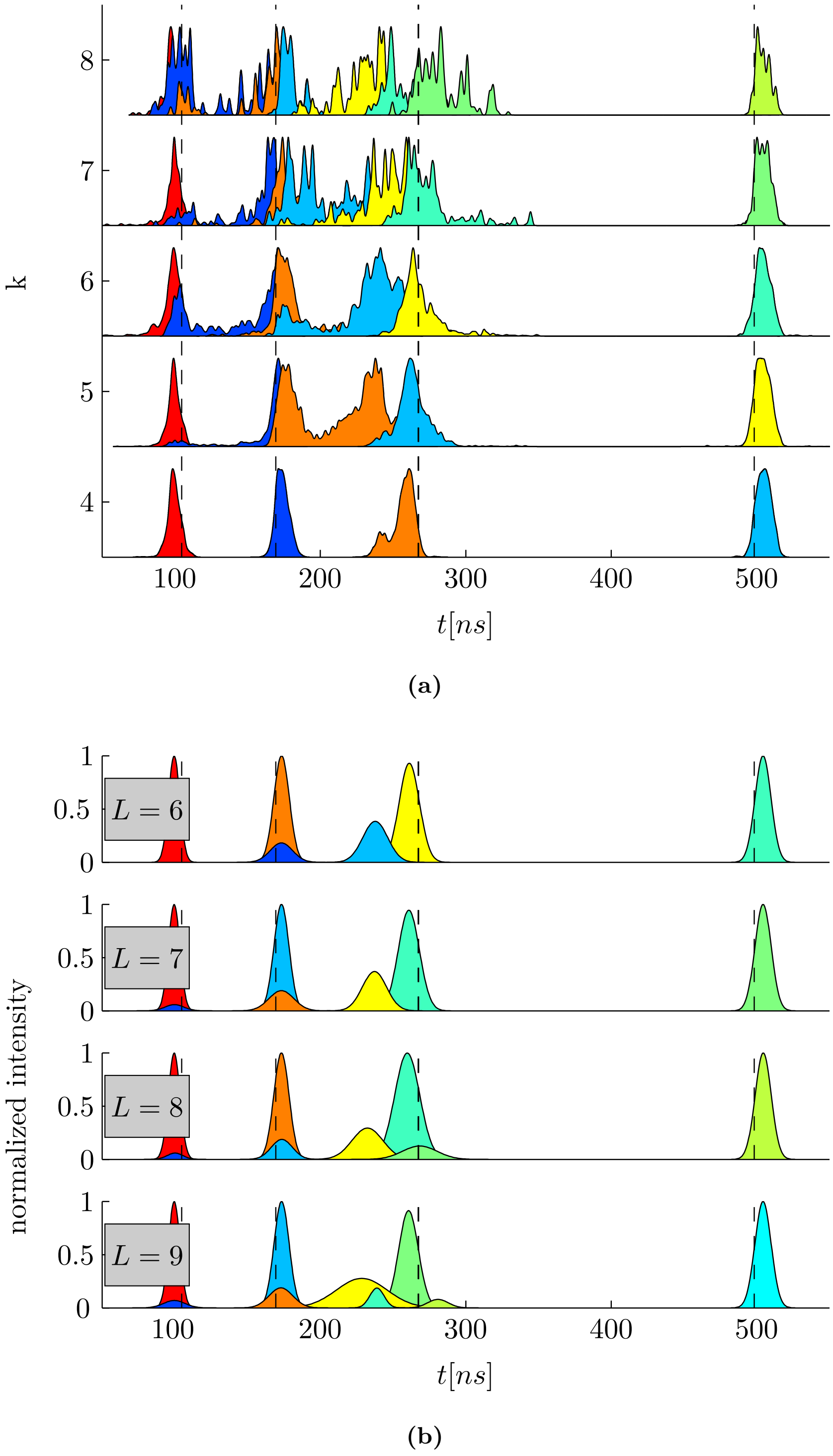


Figure 4.22 – *(a) Posterior distributions of the sorted arrival times given k. (b) Corresponding normalized pdf's of fitted Gaussian components using TAP-BHHJ with $\alpha = 0.1$ and different values of L.*

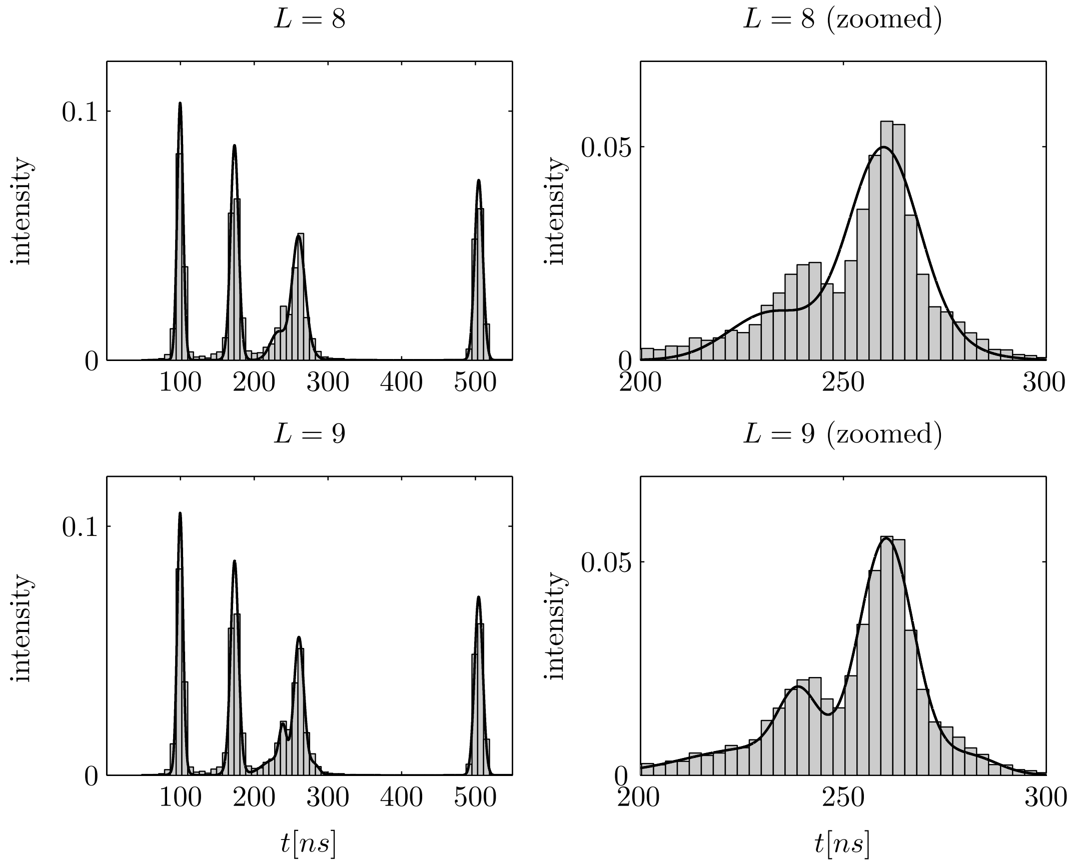


Figure 4.23 – *Histogram intensities of the observed samples along with the intensities of the fitted Gaussian components using TAP-BHHJ with $L = \{8, 9\}$ for the third example.*

4.5 Summary and Discussion

In this chapter, we studied the capability of the algorithms proposed in Chapter 2 for relabeling and summarizing the variable-dimensional posterior distributions encountered in Auger project. In Auger project, the objective is to count the number of muons and estimate their parameters in order to characterize the composite of the observed ultra-high energetic particles and to assess their origin in the universe. For this study, we have been provided with the observed data $\mathbf{n}$, i.e., the number of PE's in the bins, and the corresponding RJ-MCMC output samples by Prof. Balázs Kégl from the Laboratoire de l'Accélérateur Linéaire (LAL), Université Paris Sud 11. Here, we only concentrated on the muons' arrival times, i.e., $\boldsymbol{t}_\mu$.

In Section 4.3, we discussed two substantial methodological issues of the proposed summarizing algorithms, namely, the initialization step and the selection of the number L of Gaussian components. First, we showed that initializing the parametric model q_{η} using the "naive" initialization procedure can cause convergence issues. For example, using the naive procedure the algorithm might be initialized near a local minimum with a multimodal component that, consequently, might cause convergence issues (see Section 4.3.1). Thus, in Section 4.3.2, we proposed the "advanced" initialization procedure consisting in adding Gaussian components progressively from the residuals of the fitted model. The effectiveness of the advanced initialization procedure was shown on the first example (see, e.g., Figure 4.5).

In Section 4.3.3, we discussed two procedures for selecting an appropriate value for the number L of Gaussian components. In the first procedure, we run the algorithm with a "guess" value of L. Then, through analyzing both the distribution of the residuals and the figures showing the goodness-of-fit of the approximate posterior, we decide whether to increase L or not. For example, we recommended to increase L, provided there is a significant peak in the distribution of the residuals. This approach is similar to the forward selection procedure in the variable selection literature. The second procedure consists in eliminating components with probabilities of presence close to zero while the algorithm is running. This procedure can also be seen as the backward elimination procedure in variable selection literature. In the results presented in this chapter, we used a combination of both procedures.

We analyzed the variable-dimensional posterior distributions of three examples in this chapter. The posterior distribution of sorted arrival times given k of all three examples contained at least one large variance component with multimodal distribution. These components, on the one hand, made the summaries obtained using the BMS approach undesirable in the sense that the effects of label-switching was not removed by simply sorting the components. On the other hand, when the summarizing algorithms were used with the naive initialization procedure, because of those multimodal components, they were trapped in local minima. However, we saw that, in all three examples, using the advanced initialization procedure, the summarizing algorithm were capable of removing the label-switching effects and the resulting summaries enjoy Gaussian components with fairly reasonable variances.

In the results shown in Section 4.4, we particularly concentrated on how the goodness-of-fit of the approximate posterior distribution can be improved through choosing an appropriate value for L based on the remarks explained in Section 4.3.3. We saw that, in order to have an acceptable fitting, the chosen value of L should often be larger than k_{MAP}. In fact in the third example, the final chosen value of $L = 9$ was even larger than the maximum model visited by the Markov chain. As a result, the obtained summaries contained a few components with low probabilities of presence. Nevertheless, the delicate process of detection of muons can be carried out in a post-processing step, for example, by discarding components with probabilities of presence smaller than a certain threshold (see Section 3.4 for more information).

Conclusions and future work

In this thesis, we have addressed both computational and inferential issues related to the use of trans-dimensional Bayesian approaches for signal decomposition problems with an unknown number of components.

A substantial part of this thesis has been devoted to the inferential difficulties caused by the issue of "birth, death, and switching of components' labels" in trans-dimensional problems. This issue makes the process of summarization of variable-dimensional posterior distributions difficult. The algorithms developed so far in the literature to solve the label-switching issue are all restricted to the fixed-dimensional posteriors (see, e.g., Celeux et al., 2000 ; Stephens, 2000 ; Jasra et al., 2005 ; Sperrin et al., 2010). Hence, in variable-dimensional settings, the summarization has often been carried out by first selecting a model with the highest posterior probability (i.e, using the BMS approach) and then, applying the relabeling algorithms on the fixed-dimensional conditional posterior distribution. Moreover, we have shown that using the BMS approach results in loosing information from the discarded models and ignoring the uncertainties about the presence of components.

In Chapter 2, we have proposed a novel approach for relabeling and summarizing posterior distributions defined over union of subspaces of differing dimension. The proposed approach consists in approximating the posterior distribution by an original variable-dimensional parametric model. The proposed approach can be regarded as a continuation of the work initiated by Stephens (2000). There have been several challenges that we have solved towards obtaining an applicable approach.

The first challenge has been to develop an algorithm with reasonable computational burden, to deal with problems where the number k of components is moderate to large. More precisely, the EM-type relabeling algorithms developed in, for example, Celeux et al. (2000) ; Stephens (2000) ; Sperrin et al. (2010) ; Yao (2011), are all computationally prohibitive when $k \geq 10$, owing to the computation of an expensive summation over the latent variables in the E-step. Alternatively, we proposed SEM-type algorithms to estimate the parameters of the model. To this end, we designed an I-MH sampler to generate samples from the conditional posterior distribution of the latent variables, i.e., the allocation vectors in our problem, in the S-step. Note that Sperrin et al. (2010) has also proposed a SEM-type algorithm in a fixed-dimensional setting. But, since they work with the normalized conditional posterior distribution of the latent variables, i.e., they compute the computationally expensive summation, their algorithm cannot be applied when k is large. We have shown that the SEM-type algorithm we proposed can be efficiently used in the case where $k = 30$.

The second challenge we have encountered has been the sensitivity of maximum likelihood-type estimators derived from minimizing the KL divergence to the observed outliers. This issue has also been mentioned as future work in the recent paper of Yao (2011). To robustify the algorithms, we have proposed solutions in both the modeling and the parameter estimation stages. In the former we equipped the parametric model with a Poisson point process component to capture the observed outliers. In the latter we proposed modifications by either using robust estimators in the M-step or employing a more robust divergence measure (specifically, we have used the α-divergence proposed by Basu et al. (1998)). The resulting algorithms have desirable robustness properties.

The efficiency of the proposed approach, both for summarizing and for relabeling variable-dimensional posterior distributions, has been illustrated on two problems: joint detection and estimation of sinusoidal components observed in white Gaussian noise (Chapter 3) and joint detection and estimation of muons in the Auger project (Chapter 4). Most notably, the proposed approach 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, using the proposed approach can provide the user with more insight concerning not only the component-specific parameters but also the uncertainties about their presence in the model. The presented results have confirmed that the estimated probabilities of presence are meaningful and can be used to derive estimators with good frequentist properties.

We believe that the proposed approach will also be fruitful in other similar variable-dimensional problems, including mixture model analysis and change point detection problems.

Future work

Concerning the SEM-type algorithms we proposed for relabeling and summarizing variable-dimensional posterior distributions, we consider the following potential areas of future work.

