Non-informative priors and modelization by mixtures

Kaniav Kamary

THÈSE DE DOCTORAT
de l'Université de recherche
Paris Sciences et Lettres –
PSL Research University

préparée à l'Université
Paris – Dauphine

Lois *a priori* non-informatives et la modélisation par mélange

par Kaniav KAMARY

Ecole doctorale n°543
Spécialité : Sciences

Soutenue le 15.03.2016

Composition du Jury :

M. Christian P. ROBERT
Université Paris-Dauphine
Directeur de thèse

M. Gilles CELEUX
INRIA
Rapporteur

M. Pierre DRUILHET
Université Clermont-Ferrand 2
Rapporteur

Mme Judith ROUSSEAU
Université Paris-Dauphine
Membre du jury

Mme Anne PHILIPPE
Université de Nantes
Présidente du jury

DAUPHINE UNIVERSITÉ PARIS | PSL RESEARCH UNIVERSITY PARIS

Acknowledgments

I have been very fortunate throughout my thesis to be supported by internationally renowned researchers. Now I have a great opportunity to say a big thank you to persons who contributed to my professional and human formation in the doctoral stage.

I would first like to thank my PhD supervisor, Christian P. Robert, for all the support he has given me throughout my time as a student. By his scientific personality and his rigor, he deeply influenced me and made me more and more passionate about my work. Numerous advices in my researches, always very judicious and confidence he has manifested me, encouraged me to improve myself and to progress every day.

My deep gratitude is for Hossein Bevrani who supports me as a professor for the license and Master and advisor for the PhD. I will always be grateful for his kindness, his confidence and all the time he gave me without ever counting.

During my PhD I had the chance to work with researchers like Kerrie Mengersen, Judith Rousseau and Kate Lee as my collaborators and I want to thank them for their support and for agreeing to work with me. I think in particular of working with Kate Lee which was really rewarding.

I would like to warmly thank the team of the Kurdish Institute of Paris, Campus France and French government for their financial support and for trusting me.

I am also very grateful to have been a PhD student at CEREMADE, université Paris-Dauphine, which always provided me with a lively and intellectually stimulating environment to work. I would like to thank all the faculty, Robin Ryder and particularly director Olivier Glass.

I am very grateful to my examiners, Gilles Celeux, Pierre Druilhet, Judith Rousseau and Anne Philippe, for taking the time to give me such useful feedback.

I would like to thank all the people I have met and worked with during my time at université Paris-Dauphine. It is not possible to thank everyone here but I would particularly like to mention Sofia Tsepletidou, Roxana Dumitrescu, Viviana Letizia and Nicolas Baradel ... for their scientific input, discussions, and friendship over the years.

Lastly, I would like to thank my husband Bewar, my parents and my sisters who always believed in me and always offered me unconditional love and also my friends Ayoub Moradi, Serwa Khoramdel and Fateme Mohamadipoor who have supported me all these years, even from thousands of miles away.

Résumé

L'une des grandes applications de la statistique est la validation et la comparaison de modèles probabilistes au vu des données. Cette branche des statistiques a été développée depuis la formalisation de la fin du 19ième siècle par des pionniers comme Gosset, Pearson et Fisher. Dans le cas particulier de l'approche bayésienne, la solution à la comparaison de modèles est le facteur de Bayes, rapport des vraisemblances marginales, quelque soit le modèle évalué. Cette solution est obtenue par un raisonnement mathématique fondé sur une fonction de coût.

Ce facteur de Bayes pose cependant problème et ce pour deux raisons. D'une part, le facteur de Bayes est très peu utilisé du fait d'une forte dépendance à la loi a priori (ou de manière équivalente du fait d'une absence de calibration absolue). Néanmoins la sélection d'une loi a priori a un rôle vital dans la statistique bayésienne et par consequent l'une des difficultés avec la version traditionnelle de l'approche bayésienne est la discontinuité de l'utilisation des lois a priori impropres car ils ne sont pas justifiées dans la plupart des situations de test. La première partie de cette thèse traite d'un examen général sur les lois a priori non informatives, de leurs caractéristiques et montre la stabilité globale des distributions a posteriori en réévaluant les exemples de [Seaman III 2012].

Le second problème, indépendant, est que le facteur de Bayes est difficile à calculer à l'exception des cas les plus simples (lois conjuguées). Une branche des statistiques computationnelles s'est donc attachée à résoudre ce problème, avec des solutions empruntant à la physique statistique comme la méthode du path sampling de [Gelman 1998] et à la théorie du signal. Les solutions existantes ne sont cependant pas universelles et une réévaluation de ces méthodes suivie du développement de méthodes alternatives constitue une partie de la thèse. Nous considérons donc un nouveau paradigme pour les tests bayésiens d'hypothèses et la comparaison de modèles bayésiens en définissant une alternative à la construction traditionnelle de probabilités a posteriori qu'une hypothèse est vraie ou que les données proviennent d'un modèle spécifique. Cette méthode se fonde sur l'examen des modèles en compétition en tant que composants d'un modèle de mélange. En remplaçant le problème de test original avec une estimation qui se concentre sur le poids de probabilité d'un modèle donné dans un modèle de mélange, nous analysons la sensibilité sur la distribution a posteriori conséquente des poids pour divers modélisation préalables sur les poids et soulignons qu'un intérêt important de l'utilisation de cette perspective est que les lois a priori impropres génériques sont acceptables, tout en ne mettant pas en péril la convergence. Pour cela, les méthodes MCMC comme l'algorithme de Metropolis-Hastings et l'échantillonneur de Gibbs et des approximations de la probabilité par des méthodes empiriques sont utilisées. Une autre caractéristique de cette variante facilement mise en oeuvre est que les vitesses de convergence de la partie postérieure de la moyenne du poids et de probabilité a posteriori correspondant sont assez similaires à la solution bayésienne classique.

Dans la dernière partie de la thèse, nous sommes intéressés à la construction d'une analyse bayésienne de référence pour mélanges de gaussiennes par la création

d'une nouvelle paramétrisation centrée sur la moyenne et la variance de ces modèles, ce qui nous permet de développer une loi a priori non-informative pour les mélanges avec un nombre arbitraire de composants. Nous démontrons que la distribution postérieure associée à ce préalable est propre et fournissons des implémentations MCMC qui exhibent l'échangeabilité attendu. L'analyse repose sur des méthodes MCMC comme l'algorithme de Metropolis-within-Gibbs, Adaptive MCMC et l'algorithme de "Parallel Tempering". Cette partie de la thèse est suivie par une package R nommée **Ultimixt** qui met en œuvre une description de notre analyse bayésienne générique de mélanges de gaussiennes unidimensionnelles obtenues par une paramétrisation moyenne–variance du modèle. **Ultimixt** peut être appliqué à une analyse bayésienne des mélanges gaussiennes avec un nombre arbitraire de composants, sans avoir besoin de définir la loi a priori.
Mots clés: Distribution de mélange, Loi a priori non-informative, Analyse bayésienne, A priori impropre, Choix du modèle bayésien, Méthodes de MCMC.

Abstract

One of the major applications of statistics is the validation and comparing probabilistic models given the data. This branch statistics has been developed since the formalization of the late 19th century by pioneers like Gosset, Pearson and Fisher. In the special case of the Bayesian approach, the comparison solution of models is the Bayes factor, ratio of marginal likelihoods, whatever the estimated model. This solution is obtained by a mathematical reasoning based on a loss function.

Despite a frequent use of Bayes factor and its equivalent, the posterior probability of models, by the Bayesian community, it is however problematic in some cases. First, this rule is highly dependent on the prior modeling even with large datasets and as the selection of a prior density has a vital role in Bayesian statistics, one of difficulties with the traditional handling of Bayesian tests is a discontinuity in the use of improper priors since they are not justified in most testing situations. The first part of this thesis deals with a general review on non-informative priors, their features and demonstrating the overall stability of posterior distributions by reassessing examples of [Seaman III 2012].

Beside that, Bayes factors are difficult to calculate except in the simplest cases (conjugate distributions). A branch of computational statistics has therefore emerged to resolve this problem with solutions borrowing from statistical physics as the path sampling method of [Gelman 1998] and from signal processing. The existing solutions are not, however, universal and a reassessment of the methods followed by alternative methods is a part of the thesis. We therefore consider a novel paradigm for Bayesian testing of hypotheses and Bayesian model comparison. The idea is to define an alternative to the traditional construction of posterior probabilities that a given hypothesis is true or that the data originates from a specific model which is based on considering the models under comparison as components of a mixture model. By replacing the original testing problem with an estimation version that focus on the probability weight of a given model within a mixture model, we analyze the sensitivity on the resulting posterior distribution of the weights for various prior modelings on the weights and stress that a major appeal in using this novel perspective is that generic improper priors are acceptable, while not putting convergence in jeopardy. MCMC methods like Metropolis-Hastings algorithm and the Gibbs sampler are used. From a computational viewpoint, another feature of this easily implemented alternative to the classical Bayesian solution is that the speeds of convergence of the posterior mean of the weight and of the corresponding posterior probability are quite similar.

In the last part of the thesis we construct a reference Bayesian analysis of mixtures of Gaussian distributions by creating a new parameterization centered on the mean and variance of those models itself. This enables us to develop a genuine non-informative prior for Gaussian mixtures with an arbitrary number of components. We demonstrate that the posterior distribution associated with this prior is almost surely proper and provide MCMC implementations that exhibit the expected component exchangeability. The analyses are based on MCMC methods as

the Metropolis-within-Gibbs algorithm, adaptive MCMC and the Parallel tempering algorithm. This part of the thesis is followed by the description of **R** package named **Ultimixt** which implements a generic reference Bayesian analysis of unidimensional mixtures of Gaussian distributions obtained by a location-scale parameterization of the model. This package can be applied to produce a Bayesian analysis of Gaussian mixtures with an arbitrary number of components, with no need to specify the prior distribution.

Keywords: Mixture distribution, Non-informative prior, Bayesian analysis, Improper prior, Bayesian model choice, MCMC methods.

Contents

Chapter 1

General Introduction

1.1 Overview

In several areas of science, statistics is a powerful tool to analyze data both from controlled experiments such as natural sciences and from observational studies, mainly in the human sciences. Basically, a researcher expects to find methods which can provide means to judge a population from a subset of it, named a sample. Statistics has developed many different theories to be applied in different situations, and all of them have a characteristic in common and that is, given the uncertainty, they try to find the best strategy to answer scientists' queries.

In order to apply statistics to a problem, it is a common practice to start with a population or process to be studied. When the entire population is not available and only samples are studied, the inferential statistics is needed. These inferences can take the form of testing hypotheses, estimation, regression analysis, prediction and some other technics that have been recently developed such as spatial data and data mining. Furthermore, statistical inference defines random samples and describes the population being examined by a probability distribution that may have unknown parameters. Indeed, the main purpose of statistical theories is to infer properties about the probability distribution of the population of interest using observations. To do so, different paradigms of statistical inference have become established. Bayesian inference is considered as an important statistical technique especially in mathematical statistics because of its application in science beside a wide range of activities such as engineering, philosophy, medicine, sport, and law. In this thesis, we focus on the Bayesian inference. Although the original Bayesian theory was settled in the 18^{th} century, due to various previous computational difficulties, only in the last 30 years, the Bayesian method has grown substantially. This growth in research and applications of Bayesian methods refers to the 1980s which mostly attributed to the discovery of Markov Chain Monte Carlo methods which removed many of the computational problems.

This thesis consists of four general parts which are briefly introduced in the following sections.

1.2 Prior distribution

The Bayesian theory deals with probability statements which are conditional on the observed value and this conditional feature introduces the main difference between Bayesian and classical inferences. Despite the differences between these statistical

methods, in many simple analysis we get superficially similar conclusions from the two approaches. A Bayesian statistical inference is based on a prior probability distribution of an uncertain quantity that expresses one's beliefs about this quantity before some evidence is taken into account. In other words, a prior distribution is the distribution of this uncertain quantity, named parameter, before any data is observed. Once this prior distribution is set, Bayesian inference is straightforward in terms of minimizing posterior losses, computing higher posterior density or finding the predictive distribution [Robert 2001]. But in general, a prior distribution is not easy to precisely find out and most of critics of the Bayesian analysis focussed on the choice of the prior distributions. Furthermore, different perspectives are available to choose a prior while the impact of this choice on the resulting posterior inference should not be omitted even in the case it is negligible. The main point here is about the existence of a prior or the determination of an exact or even a parametrized distribution for the prior on the parameter, which is never unique.

However, the prior plays a fundamental role in drawing Bayesian inference because of its exploitation combined with the probability distribution of data to yield the posterior distribution. Bayesian inference is fundamentally based on the posterior distribution which is used for future inference and decisions involving the parameter. [Gelman 2002] pointed out the assessment of the information that can be included in prior distributions and the properties of the resulting posterior distributions, as key issues in setting a prior. He also mentioned prior distributions as the key part of Bayesian inference and classified them to three categories: Non-informative priors, highly informative and moderately informative hierarchical prior distributions.

In fact, the existence of fairly precise scientific or lack of information about the parameter of interest leads to two classes of priors: Informative or subjective prior, and non-informative or objective priors. One method of determining the prior is a subjective evaluation of the prior probability that can be done by using past experiments of the same problem that is considered as an approximation to the real prior distribution [Robert 2001]. Another methods are based on the maximum entropy developed in [Jaynes 1980, Jaynes 1983] and as well as parametric approximations for priors resulting from restricting the choice of prior to a parametrized density and characterize the corresponding parameters using classical methods [Robert 2001]. Finally, other techniques such as empirical and hierarchical Bayes incorporate uncertainty about the prior distribution (for details see [Robert 2001]). All these methods depend on the availability of the information on the parameter of interest. In the case of limited prior input, conjugate priors can be used to construct the prior distribution, which originated in [Raiffa 1961] and even if this choice may influence the resulting Bayesian inference, conjugate priors are not considered as part of the non-informative prior class [Robert 2001]. The most popular conjugate priors are related to the distributions associated with the exponential families which are called natural conjugate priors [Robert 2001]. This family of distributions is the only case where conjugate priors are guaranteed to exist. Despite the advantages such as being easy to deal with in both cases mathematically and computationally, the con-

jugate priors are not away from criticism. One reason is that these distributions are overly restrictive and also they are not necessarily considered as the most robust prior distributions.

The non-informative priors are requested when no information about the parameter is available. While informative priors are far from enough to allow hopes of achieving, the use of non-informative priors also underwent vary criticisms because of their influences on the relative posterior distribution. Laplace's prior is the simplest and oldest non-informative prior that is based on the principle of indifference by assigning equal probabilities to all possibilities. This prior was criticized because it results in improper resulting distributions in the case where the parameter space is infinite. This is not always a serious problem since it may lead to proper posteriors. However, the use of improper non-informative priors may also cause problem such as the marginalization paradox shown by [Stone 1972]. Some others are the possible inadmissibility of resulting Bayes estimators, Stein's paradox [Syversveen 1998] and in addition to the possibility of resulting improper posteriors [Kass 1996], considering equal probabilities for possible events is not coherent under partitioning as pointed out by [Robert 2001]. Another issue is the lack of invariance under the reparametrization of the parameter. The invariance of a prior is necessary when more than one inference about the parameter is needed. The best solution for obtaining invariant non-informative priors was represented by Jeffreys' distributions [Jeffreys 1939] where the information matrix of the sampling model is turned into a prior distribution. Jeffreys' prior is most often improper which means that it does not integrate to a finite value. Another method that was initially described by [Bernardo 1979] and further developed by [Berger 1979] is the reference prior. The advantages of this method compared with Jeffreys' method appear in the case of multidimensional problems [Syversveen 1998]. Some other methods have been also suggested by [Box 2011, Rissanen 2012, Welch 1963].

Since there is no best prior that one should use, research aims at acceding a prior so that posterior distribution is well behaved and proper while all available information about the parameter is taken into account. Recently, due to theoretical developments on sensitivity analysis, the dependence of posteriors on prior distributions can be checked by methods such as comparing posterior inferences under different reasonable choices of prior distribution. The first part of this thesis deals with selecting non-informative priors based on a critical review of [Seaman III 2012] and the main result of this work is to show that the Bayesian data analysis remains stable under different choices of non-informative prior distributions. A related paper was published in the journal of *Applied and Computational Mathematics* in July 21, 2014.

In the literature we can find a lot of theoretical and applied overviews of Bayesian statistics about the uses of non-informative priors (see [Bernardo 1994, Carlin 1996, Gelman 2013a]). A variety of methods of driving non-informative priors have been covered by [Yang 1996]. He also listed known properties of these prior distributions. Despite the wide application of non-informative priors by Bayesian community, the handling of non-informative Bayesian testing is mostly unresolved. In the following

section, we briefly address hypotheses testing and related concepts.

1.3 Bayesian model choice

As mentioned at the beginning of this chapter, among many other types of statistical inference, hypotheses testing or equivalently model selection techniques are widely applied for data analysis. Statistical hypothesis tests define a procedure of controlling the probability of incorrectly deciding that a so-called null hypothesis is false.

Differences among statistical paradigms such as frequency-based or Bayesian methods are generally much more pronounced in model checking and selection than in fitting. In a Bayesian paradigm the typical method for comparing two models involves the Bayes factors or the posterior probability of the models which are based on a specification of both likelihood and prior distribution and both are compared together. Unlike standard frequency-based methods both Bayes factors and posterior probability treat the models under comparison essentially symmetrically. However, from both classical and Bayesian points of view, model selection is the problem in which we have to choose between some models on the basis of observed data but the Bayesian model comparison based on the Bayes factors does not depend on the parameters because of the integration over all parameters in each model. On the other hand, the use of Bayes factors has the advantage of automatically including a penalty for too much model structure [Kass 1995].

The literature on Bayesian model choice is considerable by now and one of the earlier, reasonably thorough reviews, appears in [Gelfand 1992]. The Bayes factors have also been the subject of much discussion in the literature in recent years and one of the comprehensive review of Bayes factors, their computation and usage in Bayesian hypothesis testing goes back to 1995 by [Kass 1995] who proposed this criterion as a solution for the comparison of models problem. However, the decision based on the Bayes factors requires a zero-one loss and [Kadane 1980] shows that these criterions are sufficient if and only if a zero-one loss obtains. Many other works on Bayesian model selection, Bayes factors and their features can be found in [Good 1950, Berger 1996].

Because of the difficulties caused by prior specification, the Bayesian approach to test hypotheses is not always straightforward especially in the case of an absolute lack of information. In fact, the use of non-informative prior distributions for testing hypotheses is delicate because of the sensitivity of Bayes factors to the choice of the prior. The typical strategy of using non-informative prior distributions with large variances clearly affects the Bayes factors [Robert 2001]. Furthermore, improper prior distributions result in improper prior predictive distributions and undefined Bayes factors. Among some other difficulties caused by Bayes factors that will be addressed in Chapter 4, a principal drawback from which both criterions, Bayes factors and posterior probability of models, suffer is that they can be difficult to compute. In all but the simplest cases, Bayes factors must be evaluated numerically

using methods such as importance sampling, bridge sampling and reversible jump Markov Chain Monte Carlo [Green 1995]. Another method has also been recently produced by [O'Neill 2014] for computing Bayes factors that avoids the need to use reversible jump approaches. [O'Neill 2014] show that Bayes factors for the models can be expressed in terms of the posterior means of the mixture probabilities, and thus estimated from the MCMC output. In the other hand, one solution in the case that the likelihood is not available or too costly to evaluate numerically, is the approximate Bayesian computation. Some of related works can be found in [Csilléry 2010, Toni 2010, Rattan 2013] for instance. Other proposals have been made to solve particular problems with the ordinary Bayes factor such as intrinsic Bayes factors [Berger 1996] with further modifications such as the trimmed and median variants, fractional Bayes factors [O'Hagan 1995] and posterior Bayes factors [Aitkin 1991]. Consideration of Bayes factors also leads to two of the more common criteria used for model selection such as the Bayes Information Criterion (BIC) or Schwartz's criterion that provides a cursory first-order approximation to the Bayes factor [Robert 2001] and the Akaike Information Criterion (or AIC) [Akaike 1973]. A Bayesian alternative to both BIC and AIC based on the deviance has been developed by [Spiegelhalter 1998] which takes into account the prior information.
Because the existing solutions are not, however, universal in the second part of this thesis our focus is towards addressing the difficulties with the traditional handling of Bayesian model selection using Bayes factors by proposing a method which goes some way to removing these complications. The key idea is to consider a mixture model whose components are the competing models of interest and the traditional method for the model choice is replaced by a kind of Bayesian estimation problem that focuses on the probability weight of the mixture model. The method includes a novel strategy of reparametrizing the competing models towards common meaning parameters in all models, that allows for using the non-informative priors at least on the common parameters. Two substantial advantages of our method are the usability of the non-informative priors for Bayesian model choice and the other is that due to the standard MCMC algorithms, the Bayesian estimation of the model is straightforward and there is no need to compute the marginal likelihoods. A related paper was submitted for publication.

The third part of this thesis focuses on the parametrization of the mixture distributions. In the following we briefly introduce the motivation of this work.

1.4 Mixture distributions

The earliest study about the mixture models was done by [Pearson 1894] who investigated the estimation of parameters in the finite mixture model by the use of the method of moments. In 1894, [Pearson 1894] studied the dissection of asymptotic and symmetric frequency curves into two components of normal distributions. Many other papers have appeared related to the problem of statistical inference about the parameters and probabilistic properties of these densities. Since this early work,

finite mixture models have been widely used in many disciplines and there is a large body of literature on these distributions. For example in biology it is often desired to measure certain characteristics in natural populations of some particular species when the distribution of such characteristics may vary markedly with age of the individuals. Since age is difficult to ascertain in samples from populations, the biologist is dealing with a mixture of distributions and the mixing in this case is done over a parameter depending on the unobservable variate, age. Some other applications can be found in astronomy, ecology, genetics and so on due to the feature that they are easily applied to the data set in which two or more subpopulations are mixed together. In statistical applications, the mixture of densities can be used to approximate some parameters associated with a density.

The finite mixture models have also enjoyed intensive attentions over the recent years from both practical and theoretical viewpoints due to their flexibility in modeling. Some basic properties of mixtures were studied by [Robbins 1948] and [Robbins 1961] initiated the study of identifiability problem. Despite the popularity of mixtures, model estimation can be difficult when the number of components is unknown. In 1966, [Hasselblad 1966] first considered the estimation problem of mixtures by the method of maximum likelihood. [Rolph 1968] first considered Bayes estimation of the mixture parameters in the special case where the observations from the mixture population are restricted to the positive integers. In the framework of the Bayesian approach, one needs to assume that a prior distribution on component parameters is available. As summarized in [Frühwirth-Schnatter 2006], there are two main reasons why people may be interested in using the Bayesian method in finite mixture models. Firstly, including a suitable prior distribution or the parameters in the framework of the Bayesian approach may avoid spurious modes when maximizing the log-likelihood function. Secondly, when the posterior distribution for the unknown parameters is available, the Bayesian method can yield valid inference without relying on asymptotic normality. This is an advantage of the Bayesian method for estimating the parameters of a mixture distribution without the need of sample sizes very large. As mentioned before, the use of the conjugate prior produces the posterior distribution that may belong to the tractable distribution family. However, because of the complexity of mixtures, it is impossible to find a conjugate prior for the component parameters. While the posterior distributions derived from the mixture models are non standard, MCMC methods are used to generate samples from these complex distributions [Marin 2006, Frühwirth-Schnatter 2006]. Because the main idea of Bayesian estimation using MCMC methods followed by realizing a mixture model is considered as a special case of incomplete data problem with the missing component indicator variables, the problem with conjugate priors no longer poses serious obstacles to the application of Bayesian method.

In fact, the Bayesian estimators in mixture models are always well defined as long as priors are proper. Furthermore, the unidentifiability may be resolved by well defining the parameter space or using informative priors on parameters. However, in the case where no information is available for the component parameters, the choice of the prior is more delicate. [Marin 2006] demonstrates that specifying improper

prior to the component parameters results in improper posterior distribution that prohibits this kind of prior to be used for mixtures. In addition, non-informative priors assigned to the parameter of a specific component can also lead to identifiability problems. Because if each component has its own prior parameters and few observations are allocated to this component, there will be no information at all to estimate the parameter and in the case of Gibbs sampling, the sampler gets trapped in a local mode corresponding to this component.

This problem of non-identifiability in the posterior distribution can also be due to an overfitting phenomenon. Basically this happens when some components have weights equal to zero or merged together [Frühwirth-Schnatter 2006]. A full discussion about how over fitted mixtures behave can be found in [Rousseau 2011] who proved that the posterior behavior of overfitted mixtures generally depends on both the choice of the prior on the weights and the number of free parameters. [van Havre 2015] treated the issues such as non-identifiability due to overfitting, label switching and also the problem of lack of mixing caused by applying standard MCMC sampling techniques when the posterior contains multiple well separated modes.

Given the difficulty with non-informative priors, one solution is to use proper priors with the prior parameters chosen such that the prior is suitably weakly informative priors [Richardson 1997]. This method is not always applicable because of the problem of multiple prior specifications. Another method proposed by [Diebolt 1994] is to use an improper prior under the condition of forcing each component to always have a minimal number of data points assigned to it. A related work has been recently developed by [Stoneking 2014] which does not result in any data dependence of the priors.

In the third part of this thesis we define a novel reparametrisation for the mixture of distributions based on the mean and standard deviation of the mixture itself, namely global parameters. The main feature of our method is that the non-informative prior distribution can be used on the global parameters of the mixture while the resulting posterior distribution is proper. A related paper was submitted for publication.

The reparametrized mixture model will be fitted with our **R** package named `Ultimixt`. `Ultimixt` provides the functionality for estimating reparametrized Gaussian mixture models with MCMC methods. The last part of this thesis pertains to the description of the implementation and the functions of `Ultimixt`. This package can accurately compute the posterior estimate of the parameters of reparametrized univariate Gaussian mixture distribution beside having the ability of graphically summarizing the posterior results.

Chapter 2

Reflecting about Selecting Noninformative Priors

Joint work with Christian P. Robert

Abstract

Following the critical review of [Seaman III 2012], we reflect on what is presumably the most essential aspect of Bayesian statistics, namely the selection of a prior density. In some cases, Bayesian inference remains fairly stable under a large range of noninformative prior distributions. However, as discussed by [Seaman III 2012], there may also be unintended consequences of a choice of a noninformative prior and, these authors consider this problem ignored in Bayesian studies. As they based their argumentation on four examples, we reassess these examples and their Bayesian processing via different prior choices. Our conclusion is to lower the degree of worry about the impact of the prior, exhibiting an overall stability of the posterior distributions. We thus consider that the warnings of [Seaman III 2012], while commendable, do not jeopardize the use of most noninformative priors.

Keywords: Induced prior, Logistic model, Bayesian methods, Stability, Prior distribution

2.1 Introduction

The choice of a particular prior for the Bayesian analysis of a statistical model is often seen more as an art than as a science. When the prior cannot be derived from the available information, it is generally constructed as a noninformative prior. This derivation is mostly mathematical and, even though the corresponding posterior distribution has to be proper and hence constitutes a correct probability density, it nonetheless leaves the door open to criticism. The focus of this note is the paper by [Seaman III 2012], where the authors consider using a particular noninformative distribution as a problem in itself, often bypassed by users of these priors: "if parameters with diffuse proper priors are subsequently transformed, the resulting induced priors can, of course, be far from diffuse, possibly resulting in unintended influence on the posterior of the transformed parameters" (p.77). Using the inexact argument that most problems rely on MCMC methods and *hence* require proper priors, the authors restrict the focus to those priors.

In their critical study, [Seaman III 2012] investigate the negative side effects of some specific prior choices related with specific examples. Our note aims at

re-examining their investigation and at providing a more balanced discussion on these side effects. We first stress that a prior is considered as *informative* by [Seaman III 2012] "to the degree it renders some values of the quantity of interest more likely than others" (p.77), and with this definition, when comparing two priors, the prior that is more informative is deemed preferable. In contrast with this definition, we consider that an *informative* prior expresses specific, definite (prior) information about the parameter, providing quantitative information that is crucial to the estimation of a model through restrictions on the prior distribution [Robert 2007]. However, in most practical cases, a model parameter has no substance *per se* but instead calibrates the probability law of the random phenomenon observed therein. The prior is thus a tool that summarizes the information available on this phenomenon, as well as the uncertainty within the Bayesian structure. Many discussions can be found in the literature on how appropriate choices between the prior distributions can be decided. In this case, robustness considerations also have an important role to play [Lopes 2011, Stojanovski 2011]. This point of view will be obvious in this note as, e.g., in processing a logistic model in the following section. Within the sole setting of the examples first processed in [Seaman III 2012], we do exhibit a greater stability in the posterior distributions through various noninformative priors.