Online selection of L: In Section 4.3, we proposed the advanced initialization procedure whereby, after selecting the maximum number of components L_{max}, the parametric model has been built from scratch; that is, the Gaussian components were added to it progressively until $L = L_{max}$. We selected L_{max} by inspecting the posteriors of k and sorted component-specific parameters and called it a guess value. Then, we used the residuals of the fitted model to see whether the chosen value of L had been "appropriate". As a future work, one can develop a SEM-type algorithm in which the number L of components would be selected online. The distribution of the residuals and the probabilities of presence, for instance, could be used to decide when to increase or decrease L.

Theoretical convergence analysis of SEM-type algorithms: Although in the presented results of Chapter 3 and Chapter 4, we have empirically assessed the convergence of the proposed SEM-type algorithms, we did not provide mathematical convergence results. The main issues refraining us from using the results corresponding to the usual SEM algorithm in the literature (see, e.g., Nielsen, 2000a) are the correlated observed samples, i.e., the samples generated by the RJ-MCMC sampler, and the I-MH sampler used to draw the latent variables (i.e., the allocation vectors). Moreover, recall that we recommended to use the robustified SEM-type algorithms, i.e., TAP-KL2 and TAP-BHHJ. Providing convergence results for TAP-KL2 is expected to be harder than for TAP-BHHJ, however, as in the M-step of TAP-KL2 the KL divergence is not exactly minimized.

Adaptive RJ-MCMC sampler: One of the most appealing perspectives of the approach we proposed for relabeling and summarizing variable-dimensional posterior distributions is to design adaptive or automatic RJ-MCMC samplers for the problems where the posterior distribution is invariant to the permutation of components labels (and thus exhibits label-switching).

In the literature, there have been a few attempts at improving the performance of the RJ-MCMC sampler by learning the parameters of the between-models moves' proposal distributions (see, e.g., Brooks et al., 2003 ; Green, 2003 ; Hastie, 2005 ; Fan et al., 2009 ; Hastie and Green, 2011). The proposed approaches so far, however, cannot be applied to the problems with label-switching. To deal with such problems, Bardenet et al. (2012) has developed an RJ-MCMC sampler whereby only the with-in model moves has been adapted.

Since the fitted parametric model is a close approximation of the target posterior distribution, we believe that it can be used to design both with-in model and between models proposal distributions and the resulting adaptive RJ-MCMC sampler would enjoy a promising performance on signal decomposition and similar problems.

Appendix A

Optimization of the BHHJ $\boldsymbol{\alpha}$-divergence

A.1 Computation of the integral term in the BHHJ $\boldsymbol{\alpha}$-divergence

A.1.1 Option 1: Monte Carlo approximation

Let $P_{\boldsymbol{\eta}}^{\mathbf{z}}$ denote the marginal distribution of $\mathbf{z}$ when $(\boldsymbol{x}, \mathbf{z}) \sim Q_{\boldsymbol{\eta}}^{\mathcal{L}}$ and let $\tilde{\mathbf{z}}^{(1)}, \ldots, \tilde{\mathbf{z}}^{(\tilde{M})}$ denote iid draws from $P_{\boldsymbol{\eta}}^{\mathbf{z}}$. Then, the integral (2.31) can be approximated as

$$\int_{\mathbb{X}_{\mathcal{L}}} q_{\boldsymbol{\eta}}^{1+\alpha} \, \mathrm{d}\rho_{\mathcal{L}} \;\approx\; \frac{1}{\tilde{M}} \sum_{m=1}^{\tilde{M}} q_{\boldsymbol{\eta}}^{\alpha}(\tilde{\mathbf{z}}^{(m)}) \prod_{j=1}^{\tilde{k}^{(m)}} \tilde{q}_{\boldsymbol{\eta}}^{\alpha}(\tilde{\mathbf{z}}_j^{(m)}) \,,$$

where $\tilde{k}^{(m)}$ denote the length of $\tilde{z}^{(m)}$. The criterion $\hat{\mathfrak{J}}_M^{\alpha}(\boldsymbol{\eta})$ defined in (2.29) can be approximated as

$$\begin{aligned}\hat{\mathfrak{J}}_M^{\alpha}(\boldsymbol{\eta}) \;\approx\; & \frac{1}{\tilde{M}} \sum_{m=1}^{\tilde{M}} q_{\boldsymbol{\eta}}^{\alpha}(\tilde{\mathbf{z}}^{(m)}) \prod_{j=1}^{\tilde{k}^{(m)}} \tilde{q}_{\boldsymbol{\eta}}^{\alpha}(\tilde{\mathbf{z}}_j^{(m)}) \\ & - \left(1 + \frac{1}{\alpha}\right) \cdot \frac{1}{M} \sum_{i=1}^{M} q_{\boldsymbol{\eta}}^{\alpha}(\mathbf{z}^{(i)}) \prod_{j=1}^{k^{(i)}} q_{\boldsymbol{\eta}}^{\alpha}(\boldsymbol{x}_j^{(i)} \,|\, \mathbf{z}_j^{(i)}) \,. \end{aligned} \tag{A.1}$$

Note that we cannot re-use the allocation vectors $\mathbf{z}^{(i)}$ of the vector of the observed samples $\boldsymbol{x}^{(i)}$, with $i = 1, \ldots, M$, in the first term; indeed, they don't have the correct marginal distribution. For computing partial derivatives of (A.1), we can proceed as in Section 2.5.3.

A.1.2 Option 2: "exact" computation

Let us consider once more the first term in the BHHJ α-divergence:

$$\begin{aligned}
\int_{\mathbb{X}_{\mathcal{L}}} q_{\boldsymbol{\eta}}^{1+\alpha}\,\mathrm{d}\rho_{\mathcal{L}} &= \sum_{k\geq 0}\sum_{\mathbf{z}\in\mathcal{Z}} q_{\boldsymbol{\eta}}^{1+\alpha}(\mathbf{z}) \prod_{j=1}^{k} \tilde{q}_{\boldsymbol{\eta}}^{\alpha}(\mathbf{z}_j)\\
&= \sum_{k\geq 0}\sum_{\mathbf{z}\in\mathcal{Z}} q_{\boldsymbol{\eta}}^{1+\alpha}(\mathbf{z})\ |\boldsymbol{\Theta}|^{-\alpha n_{L+1}} \prod_{l=1}^{L} \tilde{q}_{\boldsymbol{\eta}}^{\alpha n_l}(l)\\
&= \sum_{b\geq 0}\ \sum_{\boldsymbol{n}\in\{0,1\}^L} C_{b+|\boldsymbol{n}|}^{b}\ |\boldsymbol{n}|!\ q_{\boldsymbol{\eta}}^{1+\alpha}(\mathbf{z})\ |\boldsymbol{\Theta}|^{-\alpha b} \prod_{l=1}^{L} \tilde{q}_{\boldsymbol{\eta}}^{\alpha n_l}(l)\,,
\end{aligned}$$

where $|\boldsymbol{n}| = \sum_{l=1}^{L} n_l$ and $\mathbf{z}$ denotes any allocation vector with $n_{L+1} = b$ for the Poisson point process component and $\boldsymbol{n} = (n_1, \ldots, n_L) \in \{0,1\}^L$ for the other Gaussian components (the order does not matter). Using expression of $q_{\boldsymbol{\eta}}(\mathbf{z})$ (2.8), we obtain

$$\begin{aligned}
\int_{\mathbb{X}_{\mathcal{L}}} q_{\boldsymbol{\eta}}^{1+\alpha}\,\mathrm{d}\rho_{\mathcal{L}} &= \sum_{b\geq 0}\ \sum_{\boldsymbol{n}\in\{0,1\}^L} \frac{C_{b+|\boldsymbol{n}|}^{b}\ |\boldsymbol{n}|!}{((b+|\boldsymbol{n}|)!)^{1+\alpha}}\ e^{-\lambda(1+\alpha)}\ \left(\lambda^{1+\alpha}\,|\boldsymbol{\Theta}|^{-\alpha}\right)^{b}\\
&\qquad \prod_{l=1}^{L} \left(\pi_l^{1+\alpha}\ \tilde{q}_{\boldsymbol{\eta}}^{\alpha}(l)\right)^{n_l}\ \left((1-\pi_l)^{1+\alpha}\right)^{1-n_l}\,,
\end{aligned}$$

and thus

$$\int_{\mathbb{X}_{\mathcal{L}}} q_{\boldsymbol{\eta}}^{1+\alpha}\,\mathrm{d}\rho_{\mathcal{L}} = e^{-\lambda(1+\alpha)} \sum_{\boldsymbol{n}\in\{0,1\}^L} \phi\left(|\boldsymbol{n}|\right) \prod_{l=1}^{L} \psi_{n_l,l} \tag{A.2}$$

with

$$\begin{aligned}
\phi(m) &= \sum_{t\geq 0} \frac{u^t}{t!\,((t+m)!)^{\alpha}}\,, \qquad u = \lambda^{1+\alpha}\,|\boldsymbol{\Theta}|^{-\alpha}\,,\\
\psi_{0,l} &= (1-\pi_l)^{1+\alpha} \quad \text{and} \quad \psi_{1,l} = \pi_l^{1+\alpha}\ \tilde{q}_{\boldsymbol{\eta}}^{\alpha}(l)\,.
\end{aligned}$$

Remark A.1. A few remarks about Equation (A.2) :

i) The series $\phi(m)$ converges very fast (faster than the exponential series) and therefore can be computed very precisely using a small number of terms. Moreover, only L+1 values of this functions are required ($m = 0, 1, \ldots, L$).

ii) The sum has 2^L terms : it is no longer infinite and can be implemented efficiently. In a naive implementation, both the computation time and the memory requirement grow exponentially with L.

iii) Taking advantage of the special structure of this sum, we could devise a recursive implementation that has a linear memory requirement and a slightly better—but still exponential—computational cost. But for larger values of L the Monte Carlo approximation is a more feasible approach.

Appendix B

Prior specification for the detection and estimation of sinusoids in Gaussian white noise

This appendix contains two papers addressing the issue of the prior specification over the signal-to-noise ratio hyperparameter, i.e., δ^2, in the problem of joint Bayesian detection and estimation of sinusoidal components in white Gaussian noise:

i) Alireza Roodaki, Julien Bect, and Gilles Fleury. An empirical Bayes approach for joint Bayesian model selection and estimation of sinusoids via reversible jump MCMC. In: *European signal Processing Conference (EUSIPCO'10), Aalborg, Denmark*, 2010.

ii) Alireza Roodaki, Julien Bect, and Gilles Fleury. On the joint Bayesian model selection and estimation of sinusoids via reversible jump MCMC in low SNR situations. In: 10^{th} *International Conference on Information Sciences, Signal Processing and their Applications (ISSPA'10) Kuala Lumpur, Malaysia*, 2010.

Assigning a weakly-informative conjugate Inverse Gamma prior, i.e., $\mathcal{IG}(\alpha_{\delta^2} = 2, \beta_{\delta^2})$, over δ^2, as recommended in Andrieu and Doucet (1999), the results provided in the above papers reveal that the value of its scale parameter has a significant influence on 1) the mixing rate of the Markov chain and 2) the posterior distribution of the number k of components. In i), we investigated an Empirical Bayes approach to select an appropriate value for this hyperparameter in a data-driven way. In ii), we took a different approach and used a truncated Jeffreys prior. However, both approaches failed in low SNR situations, while in high SNR situations the sensitivity to β_{δ^2} is negligible. In Section 3.2.5 of the present document, we propose instead to assess the sensitivity of the posterior distribution to β_{δ^2} using an SMC sampler, following an idea of Bornn et al. (2010).

ON THE JOINT BAYESIAN MODEL SELECTION AND ESTIMATION OF SINUSOIDS VIA REVERSIBLE JUMP MCMC IN LOW SNR SITUATIONS

Alireza Roodaki, Julien Bect and Gilles Fleury

Department of Signal Processing and Electronic Systems,
SUPELEC, Gif-sur-Yvette, France.

ABSTRACT

This paper addresses the behavior in low SNR situations of the algorithm proposed by Andrieu and Doucet (IEEE T. Signal Proces., 47(10), 1999) for the joint Bayesian model selection and estimation of sinusoids in Gaussian white noise. It is shown that the value of a certain hyperparameter, claimed to be weakly influential in the original paper, becomes in fact quite important in this context. This robustness issue is fixed by a suitable modification of the prior distribution, based on model selection considerations. Numerical experiments show that the resulting algorithm is more robust to the value of its hyperparameters.

Index Terms— Bayesian model selection; reversible jump MCMC; prior calibration; Bayesian sensitivity analysis; spectral analysis.

1. INTRODUCTION

Detection and separation of signals in low SNR conditions has many applications in various fields such as communication, radar and sonar—to name but a few. Moreover, sinusoids are one of the most common kind of signals used in these applications. The problem of joint detection and estimation of sinusoids in low SNR situations, assuming unknown number of components, is therefore of general importance.

A fully Bayesian algorithm based on Reversible Jump Markov Chain Monte Carlo (RJ-MCMC) technique [1] for handling this problem, not specifically in low SNR situations, has been proposed in [2]. This algorithm, of course with appropriate modifications, has been used for other applications such as polyphonic signal analysis [3], array signal processing [4], and nuclear emission spectra analysis [5]. However, to the best of our knowledge, the behavior of this algorithm in low SNR situations has never been studied. To present the problem more explicitly, in the following we will introduce the notations used in the algorithm.

Let $\mathbf{y} = (y_1,\, y_2,\, \ldots,\, y_N)^t$ be a vector of N independent observations. Based on the model $\mathcal{M}_k$ (for $k = 0,\, 1,\, \ldots,\, k_{\text{max}}$), $\mathbf{y}$ can be represented by summation of k sinusoids together with a white Gaussian noise. Defining the $N \times 2k$ matrix containing the sinusoids with different radial frequencies, $\mathbf{D}_k$, as below

$$\mathbf{D}_k(i{+}1, 2j{-}1) \triangleq \cos(\omega_{j,k} i), \mathbf{D}_k(i{+}1, 2j) \triangleq \sin(\omega_{j,k} i)$$

for $i = 0, \ldots, N-1$ and $j = 1, \ldots, k$, one can write the normal linear regression model for the current problem with k components:

$$\mathbf{y} = \mathbf{D}_k.\mathbf{a}_k + \mathbf{n},$$

where $\mathbf{n}$ is the white Gaussian noise of variance σ^2. The unknown parameters are assumed to be the number of components k and $\boldsymbol{\theta}_k = \{\mathbf{a}_k, \boldsymbol{\omega}_k, \sigma^2\}$.