The plan of the note is as follows: we first provide a brief review of noninformative priors in Section 4.2. In Section 2.3, we propose a Bayesian analysis of a logistic model (Seaman III et al.'s (2012) first example) by choosing the normal distribution $N(0, \sigma^2)$ as the regression coefficient prior. We then compare it with a g-prior, as well as flat and Jeffreys' priors, concluding to the stability of our results. The next sections cover the second to fourth examples of [Seaman III 2012], modeling covariance matrices, treatment effect in biomedical studies, and a multinomial distribution. When modeling covariance matrices, we compare two default priors for the standard deviations of the model coefficients. In the multinomial setting, we discuss the hyperparameters of a Dirichlet prior. Finally, we conclude with the argument that the use of noninformative priors is reasonable within a fair range and that they provide efficient Bayesian estimations when the information about the parameter is vague or very poor.

2.2 Noninformative priors

As mentioned above, when prior information is unavailable and if we stick to Bayesian analysis, we need to resort to one of the so-called *noninformative priors*. Since we aim at a prior with minimal impact on the final inference, we define a noninformative prior as a statistical distribution that expresses vague or general information about the parameter in which we are interested. In constructive terms, the first rule for determining a noninformative prior is the principle of indifference, using uniform distributions which assign equal probabilities to all possibilities [Laplace 1820]. This distribution is however not invariant under reparametrization

,(see [Berger 1980, Robert 2007] for references). If the problem does not allow for an invariance structure, Jeffreys' priors [Jeffreys 1939], then reference priors, exploit the probabilistic structure of the problem under study in a more formalized way. Other methods have been advanced, like the little-known data-translated likelihood of [Box 2011], maxent priors [Jaynes 2003], minimum description length priors [Rissanen 2012] and probability matching priors [Welch 1963].

[Bernardo 2009] envision noninformative priors as a mere mathematical tool, while accepting their feature of minimizing the impact of the prior selection on inference: "Put bluntly, data cannot ever speak entirely for themselves, every prior specification has some informative posterior or predictive implications and *vague* is itself much too vague an idea to be useful. There is no "objective" prior that represents ignorance" (p.298). There is little to object against this quote since, indeed, prior distributions can never be quantified or elicited exactly, especially when no information is available on those parameters. Hence, the concept of "true" prior is meaningless and the quantification of prior beliefs operates under uncertainty. As stressed by [Berger 1994], noninformative priors enjoy the advantage that they can be considered to provide robust solutions to relevant problems even though "the user of these priors should be concerned with robustness with respect to the class of reasonable noninformative priors" (p.59).

2.3 Example 1: Bayesian analysis of the logistic model

The first example in [Seaman III 2012] is a standard logistic regression modeling the probability of coronary heart disease as dependent on the age x by

$$\rho(x) = \frac{\exp(\alpha + \beta x)}{1 + \exp(\alpha + \beta x)}. \tag{2.1}$$

First we recall the original discussion in [Seaman III 2012] and then run our own analysis by selecting some normal priors as well as the g-prior, the flat prior and Jeffreys' prior.

2.3.1 Seaman et al.'s (2012) analysis

For both parameters of the model (2.1), [Seaman III 2012] chose a normal prior $N(0, \sigma^2)$. A first surprising feature in this choice is to opt for an *identical* prior on both intercept and slope coefficients, instead of, e.g., a g-prior (discussed in the following) that would rescale each coefficient according to the variation of the corresponding covariate. Indeed, since x corresponds to age, the second term βx in the regression varies 50 times more than the intercept. When plotting logistic cdf's induced by a few thousands simulations from the prior, those cumulative functions mostly end up as constant functions with the extreme values 0 and 1. This behavior is obviously not particularly realistic since the predicted phenomenon is the occurrence of coronary heart disease. Under this minimal amount of information,

the prior is thus using the wrong scale: the simulated cdfs should have a reasonable behavior over the range $(20, 100)$ of the covariate x. For instance, it should focus on a -5 log-odds ratio at age 20 and a $+5$ log-odds ratio at 100, leading to the comparison pictured in Figure 2.1 (left versus right). Furthermore, the fact that the coefficient of x may be negative also bypasses a basic item of information about the model and answers the later self-criticism in [Seaman III 2012] that the prior probability that the ED50 is negative is 0.5. Using instead a flat prior would answer the authors' criticisms about the prior behavior, as we now demonstrate.

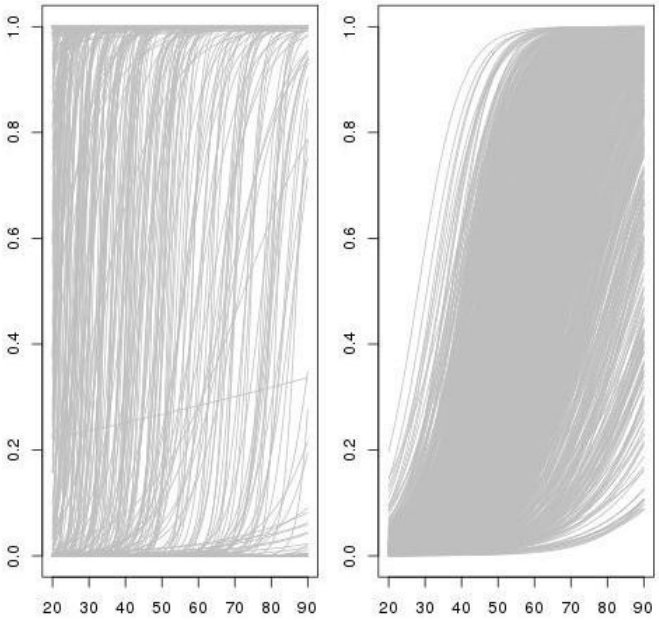


Figure 2.1: Logistic cdfs across a few thousand simulations from the normal prior, when using the prior selected by [Seaman III 2012] (left) and the prior defined as the G-prior(right)

We stress that [Seaman III 2012] produce no further justification for the choice of the prior variance $\sigma^2 = 25^2$, other than there is no information about the model parameters. This is a completely arbitrary choice of prior, arbitrariness that does have a considerable impact on the resulting inference, as already discussed. [Seaman III 2012] further criticized the chosen prior by comparing both posterior mode and posterior mean derived from the normal prior assumption with the MLE. If the MLE is the golden standard there then one may wonder about the relevance of a Bayesian analysis! When the sample size N gets large, most simple Bayesian analyses based on noninformative prior distributions give results similar to standard non-Bayesian approaches [Gelman 2013a]. For instance, we can often interpret classical point estimates as exact or approximate posterior summaries based on some implicit full probability model. Therefore, as N increases, the influence of the prior on posterior inferences decreases and, when N goes to infinity, most priors lead to the same inference. However, for smaller sample sizes, it is inappropriate to

$\sigma = 10$			
$\hat{\alpha}$		$\hat{\beta}$	
mean	s.d	mean	s.d
3.482	11.6554	-0.0161	0.0541
$\sigma = 25$			
18.969	24.119	-0.0882	0.1127
$\sigma = 100$			
137.63	64.87	-0.6404	0.3019
$\sigma = 900$			
237.2	86.12	-1.106	0.401

Table 2.1: Posterior estimates of the logistic parameters using a normal prior when $\sigma = 10, 25, 100, 900$

summarize inference about the parameter by one value like the mode or the mean, especially when the posterior distribution of the parameter is more variable or even asymmetric.

The dataset used here to infer on (α, β) is the Swiss banknote benchmark (available in R). The response variable y indicates the state of the banknote, i.e. whether the bank note is genuine or counterfeit. The explanatory variable is the bill length. This data yields the maximum likelihood estimates $\tilde{\alpha} = 233.26$ and $\tilde{\beta} = -1.09$. To check the impact of the normal prior variance, we used a random walk Metropolis-Hastings algorithm as in [Marin 2007] and derived the estimators reproduced in Table 2.1. We can spot definitive changes in the results that are caused by moves in the coefficient σ, hence concluding to the clear sensitivity of the posterior to the choice of hyperparameter σ (see also Figure 2.2).

2.3.2 Larger classes of priors

Normal priors are well-know for their lack of robustness (see e.g. [Berger 1994]) and the previous section demonstrates the long-term impact of σ. However, we can limit variations in the posteriors, using the g-priors of [Zellner 1986],

$$\alpha, \beta \mid X \sim N_2(0, g(X^T X)^{-1}). \tag{2.2}$$

where the prior variance-covariance matrix is a scalar multiple of the information matrix for the linear regression. This coefficient g plays a decisive role in the analysis, however large values of g imply a more diffuse prior and, as shown e.g. in [Marin 2007], if the value of g is large enough, the Bayes estimate stabilizes. We will select g as equal to the sample size 200, following [Liang 2008], as it means that the amount of information about the parameter is equal to the amount of information contained in one single observation.

A second reference prior is the flat prior $\pi(\alpha, \beta) = 1$. And Jeffreys' prior constitutes our third prior as in [Marin 2007]. In the logistic case, Fisher's information matrix is $\mathbf{I}(\alpha, \beta, X) = X^T W X$, where $X = \{x_{ir}\}$ is the design matrix,

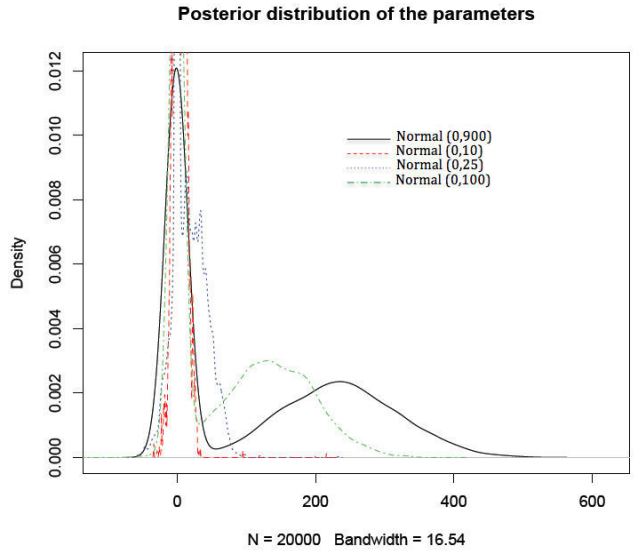


Figure 2.2: Posterior distributions of the logistic parameter α when priors are $N(0, \sigma)$ for $\sigma = 10, 25, 100, 900$, based on 10^4 MCMC simulations.

$W = \text{diag}\{m_i\pi_i(1-\pi_i)\}$ and m_i is the binomial index for the ith count [Firth 1993]. This leads to Jeffreys' prior $\{\det(\mathbf{I}(\alpha, \beta, X))\}^{\frac{1}{2}}$, proportional to

$$\left[\sum_{i=1}^{n} \frac{\exp(\alpha+\beta x_i)}{\{1+\exp(\alpha+\beta x_i)\}^2} \sum_{i=1}^{n} \frac{x_i^2 \exp(\alpha+\beta x_i)}{\{1+\exp(\alpha+\beta x_i)\}^2} - \left\{\sum_{i=1}^{n} \frac{x_i \exp(\alpha+\beta x_i)}{\{1+\exp(\alpha+\beta x_i)\}^2}\right\}^2\right]^{\frac{1}{2}}$$

This is a nonstandard distribution on (α, β) but it can be easily approximated by a Metropolis-Hastings algorithm whose proposal is the normal Fisher approximation of the likelihood, as in [Marin 2007].

Bayesian estimates of the regression coefficients associated with the above three noninformative priors are summarized in Table 2.2. Those estimates vary quite moderately from one choice to the next, as well as relatively to the MLEs and to the results shown in Table 2.1 when $\sigma = 900$. Figure 2.3 is even more definitive about this stability of Bayesian inferences under different noninformative prior choices.

2.4 Example 2: Modeling covariance matrices

The second choice of prior criticized by [Seaman III 2012], was proposed by [Barnard 2000] for the modeling of covariance matrices. However the paper falls short of demonstrating a clear impact of this prior modeling on posterior inference. Furthermore the adopted solution of using another proper prior resulting in a "wider" dispersion requires a prior knowledge of how wide is wide enough. We thus run Bayesian analyses considering prior beliefs specified by both [Seaman III 2012] and [Barnard 2000].

g-prior			
$\hat{\alpha}$		$\hat{\beta}$	
mean	s.d	mean	s.d
237.63	88.0377	-1.1058	0.4097
Flat prior			
236.44	85.1049	-1.1003	0.3960
Jeffreys' prior			
237.24	87.0597	-1.1040	0.4051

Table 2.2: Posterior estimates of the logistic parameters under a g-prior, a flat prior and Jeffreys' prior for the banknote benchmark. Posterior means and standard deviations remain quite similar under all priors. All point estimates are averages of MCMC samples of size 10^4.

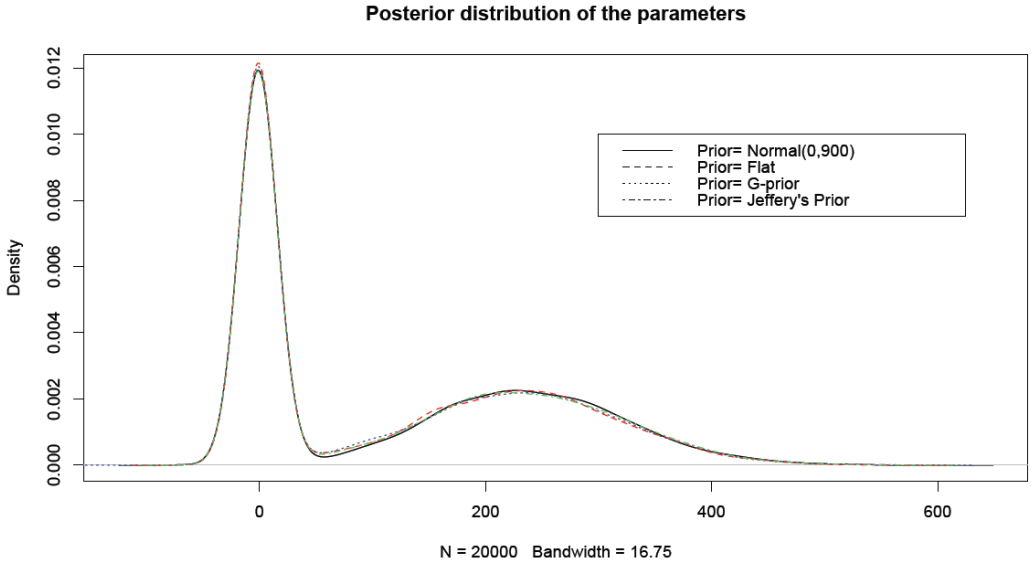


Figure 2.3: Posterior distributions of the parameters of the logistic model when the prior is $N(0, 900^2)$, g-prior, flat prior and Jeffreys' prior, respectively. The estimated posterior distributions are based on 10^4 MCMC iterations.