As in many Bayesian model selection approaches for normal linear regression problem, the well-known conditionally conjugate g-prior [6, 7, 8], which provides tractable computations, has been assigned as a prior over the amplitudes in the model proposed in [2]. The g-prior is a zero mean multivariate normal distribution with $\sigma^2/g(\mathbf{D}_k^t \mathbf{D}_k)^{-1}$ as its covariance matrix. The variable called g controls the expected size of the amplitudes. This parameter has been substituted by δ^{-2} in [2] and δ^2 has been called the Expected SNR (ESNR).

Owing to the influence of the ESNR on the performance of the algorithm, particularly in the Bayesian model selection part, several approaches for setting or estimating it have been proposed in the variable selection literature; see [7, 8, 9] and references therein. To keep the Fully Bayesian spirit, a vague conjugate Inverse-Gamma ($\mathcal{IG}$) prior has been assigned over ESNR in [2], i.e. $p\left(\delta^2|\alpha_{\delta^2}, \beta_{\delta^2}\right) = \mathcal{IG}\left(\,\cdot\,|\alpha_{\delta^2}, \beta_{\delta^2}\right)$. Although it was mentioned that the performance of the proposed algorithm is not sensitive to the value of the scale parameter β_{δ^2}, our experiments have shown that this parameter becomes influential when dealing with low SNR signals.

The structure of this article is as follows. Section 2 briefly recalls the Bayesian algorithm proposed in [2]. Section 3 discusses first the "dimensionality penalty" induced by the hyperparameter δ^2 and then the effect of β_{δ^2} on the posterior distribution of k and δ^2. Section 4 discusses solutions to the problem of choosing β_{δ^2}: since the usual data-driven approaches fail in low SNR situations, we propose to use a truncated Jeffrey prior instead. Section 5 presents numerical results that support the proposed method and discusses its sensitivity to the lower bound δ^2_{min} of the truncated prior. Finally, Section 6 concludes the article and addresses possible future works.

2. BAYESIAN FRAMEWORK

The full joint distribution of the observed signal and the unknown parameters, in the model proposed by [2], has the following hierarchical structure:

$$\begin{aligned} p(\mathbf{y}, k, \boldsymbol{\theta}_k, \delta^2) = p(\mathbf{y}\,|\,k, \boldsymbol{\theta}_k)\, p(\boldsymbol{\theta}_k\,|\,k, \delta^2) \\ \times\, p(k)\, p(\delta^2). \end{aligned} \tag{1}$$

2.1. Prior distributions

As proposed by [2], the prior over k is a Poisson distribution with mean Λ, truncated to $\{0, 1, \ldots, k_{\max}\}$. Conditional on k, the ω_k's are independent and identically distributed, with a uniform distribution on $(0, \pi)$. The noise variance σ^2 is endowed with Jeffrey's uninformative prior, i.e. $p(\sigma^2) \propto 1/\sigma^2$, where the symbol $\propto$ denotes proportionality.

Furthermore, they have suggested to assign a conjugate $\mathcal{IG}(\alpha_{\delta^2}, \beta_{\delta^2})$ prior over ESNR and to set α_{δ^2} to two for having an infinite variance. However, as it can be seen in Figure 1, the posterior over δ^2 is severely sensitive to the value of β_{δ^2}.

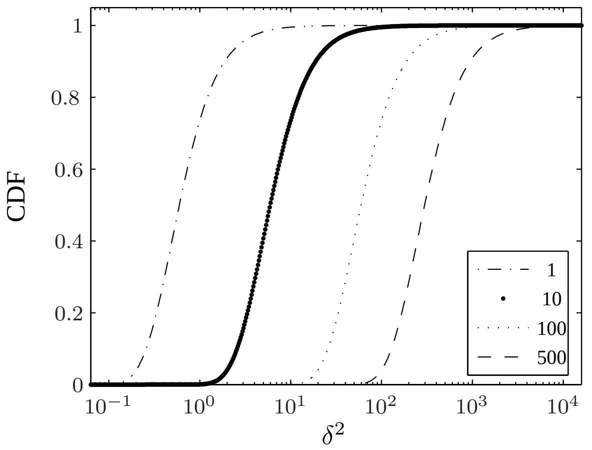


Fig. 1: CDFs of priors over δ^2 for different values of β_{δ^2}.

The hyperparameter Λ has been assigned in [2] a Gamma prior, i.e. $p(\Lambda) = \mathcal{G}(\alpha_\Lambda, \beta_\Lambda)$, with $\alpha_\Lambda \approx \frac{1}{2}$ as a shape parameter and $\beta_\Lambda \approx 0$ as a scale parameter. This is equivalent to using a negative binomial prior over k that puts more emphasis on small values. In this paper, in order to have an almost flat prior over k, the parameter α_Λ is set to a value close to 1.

2.2. Sampling structure

Based on (1) and Bayes Theorem, after simply integrating $\mathbf{a}_k$ and σ^2 out, the joint posterior distribution of k and $\boldsymbol{\omega}_k$, up to a normalizing constant, can be written as

$$p\left(k, \boldsymbol{\omega}_k, \delta^2, \Lambda \,|\, \mathbf{y}\right) \propto (\mathbf{y}^t \mathbf{P}_k \mathbf{y})^{-N/2} \frac{\Lambda^k \pi^{-k}}{k!\,(\delta^2+1)^k} \times \mathbb{1}_{(0,\pi)^k}(\boldsymbol{\omega}_k)\, p(\delta^2)\, p(\Lambda)\,, \quad (2)$$

with

$$\mathbf{P}_k = \mathbf{I}_N - \frac{\delta^2}{1+\delta^2} \mathbf{D}_k \left(\mathbf{D}_k^t \mathbf{D}_k\right)^{-1} \mathbf{D}_k^t. \quad (3)$$

In the following, different steps for sampling from the above distribution are briefly described. For more detailed expressions, please refer to [1, 2].

The sampler consists of a Metropolis-Hastings (MH) move for the target density (2), which updates the values of k and $\boldsymbol{\omega}_k$, followed by a sequence of Gibbs moves to update δ^2 and Λ. The proposal kernel, in the MH step, is a mixture of within-model moves, which update the radial frequencies without changing k, and between-models moves, which change the value of k by adding or removing a component (so-called birth/death move). The Gibbs move for δ^2 if performed by demarginalization of σ^2 and $\mathbf{a}_k$ and then sampling from the "uncollapsed" posterior of δ^2.

Except for a modification in the birth/death ratio, the moves implemented in our sampler are the same as in [2]. In the birth move, after proposing a new component by sampling its radial frequency from $U(0, \pi)$, it is *randomly* located among the previous components. Then, the move is accepted with probability $\alpha_{birth} = \min\{1, r_{birth}\}$, where

$$r_{birth} = \left(\frac{\mathbf{y}^t \mathbf{P}_{k+1} \mathbf{y}}{\mathbf{y}^t \mathbf{P}_k \mathbf{y}}\right)^{-N/2} \frac{1}{1+\delta^2}. \quad (4)$$

One should note that the birth ratio (4) differs from the one reported in [2] by a multiplicative factor of $1/(k+1)$. A similar mistake for a similar algorithm has been found in the field of genetics [10]. Note that using the ratio given in [2] amounts to changing the prior distribution on k. This issue will be dealt with in greater detail in a forthcoming paper. In the meantime, the reader is referred to [11] for more information on the role of permutations and sorting in the computation of RJ-MCMC ratios.

3. SENSITIVITY TO THE VALUE OF β_{δ^2}

In this section, the effect of β_{δ^2} on the performance of the algorithm in low SNR situations is discussed.

To better understand the importance of β_{δ^2}, the role of δ^2 will be discussed first, following the ideas introduced in [9, 12] to make a connection between Bayesian algorithms and model selection criteria. Let us assume, for the sake of simplicity, a flat prior over the number of components. Then, the log-posterior can be written as

$$\log p\left(k, \boldsymbol{\omega}_k \,|\, \mathbf{y}, \delta^2\right) = -\frac{N}{2} \log\left(\mathbf{y}^t \mathbf{P}_k \mathbf{y}\right) - F \cdot k + C, \quad (5)$$

where $F = \log\left(\pi\left(1+\delta^2\right)\right)$ and C is a constant which does not depend on k and $\boldsymbol{\omega}_k$. F can be interpreted as a dimensionality penalty, which penalizes complex models. Therefore, large values of δ^2, which result in large values of F, cause the algorithm to neglect small components with respect to the noise. Conversely, "small" values of δ^2 result in an algorithm which does not penalize enough "small" components and leads to overfitting.

In addition to—and partly because of—its role in the model selection properties of the algorithm, the value of δ^2 has a strong influence on the behavior of the resulting algorithm. For low values of δ^2, the Markov chain has to visit much more often regions of the state space corresponding to high values of k, where the algorithmic complexity of running the chain is much higher. For high values of δ^2, the posterior distribution has sharper peaks and valleys, which makes it much more difficult for the chain to explore, resulting in a slower convergence rate.

Turning to the role of β_{δ^2}, first, one should note that the $\mathcal{IG}$ prior used in [2], although chosen to be weakly informative, is not really "vague" (see Figure 1). In fact, it has a mode at $\beta_{\delta^2}/(\alpha_{\delta^2}+1)$. By changing its scale parameter the behavior of the algorithm can be controlled just like changing the values of δ^2 itself, esp. in the low SNR situations where likelihood does not provide much information about δ^2. Figure 2 displays the sensitivity of the posteriors of k and δ^2 to the hyperparameter β_{δ^2} in an experiment of signal detection under $\mathcal{M}_1$ with $\mathrm{SNR} = -1\,\mathrm{dB}$, which is not very low. In this study, SNR is defined as $\|\mathbf{D}_k\mathbf{a}_k\|^2 / \left(N\sigma^2\right)$. It can be seen in this figure that the posterior of δ^2 is moving to the right by increasing the value of β_{δ^2}. Moreover, if one is interested in model selection based on the maximum of the posterior of the number of components, i.e. $\arg\max_{k\in\{0,\cdots,k_{\max}\}} p(k\,|\,\mathbf{y})$, the selected models under $\beta_{\delta^2}=1$, $\beta_{\delta^2}=10$, and $\beta_{\delta^2}=100$ would be $\mathcal{M}_2$, $\mathcal{M}_2$, and $\mathcal{M}_1$, respectively. The differences in the results for Bayesian model averaging (not shown in this paper) are even more important.

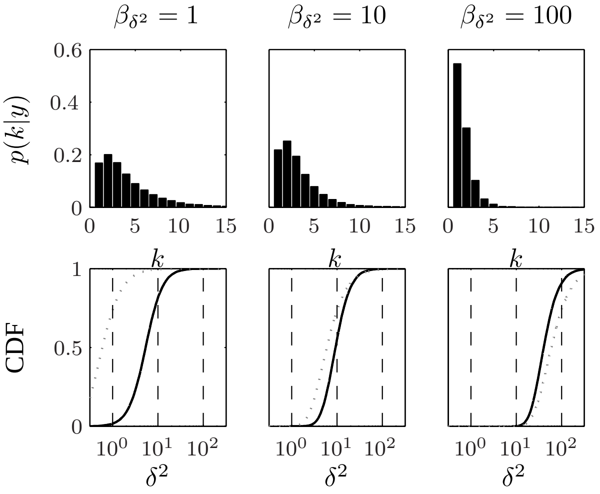


Fig. 2: The posteriors of k and δ^2 under the experiment of signal detection with $\mathrm{SNR} = -1\,\mathrm{dB}$ and different values of β_{δ^2}. In the second row, the gray dotted lines show the prior and the black lines show the posterior of δ^2. The length of the chain was set to 100k, with a burn-in period of 20k samples.

4. PROPOSED METHODS

In the following possible methods for either estimating a reasonable value for β_{δ^2} from the observed data or stabilizing the algorithm by modifying the prior are introduced.

4.1. Data-driven methods

In order to estimate a proper value for β_{δ^2} the first two approaches that may come to mind are the Fully Bayesian and the Empirical Bayes (EB) methods. The former one is constructed by assigning a vague conjugate Gamma prior over β_{δ^2}, that is, $\beta_{\delta^2} \sim \mathcal{G}(a,b)$. Then, one can update it by performing a Gibbs move with $\mathcal{G}(a+\alpha_{\delta^2}, b+\delta^{-2})$ as proposal distribution. On the other hand, the EB method is a data-driven approach in which the marginal likelihood of the parameter given the data, i.e. $p\left(\mathbf{y}\,|\,\beta_{\delta^2}\right)$, is maximized. This idea has been used in [7, 9, 12] for estimating δ^2. However, since in this problem, $p\left(\mathbf{y}\,|\,\beta_{\delta^2}\right)$ does not exist in closed form, one should use Monte Carlo methods to estimate β_{δ^2} as in [13].

4.2. Using a truncated Jeffrey prior over δ^2

The idea of using an improper Jeffrey prior over ESNR, which provides a flat prior over the $\log\left(\delta^2\right)$ in contrary to the current prior, has been mentioned in [2] but it is not used as $\delta^2 = 0$ would become an absorbing state of the Markov chain. Here, we propose to truncate the Jeffrey prior using a lower bound δ^2_{min} and an upper bound δ^2_{max}. The sensitivity of the algorithm to δ^2_{max} can be reduced by setting it to a large value, say 10000. However, choosing the value of the lower bound is less trivial, since it controls the minimal dimensionality penalty induced by the prior; a numerical sensitivity analysis will be carried out in the next section.