2.4.1 Setting

The multivariate regression model of [Barnard 2000] is

$$Y_j \mid X_j, \beta_j, \tau_j \sim N(X_j\beta_j, \tau_j^2 I_{n_j}), \quad j = 1, 2, \ldots, m. \tag{2.3}$$

where Y_j is a vector of n_j dependent variables, X_j is an $n_j \times k$ matrix of covariate variables, and β_j is a k-dimensional parameter vector. For this model, [Barnard 2000] considered an iid normal distribution as the prior

$$\beta_j \sim N(\bar{\beta}, \Sigma)$$

conditional on $\bar{\beta}, \Sigma$ where $\bar{\beta}, \tau_j^2$ for $j = 1, 2, \ldots, m$ are independent and follow a normal and inverse-gamma priors, respectively. Assuming that $\bar{\beta}, \tau_j^2$'s and Σ are a priori independent, [Barnard 2000] firstly provide a full discussion on how to choose a prior for Σ because it determines the nature of the shrinkage of the posterior of the individual β_j is towards a common target. The covariance matrix Σ is defined as a diagonal matrix with diagonal elements S, multiplied by a $k \times k$ correlation matrix R,

$$\Sigma = \text{diag}(S) R \text{diag}(S) \,.$$

Note that S is the $k \times 1$ vector of standard deviations of the β_js, $(S_1, \ldots, S_k)$. [Barnard 2000] propose lognormal distributions as priors on S_j. The correlation matrix could have (1) a joint uniform prior $p(R) \propto 1$, or (2) a marginal prior obtained from the inverse-Wishart distribution for Σ which means $p(R)$ is derived from the integral over $S_1, \ldots, S_k$ of a standard inverse-Wishart distribution. In the second case, all the marginal densities for r_{ij} are uniform when $i \neq j$ [Barnard 2000].

Considering the case of a single regressor, i.e. $k = 2$, [Seaman III 2012] chose a different prior structure, with a flat prior on the correlations and a lognormal prior with means 1 and -1, and standard deviations 1 and 0.5 on the standard deviations of the intercept and slope, respectively. Simulating from this prior, they concluded at a high concentration near zero. They then suggested that the lognormal distribution should be replaced by a gamma distribution $G(4, 1)$ as it implies a more diffuse prior. The main question here is whether or not the induced prior is more diffuse should make us prefer gamma to lognormal as a prior for S_j, as discussed below.

2.4.2 Prior beliefs

First, Barnard et al.'s (2000) basic modeling intuition is "that each regression is a particular instance of the same type of relationship" (p.1292). This means an exchangeable prior belief on the regression parameters. As an example, they suppose that m regressions are similar models where each regression corresponds to a different firm in the same industry branch. Exploiting this assumption, when β_j has a normal prior like $\beta_{ij} \sim N(\bar{\beta}_i, \sigma_i^2)$, $j = 1, 2, \ldots, m$, the standard deviation of β_{ij} $(S_i = \sigma_i)$ should be small as well so "that the coefficient for the ith explanatory variable is similar in the different regressions" (p.1293). In other words, S_i concentrated on small values implies little variation in the ith coefficient. Toward this goal, [Barnard 2000] chose a prior concentrated close to zero for the standard deviation of the slope so that the posterior of this coefficient would be shrunken together across the regressions. Based on this basic idea and taking tight priors on Σ for $\beta_j, j = 1, \ldots, m$, they investigated the shrinkage of the posterior on β_j as well as the degree of similarity of the slopes. Their analysis showed that a standard deviation prior that is more concentrated on small values results in substantial shrinkage in the coefficients relative to other prior choices.

Consider for instance the variation between the choices of lognormal and gamma distributions as priors of S_2, standard deviation of the regression slope. Figure 2.4

compares the lognormal prior with mean -1 and standard deviation 0.5 and the gamma distribution $G(4, 1)$.

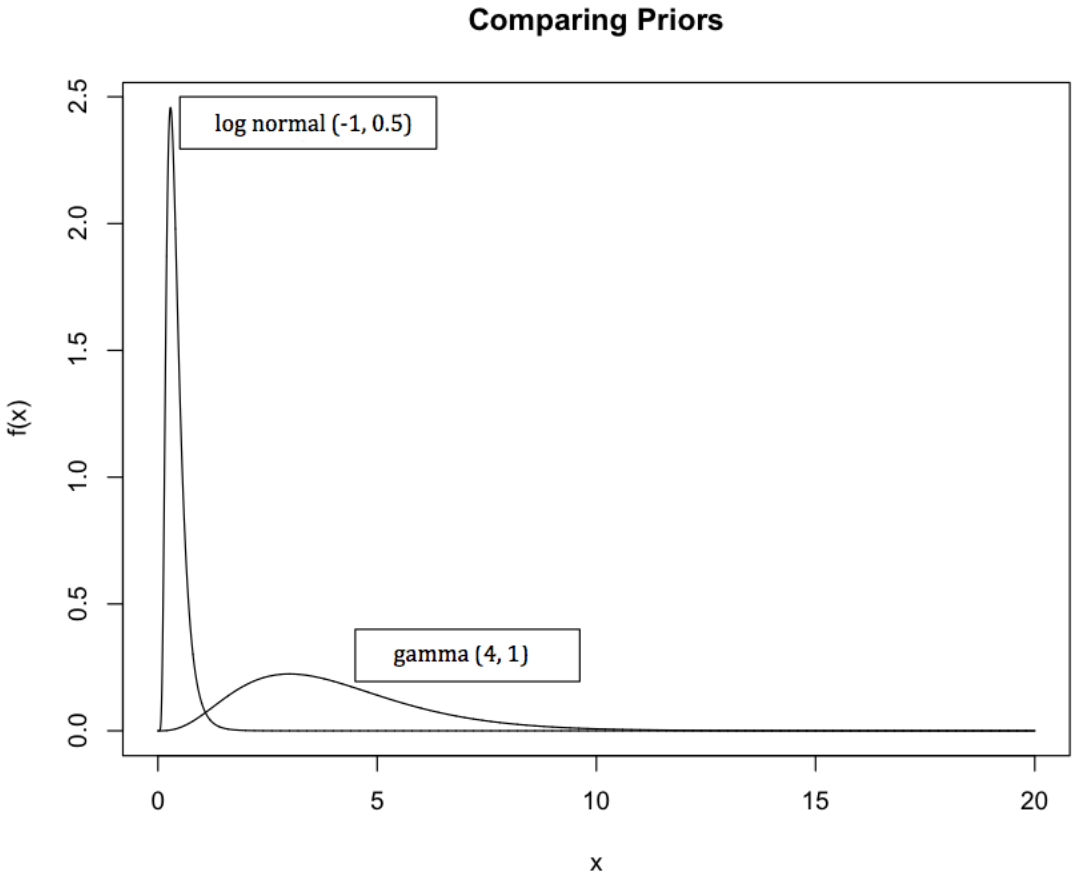


Figure 2.4: Comparison of lognormal and gamma priors for the standard deviation of the regression slope.

In this case, most of the mass of the lognormal prior is concentrated on values close to zero whereas the gamma prior is more diffuse. The $10, 50, 90$ percentiles of $LN(-1, 0.5)$ and $G(4, 1)$ are $0.19, 0.37, 0.7$ and $1.74, 3.67, 6.68$, respectively. Thus, choosing $LN(-1, 0.5)$ as the prior of S_2 is equivalent to believe that values of β_2 in the m regressions are much closer together than the situation where we assume $S_2 \sim G(4, 1)$. To assess the difference between both prior choices on S_2 and their impact on the degree of similarity of the regression coefficients, we resort to a simulated example, similar to [Barnard 2000], except that $m = 4$ and $n_j = 36$.

The explanatory variables are simulated standard normal variates. We also take $\tau_j \sim IG(3, 1)$ and $\bar{\beta} \sim N(0, 1000I)$. The prior for Σ is such that $\pi(R) \propto 1$ and we run Seaman et al.'s (2012) analyses under $S_2 \sim LN(-1, 0.5)$ and $S_2 \sim G(4, 1)$.

$S_i \sim LN(-1, 0.5)$								
Regression 1			Regression 2		Regression 3		Regression 4	
Estimate	mean	s.d	mean	s.d	mean	s.d	mean	s.d
Intercept	16.74	0.17	16.72	0.17	16.79	1.09	16.82	0.69
Slope	-9.27	0.42	-9.47	0.25	-9.66	0.98	-9.63	0.45
$S_i \sim G(4, 1)$								
Regression 1			Regression 2		Regression 3		Regression 4	
Estimate	mean	s.d	mean	s.d	mean	s.d	mean	s.d
Intercept	16.73	0.23	16.73	0.22	16.85	0.37	16.76	0.32
Slope	-9.30	0.30	-9.47	0.34	-9.73	0.23	-9.64	0.80

Table 2.3: Posterior estimations of regression coefficients when their standard deviations are distributed as $LN(-1, 0.5)$ and $G(4, 1)$.

$S_i \sim LN(-1, 0.5)$								
Regression 1			Regression 2		Regression 3		Regression 4	
Estimate	mean	s.d	mean	s.d	mean	s.d	mean	s.d
S_1	0.43	0.27	0.44	0.26	0.42	0.26	0.41	0.24
S_2	0.42	0.27	0.43	0.25	0.42	0.25	0.43	0.32
$S_i \sim G(4, 1)$								
Regression 1			Regression 2		Regression 3		Regression 4	
Estimate	mean	s.d	mean	s.d	mean	s.d	mean	s.d
S_1	2.31	1.28	2.33	1.29	2.29	1.29	2.29	1.26
S_2	2.32	1.29	2.23	1.28	2.25	1.23	2.30	1.26

Table 2.4: Posterior estimations standard deviations of the regression coefficients when their priors are distributed as $LN(-1, 0.5)$ versus $G(4, 1)$.

2.4.3 Comparison of posterior outputs

As seen in Tables 2.3 and 2.4, respectively. The differences between the regression estimates are quite limited from one prior to the next, while the estimates of the standard deviations vary much more. In the lognormal case, the posterior of S_i is concentrated on smaller values relative to the gamma prior. Figure 2.5 displays the posterior distributions of those parameters. The impact of the prior choice is quite clear on the standard deviations. Therefore, since the posteriors of both intercepts and slopes for all four regressions are centered in $(16.5, 17)$ and $(-10, -9)$, respectively, we can conclude at the stability of Bayesian inferences on β_j when selecting two different prior distributions on S_j. That the posteriors on the S_i's differ is in fine natural since those are hyperparameters that are poorly informed by the data, thus reflecting more the modeling choices of the experimenter.

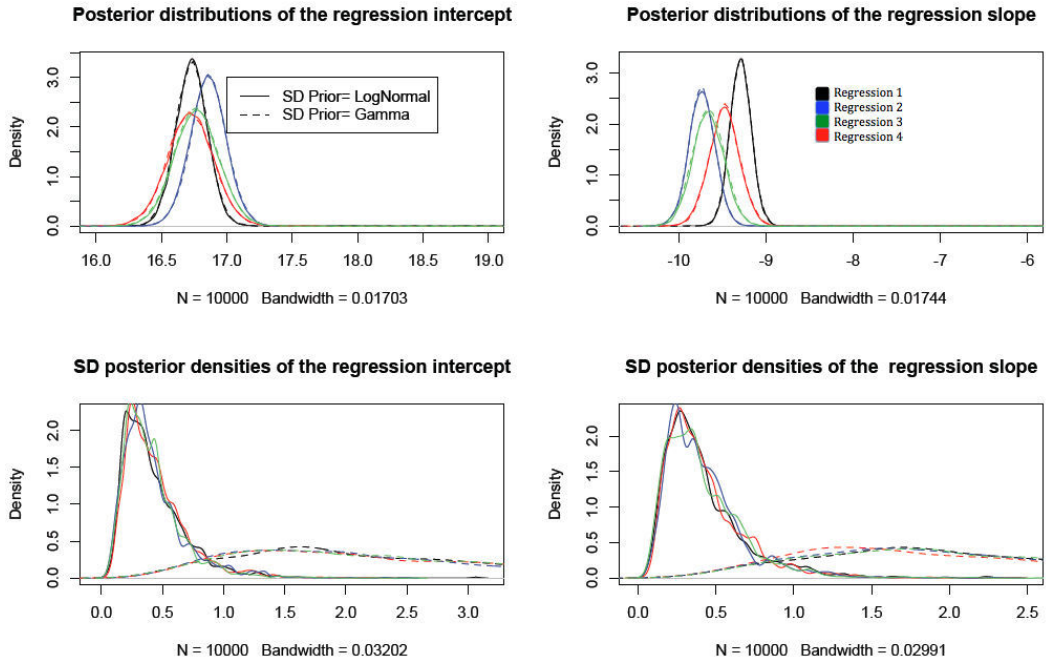


Figure 2.5: Estimated posterior densities of the regression intercept (top left), slope (top right), standard deviation of the intercept (down left) and standard deviation of the slope (down right), respectively for 4 different normal regressions. All estimates based on 10^5 iterations simulated from Metropolis-withing-Gibbs algorithm.

2.5 Examples 3 and 4: Prior choices for a proportion and the multinomial coefficients

This section considers more briefly the third and fourth examples of [Seaman III 2012]. The third example relates to a treatment effect analyzed by [Cowles 2002] and the fourth one covers a standard multinomial setting.

2.5.1 Proportion of treatment effect captured

In [Cowles 2002] two models are compared for surrogate endpoints, using a link function g that either includes the surrogate marker or not. The quantity of interest is a proportion of treatment effect captured: it is defined as PTE $\equiv 1 - \beta_1/\beta_{R,1}$, where $\beta_1, \beta_{R,1}$ are the coefficients of an indicator variable for treatment in the first and second regression models under comparison, respectively. [Seaman III 2012] restricted this proportion to the interval $(0, 1)$ and under this assumption they proposed to use a generalized beta distribution on $\beta_1, \beta_{R,1}$ so that PTE stayed within $(0, 1)$.

We find this example most intriguing in that, even if PTE could be turned into a meaningful quantity (given that it depends on parameters from different models), the criticism that it may take values outside $(0, 1)$ is rather dead-born since it suffices to impose a joint prior that ensures the ratio stays within $(0, 1)$. This actually is

the solution eventually proposed by the authors. If we have prior beliefs about the parameter space (which depends on $\beta_1/\beta_{R,1}$ in this example) the prior specified on the quantity of interest should integrate these beliefs. In the current setting, there is seemingly no prior information about $(\beta_1, \beta_{R,1})$ and hence imposing a prior restriction to $(0, 1)$ is not a logical specification. For instance, using normal priors on β_1 and $\beta_{R,1}$ lead to a Cauchy prior on $\beta_1/\beta_{R,1}$, which support is not limited to $(0, 1)$. We will not discuss this rather artificial example any further.

2.5.2 Multinomial model and evenness index

The final example in [Seaman III 2012] deals with a measure called *evenness index* $H(\theta) = -\sum \theta_i \log(\theta_i)/\log(K)$ that is a function of a vector θ of proportions θ_i, $i = 1, \ldots, K$. The authors assume a Dirichlet prior on θ with hyperparameters first equal to 1 then to 0.25. For the transform $H(\theta)$, Figure 2.6 shows that the first prior concentrates on $(0.5, 1)$ whereas the second does not. Since there is nothing special about the uniform prior, re-running the evaluation with the Jeffreys prior reduces this feature, which anyway is a characteristic of the prior distribution, not of a posterior distribution that would account for the data. The authors actually propose to use the $\text{Dir}(1/4, 1/4, \ldots, 1/4)$ prior, presumably on the basis that the induced prior on the evenness is then centered close to 0.5. If we consider the more generic $\text{Dir}(\gamma_1, \ldots, \gamma_K)$ prior, we can investigate the impact of the γ_i's when they move from 0.1 to 1. In Figure 2.6, the induced priors on $H(\theta)$ indeed show a decreasing concentration of the posterior on $(0.5, 1)$ as γ_i decreases towards zero. To further the comparison, we generated datasets of size $N = 50, 100, 250, 1000, 10,000$. Figure 2.7 shows the posteriors associated with each of the four Dirichlet priors for these samples, including modes that are all close to 0.4 when $N = 10^4$. Even for moderate sample sizes like 50, the induced posteriors are almost similar. When the sample size is 50, Table 2.5 shows there is some degree of variation between the posterior means, even though, as expected, this difference vanishes when the sample size increases.

Note that, while Dirichlet distributions are conjugate priors, hence potentially lacking in robustness, Jeffreys's prior is a special case corresponding to $\gamma_i = 1/K$ (here K is equal to 8). Figure 2.8 reproduces the transform of Jeffreys' prior for the evenness index (left) and the induced posterior densities for the same values of N. Since it is a special case of the above, the same features appear. A potential alternative we did not explore is to set a non-informative prior on the hyperparameters of the Dirichlet distribution.

2.6 Conclusion

In this note, we have reassessed the examples supporting the critical review of [Seaman III 2012], mostly showing that off-the-shelf noninformative priors are not suffering from the shortcomings pointed out by those authors. Indeed, according to

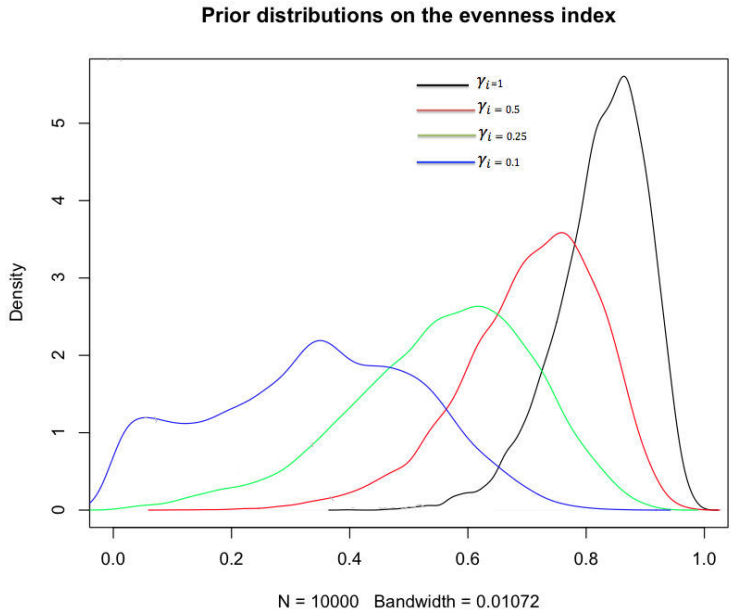


Figure 2.6: Priors induced on the evenness index: Four Dirichlet prior are assigned to θ with hyperparameters all equal to $0.1, 0.25, 0.5, 1$, based on 10^4 simulations.

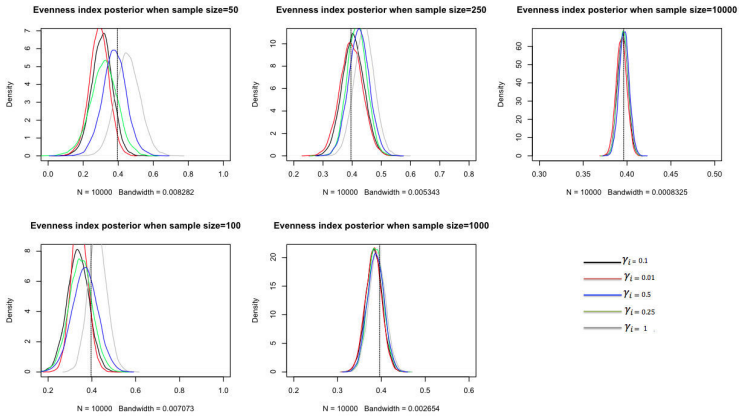


Figure 2.7: Estimated posterior densities of $H(\theta)$ considering sample sizes of $50, 100, 250, 1000, 10,000$. They correspond to the priors on θ shown in Figure 2.6 and are based on 10^4 posterior simulations. The vertical line indicates the mode of all posteriors when sample size is large enough.

the outcomes produced therein, those noninformative priors result in stable posterior inferences and reasonable Bayesian estimations for the parameters at hand. We thus consider the level of criticism found in the original paper rather unfounded, as it either relies on a highly specific choice of a proper prior distribution or on bypassing basic prior information later used for criticism. The paper of [Seaman III 2012] concludes with recommendations for prior checks. These recommendations are mostly sensible if mainly expressing the fact that some prior information is almost always

Sample size	50	100	250	1000	10,000
Dirichlet prior when $\gamma_i = 0.1$					
Posterior mean	0.308	0.336	0.403	0.383	0.395
Dirichlet prior when $\gamma_i = 0.25$					
Posterior mean	0.317	0.438	0.417	0.387	0.396
Dirichlet prior when $\gamma_i = 0.5$					
Posterior mean	0.378	0.368	0.423	0.387	0.397
Dirichlet prior when $\gamma_i = 1$					
Posterior mean	0.454	0.425	0.441	0.390	0.396
Jeffreys' prior: $\gamma_i = 0.125$					
Posterior mean	0.413	0.411	0.406	0.390	0.396
Posterior s.d	0.058	0.057	0.037	0.018	0.006

Table 2.5: Posterior means of $H(\theta)$ for the priors shown in Figure 2.6 and Jeffreys' prior on θ for sample sizes $50, 100, 250, 1000, 10,000$.

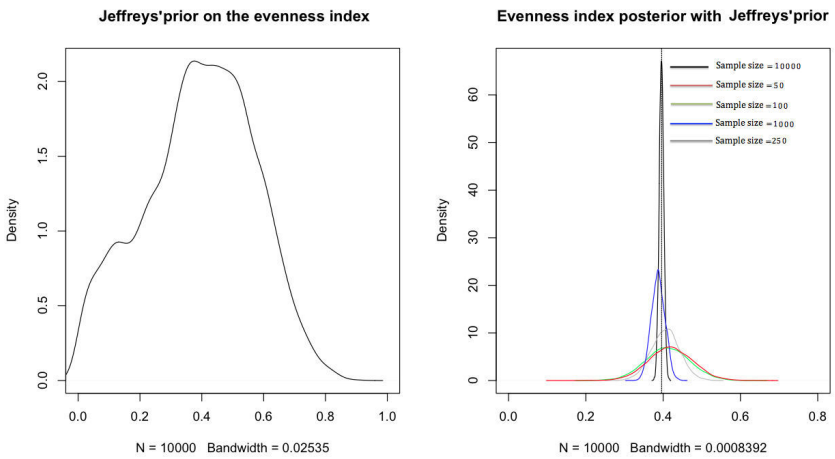


Figure 2.8: Jeffreys' prior and estimated posterior densities of $H(\theta)$ considering sample sizes $50, 100, 250, 1000, 10,000$. The posterior distributions are based on 10^4 posterior draws. The vertical line indicates the mode of the posterior density when the sample size is 10^4.

available on some quantities of interest. Our sole point of contention is the repeated and recommended reference to MLE, if only because it implies assessing or building the prior from the data. The most specific (if related to the above) recommendation is to use conditional mean priors as exposed by [Christensen 2011]. For instance, in the first (logistic) example, this meant putting a prior on the cdfs at age 40 and age 60. The authors picked a uniform in both cases, which sounds inconsistent with the presupposed shape of the probability function.

In conclusion, we find there is nothing pathologically wrong with either the paper of [Seaman III 2012] or the use of "noninformative" priors! Looking at induced priors on more intuitive transforms of the original parameters is a commendable suggestion, provided some intuition or prior information is already available on those. Using

a collection of priors including reference or invariant priors helps as well towards building a feeling about the appropriate choice or range of priors and looking at the dataset induced by simulating from the corresponding predictive cannot hurt.

Chapter 3

Supplementary material: Reflecting about Selecting Noninformative Priors

This chapter contains the statistical tools, computational details and some more data analyses related to the examples studied in the chapter 2.

3.1 Example 1

The first example of Chapter 2 is about the Bayesian analysis of the logistic model. The non standard posterior distributions resulted by assigning different non-informative priors to the parameters of the model are given as follows

- ✓ for a flat prior $\pi(\alpha, \beta) = 1$:

$$f(\alpha, \beta|\rho, x) = \exp\left(\sum_{i=1}^{n} \rho_i(\alpha+\beta x_i)\right) \Big/ \prod_{i=1}^{n}(1+\exp(\alpha+\beta x_i))$$

- ✓ for g-prior $\alpha, \beta|X \sim \mathcal{N}(0, g(X^TX)^{-1})$:

$$f(\alpha, \beta|\rho, x) = |X^TX|^{1/2} \exp\left(-g/2\binom{\alpha}{\beta}^T X^TX\binom{\alpha}{\beta} + \sum_{i=1}^{n} \rho_i(\alpha+\beta x_i)\right) \Big/ 2\pi\sqrt{g}\prod_{i=1}^{n}(1+\exp(\alpha+\beta x_i))$$

- ✓ and for the Jeffrey's prior, the log-likelihood of the logistic model is given by

$$\ell(\alpha, \beta) = \sum_{i=1}^{n} (\rho_i(\alpha + \beta xi) - \ln(1 + \exp(\alpha + \beta x_i)))$$

The second derivate of ℓ with respect to α and β is

$$\frac{\partial^2 \ell(\alpha, \beta)}{\partial \alpha^2} = -\sum_{i=1}^{n} \exp(\alpha+\beta x_i) \Big/ (1+\exp(\alpha+\beta x_i))^2$$

$$\frac{\partial^2 \ell(\alpha, \beta)}{\partial \beta^2} = -\sum_{i=1}^{n} x_i^2 \exp(\alpha+\beta x_i) \Big/ (1+\exp(\alpha+\beta x_i))^2$$

$$\frac{\partial^2 \ell(\alpha, \beta)}{\partial \alpha \partial \beta} = -\sum_{i=1}^{n} x_i \exp(\alpha+\beta x_i) \Big/ (1+\exp(\alpha+\beta x_i))^2$$

and the matrix of Fisher information can be written as following:

$$\mathbf{I}(\alpha,\beta,x)=\begin{pmatrix} -\sum_{i=1}^{n}\frac{\exp(\alpha+\beta x_i)}{(1+\exp(\alpha+\beta x_i))^2} & -\sum_{i=1}^{n}\frac{x_i\exp(\alpha+\beta x_i)}{(1+\exp(\alpha+\beta x_i))^2} \\ -\sum_{i=1}^{n}\frac{x_i\exp(\alpha+\beta x_i)}{(1+\exp(\alpha+\beta x_i))^2} & -\sum_{i=1}^{n}\frac{x_i^2\exp(\alpha+\beta x_i)}{(1+\exp(\alpha+\beta x_i))^2} \end{pmatrix}$$

✓ The invariant Jeffreys prior computed from $\sqrt{det(\mathbf{I}(\alpha,\beta,x))}$ yields the following posterior distribution for α and β

$$f(\alpha,\beta|\rho,x)=\sqrt{\sum_{i=1}^{n}\frac{\exp(\alpha+\beta x_i)}{(1+\exp(\alpha+\beta x_i))^2}\sum_{i=1}^{n}\frac{x_i^2\exp(\alpha+\beta x_i)}{(1+\exp(\alpha+\beta x_i))^2}-\{\sum_{i=1}^{n}\frac{x_i\exp(\alpha+\beta x_i)}{(1+\exp(\alpha+\beta x_i))^2}\}^2}$$
$$\times\frac{\exp\left(\sum_{i=1}^{n}\rho_i(\alpha+\beta x_i)\right)}{\prod_{i=1}^{n}(1+\exp\left(\alpha+\beta x_i\right))}.$$

We can sample from the posterior distributions above using the Metropolis-Hastings algorithm for each prior specification in which the proposal distribution is a random walk multivariate normal distribution based on the maximum likelihood estimate as starting value and the asymptotic covariance matrix of the maximum likelihood estimate as the covariance matrix of the proposal. The implementation in **R** can be found in [Kamary 2016a].

We run the Metropolis-Hastings algorithm with 10^4 iterations for the bank dataset by considering the bill length as the explanatory variable, and we test three different proposal scales, $\tau=0.1,1,5$. As shown in Figures 3.1, 3.2 and 3.3, for all values of τ, the chains simulated for α and β are convergent to the target distribution and able to move around the normal range with decreasing autocorrelations. However, in the case where $\tau=5$, the acceptance rate is low and the histograms of the output are far from the target distribution even after 10^4 iterations. The autocorrelation graph for $\tau=1$ decreases quicker than the cases where $\tau=0.1,5$. By comparing the raw sequences and the autocorrelation graphs provided by three algorithms above and also the corresponding acceptance rates, the best mixing behavior is related to $\tau=1$.

By comparing the plots shown in Figures 3.1, 3.2 and 3.3, despite the fact that three different non informative priors were assigned to the parameters of the logistic model, there is no visible difference between the posterior draws.