5. SIMULATION RESULTS AND DISCUSSION

In this section, we study the performance of the proposed solutions for reducing the sensitivity of the Bayesian algorithm to the prior over δ^2. Simulations are carried out with the observed signal of length $N = 64$. In this paper, the problem of signal detection in low SNR situation is considered. The parameters of the single sinusoid are as follows: $\omega_{1,1} = 0.2\pi$, $-\arctan(a_{s_1}/a_{c_1}) = \pi/3$, and $a^2_{s_1} + a^2_{c_1} = 20$. The length of chain in all simulations was 100k, with a burn-in period of 20k samples.

The data-driven approaches estimate a reasonable value for the hyperparameter β_{δ^2} in high SNR situations but do not perform satisfactorily in low SNR situations. In fact, in these situations, our numerical experiments showed that β_{δ^2} is estimated to be very close to 0, which imposes too small δ^2, using both methods. It has also been reported in [7] that the EB method tends to estimate δ^2 as 0 under the null model in a similar framework.

On the other hand, in the case of using a truncated Jeffrey prior over δ^2, the value of δ^2_{min} determines the minimal dimensionality penalty. One should note that, a reasonable range of values for the lower bound is restricted, since having a high minimal penalty is not suitable. Moreover, setting δ^2_{min} to a large value might cause convergence issues. Thus, up to now, we have translated the problem of estimating a proper value for the hyperparameter β_{δ^2} to the problem of finding a reasonable value for δ^2_{min}. In the sequel, the sensitivity of the algorithm to the variations of this parameter is studied.

Figure 3 shows the posterior distributions for k and δ^2 for the same observed signal as Figure 2. As depicted in this figure, the sensitivity of the algorithm to the variations of δ^2_{min} is much less than that of β_{δ^2}. In fact no matter what the value of δ^2_{min} is, the model $\mathcal{M}_1$ would be selected based on the MAP of k. For further studying the sensitivity of the algorithm to the parameter δ^2_{min}, the probabilities of selected models based on $\arg\max p(k\,|\,\mathbf{y})$ in 100 realizations of the sampler for different values of SNR were estimated. Figure 4 shows the sensitivity of the algorithm to this parameter for the cases of $\mathrm{SNR} = -3\,\mathrm{dB}$ and $\mathrm{SNR} = -4\,\mathrm{dB}$. In this figure, the algorithm was run with $\delta^2_{\text{min}} = 0.5$. The probabilities for other values of δ^2_{min} were obtained using importance sampling. This method has already been used for the sensitivity analysis of Bayesian algorithms to their priors; see for instance [14]. It can be concluded from figure 4 that the probabilities are not very sensitive to the choice of δ^2_{min}. However, as the value of the lower bound increases, P_2 decreases while P_0 increases: this was predictable, as δ^2_{min} controls the minimal dimensionality penalty.

6. CONCLUSION

The main contribution of this paper has been to explain the lack of robustness, in low SNR situations, of the algorithm proposed in [2] and to propose solutions for fixing it. Simulation results showed that a truncated Jeffrey prior over δ^2 significantly improves the performance of the sampler in situations where the usual data-driven approaches (Empirical Bayes and Fully Bayes) fail. Sensitivity analyses, which are efficiently carried out using importance sampling, reveal that the resulting algorithm is rather robust to variations of the lower bound δ^2_{min} in a reasonable range. A natural direction for future work would be to propose a data-driven approach for the automatic selection of this threshold and to assess more systematically the performances of this algorithm.

References

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

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

[3] M. Davy, S. J. Godsill, and J. Idier, "Bayesian analysis of polyphonic western tonal music," *J. Acoust. Soc. Am.*, vol. 119, pp. 2498–2517, 2006.

[4] J. R. Larocque and J. P. Reilly, "Reversible jump MCMC for joint detection and estimation of sources in coloured noise," *IEEE T. Signal Proces.*, vol. 50, pp. 231–240, 2000.

[5] S. Gulam Razul, W. Fitzgerald, and C. Andrieu, "Bayesian model selection and parameter estimation of nuclear emission spectra using RJMCMC," *Nucl. Instrum. Meth. A*, vol. 497, no. 2-3, pp. 492–510, 2003.

[6] A. Zellner, "On assessing prior distributions and Bayesian regression analysis with g-prior distributions," *Bayesian Inference and Decision Techniques: Essays in Honor of Bruno de Finetti, (eds. P. K. Goel and A. Zellner)*, pp. 233–243, 1986.

[7] F. Liang, R. Paulo, G. Molina, M. Clyde, and J. Berger, "Mixtures of g-priors for Bayesian variable selection," *J. Am. Stat. Assoc.*, vol. 103, no. 481, pp. 410–423, 2008.

[8] C. Fernández, E. Ley, and M. Steel, "Benchmark priors for Bayesian model averaging," *J. Econometrics*, vol. 100, no. 2, pp. 381–427, 2001.

[9] E. I. George and D. P. Foster, "Calibration and empirical Bayes variable selection," *Biometrika*, vol. 87, no. 4, pp. 731–747, 2000.

[10] J. Jannink and R. Fernando, "On the Metropolis-Hastings acceptance probability to add or drop a quantitative trait locus in Markov chain Monte Carlo-based Bayesian analyses," *Genetics*, vol. 166, no. 1, pp. 641–643, 2004.

[11] S. Richardson and P. J. Green, "On Bayesian analysis of mixtures with an unknown number of components," *J. Roy. Stat. Soc. B Met.*, vol. 59, no. 4, pp. 731–792, 1997.

[12] W. Cui and E. I. George, "Empirical Bayes vs. fully Bayes variable selection," *J. Stat. Plan. Infer.*, vol. 138, no. 4, pp. 888–900, 2008.

[13] R. A. Levine and G. Casella, "Implementations of the Monte Carlo EM algorithm," *J. Comput. Graph. Stat.*, pp. 422–439, 2001.

[14] J. Besag, P. Green, D. Higdon, and K. Mengersen, "Bayesian computation and stochastic systems (with discussion)," *Statistical Science*, vol. 10, no. 1, pp. 3–41, 1995.

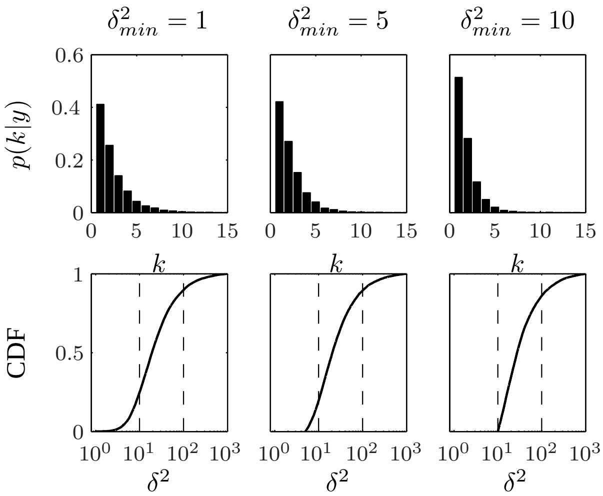
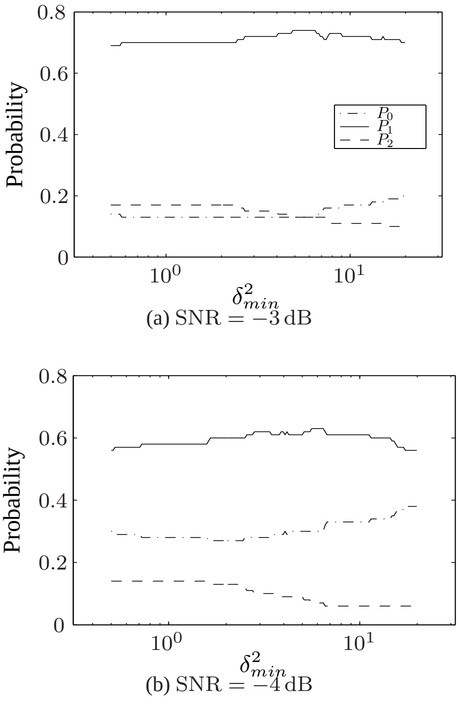


Fig. 3: The posteriors of k and δ^2 under the experiment of signal detection with $\text{SNR} = -1\,\text{dB}$ and different values of δ^2_{min}.



(a) $\text{SNR} = -3\,\text{dB}$

(b) $\text{SNR} = -4\,\text{dB}$

Fig. 4: Probabilities of $\arg\max p(k \mid \mathbf{y}) = 0$, $\arg\max p(k \mid \mathbf{y}) = 1$, and $\arg\max p(k \mid \mathbf{y}) \geq 2$ are denoted, respectively, by P_0, P_1, and P_2 in 100 realization of the algorithm using $\delta^2_{\text{min}} = 0.5$. The probabilities for other values of δ^2_{min}, i.e. $\delta^2_{\text{min}} \in (0.5, 20]$, are estimated using the importance sampling method.

AN EMPIRICAL BAYES APPROACH FOR JOINT BAYESIAN MODEL SELECTION AND ESTIMATION OF SINUSOIDS VIA REVERSIBLE JUMP MCMC

Alireza Roodaki, Julien Bect, and Gilles Fleury

E3S — SUPELEC Systems Sciences
Dept. of Signal Processing and Electronic Systems, SUPELEC, Gif-sur-Yvette, France.
Email: {alireza.roodaki, julien.bect, gilles.fleury}@supelec.fr

ABSTRACT

This paper addresses the sensitivity of the algorithm proposed by Andrieu and Doucet (IEEE Trans. Signal Process., 47(10), 1999), for the joint Bayesian model selection and estimation of sinusoids in white Gaussian noise, to the values of a certain hyperparameter claimed to be weakly influential in the original paper. A deeper study of this issue reveals indeed that the value of this hyperparameter (the scale parameter of the expected signal-to-noise ratio) has a significant influence on 1) the mixing rate of the Markov chain and 2) the posterior distribution of the number of components. As a possible workaround for this problem, we investigate an Empirical Bayes approach to select an appropriate value for this hyperparameter in a data-driven way. Marginal likelihood maximization is performed by means of an importance sampling based Monte Carlo EM (MCEM) algorithm. Numerical experiments illustrate that the sampler equipped with this MCEM procedure provides satisfactory performances in moderate to high SNR situations.

1. INTRODUCTION

In this paper, we address the problem of detection and estimation of sinusoids in white Gaussian noise, assuming that the number of component is unknown. A fully Bayesian algorithm, based on the Reversible Jump Markov Chain Monte Carlo (RJ-MCMC) technique [8, 9], has been proposed for this problem in [1]. Similar algorithms have also been used for other applications such as polyphonic signal analysis [3], array signal processing [12], and nuclear emission spectra analysis [10]. However, to the best of our knowledge, the sensitivity of the algorithm to the value of its hyperparameters has never been clearly discussed.

Let $\mathbf{y} = (y_1, y_2, \ldots, y_N)^t$ be a vector of N observations of an observed signal. We consider the finite family of embedded models $\{\mathscr{M}_k,\ 0 \le k \le k_{\max}\}$, where $\mathscr{M}_k$ assumes that $\mathbf{y}$ can be written as a linear combination of k sinusoids observed in white Gaussian noise. Let $\boldsymbol{\omega}_k = \left(\omega_{1,k}, \ldots, \omega_{k,k}\right)$ be the vector of radial frequencies in model $\mathscr{M}_k$, and let $\mathbf{D}_k$ be the corresponding $N \times 2k$ design matrix defined by

$$\mathbf{D}_k(i+1, 2j-1) \triangleq \cos(\omega_{j,k} i), \quad \mathbf{D}_k(i+1, 2j) \triangleq \sin(\omega_{j,k} i)$$

for $i = 0, \ldots, N-1$ and $j = 1, \ldots, k$. Then the observed signal $\mathbf{y}$ follows under $\mathscr{M}_k$ a normal linear regression model:

$$\mathbf{y} = \mathbf{D}_k.\mathbf{a}_k + \mathbf{n},$$

where $\mathbf{n}$ is a white Gaussian noise with variance σ^2. The unknown parameters are assumed to be the number of components k and $\boldsymbol{\theta}_k = \{\mathbf{a}_k, \boldsymbol{\omega}_k, \sigma^2\}$.

Assuming that no (or little) information is available about the vector of amplitudes $\mathbf{a}_k$, the conditionally conjugate g-prior is usually recommended as a default prior in the Bayesian variable selection literature [14, 21]. Under this prior, the distribution of $\mathbf{a}_k$ conditionally to σ^2, k and $\boldsymbol{\omega}_k$ is Gaussian with $\sigma^2/g\ (\mathbf{D}_k^t \mathbf{D}_k)^{-1}$ as its covariance matrix, where g is a positive parameter. Following [1], a zero-mean g-prior for $\mathbf{a}_k$ will be used in this paper. Our results, however, are likely to remain relevant for any covariance matrix of the form $\sigma^2/g\ \Sigma_k$ (with Σ_k possibly depending on k and $\boldsymbol{\omega}_k$).

The parameter $\delta^2 = 1/g$, called the Expected SNR (ESNR), controls the expected size of the amplitudes. Owing to its influence on the performance of the algorithm, and assuming again that no (or little) information is available, the hyperparameter δ^2 is given in [1] a conjugate inverse gamma prior with parameters α_{δ^2} and β_{δ^2}, that we denote by $\mathscr{IG}(\alpha_{\delta^2}, \beta_{\delta^2})$. Such a hierarchical Bayes approach is usually hoped to increase the robustness of the statistical analysis; see [18, Section 10.2] for more information. The first parameter is set to $\alpha_{\delta^2} = 2$, in order to have an heavy-tailed "weakly informative" prior (with infinite variance). It is claimed in [1, Section V.D] that the value of β_{δ^2} has a weak influence on the performance of the algorithm.