3.2 Example 2

Bayesian inference of multivariate regression model 2.3 using the prior modeling defined in section 2.4 derives the following joint posterior probability

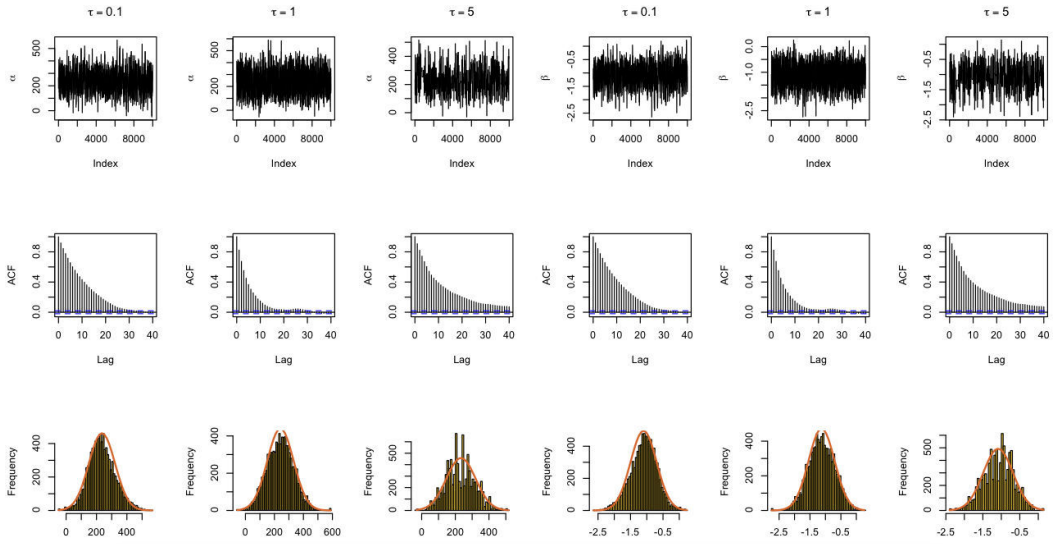


Figure 3.1: Simulation of posterior distribution of α and β with a multivariate normal random walk when the proposal scale τ takes values $0.1, 1, 5$ and a flat prior is assigned to α and β. From top to bottom: Sequence of 10^4 iterations; Empirical autocorrelation; Histogram of the last 9000 iterations compared with the target density.

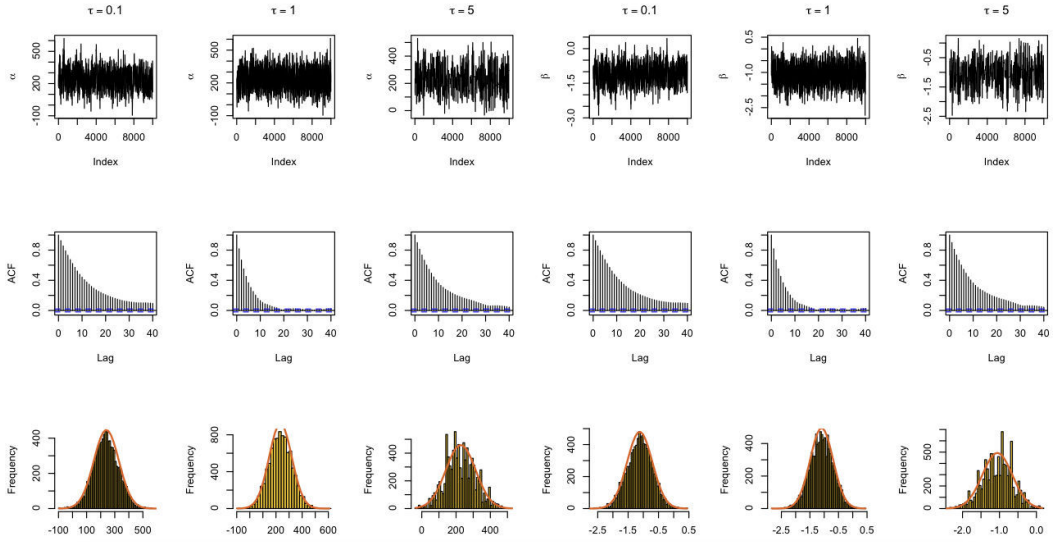


Figure 3.2: Simulation of posterior distribution of α and β with a multivariate normal random walk when the proposal scale τ takes values $0.1, 1, 5$ and a g-prior is assigned to α and β. From top to bottom: Sequence of 10^4 iterations; Empirical autocorrelation; Histogram of the last 9000 iterations compared with the target density.

$$\begin{aligned}\pi(\beta_j, \tau_j^2, \bar{\beta}, S, R|Y_j, X_j) &\propto \ell(\beta_j, \tau_j^2, \bar{\beta}|Y_j, X_j) \times \pi(\beta_j|\bar{\beta}, S, R)\pi(\tau_j^2)\pi(\bar{\beta})\pi(S, R) \\ &\propto {}^{1}\!/\!_{(\tau_j^2)^{n_j/2-a-1}} \exp\left(-{}^{(Y_j - X_j\beta_j)^T(Y_j - X_j\beta_j)}\!/\!_{2\tau_j^2}\right) \\ &\times |\text{diag}(S)R\text{diag}(S)|^{-1/2} \exp\left(-{}^{(\beta_j-\bar{\beta})^T(\text{diag}(S)R\text{diag}(S))^{-1}(\beta_j-\bar{\beta})}\!/\!_{2}\right) \\ &\times \exp({}^{-b}\!/\!_{\tau_j^2} - {}^{\bar{\beta}^T\bar{\beta}}\!/\!_{2000})\pi(S)\end{aligned}$$

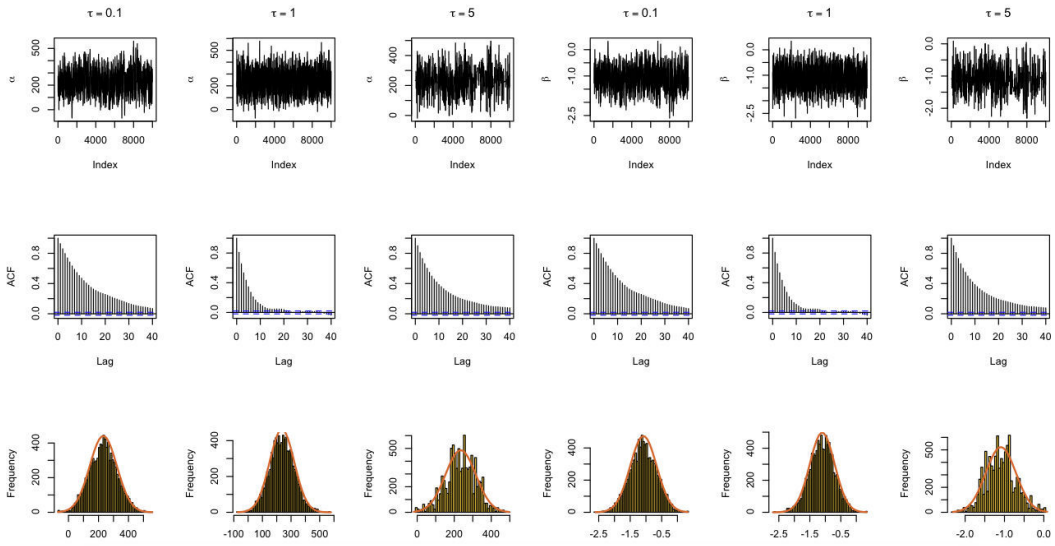


Figure 3.3: Simulation of posterior distribution of α and β with a multivariate normal random walk when the proposal scale τ takes values $0.1, 1, 5$ and a Jeffreys prior is assigned to α and β. From top to bottom: Sequence of 10^4 iterations; Empirical autocorrelation; Histogram of the last 9000 iterations compared with the target density.

that is resulted under the assumption of independence between the parameter S and R and $\pi(R) \propto 1$. The conditional posterior distribution of the parameter β_j can be obtained as following

$$\begin{aligned}\pi(\beta_j|\tau_j^2,\bar{\beta},S,R,Y_j,X_j) &\propto \exp\left(-(Y_j-X_j\beta_j)^T(Y_j-X_j\beta_j)/2\tau_j^2 - (\beta_j-\bar{\beta})^T(\text{diag}(S)R\text{diag}(S))^{-1}(\beta_j-\bar{\beta})/2\right)\\ &\propto \exp\left(-(Y_j-X_j\beta_j)^T(\tau_j^2\mathbb{I}_k)^{-1}(Y_j-X_j\beta_j)-(\beta_j-\bar{\beta})^T(\text{diag}(S)R\text{diag}(S))^{-1}(\beta_j-\bar{\beta})/2\right)\end{aligned}$$

where $\mathbb{I}_k$ is the identity matrix of size k. If we replace $\text{diag}(S)R\text{diag}(S)$ by Σ, dropping multiplicative terms that do not involve β_j gives

$$\pi(\beta_j|\tau_j^2,\bar{\beta},S,R,Y_j,X_j) \propto \exp(-1/2\begin{pmatrix}X_j\beta_j - Y_j\\ \beta_j-\bar{\beta}\end{pmatrix}^T\begin{pmatrix}\tau_j^2\mathbb{I}_k & 0\\ 0 & \Sigma\end{pmatrix}^{-1}\begin{pmatrix}X_j\beta_j - Y_j\\ \beta_j-\bar{\beta}\end{pmatrix})$$

or

$$\pi(\beta_j|\tau_j^2,\bar{\beta},S,R,Y_j,X_j) \propto \exp(-1/2(\begin{pmatrix}X_j\\ \mathbb{I}_k\end{pmatrix}\beta_j-\begin{pmatrix}Y_j\\ \bar{\beta}\end{pmatrix})^T\begin{pmatrix}\tau_j^2\mathbb{I}_k & 0\\ 0 & \Sigma\end{pmatrix}^{-1}(\begin{pmatrix}X_j\\ \mathbb{I}_k\end{pmatrix}\beta_j-\begin{pmatrix}Y_j\\ \bar{\beta}\end{pmatrix}))$$

As shown in [Christensen 2011], if we define

$$\tilde{\beta_j} = \left(\tau_j^{-2}X_j^TX_j + \Sigma^{-1}\right)^{-1}(\tau_j^{-2}X_j^TY_j + \Sigma^{-1}\bar{\beta})$$

we can rewrite the posterior density as

$$\pi(\beta_j|\tau_j^2,\bar{\beta},S,R,Y_j,X_j) \propto \exp\left(-1/2(\beta_j-\tilde{\beta}_j)^T\left(\tau_j^{-2}X_j^TX_j+\Sigma^{-1}\right)(\beta_j-\tilde{\beta}_j)\right)$$

which implies a multivariate Gaussian distribution as following

$$\beta_j|\tau_j^2,\bar{\beta},S,R,Y_j,X_j \sim \mathcal{N}\left(\tilde{\beta}_j,\left(\tau_j^{-2}X_j^TX_j+\Sigma^{-1}\right)^{-1}\right).$$

The posterior density of the parameter τ_j^2 given $\beta_j,\bar{\beta},S,R,Y_j,X_j$ is inverse gamma distribution

$$\tau_j^2|\beta_j,\bar{\beta},S,R,Y_j,X_j \sim \mathcal{IG}\left(n_j+2a/2,\,(Y_j-X_j\beta_j)^T(Y_j-X_j\beta_j)+2b/2\right)$$

when the prior distribution is supposed to be $\mathcal{IG}(a,b)$. To obtain the full conditional for $\bar{\beta}$, we can write

$$\begin{aligned}\bar{\beta}|\beta_j,\tau_j^2,S,R,Y_j,X_j &\propto \exp\left(-(\beta_j-\bar{\beta})^T(\text{diag}(S)R\text{diag}(S))^{-1}(\beta_j-\bar{\beta})/2-\bar{\beta}^T\bar{\beta}/2\right)\\ &\propto \exp(-1/2\begin{pmatrix}\bar{\beta}-\beta_j\\ \bar{\beta}-0\end{pmatrix}^T\begin{pmatrix}\Sigma & 0\\ 0 & 1000\mathbb{I}_k\end{pmatrix}^{-1}\begin{pmatrix}\bar{\beta}-\beta_j\\ \bar{\beta}-0\end{pmatrix})\end{aligned}$$

which implies a multivariate Gaussian distribution with the following form

$$\bar{\beta}|\beta_j,\tau_j^2,S,R,Y_j,X_j \sim \mathcal{N}\left(\beta_0,(\Sigma^{-1}+1/1000\mathbb{I}_k)^{-1}\right) \tag{3.1}$$

where $\beta_0=(\Sigma^{-1}+1/1000\mathbb{I}_k)^{-1}\Sigma^{-1}\beta_j$. For the correlation matrix R, the full conditional density will therefore be

$$R|\beta_j,\bar{\beta},\tau_j^2,S,Y_j,X_j \propto |\text{diag}(S)R\text{diag}(S)|^{-1/2}\exp\left(-(\beta_j-\bar{\beta})^T(\text{diag}(S)R\text{diag}(S))^{-1}(\beta_j-\bar{\beta})/2\right).$$

When the prior of the standard deviations, S, is log normal $\mathcal{LN}(\mu,\sigma)$, the conditional posterior is given by

$$\begin{aligned}S|\beta_j,\bar{\beta},\tau_j^2,R,Y_j,X_j \propto\ & |\text{diag}(S)R\text{diag}(S)|^{-1/2}\exp\left(-(\beta_j-\bar{\beta})^T(\text{diag}(S)R\text{diag}(S))^{-1}(\beta_j-\bar{\beta})/2\right)\\ &\times\prod_{j=1}^{k}1/s_j\exp(-(ln(s_j)-\mu)^2/2\sigma^2)\end{aligned}$$

and in the case where gamma $\mathcal{G}(\delta,\zeta)$ prior is placed on the elements of S, we will have

$$\begin{aligned}S|\beta_j,\bar{\beta},\tau_j^2,R,Y_j,X_j \propto\ & |\text{diag}(S)R\text{diag}(S)|^{-1/2}\exp\left(-(\beta_j-\bar{\beta})^T(\text{diag}(S)R\text{diag}(S))^{-1}(\beta_j-\bar{\beta})/2\right)\\ &\times\prod_{j=1}^{k}s_j^{\delta-1}\exp(-s_j/\zeta)\end{aligned}$$

The analyses of this example in Chapter 2 are based on the Metropolis-within-Gibbs algorithm in which the correlations $r_{ij}; i \neq j$ are independently simulated from uniform proposal distributions. For both prior specifications of S, log normal and gamma distributions, the implementation in **R** can be seen in [Kamary 2016a].