The contribution of this paper, which can be seen as a continuation of [1], is twofold. First, on the basis of extensive numerical experiments, we argue that the value of β_{δ^2} can have a strong influence on 1) the mixing rate of the Markov chain and 2) the posterior distribution of the number of components. Second, instead of using a fixed value for the hyperparameter β_{δ^2}, we investigate the capability of an Empirical Bayes (EB) approach to estimate it from the data, in the spirit of the approach used in [2, 6] to estimate δ^2. More precisely, since the marginal likelihood of β_{δ^2} is not available in closed form, we implement an Importance Sampling (IS) based Monte Carlo Expectation Maximization (MCEM) algorithm [13, 20] to maximize it numerically.

The paper is outlined as follows. Section 2 recalls the hierarchical Bayesian model and the RJ-MCMC sampler proposed in [1]. Section 3 discusses the influence of β_{δ^2} on both the mixing rate of the Markov chain and the posterior distribution of the number k of components. Section 4 explains the fundamentals of the MCEM algorithm, which is used for estimating β_{δ^2}. Section 5 presents the results of our numerical experiments and discusses the pros and cons of the Empirical Bayes approach in estimating β_{δ^2}. Finally, Section 6 concludes the paper and gives directions for future work.

2. BAYESIAN FRAMEWORK

This section describes the prior distribution and the RJ-MCMC sampler considered in this paper, following [1] unless explicitly stated otherwise.

2.1 Prior distributions

The joint prior distribution of the unknown parameters is chosen to have the following hierarchical structure:

$$\begin{aligned} p\left(k, \boldsymbol{\theta}_k, \delta^2\right) \;=\; & p(\mathbf{a}_k \mid k, \boldsymbol{\omega}_k, \sigma^2, \delta^2)\ p(\boldsymbol{\omega}_k \mid k) \\ & \times p(k)\ p(\sigma^2)\ p(\delta^2)\,. \end{aligned} \tag{1}$$

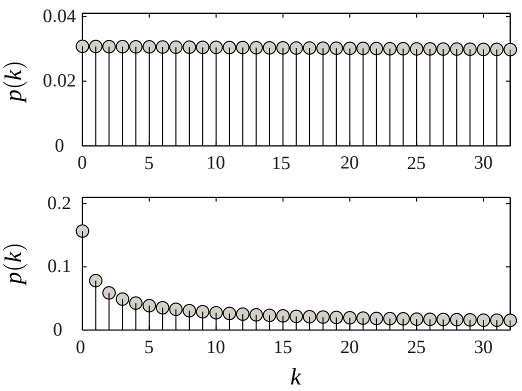


Figure 1: Truncated negative binomial prior on k corresponding to $\alpha_\Lambda = 1.0$ (upper plot) and $\alpha_\Lambda = 0.5$ (lower plot), with $k_{\max} = 32$ and $\beta_\Lambda = 0.001$.

The conditional distribution of $\mathbf{a}_k$ is the g-prior distribution already described in the introduction. Conditional on k, the components of $\boldsymbol{\omega}_k$ are independent and identically distributed, with a uniform distribution on $(0,\pi)$. The noise variance σ^2 is endowed with Jeffrey's improper prior, i.e. $p(\sigma^2) \propto 1/\sigma^2$, where the symbol $\propto$ denotes proportionality.

The prior distribution of k is defined in [1] in two steps, following once again the hierarchical Bayes philosophy. First, k is given a Poisson distribution with mean Λ, truncated to $\{0,1,\ldots,k_{\max}\}$. Then, to increase the robustness of the inference in a context of weak prior information on k, the hyperparameter Λ is given a conjugate Gamma prior, with shape parameter $\alpha_\Lambda \approx \frac{1}{2}$ and scale parameter $\beta_\Lambda \approx 0$. This is equivalent to using for k a (truncated) negative binomial prior[1] that puts a strong emphasis on small values. In this paper, we set $\alpha_\Lambda = 1$ in order to have an almost flat prior for k over $\{0,\ldots,k_{\max}\}$; see Figure 1 for a comparison of the two prior distributions.

2.2 Sampling structure

The hierarchical structure and prior distributions just described make it possible to integrate parameters $\mathbf{a}_k$ and σ^2 out of the posterior distribution analytically. This *marginalization* step [17] yields the following marginal posterior distribution:

$$p\left(k,\boldsymbol{\omega}_k,\delta^2,\Lambda\,|\,\mathbf{y}\right) \propto (\mathbf{y}^t\mathbf{P}_k\mathbf{y})^{-N/2}\,\frac{\Lambda^k\pi^{-k}}{k!\,(\delta^2+1)^k} \times p(\delta^2)\,p(\Lambda)\,\mathbb{1}_{(0,\pi)^k}(\boldsymbol{\omega}_k)\,, \qquad (2)$$

with
$$\mathbf{P}_k = \mathbf{I}_N - \frac{\delta^2}{1+\delta^2}\,\mathbf{D}_k\left(\mathbf{D}_k^t\mathbf{D}_k\right)^{-1}\mathbf{D}_k^t$$

when $k \geq 1$ and $\mathbf{P}_0 = \mathbf{I}_N$.

The joint posterior distribution (2) is the target distribution of the RJ-MCMC sampler. In the following, different steps for sampling from the target distribution are briefly described. For more detailed expressions please refer to [1, 8].

The RJ-MCMC sampler, that leaves the target density (2) invariant, consists of a Metropolis-Hastings (MH) move for updating the value of k and $\boldsymbol{\omega}_k$, followed by a sequence of Gibbs moves to update δ^2 and Λ. (The conditional distribution of δ^2 given k, $\boldsymbol{\omega}_k$, Λ and $\mathbf{y}$ is sampled from by first *demarginalizing* [17] σ^2 and $\mathbf{a}_k$ and then sampling from the full conditional distribution.)

Since the problem under consideration is trans-dimensional, the proposal distribution for the MH move updating k and $\boldsymbol{\omega}_k$ is in fact a mixture of proposal distributions performing within-model moves (updating radial frequencies without changing k) and between-models moves ("birth" and "death" moves, which respectively add and remove components). Except for a modification described below, the moves implemented in our sampler are the same as in [1].

2.3 Correction of the birth ratio in [1]

In the birth move proposed in [1], and also used in this paper, the insertion of a new sinusoid is proposed as follows: first a new radial frequency is sampled from the uniform distribution on $(0,\pi)$ and, then, it is inserted at a random location[2] among the existing ones. According the theory of RJ-MCMC samplers [8] and using the same proportion of birth and death moves as in [1], the move is accepted with probability $\alpha_{\text{birth}} = \min\{1, r_{\text{birth}}\}$, where

$$r_{\text{birth}} = \left(\frac{\mathbf{y}^t\mathbf{P}_{k+1}\mathbf{y}}{\mathbf{y}^t\mathbf{P}_k\mathbf{y}}\right)^{-N/2}\,\frac{1}{1+\delta^2}\,. \qquad (3)$$

One should note that the birth ratio computed in [1] differs from (3) by a $1/(k+1)$ factor. A similar mistake in computing RJ-MCMC ratios has been reported in the field of genetics [11]. Note that this additional factor is equivalent to using a different prior distribution over k. A detailed justification of (3) will be provided in a forthcoming paper.

3. SENSITIVITY OF THE ALGORITHM TO β_{δ^2}

This section first reviews related work concerning the role of δ^2 in the Bayesian variable selection literature, and then proceeds to describing the role of β_{δ^2} in the present problem.

3.1 Review of related work in Bayesian variable selection

It has been highlighted in the variable selection literature that the parameter δ^2, which controls the expected relative size of the amplitudes with respect to σ, implicitly defines a "dimensionality penalty" from the model selection point of view [2, 6]. Indeed, considering that $p(k)$ is approximately constant for $k \in [0, k_{\max}]$, we have

$$\log p\left(k,\boldsymbol{\omega}_k\,|\,\mathbf{y},\delta^2\right) \approx -\frac{N}{2}\log\left(\mathbf{y}^t\mathbf{P}_k\mathbf{y}\right) - F\cdot k + C, \qquad (4)$$

where $F = \log\left(\pi\left(1+\delta^2\right)\right)$ and C is a constant which does not depend on k and $\boldsymbol{\omega}_k$. F can be interpreted as a dimensionality penalty, which penalizes complex models. Thus, δ^2 plays the role of a regularization parameter, "large" values of which favor sparse signal representations at the expense of detection sensitivity. Conversely, "small" values of δ^2 typically lead to the selection of overfitting models (i.e., in terms of detection performance, false positives).

In the Bayesian variable selection literature, many researchers have tried to either set an appropriate fixed value to δ^2 or estimate it using different approaches. In [4], several fixed values for δ^2 are compared in a model averaging framework, and $\delta^2 = \max\{N, p^2\}$ is recommended as a default ("benchmark") value, where p denotes the number of variables. Several approaches for the estimation of δ^2, both EB or fully Bayesian, have been proposed and compared

[1] Indeed, the marginal prior distribution of k is given by

$$p(k) = \frac{\Gamma(k+\alpha_\Lambda)}{\Gamma(\alpha_\Lambda)\,k!}\left(\frac{\beta_\Lambda}{\beta_\Lambda+1}\right)^{\alpha_\Lambda}\left(\frac{1}{\beta_\Lambda+1}\right)^k,$$

which is a negative binomial distribution. See, e.g., [5, Section 2.7 and 17.2], where the negative binomial distribution is advocated as a robust alternative to the Poisson distribution.

[2] Note that the same ratio would be obtained if the radial frequency were sorted instead [16].

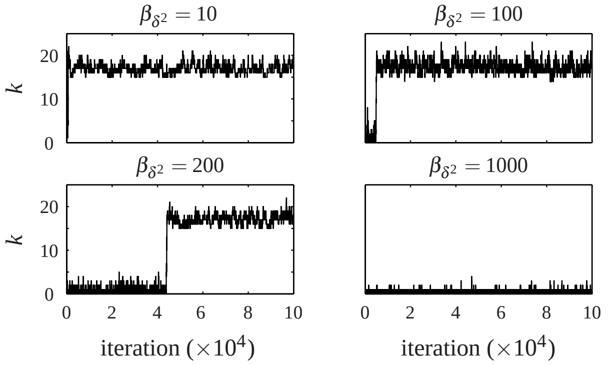


Figure 2: Mixing of the chain for different values of β_{δ^2}. The true model is $\mathscr{M}_{15}$, and the sampler is initialized in $\mathscr{M}_0$.

in [2, 6, 14]. It is concluded in [2] that the Maximum Marginal Likelihood (MML) approach is superior to the others (in terms of mean square error), but the conclusions of [14]—in a slightly different setting—suggest that some fully Bayesian approaches can perform just as well.

3.2 Role of β_{δ^2}

Our numerical experiments have revealed that the value of β_{δ^2} can have a significant influence on 1) the posterior distribution of the number of components and 2) the convergence rate of the Markov chain.

The former fact can be understood in light of Section 3.1 where the role of δ^2 as a dimensionality penalty has been highlighted. Indeed, since β_{δ^2} is a scale parameter for the prior distribution of δ^2, it can be expected that, probably to a lesser extent, β_{δ^2} should play a similar role. In other words, high values of β_{δ^2} are expected to favor sparse solutions, with a risk of omitting low SNR components, whereas low values of β_{δ^2} are expected to allow solutions with many components (high values of k). This point will be further discussed in Section 5 on the basis of numerical results.

Let us now discuss the influence of β_{δ^2} on the mixing of the sampler. We have found that large values of β_{δ^2} lead to a sampler that has severe mixing issues and often gets trapped in local modes of the target distribution. This issue is illustrated in Figure 2, which shows the mixing of the chain for different values of β_{δ^2} in a case where the true model is $\mathscr{M}_{15}$, the number of samples $N = 64$, and the sampler is initialized in $\mathscr{M}_0$. The mixing issue of the chain when $\beta_{\delta^2} > 100$ is highlighted in this figure, which causes the sampler to get stuck for many iterations at a local mode. In fact, when $\beta_{\delta^2} = 1000$ the sampler cannot escape from the local mode after 100k iterations. This convergence issue might similarly happen when the true signal is near null model and the sampler is initialized near full model. So, for large values of β_{δ^2}, the algorithm is sensitive to the initialized state. On the other hand, too small values of β_{δ^2} which corresponds to assuming low ESNR, would cause the algorithm to explore many regions of low probability of the space in low SNR situations which can be really computationally expensive and causes convergence problems.

A possible solution to the mixing issue would be to use a combination of simulated annealing and MCMC sampler as is done, for example, in [7]. In the next section we follow a different path and use an EB approach to estimate β_{δ^2} from the data.

4. IMPORTANCE SAMPLING BASED MCEM ALGORITHM

Hierarchical models are commonly used in Bayesian model (or variable) selection problems. However, this hierarchy should stop at some point with all remaining parameters assumed fixed. Then, based on some prior beliefs, these parameters can be set. However, for some parameters which no information is provided beforehand, rather than setting them to a fixed value, the EB approach uses the observed data to estimate them. It avoids using arbitrary choices which may be at odds with the observed data.

In this method, one tries to estimate β_{δ^2} such that the marginal likelihood is maximized. In other words,

$$\hat{\beta}_{\delta^2} = \mathrm{argmax}_{\beta_{\delta^2}}\, p(\mathbf{y}|\beta_{\delta^2}).$$

This is similar to MML method proposed in [6] for estimating δ^2. The maximum likelihood may be easier to compute when the data is augmented by a set of latent variables, $\mathbf{u}$ say. These latent variables, in our case, are $\{\boldsymbol{\omega}_k, k, \delta^2, \Lambda\}$. Then, one can use the EM algorithm that entails, at iteration $r+1$, an E-step for computing the expected log-likelihood

$$Q(\beta_{\delta^2}|\hat{\beta}^r_{\delta^2}) = E_{\hat{\beta}^{(r)}_{\delta^2}}\left\{\ln p(\mathbf{y},\mathbf{u}|\beta_{\delta^2})|\mathbf{y}\right\} \tag{5}$$

and, an M-step, for maximization of $Q(\beta_{\delta^2}|\hat{\beta}^r_{\delta^2})$ over β_{δ^2} in order to obtain the MLE of it, $\hat{\beta}^{r+1}_{\delta^2}$.

However, in our case, computing the E-step is not possible analytically. Therefore, here, we propose to use Monte Carlo approximation of (5), which is called MCEM [13, 15], by simulating samples from $p(\mathbf{u}|\mathbf{y},\hat{\beta}^r_{\delta^2})$. Moreover, the Monte Carlo estimation of (5) can be implemented in a more efficient way using the idea of Importance Sampling (IS). As is explained in [13, 15], in this framework, samples are just generated from $p(\mathbf{u}|\mathbf{y},\hat{\beta}^0_{\delta^2})$, where $\hat{\beta}^0_{\delta^2}$ is the initial value. Then, for m number of generated samples, the E-step can be written as

$$Q(\beta_{\delta^2}|\hat{\beta}^r_{\delta^2}) = \sum_{t=1}^{m} w_t \ln p(\mathbf{y},u_t|\beta_{\delta^2}) \Big/ \sum_{t=1}^{m} w_t \tag{6}$$

where

$$w_t = \frac{p(u_t|\mathbf{y},\beta^{(r)}_{\delta^2})}{p(u_t|\mathbf{y},\beta^{(0)}_{\delta^2})}$$

are the weights which in our case would simplify to

$$w_t = \left(\frac{\beta^{(r)}_{\delta^2}}{\beta^{(0)}_{\delta^2}}\right)^{\alpha_{\delta^2}} \exp\left(-\frac{\beta^{(r)}_{\delta^2} - \beta^{(0)}_{\delta^2}}{\delta^2_t}\right).$$

Since the RJ-MCMC sampler introduced in Section 2 can easily generate m samples from $p(\mathbf{u}|\mathbf{y},\hat{\beta}^0_{\delta^2})$, these samples can be used to perform the IS based MCEM procedure. So, in each MCEM iteration, a batch of m samples is generated from the RJ-MCMC sampler in order to compute (6). The computationally efficient point of this procedure is that once the IS based MCEM algorithm is stopped, the generated samples are not discarded. They can be used to generate the desired posterior distribution of the unknown parameters by using the importance weights.

However, one should note that this procedure is sensitive to the value of $\hat{\beta}^0_{\delta^2}$. In order to reduce the variations of w_t, it is proposed in [13] to run a few burn-in iterations using a simple MCEM method without importance reweighting.

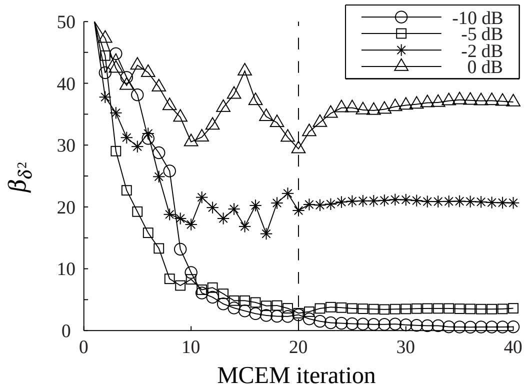


Figure 3: Estimated values of β_{δ^2} using the IS-based MCEM algorithm. The signal is generated under $\mathscr{M}_1$ with $N = 64$, $\omega_{1,1} = 0.2\pi$, for several values of the SNR (see legend). The vertical line indicates the burn-in period.

5. SIMULATION RESULTS AND DISCUSSION

In this section, we will investigate the capability of the IS based MCEM algorithm for assessing β_{δ^2} in different situations. Moreover, we will compare the performance of the sampler with several fixed values of β_{δ^2}. Simulations are performed on two different sample sizes $N = 64$ and $N = 256$ generated according to $\mathscr{M}_1$ with different SNRs. The SNR is defined as

$$\mathrm{SNR} \triangleq \frac{\|\mathbf{D}_k \mathbf{a}_k\|^2}{N\sigma^2}.$$

The parameters of the single sinusoid are as follows: $\omega_{1,1} = 0.2\pi$, $-\arctan(a_{2,1}/a_{1,1}) = \pi/3$, and $a_{11}^2 + a_{2,1}^2 = 20$.

In the IS based MCEM algorithm, first, 20 burn-in iterations with $m = 100$ samples were carried out. Then, the 20 IS based MCEM procedure iterations with $m = 5000$ were performed to estimate β_{δ^2}. So, finally, in addition to an approximate estimate of β_{δ^2}, 100k samples from the RJ-MCMC sampler are obtained and can be used to produce the posterior distributions of the unknown parameters, of course by using the importance weights. Figure 3 shows the performance of the IS based MCEM algorithm in estimating the value of β_{δ^2} for different observed signals. This relation between the value of β_{δ^2} and SNR, that is illustrated in figure 3, is remarkably consistent with expectations. It is worthwhile to note that variation of the estimated values of β_{δ^2} is substantially reduced after the burn-in period, as it is shown in figure 3, which illustrates the convergence of the algorithm.

Table 1 presents the probabilities of $\arg\max p(k|\mathbf{y})$ in 100 realizations of the algorithms. In each realization, 100k samples were generated and the first 20k samples were discarded as the burn-in period. The results are presented for different fixed values of β_{δ^2} together with the results obtained by applying the IS based MCEM algorithm for estimating β_{δ^2}.

First, let us consider the case of fixed β_{δ^2}. From the results presented in Table 1, it can be concluded that the value of β_{δ^2} has a strong influence on the posterior distribution of the number of components. Indeed choice of β_{δ^2} would become more critical as the SNR decreases. Though the sampler produces reasonable results for a wide range of values of β_{δ^2}, i.e. $10 \leq \beta_{\delta^2} \leq 1000$, in high SNR situations (not shown here), the behavior of the sampler significantly varies by changing the value of this parameter in low SNR situations. For instance, when SNR $= -5$ dB, while the probability of detecting one component is almost the same for the mentioned interval, setting $\beta_{\delta^2} = 10$ provides a sampler which overestimates the number of components. On the other hand, larger values of β_{δ^2} leads to a sampler that underestimates the number of components. According to the obtained results, choosing a very small value for β_{δ^2}, one say, is not suitable. For the values of SNR < 0 dB, it makes convergence problems for the sampler by accepting most of proposed birth or death moves. More precisely, it leads to a sampler which explores all possible regions, even low probable ones, which would be really computationally expensive when $k_{\max}$ is large. However, one should note that for all simulations the samplers were initialized near null model, otherwise for values of $\beta_{\delta^2} > 100$ the results would definitely changed. In the case that $N = 256$, the sensitivity of the sampler to the choice of β_{δ^2} is less critical. This may be caused by the fact that the observed signal is more informative in this case. Finally, a fixed value of $\beta_{\delta^2} \in [50, 100]$ provides a sampler with more reasonable performance for most values of SNR.

Turning to the results of the EB approach used here to automatically estimate the value of β_{δ^2} from the data, it can be seen from the table that the sampler equipped with the IS-based MCEM algorithm has a quite satisfactory behavior in moderate to high SNR situations (0 dB, -2 dB, and even -5 dB for $N = 256$). However, it is clear that the algorithm fails to select an appropriate value for β_{δ^2} in low SNR situations (-10 dB, and -5 dB for $N = 64$): the selected value is typically much too small, leading to severe overfitting. A similar behavior is observed in experiments under the null model $\mathscr{M}_0$ (not shown here).

In fact, based on Table 1, it seems that using $\beta_{\delta^2} = 50$ gives, in all the situations considered here, results that are similar to or better than the results of the EB approach. Additional experimental results under various configurations and sample sizes are required, however, to issue a general recommendation regarding the choice of an appropriate fixed value for β_{δ^2} (possibly depending on N) and, also, to confirm the capability of the EB approach to automatically select such a value in moderate to high SNR situations.

6. CONCLUSION

In this paper, first, the sensitivity of the RJ-MCMC algorithm proposed in [1] for detection and estimation of sinusoids to the hyperparameter β_{δ^2} has been investigated. Then, an IS-based MCEM algorithm has been used to estimate this parameter given the data, following an empirical Bayes (EB) approach. The IS-based MCEM method has proved able to automatically estimate an appropriate value for β_{δ^2} in moderate to high SNR situations.

The main limitation of the EB approach is that it cannot estimate a proper value for β_{δ^2} in very low SNR situations. This limitation was, however, predictable as in such cases the observed signal carries very little information about the parameter of interest. To overcome this limitation and avoid the problem of choosing a *scale* for $p(\delta^2)$, a truncated Jeffrey prior has been proposed in [19] and very promising results have been obtained.

As mentioned in Section 1, this model and RJ-MCMC sampler have also been used in other applications such as polyphonic signal analysis [3], array signal processing [12], and nuclear emission spectra analysis [10]. The contributions of this paper are likely to be useful in these applications as well.

REFERENCES

[1] C. Andrieu and A. Doucet. Joint Bayesian model selection and estimation of noisy sinusoids via reversible jump MCMC. *IEEE Trans. Signal Process.*, 47(10):2667–2676, 1999.

[2] W. Cui and E. I. George. Empirical Bayes vs. fully Bayes variable selection. *J. Stat. Plann. Inference*, 138(4):888–900, 2008.

[3] M. Davy, S. J. Godsill, and J. Idier. Bayesian analysis of polyphonic western tonal music. *J. Acoust. Soc. Am.*, 119:2498–2517, 2006.

N	β_{δ^2}	P_0	P_1	P_2	P_3	P_4
64	**1**	0.25	0.04	0.04	0.03	0.64
	10	0.64	0.13	0.05	0.02	0.16
	50	0.81	0.09	0.00	0.00	0.10
	100	0.87	0.11	0.00	0.01	0.01
	1000	0.97	0.02	0.01	0.00	0.00
	EB	0.05	0.04	0.02	0.06	0.83
256	**1**	0.01	0.05	0.16	0.18	0.60
	10	0.08	0.45	0.25	0.12	0.10
	50	0.18	0.76	0.04	0.02	0.00
	100	0.22	0.73	0.05	0.00	0.00
	256	0.35	0.63	0.02	0.00	0.00
	1000	0.48	0.51	0.01	0.00	0.00
	EB	0.00	0.22	0.16	0.12	0.50

N	β_{δ^2}	P_0	P_1	P_2	P_3	P_4
64	**1**	0.03	0.09	0.13	0.06	0.69
	10	0.09	0.56	0.12	0.07	0.16
	50	0.27	0.57	0.11	0.00	0.05
	100	0.31	0.60	0.08	0.00	0.01
	1000	0.54	0.45	0.01	0.00	0.00
	EB	0.01	0.22	0.25	0.12	0.42
256	**1**	0.00	0.71	0.22	0.05	0.02
	10	0.00	0.79	0.18	0.01	0.02
	50	0.00	0.92	0.06	0.00	0.02
	100	0.00	0.93	0.07	0.00	0.00
	256	0.00	0.99	0.00	0.01	0.00
	1000	0.00	0.99	0.01	0.00	0.00
	EB	0.00	0.92	0.05	0.02	0.01

N	β_{δ^2}	P_0	P_1	P_2	P_3	P_4
64	**1**	0.00	0.32	0.32	0.14	0.22
	10	0.00	0.68	0.23	0.07	0.02
	50	0.02	0.84	0.10	0.02	0.02
	100	0.01	0.93	0.04	0.01	0.01
	1000	0.02	0.97	0.01	0.00	0.00
	EB	0.00	0.69	0.22	0.04	0.05
256	**1**	0.00	0.89	0.10	0.01	0.00
	10	0.00	0.95	0.05	0.00	0.00
	50	0.00	0.95	0.04	0.00	0.01
	100	0.00	0.95	0.05	0.00	0.00
	256	0.00	1.00	0.00	0.00	0.00
	1000	0.00	1.00	0.00	0.00	0.00
	EB	0.00	0.94	0.04	0.02	0.00

N	β_{δ^2}	P_0	P_1	P_2	P_3	P_4
64	**1**	0.00	0.72	0.17	0.07	0.04
	10	0.00	0.86	0.08	0.05	0.01
	50	0.00	0.87	0.11	0.02	0.00
	100	0.00	0.95	0.05	0.00	0.00
	1000	0.00	0.98	0.02	0.00	0.00
	EB	0.00	0.88	0.09	0.02	0.01
256	**1**	0.00	0.91	0.09	0.00	0.00
	10	0.00	0.95	0.05	0.00	0.00
	50	0.00	0.98	0.02	0.00	0.00
	100	0.00	0.94	0.06	0.00	0.00
	256	0.00	0.98	0.02	0.00	0.00
	1000	0.00	1.00	0.00	0.00	0.00
	EB	0.00	0.98	0.02	0.00	0.00

Table 1: Probability of $\arg\max p(k|\mathbf{y}) = 0$, $\arg\max p(k|\mathbf{y}) = 1$, $\arg\max p(k|\mathbf{y}) = 2$, $\arg\max p(k|\mathbf{y}) = 3$, and $\arg\max p(k|\mathbf{y}) \geq 4$, are denoted, respectively, by P_0, P_1, P_2, P_3, and P_4. The value of the SNR is respectively -10 dB (top-left), -5 dB (top-right), -2 dB (bottom-left) and 0 dB (bottom-right). These probabilities have been estimated based on the output of 100 runs of the algorithm under $\mathscr{M}_1$ with two different sample sizes ($N = 64$ and $N = 256$). The length of the chain was set to 100k, with a burn-in period of 20k samples. Results are presented for several fixed values of β_{δ^2} and for the IS-based MCEM algorithm.

[4] C. Fernández, E. Ley, and M. Steel. Benchmark priors for Bayesian model averaging. *J. Econometrics*, 100(2):381–427, 2001.

[5] A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin. *Bayesian Data Analysis (second edition)*. Chapman & Hall / CRC, 2004.

[6] E. I. George and D. P. Foster. Calibration and empirical Bayes variable selection. *Biometrika*, 87(4):731–747, 2000.

[7] R. Gramacy, R. Samworth, and R. King. Importance tempering. *Stat. Comput.*, 20:1–7, 2010.