In the algorithm, the parameter β_j and $\Sigma = \text{diag}(S)R\text{diag}(S)$ are initialized by the maximum likelihood estimate and the asymptotic covariance matrix of the maximum likelihood estimate and the other parameters are started from a random value simulated from their prior distributions. The proposal distribution of logarithm function of S is a random walk multivariate normal distribution with the possibility of calibrating the proposal scale. As discussed in Chapter 2, replacing log normal distribution assigned to the hyper parameter S by gamma distribution does not influence the conditional posterior distribution of the regression intercept and slope. Here, we deal with the convergence of the simulated samples obtained by implementing the Metropolis-within-Gibbs algorithm and also the impact of both prior choices on Bayesian inference of the other parameters. To do so, we simulate $n_1 = 36$ data points from normal regression model with an explanatory variable and run the program over 10^4 iterations when the scale of the proposal distribution of s_j is equal to 1.

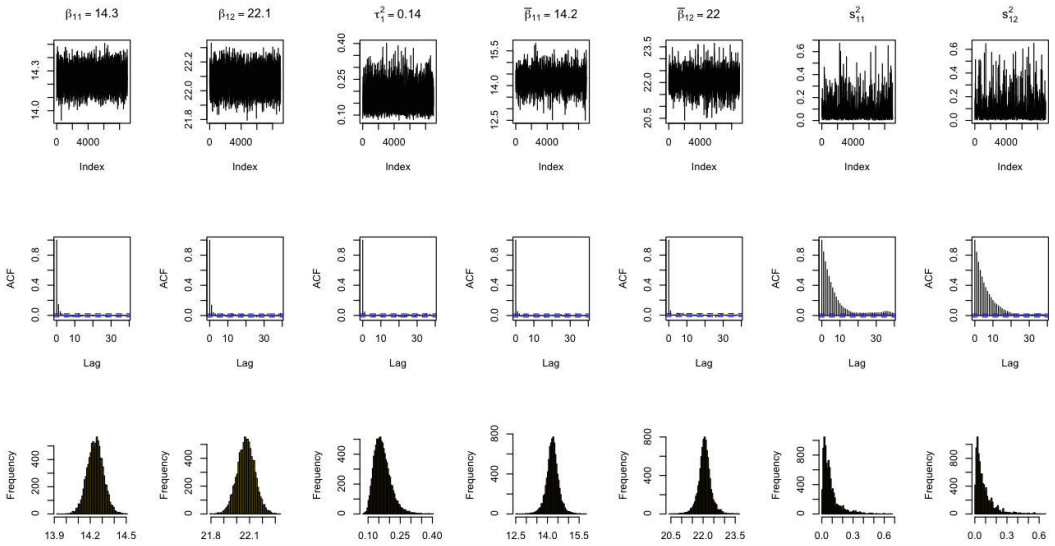


Figure 3.4: Simulation of posterior distribution of the parameters of the normal regression model when standard deviations of the intercept and slope have log normal distribution. From top to bottom: Sequence of last 9000 iterations; Empirical autocorrelation; Histograms. True values of the parameters are indicated at the top of sequence plots.

Figures 3.4 and 3.5 give an assessment of the convergence of the algorithm and show that for both priors, log normal and gamma distributions, the distributions of the chains visually cover the whole support of the target distribution with sufficient regularity for 10^4 MCMC iterations. The autocorrelation plots show high mixing behavior of the chains and from the histograms, the distributions of the generated

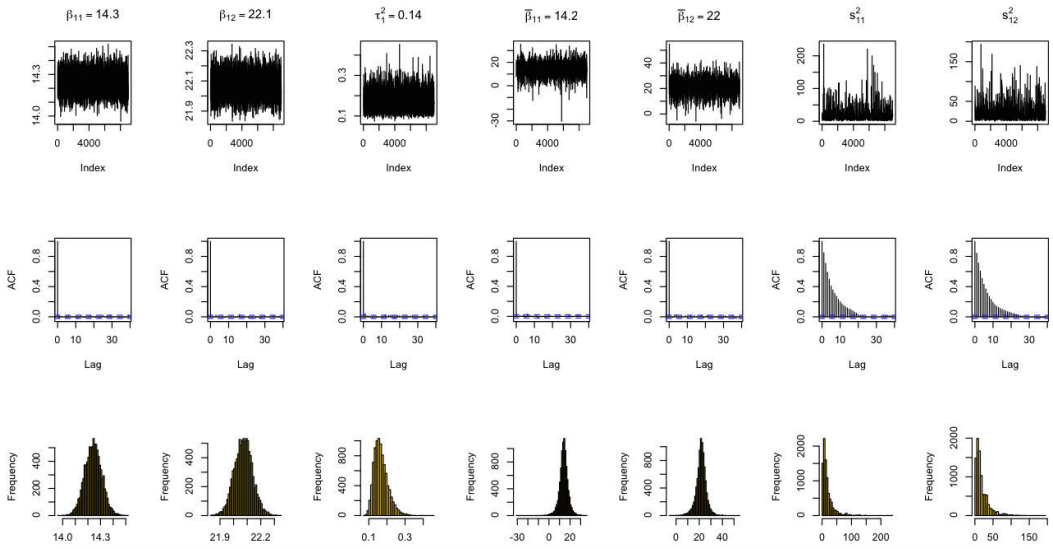


Figure 3.5: Simulation of posterior distribution of the parameters of the normal regression model when standard deviations of the intercept and slope have gamma distribution. From top to bottom: Sequence of last 9000 iterations; Empirical autocorrelation; Histograms. True values of the parameters are indicated at the top of sequence plots.

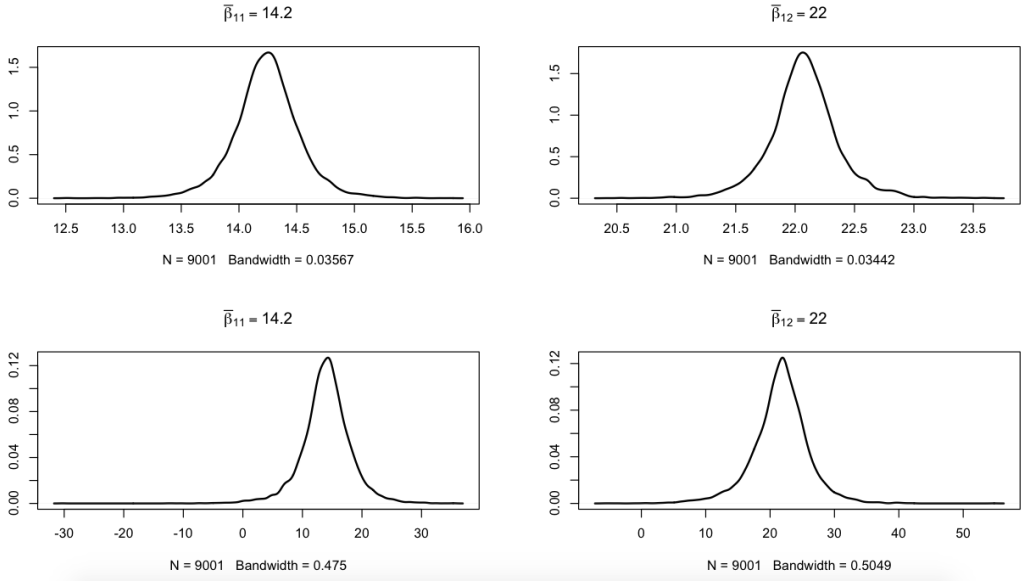


Figure 3.6: Empirical density of simulated draws from conditional posterior density of $\bar{\beta}_{11}$ and $\bar{\beta}_{12}$ based on last 9000 iterations when (*Top*) log normal prior and (*Bottom*) gamma prior are assigned to the hyper parameter s_j's. True values of the parameters are indicated at the top of each graph.

samples are slightly concentrated over the true values. By comparing the range of the histograms of $\bar{\beta}_{1j}$ and s_{1j} in Figure 3.4 with those of Figure 3.5, we can see that when the distribution of $s_{1j}; j = 1, 2$ is tightened up near zero (which corresponds to

the output of the simulation run in the case of log normal prior) the chains simulated for $\bar{\beta}_{1j}$ have a lot of density over a narrow interval near the true values while the distribution of $\bar{\beta}_{1j}$ spreads out over a wide range in the case of gamma prior.

This impact of the prior choices for the standard deviations on the posterior distribution of $\bar{\beta}$ becomes more visible when comparing the empirical density plotted in Figure 3.6. In both cases, the empirical densities are centered over the true values of the parameter $\bar{\beta}$ while the dispersion of two cases is impacted by the change in the prior choice for s_j's. The main reason for this effect is that as shown in (3.1), the conditional posterior distribution of $\bar{\beta}$ depends on Σ and so on S. A stretched or squeezed prior choice allocated to s_j's influences the posterior results of standard deviation S and therefore impacts those of $\bar{\beta}$.

Chapter 4

Testing hypotheses as a mixture estimation model

Joint work with Kerrie Mengersen, Christian P. Robert and Judith Rousseau

Abstract

We consider a novel paradigm for Bayesian testing of hypotheses and Bayesian model comparison. Our alternative to the traditional construction of posterior probabilities that a given hypothesis is true or that the data originates from a specific model is to consider the models under comparison as components of a mixture model. We therefore replace the original testing problem with an estimation one that focus on the probability weight of a given model within a mixture model. We analyze the sensitivity on the resulting posterior distribution on the weights of various prior modeling on the weights. We stress that a major appeal in using this novel perspective is that generic improper priors are acceptable, while not putting convergence in jeopardy. Among other features, this allows for a resolution of the Lindley–Jeffreys paradox. When using a reference Beta $\mathcal{B}(a_0, a_0)$ prior on the mixture weights, we note that the sensitivity of the posterior estimations of the weights to the choice of a_0 vanishes with the sample size increasing and advocate the default choice $a_0 = 0.5$, derived from Rousseau and Mengersen (2012). Another feature of this easily implemented alternative to the classical Bayesian solution is that the speeds of convergence of the posterior mean of the weight and of the corresponding posterior probability are quite similar.

Keywords: Noninformative prior, Mixture of distributions, Bayesian analysis, testing statistical hypotheses, Dirichlet prior, Posterior probability

4.1 Introduction

While a if not the central problem of statistical inference and a dramatically differentiating feature between classical and Bayesian paradigms [Neyman 1933, Berger 1987, Casella 1987, Gigerenzer 1991, Berger 2003a, Mayo 2006, Gelman 2008], the handling of hypothesis testing by Bayesian theory is wide open to controversy and divergent opinions, even within the Bayesian community [Jeffreys 1939, Bernardo 1980, Berger 1985, Aitkin 1991, Berger 1992, De Santis 1997, Bayarri 2007, Christensen 2011,

Johnson 2010, Gelman 2013a, Robert 2014]. In particular, the handling of the non-informative Bayesian testing case is mostly unresolved and has produced much debate, witness the specific case of the Lindley or Jeffreys–Lindley paradox [Lindley 1957, Shafer 1982, DeGroot 1982, Robert 1993, Lad 2003, Spanos 2013, Sprenger 2013, Robert 2014].

Bayesian model selection is understood here as the comparison of several potential statistical models towards the selection of the model that fits the current data the "best". For instance, [Christensen 2011] consider this is a decision issue that pertains to testing, while [Robert 2001] expressed it as a model index estimation setting and [Gelman 2013a] do not agree about the decisional aspect. A mostly accepted perspective is however that Bayesian model selection does not primarily seek to identify which model is "true" (if any), but rather to indicate which model fits the data better given all the available information. As discussed in the Bayesian literature (see, e.g. [Berger 1992, Madigan 1994, Balasubramanian 1997, MacKay 2002, Consonni 2013]), tools like the Bayes factor [Jeffreys 1939] naturally include a penalization factor addressing model complexity, penalization mimicked by approximations like the Bayes Information (BIC) and the Deviance Information (DIC) criteria [Schwarz 1978, Csiszár 2000, Spiegelhalter 2002, Plummer 2008]. Posterior predictive tools have been successfully advocated in [Gelman 2013a], even though they can be criticized for multiple uses of the (same) data.

Let us recall very briefly (referring to [Berger 1985, Robert 2001]) that the standard Bayesian approach to testing is to consider two families of models, one for each of the hypotheses under comparison,

$$\mathfrak{M}_1 : \; x \sim f_1(x|\theta_1)\,, \; \theta_1 \in \Theta_1 \quad \text{and} \quad \mathfrak{M}_2 : \; x \sim f_2(x|\theta_2)\,, \; \theta_2 \in \Theta_2\,,$$

and to associate with each of those models a prior distribution,

$$\theta_1 \sim \pi_1(\theta_1) \quad \text{and} \quad \theta_2 \sim \pi_2(\theta_2)\,,$$

in order to compare the marginal likelihoods

$$m_1(x) = \int_{\Theta_1} f_1(x|\theta_1)\,\pi_1(\theta_1)\,\mathrm{d}\theta_1 \quad \text{and} \quad m_2(x) = \int_{\Theta_2} f_2(x|\theta_2)\,\pi_1(\theta_2)\,\mathrm{d}\theta_2$$

either through the *Bayes factor* or through the posterior probability, respectively:

$$\mathfrak{B}_{12} = \frac{m_1(x)}{m_2(x)}, \quad \mathbb{P}(\mathfrak{M}_1|x) = \frac{\omega_1 m_1(x)}{\omega_1 m_1(x) + \omega_2 m_2(x)};$$

the latter depends on the prior weights ω_i of both models. Both testing and model selection are thus expressed as a comparison of models. The Bayesian decision step proceeds by comparing the Bayes factor $\mathfrak{B}_{12}$ to the threshold value of one or comparing the posterior probability $\mathbb{P}(\mathfrak{M}_1|x)$ to a bound derived from a 0–1 loss function (or a "golden" bound like $\alpha = 0.05$ inspired from frequentist practice [Berger 1987, Berger 1997, Berger 1999, Berger 2003a, Ziliak 2008]. As a general rule, when comparing more than two models, the model with the highest posterior

probability is the one selected, but this rule is highly dependent on the prior modeling, even with large datasets, which makes it hard to promote as the default solution in practical studies.