[8] P. J. Green. Reversible jump MCMC computation and Bayesian model determination. *Biometrika*, 82(4):711–732, 1995.

[9] P. J. Green. Trans-dimensional Markov chain Monte Carlo. In P. J. Green, N. L. Hjort, and S. Richardson, editors, *Highly Structured Stochastic Systems*, pages 179–198. O.U.P., 2003.

[10] S. Gulam Razul, W. Fitzgerald, and C. Andrieu. Bayesian model selection and parameter estimation of nuclear emission spectra using RJMCMC. *Nucl. Instrum. Meth. A*, 497(2-3):492–510, 2003.

[11] J. Jannink and R. Fernando. On the Metropolis-Hastings acceptance probability to add or drop a quantitative trait locus in Markov chain Monte Carlo-based Bayesian analyses. *Genetics*, 166(1):641–643, 2004.

[12] J. R. Larocque and J. P. Reilly. Reversible jump MCMC for joint detection and estimation of sources in coloured noise. *IEEE Trans. Signal Process.*, 50:231–240, 2000.

[13] R. A. Levine and G. Casella. Implementations of the Monte Carlo EM algorithm. *J. Comput. Graph. Stat.*, pages 422–439, 2001.

[14] F. Liang, R. Paulo, G. Molina, M. Clyde, and J. Berger. Mixtures of g-priors for Bayesian variable selection. *J. Am. Stat. Assoc.*, 103(481):410–423, 2008.

[15] F. Quintana, J. Liu, and G. del Pino. Monte Carlo EM with importance reweighting and its applications in random effects models. *Comput. Stat. Data An.*, 29(4):429–444, 1999.

[16] S. Richardson and P. J. Green. On Bayesian analysis of mixtures with an unknown number of components. *J. Roy. Stat. Soc. B Met.*, 59(4):731–792, 1997.

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

[18] C. P. Robert. *The Bayesian Choice (second edition)*. Springer, 2007.

[19] A. Roodaki, J. Bect, and G. Fleury. On the Joint Bayesian model selection and estimation of noisy sinusoids via reversible jump MCMC in Low SNR Situations. In *Proc. 10th Int. Conf. on Information Science, Signal Processing and their Application (ISSPA), Kuala lumpur, Malaysia*, pages 5–8, 2010.

[20] G. C. G. Wei and M. A. Tanner. A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithms. *J. Am. Stat. Assoc.*, 85(411):699–704, 1990.

[21] A. Zellner. On assessing prior distributions and bayesian regression analysis with g-prior distributions. In P. K. Goel and A. Zellner, editors, *Bayesian Inference and Decision Techniques: Essays in Honor of Bruno de Finetti*, pages 233–243. North-Holland/Elsevier, 1986.

References

C. Andrieu and A. Doucet. Joint Bayesian model selection and estimation of noisy sinusoids via reversible jump MCMC. *IEEE Transactions on Signal Processing*, 47(10): 2667–2676, 1999.

C. Andrieu and É. Moulines. On the ergodicity properties of some adaptive MCMC algorithms. *Annals of Applied Probability*, 16(3):1462–1505, 2006.

C. Andrieu, A. Doucet, W. J. Fitzgerald, and J. M. Pérez. Bayesian computational approaches to model selection. *Nonlinear and Nonstationary Signal Processing (Cambridge, 1998)*, pages 1–41, 1998.

C. Andrieu, N. De Freitas, and A. Doucet. Reversible jump MCMC Simulated annealing for neural networks. In *Proceedings of the 16th Conference on Uncertainty in Artificial Intelligence*, pages 11–18. Morgan Kaufmann Publishers Inc., 2000.

C. Andrieu, N. De Freitas, and A. Doucet. Robust full Bayesian learning for radial basis networks. *Neural Computation*, 13(10):2359–2407, 2001a.

C. Andrieu, P. M. Djurić, and A. Doucet. Model selection by MCMC computation. *Signal Processing*, 81(1):19–37, 2001b.

C. Andrieu, E. Barat, and A. Doucet. Bayesian deconvolution of noisy filtered point processes. *IEEE Transactions on Signal Processing*, 49(1):134–146, 2002.

S. Asmussen, K. Binswanger, and B. Højgaard. Rare events simulation for heavy-tailed distributions. *Bernoulli*, pages 303–322, 2000.

Y. F. Atchadé and J. S. Rosenthal. On adaptive Markov Chain Monte Carlo algorithms. *Bernoulli*, 11(5):815–828, 2005.

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

Auger Collaboration. Properties and performance of the prototype instrument for the Pierre Auger Observatory. *Nuclear Instruments and Methods in Physics Research A*, 523:50–95, 2004.

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*, 20(1):63–79, 2011.

E. Barat and T. Dautremer. Nonparametric Bayesian estimation of x/γ-ray spectra using a hierarchical polya tree-dirichlet mixture model. In *AIP CONFERENCE PROCEEDINGS*, volume 872, page 477. IOP INSTITUTE OF PHYSICS PUBLISHING LTD, 2006.

M. M. Barbieri and J. O. Berger. Optimal predictive model selection. *The Annals of Statistics*, 32(3):870–897, 2004.

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

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.

A. Basu, I. R. Harris, N. L. Hjort, and M. C. Jones. Robust and efficient estimation by minimising a density power divergence. *Biometrika*, 85(3):549, 1998.

J. O. Berger. Robust Bayesian analysis: sensitivity to the prior. *Journal of Statistical Planning and Inference*, 25(3):303–328, 1990.

J. O. Berger, V. De Oliveira, and B. Sansó. Objective Bayesian analysis of spatially correlated data. *Journal of the American Statistical Association*, 96(456):1361–1374, 2001.

J. O. Berger, J. M. Bernardo, and D. Sun. The formal definition of reference priors. *The Annals of Statistics*, 37(2):905–938, 2009.

J. M. Bernardo and A. F. M. Smith. *Bayesian Theory.* Wiley New York, 2000.

J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith. On the development of reference priors. *Bayesian Statistics*, 4:35–60, 1992.

L. Bornn, A. Doucet, and R. Gottardo. An efficient computational approach for prior sensitivity analysis and cross-validation. *Canadian Journal of Statistics*, 38(1):47–64, 2010.

G. L. Bretthorst. *Bayesian Spectrum Analysis and Parameter Estimation*. Springer Berlin, 1988.

M. Broniatowski and A. Keziou. Parametric estimation and tests through divergences and the duality technique. *Journal of Multivariate Analysis*, 100(1):16–36, 2009.

M. Broniatowski, G. Celeux, and J. Diebolt. Reconnaissance de mélanges de densités par un algorithme d'apprentissage probabiliste. *Data analysis and informatics*, pages 359–373, 1983.

S. P. Brooks, P. Giudici, and G. O. Roberts. Efficient construction of reversible jump Markov chain Monte Carlo proposal distributions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 65(1):3–39, 2003.

K. P. Burnham and D. R. Anderson. *Model selection and multimodel inference: a practical information-theoretic approach*. Springer Verlag, 2002.

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

O. Cappé, A. Doucet, M. Lavielle, and E. Moulines. Simulation-based methods for blind maximum-likelihood filter identification. *Signal processing*, 73(1):3–25, 1999.

O. Cappé, C. P. Robert, and T. Rydén. Reversible jump, birth-and-death and more general continuous time Markov chain Monte Carlo samplers. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 65(3):679–700, 2003.

G. Casella and E. I. George. Explaining the Gibbs sampler. *American Statistician*, pages 167–174, 1992.

G. Celeux. Bayesian inference for mixtures: The label-switching problem. In *Compstat 98 (R. Payne and P. J. Green, eds.)*, pages 227–232, 1998.

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

G. Celeux and J. Diebolt. A stochastic approximation type EM algorithm for the mixture problem. *Stochastics and Stochastics reports*, 41(1):119–134, 1992. ISSN 1744-2508.

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

G. Celeux, M. E. Anbari, J. M. Marin, and C. P. Robert. Regularization in regression: comparing Bayesian and frequentist methods in a poorly informative situation. *Bayesian Analysis*, 7(1):1–26, 2012.

S. S. Chen, D. L. Donoho, and M. A. Saunders. Atomic decomposition by basis pursuit. *SIAM journal on scientific computing*, 20(1):33–61, 1999.

H. Chipman, E. I. George, R. E. McCulloch, M. Clyde, D. P. Foster, and R. A. Stine. The practical implementation of Bayesian model selection. *Lecture Notes-Monograph Series*, pages 65–134, 2001.

N. Chopin. A sequential particle filter method for static models. *Biometrika*, 89(3):539, 2002.

M. A. Clyde and E. I. George. Model uncertainty. *Statistical Science*, 19(1):81–94, 2004.

P. Congdon. *Bayesian Statistical Modeling*. John Wiley, 2006.

M. K. Cowles and B. P. Carlin. Markov chain Monte Carlo convergence diagnostics: a comparative review. *Journal of the American Statistical Association*, pages 883–904, 1996.

I. Csiszár. Information-type measures of difference of probability distributions and indirect observations. *Studia scientiarum mathematicarum Hungarica*, 2:299–318, 1967.

W. Cui and E. I. George. Empirical Bayes vs. fully Bayes variable selection. *Journal of Statistical Planning and Inference*, 138(4):888–900, 2008.

R. N. Davé and R. Krishnapuram. Robust clustering methods: a unified view. *IEEE Transactions on Fuzzy Systems*, 5(2):270–293, 1997.

M. Davy, S. J. Godsill, and J. Idier. Bayesian analysis of polyphonic western tonal music. *Journal of the Acoustical Society of America*, 119:2498–2517, 2006.

P. Del Moral, A. Doucet, and A. Jasra. Sequential Monte Carlo samplers. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(3):411–436, 2006.

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)*, 1977.

X. Descombes, MNM van Lieshout, R. Stoica, and J. Zerubia. Parameter estimation by a Markov chain Monte Carlo technique for the Candy model. In *In IEEE Workshop Statistical Signal Processing, Singapore*, 2001.

X. Descombes, F. Kruggel, G. Wollny, and H.J. Gertz. An object-based approach for detecting small brain lesions: application to Virchow-Robin spaces. *IEEE Transactions on Medical Imaging*, 23(2):246–255, 2004.

P. Diaconis, S. Holmes, and R. M. Neal. Analysis of a nonreversible Markov chain sampler. *The Annals of Applied Probability*, 10(3):726–752, 2000.

J. Diebolt and G. Celeux. Asymptotic properties of a stochastic EM algorithm for estimating mixing proportions. *Stochastic Models*, 9(4):599–613, 1993. ISSN 1532-6349.

J. Diebolt and C. P. Robert. Estimation of finite mixture distributions through Bayesian sampling. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, pages 363–375, 1994.

P. M. Djurić. A model selection rule for sinusoids in white Gaussian noise. *IEEE Transactions on Signal Processing*, 44(7):1744–1751, 1996.

N. Dobigeon, A. O. Hero, and J. Y. Tourneret. Hierarchical Bayesian sparse image reconstruction with application to MRFM. *IEEE Transactions on Image Processing*, 18(9): 2059–2070, 2009.

L. Dou and R. J. W. Hodgson. Bayesian inference and Gibbs sampling in spectral analysis and parameter estimation. I. *Inverse problems*, 11:1069, 1995.

L. Dou and R. J. W. Hodgson. Bayesian inference and Gibbs sampling in spectral analysis and parameter estimation: II. *Inverse problems*, 12:121, 1996.

A. Doucet, S. G. Godsill, and C. Andrieu. On sequential Monte Carlo sampling methods for Bayesian filtering. *Statistics and Computing*, 10(3):197–208, 2000.

A. Doucet, N. De Freitas, and N. Gordon. *Sequential Monte Carlo Methods in Practice*. Springer Verlag, 2001.

N. R. Draper and H. Smith. *Applied Regression Analysis (Wiley Series in Probability and Statistics)*. Wiley-Interscience, second edition, 1981.

M. Elad and M. Aharon. Image denoising via sparse and redundant representations over learned dictionaries. *IEEE Transactions on Image Processing*, 15(12):3736–3745, 2006.

Y. Fan, G. W. Peters, and S. A. Sisson. Automating and evaluating reversible jump MCMC proposal distributions. *Statistics and Computing*, 19(4):409–421, 2009.

C. Fernández, E. Ley, and M. F. J. Steel. Benchmark priors for Bayesian model averaging. *Journal of Econometrics*, 100(2):381–427, 2001.

C. Fevotte and S. J. Godsill. Sparse linear regression in unions of bases via Bayesian variable selection. *IEEE Signal Processing Letters*, 13(7):441–444, 2006.

D. H. Fremlin. *Measure Theory, Volume 2 : Broad Foundations*. Torres Fremlin, 2001.

S. Frühwirth-Schnatter. *Finite Mixture and Markov Switching Models*. Springer Verlag, 2006.

H. Fujisawa and S. Eguchi. Robust estimation in the normal mixture model. *Journal of Statistical Planning and Inference*, 136(11):3989–4011, 2006.

A. E. Gelfand and A. F. M. Smith. Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, pages 398–409, 1990.

A. Gelman, G. O. Roberts, and W. Gilks. Efficient Metropolis jumping rules. *Bayesian Statistics*, 5:599–608, 1996.

A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin. *Bayesian Data Analysis*. Chapman & Hall, 2004.

S. Geman and D. Geman. Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6:721–741, 1984.

E. I. George and D. P. Foster. Calibration and empirical Bayes variable selection. *Biometrika*, 87(4):731–747, 2000.

C. J. Geyer and J. Møller. Simulation procedures and likelihood inference for spatial point processes. *Scandinavian Journal of Statistics*, 21(4):359–373, 1994.

W. R. Gilks and C. Berzuini. Following a moving target—Monte Carlo inference for dynamic Bayesian models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(1):127–146, 2001.

W. R. Gilks, S. Richardson, and D. J. Spiegelhalter. *Markov Chain Monte Carlo in Practice*. Chapman & Hall/CRC, 1996.

N. J. Gordon, D. J. Salmond, and A. F. M. Smith. Novel approach to nonlinear/non-Gaussian Bayesian state estimation. *IEE Proceedings F*, 140(2):107–113, 1993.

P. J. Green. Reversible jump MCMC computation and Bayesian model determination. *Biometrika*, 82(4):711–732, 1995.

P. J. Green. Trans-dimensional Markov chain Monte Carlo. In P. J. Green, N. L. Hjort, and S. Richardson, editors, *Highly Structured Stochastic Systems*, pages 179–198. O.U.P., 2003.

B. Grün and F. Leisch. Dealing with label switching in mixture models under genuine multimodality. *Journal of Multivariate Analysis*, 100(5):851–861, 2009.

H. Haario, E. Saksman, and J. Tamminen. An adaptive Metropolis algorithm. *Bernoulli*, pages 223–242, 2001.

C. Han and B. P. Carlin. Markov chain Monte Carlo methods for computing Bayes factors. *Journal of the American Statistical Association*, 96(455):1122–1132, 2001.

D. Hastie. *Towards automatic reversible jump Markov chain Monte Carlo*. PhD thesis, Citeseer, 2005.

D. I. Hastie and P. J. Green. Model choice using reversible jump Markov chain Monte Carlo. *Statistica Neerlandica*, 2011.

W. K. Hastings. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1):97–109, 1970.

AO Hero III. Timing estimation for a filtered Poisson process in Gaussian noise. *IEEE Transactions on Information Theory*, 37(1):92–106, 1991.

J. A. Hoeting, D. Madigan, A. E. Raftery, and C. T. Volinsky. Bayesian model averaging: a tutorial. *Statistical Science*, pages 382–401, 1999.

M. Hong, M. F. Bugallo, and P. M. Djurić. Joint model selection and parameter estimation by population Monte Carlo simulation. *IEEE Journal of Selected Topics in Signal Processing*, 4(3):526–539, 2010.

P. J. Huber and E. M. Ronchetti. *Robust Statistics (2nd Edition).* Wiley., 2009.

M. Hurn and H. Rue. High level image priors in confocal microscopy applications. *The Art and Science of Bayesian Image Analysis*, pages 36–43, 1997.

J.-L. Jannink and R. L. Fernando. On the Metropolis-Hastings acceptance probability to add or drop a quantitative trait locus in Markov chain Monte Carlo-based Bayesian analyses. *Genetics*, 166:641–643, 2004.

A. Jasra. *Bayesian inference for mixture models via Monte Carlo computation.* PhD thesis, Imperial College London, 2005.

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*, 20(1):50–67, 2005. ISSN 0883-4237.

E. T. Jaynes. Bayesian spectrum and chirp analysis. *Maximum Entropy and Bayesian Spectral Analysis and Estimation Problems*, 1, 1987.

M. C. Jones, N. L. Hjort, I. R. Harris, and A. Basu. A comparison of related density-based minimum divergence estimators. *Biometrika*, 88(3):865, 2001.

A. F. Karr. *Point Processes and their Statistical Inference (second edition).* CRC, 1991.

R. E. Kass and A. E. Raftery. Bayes factors. *Journal of the American Statistical Association*, pages 773–795, 1995.

R. E. Kass and L. Wasserman. The selection of prior distributions by formal rules. *Journal of the American Statistical Association*, pages 1343–1370, 1996.

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

B. Kégl and D. Veberic. Single muon response: Tracklength. Technical report, LAL, University of Paris-Sud / CNRS, France, 2009.

S. Kullback and R. A. Leibler. On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86, 1951.

C. Lacoste, X. Descombes, and J. Zerubia. Point processes for unsupervised line network extraction in remote sensing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1568–1579, 2005.

P. Lahiri. *Model selection*, volume 38. Lecture Notes–Monograph Series, Beachwood, OH: Institute of Mathematical Statistics, 2001.

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*, 50:231–240, 2002.

J. R. Larocque, J. P. Reilly, and W. Ng. Particle filters for tracking an unknown number of sources. *IEEE Transactions on Signal Processing*, 50(12):2926–2937, 2002.

R. A. Levine and G. Casella. Implementations of the Monte Carlo EM algorithm. *Journal of Computational and Graphical Statistics*, pages 422–439, 2001.

M. S. Lewicki and T. J. Sejnowski. Learning overcomplete representations. *Neural computation*, 12(2):337–365, 2000.

F. Liang, R. Paulo, G. Molina, M. A. Clyde, and J. O. Berger. Mixtures of g-priors for Bayesian variable selection. *Journal of the American Statistical Association*, 103(481): 410–423, 2008.

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

J. S. Liu and R. Chen. Sequential Monte Carlo methods for dynamic systems. *Journal of the American statistical Association*, pages 1032–1044, 1998.

S. G. Mallat and Z. Zhang. Matching pursuits with time-frequency dictionaries. *IEEE Transactions on Signal Processing*, 41(12):3397–3415, 1993.

S.G. Mallat. *A Wavelet Tour of Signal Processing: The Sparse Way.* Academic Press, third edition, 2009.

J. M. Marin, K. L. Mengersen, and C. P. Robert. Bayesian modelling and inference on mixtures of distributions. *Bayesian Thinking, Modeling and Computation*, 25, 2005.

R. A. Maronna, R. D. Martin, and V. J. Yohai. *Robust Statistics: Theory and Methods.* Wiley, 2006.

G. J. McLachlan and T. Krishnan. *The EM Algorithm and Extensions.* Wiley Interscience, 2008.

G. J. McLachlan and D. Peel. *Finite Mixture Models.* Wiley-Interscience, 2000.

L. Melie-García, E. J. Canales-Rodríguez, Y. Alemán-Gómez, C. P. Lin, Y. Iturria-Medina, and P. A. Valdés-Hernández. A Bayesian framework to identify principal intravoxel diffusion profiles based on diffusion-weighted MR imaging. *Neuroimage*, 42(2):750–770, 2008.

V. Melnykov and I. Melnykov. Initializing the EM algorithm in Gaussian mixture models with an unknown number of components. *Computational Statistics & Data Analysis*, 2011.

K. L. Mengersen and C. P. Robert. MCMC convergence diagnostics: a reviewww. In *Bayesian statistics 6: proceedings of the Sixth Valencia International Meeting*, volume 6, page 415. Oxford University Press, USA, 1999.

K. L. Mengersen, C. P. Robert, and M. Titterington. *Mixtures: Estimation and Applications*, volume 896. Wiley, 2011.

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*, 21 (6):1087, 1953.

S. P. Meyn and R. L. Tweedie. *Markov Chains and Stochastic Stability.* Springer Verlag London, 1993.

M. Mihoko and S. Eguchi. Robust blind source separation by beta divergence. *Neural computation*, 14(8):1859–1886, 2002.

A. J. Miller. *Subset Selection in Regression.* CRC Press, second edition, 2002.

Y. Mishchencko, J. T. Vogelstein, and L. Paninski. A Bayesian approach for inferring neuronal connectivity from calcium fluorescent imaging data. *The Annals of Applied Statistics*, 5(2B):1229–1261, 2011.

M. Miyamura and Y. Kano. Robust Gaussian graphical modeling. *Journal of Multivariate Analysis*, 97(7):1525–1550, 2006.

W. Ng, J. P. Reilly, T. Kirubarajan, and J. R. Larocque. Wideband array signal processing using MCMC methods. *IEEE Transactions on Signal Processing*, 53(2):411–426, 2005.

W. Ng, T. Chan, H. C. So, and K. C. Ho. On particle filters for landmine detection using impulse ground penetrating radar. In *5th IEEE Workshop on Sensor Array and Multichannel Signal Processing*, pages 225–228, 2008.

G. K. Nicholls. Bayesian image analysis with Markov chain Monte Carlo and coloured continuum triangulation models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 60(3):643–659, 1998.

S. F. Nielsen. The stochastic EM algorithm: estimation and asymptotic results. *Bernoulli*, 6(3):457–489, 2000a. ISSN 1350-7265.

S. F. Nielsen. On simulated EM algorithms. *Journal of Econometrics*, 96(2):267–292, 2000b. ISSN 0304-4076.

J. Nocedal and S. J. Wright. *Numerical Optimization*. Springer Verlag, 1999.

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*, 19(2):313–331, 2010.

L. Parclo. *Statistical Inference Based on Divergence Measures*. CRC, 2005.

A. Pievatolo and P. J. Green. Object restoration through dynamic polygons. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 60:609–26, 1998.

E. Punskaya, C. Andrieu, A. Doucet, and W. J. Fitzgerald. Bayesian curve fitting using MCMC with applications to signal segmentation. *IEEE Transactions on Signal Processing*, 50(3):747–758, 2002.

F. A. Quintana, J. S. Liu, and G. E. Del Pino. Monte Carlo EM with importance reweighting and its applications in random effects models. *Computational Statistics and Data Analysis*, 29(4):429–444, 1999.

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)*, 59(4):731–792, 1997.

S. Richardson and P. J. Green. Corrigendum: on Bayesian analysis of mixtures with an unknown number of components. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 60(3):661, 1998.

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)*, 59(4):758–764, 1997.

C. P. Robert. *The Bayesian Choice (second edition)*. Springer, 2007.

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

G. O. Roberts and J. S. Rosenthal. Optimal scaling for various Metropolis-Hastings algorithms. *Statistical Science*, pages 351–367, 2001.

G. O. Roberts and J. S. Rosenthal. General state space Markov chains and MCMC algorithms. *Probability Surveys*, 1:20–71, 2004.

G. O. Roberts and J. S. Rosenthal. Harris recurrence of Metropolis-within-Gibbs and trans-dimensional Markov chains. *The Annals of Applied Probability*, 16(4):2123–2139, 2006.

G. O. Roberts and J. S. Rosenthal. Examples of adaptive MCMC. *Journal of Computational and Graphical Statistics*, 18(2):349–367, 2009.

O. Rosec, J. M. Boucher, B. Nsiri, and T. Chonavel. Blind marine seismic deconvolution using statistical MCMC methods. *IEEE Journal of Oceanic Engineering*, 28(3):502–512, 2003.

D. V. Rubtsov and J. L. Griffin. Time-domain Bayesian detection and estimation of noisy damped sinusoidal signals applied to NMR spectroscopy. *Journal of Magnetic Resonance*, 188(2):367–379, 2007.

H. Rue and M. A. Hurn. Bayesian object identification. *Biometrika*, 86(3):649–660, 1999.

M. N. Schmidt and M. Mørup. Infinite non-negative matrix factorization. In 18^{th} *European Signal Processing Conference (EUSIPCO)*, Aug 2010.

A. Schuster. On the investigation of hidden periodicities with application to a supposed 26 day period of meteorological phenomena. *Terrestrial Magnetism*, 3(1):13–41, 1898.

Z. G. Shi, J. X. Zhou, H.Z. Zhao, and Q. Fu. Study on joint Bayesian model selection and parameter estimation method of GTD model. *Science in China Series F: Information Sciences*, 50(2):261–272, 2007.

M. J. Sillanpaa, D. Gasbarra, and E. Arjas. Comment on "On the Metropolis-Hastings acceptance probability to add or drop a quantitative trait locus in Markov chain Monte Carlo-based Bayesian analyses". *Genetics*, 167(2):1037, 2004.

S. A. Sisson. Transdimensional Markov chains: a decade of progress and future perspectives. *Journal of the American Statistical Association*, 100(471):1077–1090, 2005.

M. Sperrin, T. Jaki, and E. Wit. Probabilistic relabelling strategies for the label switching problem in Bayesian mixture models. *Statistics and Computing*, 20:357–366, 2010.

M. Stephens. 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)*, 59(4):768–769, 1997a.

M. Stephens. *Bayesian methods for mixture of normal distributions*. PhD thesis, D Phill Thesis. University of Oxford, Oxford., 1997b.

M. Stephens. Dealing with label switching in mixture models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, pages 795–809, 2000.

P. Stoica and Y. Selen. Model-order selection: a review of information criterion rules. *IEEE Signal Processing Magazine*, 21(4):36–47, 2004.

P. Stoica, R. L. Moses, B. Friedlander, and T. Soderstrom. Maximum likelihood estimation of the parameters of multiple sinusoids from noisy measurements. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 37(3):378–392, 1989.

R. Stoica, X. Descombes, and J. Zerubia. A Gibbs point process for road extraction from remotely sensed images. *International Journal of Computer Vision*, 57(2):121–136, 2004.

M. A. Tanner and W. H. Wong. The calculation of posterior distributions by data augmentation. *Journal of the American statistical Association*, 82(398):528–540, 1987. ISSN 0162-1459.

M. B. Thompson. Graphical comparison of MCMC performance. Technical report, Dept. of Statistics, University of Toronto. arXiv:1011.4457v1 [stat.CO], 2010.

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

L. Tierney. Markov chains for exploring posterior distributions. *The Annals of Statistics*, 22(4):1701–1728, 1994.

L. Tierney. A note on Metropolis-Hastings kernels for general state spaces. *Annals of Applied Probability*, pages 1–9, 1998.

D. A. Van Dyk and T. Park. Partially collapsed Gibbs samplers: theory and methods. *Journal of the American Statistical Association*, 103(482):790–796, 2008.

MNM Van Lieshout. *Markov Point Processes and their Applications*. Imperial College Press London, 2000.

G. C. G. Wei and M. A. Tanner. A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithms. *Journal of the American statistical Association*, 85(411):699–704, 1990.

M. West. Approximating posterior distributions by mixture. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 55(2):409–422, 1993. ISSN 0035-9246.

P. J. Wolfe, S. J. Godsill, and W. J. Ng. Bayesian variable selection and regularization for time–frequency surface estimation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 66(3):575–589, 2004.

W. Yao. Model based labeling for mixture models. *Statistics and Computing*, pages 1–11, 2011.

A. Zellner. On assessing prior distributions and Bayesian regression analysis with g-prior distributions. In P. K. Goel and A. Zellner, editors, *Bayesian Inference and Decision Techniques: Essays in Honor of Bruno de Finetti*, pages 233–243. North-Holland/Elsevier, 1986.