Some well-documented difficulties with this traditional handling of Bayesian tests and Bayesian model choices via posterior probabilities are, among others [Vehtari 2002, Vehtari 2012]:

- ✓ a tension between using posterior probabilities as justified by a binary loss function but depending on unnatural prior weights and using Bayes factors [Jeffreys 1939] that eliminate this dependence but escape as well the direct connection with the posterior distribution, unless the prior weights are integrated within the loss function [Berger 1985, Robert 2001];;
- ✓ a subsequent and delicate interpretation (or calibration) of the strength of the Bayes factor [Jeffreys 1939, Dickey 1978, Kass 1995, Lavine 1999] towards supporting a given hypothesis or model, mostly due to the fact that it is not a Bayesian decision rule (once more, unless the loss function is artificially modified to incorporate the prior weights);
- ✓ a similar difficulty with posterior probabilities, with the correlated tendency to interpret them as p-values (rather than the opposite) when they only report through a marginal likelihood ratio the respective strengths of fitting the data to both models (and nothing about the "truth" of either model);
- ✓ a long-lasting impact of the prior modeling, meaning the choice of the prior distributions on the parameter spaces of both models under comparison, despite the existence of an overall consistency proof for the Bayes factor [Berger 2003b, Rousseau 2007, McVinish 2009];
- ✓ a discontinuity in the use of improper priors since they are not justified in most testing situations [DeGroot 1970, DeGroot 1973, Robert 2001, Robert 2014], leading to many alternative if *ad hoc* solutions, where the data is either used twice [Aitkin 1991, Aitkin 2010, Gelman 2013b]; or split in artifial ways [O'Hagan 1995, Berger 1996, Berger 1998, Berger 2001];
- ✓ a binary (*accept* vs.*reject*) outcome more suited for immediate decision (if any) than for model evaluation, in connection with the use of a rudimentary binary loss function that many deem unnatural [Gelman 2013a];
- ✓ a related impossibility to ascertain simultaneous misfit (i.e., a lack of fit for both models under comparison) or to detect the presence of outliers;
- ✓ a lack of assessment of the uncertainty associated with the decision itself;
- ✓ a difficult computation of marginal likelihoods in most settings [Chen 2000, Marin 2011] with further controversies about which solution to adopt [Newton 1994, Neal 1994, Green 1995, Chib 1995, Neal 1999, Skilling 2006, Steele 2006, Chopin 2010];

- ✓ a strong dependence of the values of posterior probabilities on conditioning statistics, which in turn undermines their validity for model assessment, as exhibited in Approximate Bayesian computation (ABC) settings by [Robert 2011] and [Marin 2014];

- ✓ a temptation to create pseudo-frequentist equivalents such as q-values [Johnson 2010, Johnson 2013b, Johnson 2013a] with even less Bayesian justifications.

Rather than vainly attempting to solve those numerous issues in the light of the many attempts listed above, which clearly failed to produce a consensus, we therefore propose a paradigm shift in the Bayesian processing of hypothesis testing and of model selection, namely to adopt a completely novel perspective on this issue, perspective that provides a convergent and naturally interpretable solution, while allowing for a more extended use of improper priors. This approach relies on the simple representation of (or embedding into) the problem as a two-component mixture estimation problem where the weights are formally equal to 0 or 1. The mixture model [Frühwirth-Schnatter 2006] thus contains both models under comparison as extreme cases. This approach is inspired from the consistency result of [Rousseau 2011] on estimated overfitting mixtures, where the authors established that over-parameterised mixtures can be consistently estimated, despite the parameter standing on a (or several) boundary(ies) of the parameter space. While this mixture representation is not directly equivalent to the use of a posterior probability, i.e., the posterior estimator of the mixture weight cannot be considered as a proxy to the posterior probability value, we do not perceive this as a negative feature but rather as a new tool having the potential of a better approach to testing, with a further valuable property of not expanding the number of parameters in the model (and hence keeping in line with Occam's razor, see, e.g., [Adams 1987, Jefferys 1992, Rasmussen 2001]). Our new paradigm to Bayesian testing requires a calibration of the posterior distribution of the weight of a model, while moving from the admittedly artificial and rarely understood notion of the posterior probability of a model.

The plan of the paper is as follows: Section 4.2 provides a description of the mixture model specifically created for this setting, while Section 6.2 details the implementation issues with estimating the parameters of the mixture. Section 4.3 details at great length how the mixture approach performs in the most standard i.i.d. models. Section 4.4 demonstrates its application on a survival dataset. Section 4.5 expands [Rousseau 2011] to provide conditions on the hyperameters of the mixture model that are sufficient to achieve convergence. Section 4.6 concludes on the generic applicability of the above principle.

4.2 Testing problems as estimating mixture models

4.2.1 A new paradigm for testing

Given two classes of statistical models,

$$\mathfrak{M}_1 : \; x \sim f_1(x|\theta_1)\,, \;\; \theta_1 \in \Theta_1 \quad \text{and} \quad \mathfrak{M}_2 : \; x \sim f_2(x|\theta_2)\,, \;\; \theta_2 \in \Theta_2\,,$$

which may correspond to an hypothesis to be tested and its alternative, respectively, it is always possible to embed both models within an encompassing mixture model

$$\mathfrak{M}_\alpha : \; x \sim \alpha f_1(x|\theta_1) + (1-\alpha) f_2(x|\theta_2)\,, \;\; 0 \leq \alpha \leq 1\,. \tag{4.1}$$

Indeed, both models correspond to very special cases of the mixture model, one for $\alpha = 1$ and the other for $\alpha = 0$ (with a slight notational inconsistency in the indices).[1]

When considering a sample $(x_1, \ldots, x_n)$ from one of the two models, the mixture representation still holds at the likelihood level, namely the likelihood for each model is a special case of the weighted sum of both likelihoods. However, this is not directly appealing for estimation purposes since it corresponds to a mixture with *a single observation*. See however [O'Neill 2014] for a computational solution based upon this representation.

What we propose in this paper is to draw inference on the individual mixture representation (6.1), acting as if each observation was individually and independently[2] produced by the mixture model. While this apparently constitutes an approximation to the real (unknown) model, except in the cases when $\alpha = 0, 1$, we see several definitive advantages to this paradigm shift:

- ✓ relying on a Bayesian estimate of the weight α rather than on the posterior probability of model $\mathfrak{M}_1$ does produce an equally convergent indicator of which model is "true" (see Section 4.5), while removing the need of overwhelmingly artificial prior probabilities on model indices, ω_1 and ω_2;

- ✓ the interpretation of this estimator of α is at least as natural as handling the posterior probability, while avoiding the caricatural zero-one loss setting [DeGroot 1970, DeGroot 1973, Berger 1985]. The quantity α and its posterior distribution provide a measure of proximity to both models for the data at

[1] The choice of possible encompassing models is obviously unlimited: for instance, a Geometric mixture

$$x \sim f_\alpha(x) \propto f_1(x|\theta_1)^\alpha \, f_2(x|\theta_2)^{1-\alpha}$$

is a conceivable alternative. However, such alternatives are less practical to manage, starting with the issue of the intractable normalizing constant. Note also that when f_1 and f_2 are Gaussian densities, the Geometric mixture remains Gaussian for all values of α. Similar drawbacks can be found with harmonic mixtures.

[2] An extension to the iid case will be considered in Example 4.3.6 for linear models. Dependent observations like Markov chains can be modeled by a straightforward extension of (6.1) where both terms in the mixture are conditional on the relevant past observations.

hand, while being also interpretable as a propensity of the data to stand with (or to stem from) one of the two models. This representation further allows for alternative perspectives on testing and model choice, through the notions of predictive tools [Gelman 2013a], cross-validation [Vehtari 2002], and information indices like WAIC [Vehtari 2012];

✓ the highly problematic computation [Chen 2000, Marin 2011] of the marginal likelihoods is bypassed, standard algorithms being available for Bayesian mixture estimation [Richardson 1997, Berkhof 2003, Frühwirth-Schnatter 2006, Lee 2009];

✓ the extension to a finite collection of models to be compared is straightforward, as this simply involves a larger number of components. This approach further allows to consider all models at once rather than engaging in pairwise costly comparisons and thus to eliminate the least likely models by simulation, those being not explored by the corresponding algorithm [Carlin 1995, Richardson 1997];

✓ the (simultaneously conceptual *and* computational) difficulty of "label switching" [Celeux 2000, Stephens 2000, Jasra 2005] that plagues both Bayesian estimation and Bayesian computation for most mixture models completely vanishes in this particular context, since components are no longer exchangeable. In particular, we compute neither a Bayes factor[3] nor a posterior probability related with the substitute mixture model and we hence avoid the difficulty of recovering the modes of the posterior distribution [Berkhof 2003, Lee 2009, Rodriguez 2014]. Our perspective is solely centered on estimating the parameters of a mixture model where both components are always identifiable;

✓ the posterior distribution of α evaluates more thoroughly the strength of the support for a given model than the single figure outcome of a Bayes factor or of a posterior probability. The variability of the posterior distribution on α allows for a more thorough assessment of the strength of the support of one model against the other;

✓ an additional feature missing from traditional Bayesian answers is that a mixture model also acknowledges the possibility that, for a finite dataset, *both* models or *none* could be acceptable. This possibility will be seen in some illustrations below (Section 4.3)

✓ while standard (proper and informative) prior modeling can be painlessly reproduced in this novel setting, non-informative (improper) priors now are manageable therein, provided both models under comparison are first reparameterised towards common-meaning and shared parameters, as for instance with

[3] Using a Bayes factor to test for the number of components in the mixture (6.1) as in [Richardson 1997] would be possible. However, the outcome would fail to answer the original question of selecting between both (or more) models.

location and scale parameters. In the special case when all parameters can be made common to both models[4], the mixture model (6.1) can read as

$$\mathfrak{M}_\alpha : \; x \sim \alpha f_1(x|\theta) + (1-\alpha) f_2(x|\theta)\,, 0 \leq \alpha \leq 1\,.$$

For instance, if θ is a location parameter, a flat prior $\pi(\theta) \propto 1$ can be used with no foundational difficulty, in opposition to the testing case [DeGroot 1973, Berger 1998];

✓ continuing from the previous argument, using the *same* parameters or some *identical* parameters on both components is an essential feature of this reformulation of Bayesian testing, as it highlights the fact that the opposition between the two components of the mixture is not an issue of enjoying different parameters, but quite the opposite. As further stressed below, this or even *those* common parameter(s) is (are) nuisance parameters that need be integrated out (as they also are in the traditional Bayesian approach through the computation of the marginal likelihoods);

✓ even in the setting when the parameters of the mixture components, θ_1 and θ_2, differ, they can be integrated out by mere Monte Carlo methods;

✓ the choice of the prior model probabilities is rarely discussed in a classical Bayesian approach, even though those probabilities linearly impact the posterior probabilities and can be argued to promote the alternative of using the Bayes factor instead. In the mixture estimation setting, prior modeling only involves selecting a prior on α, for instance a Beta $\mathcal{B}(a_0, a_0)$ distribution, with a wide range of acceptable values for the hyperparameter a_0, as demonstrated in Section 4.5. While the value of a_0 impacts the posterior distribution of α, it can be argued that (a) it nonetheless leads to an accumulation of the mass near 1 or 0, i.e. to favor the most likely or the true model over the other one, and (b) a sensitivity analysis on the impact of a_0 is straightforward to carry on;

✓ in most settings, this approach can furthermore be easily calibrated by a parametric boostrap experiment providing a posterior distribution of α under each of the models under comparison. The prior predictive error can therefore be directly estimated and can drive the choice of the hyperparameter a_0, if need be.

[4] While this may sound like an extremely restrictive requirement in a traditional mixture model, let us stress here that the presence of common parameters becomes quite natural within a testing setting. To wit, when comparing two different models for the *same* data, moments like $\mathbb{E}[X^\gamma]$ are defined in terms of the observed data and hence *should* be the *same* for both models. Reparametrising the models in terms of those common meaning moments does lead to a mixture model with some and maybe *all* common parameters. We thus advise the use of a common parameterisation, whenever possible.

4.2.2 Mixture estimation

Before studying the application of the above principle to some standard examples in Section 4.3, we point out a few specificities of mixture estimation in such a particular setting. While the likelihood is a regular mixture likelihood, the fact that the weights are a priori close to the boundaries means that the usual completion approach of [Diebolt 1994] is bound to be quite inefficient as soon as the sample size grows to moderate values. More precisely, if we consider a sample $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ from (6.1) (or assumed to be from (6.1)), the completion of the sample by the latent component indicators ζ_i $(i = 1, \ldots, n)$ leads to the completed likelihood

$$\mathcal{L}(\theta, \alpha_1, \alpha_2 \mid \mathbf{x}, \zeta) = \prod_{i=1}^{n} \alpha_{\zeta_i} f(x_i \mid \theta_{\zeta_i}) = \alpha^{n_1}(1-\alpha)^{n_2} \prod_{i=1}^{n} f(x_i \mid \theta_{\zeta_i}), \tag{4.2}$$

where $(n_1, n_2) = (\sum_{i=1}^{n} \mathbb{I}_{\zeta_i=1}, \sum_{i=1}^{n} \mathbb{I}_{\zeta_i=2})$ under the constratin $n = \sum_{j=1}^{2} \sum_{i=1}^{n} \mathbb{I}_{\zeta_i=j}$. This decomposition leads to a natural Gibbs implementation [Diebolt 1994] where the latent variables ζ_i and the parameters are generated from their respective conditional distributions. For instance, under a Beta $\mathcal{B}e(a_1, a_2)$ prior, α is generated from a Beta $\mathcal{B}e(a_1 + n_1, a_2 + n_2)$.

However, while this Gibbs sampling scheme is valid from a theoretical point of view, it faces convergence difficulties in the current setting, especially with large samples, due to the prior concentration on the boundaries of $(0, 1)$ for the mixture weight α. This feature is illustrated by Figure 4.1: as the sample size n grows, the Gibbs sample of the α's shows less and less switches between the vicinity of zero and the vicinity of one. The lack of label switching for regular mixture models is well-know, see, e.g., [Celeux 2000] and [Lee 2009]. It is due to the low probability of switching all component labels ζ_i at once. This issue is simply exacerbated on the current case due to extreme values for α.

Therefore, an alternative to the Gibbs sampler is needed [Lee 2009] and we resort to a simple Metropolis-Hastings algorithm where the model parameters θ_i are generated from the respective posteriors of both models (that is, based on the entire sample) and where the mixture weight α is generated either from the prior distribution or from a random walk proposal on $(0, 1)$. It is indeed a quite rare occurrence for mixtures when we can use independent proposals. In the testing setting, the parameter θ_i can be considered independently within each model and its posterior can be based on the whole dataset. (In cases when a common parameter is used in both components, one of the two available posteriors is chosen at random at each iteration, either uniformly or based on the current value of α.) The equivalent of Figure 4.1 for this Metropolis–Hastings implementation, Figure 4.2 exhibits a clear difference in the exploration abilities of the resulting chain.

We also point out that, due to the specific pattern of the posterior distribution on α accumulating most of its weight on the endpoints of $(0, 1)$, the use of the posterior mean is highly inefficient and thus we advocate that the posterior median be instead used as the relevant estimator of α.

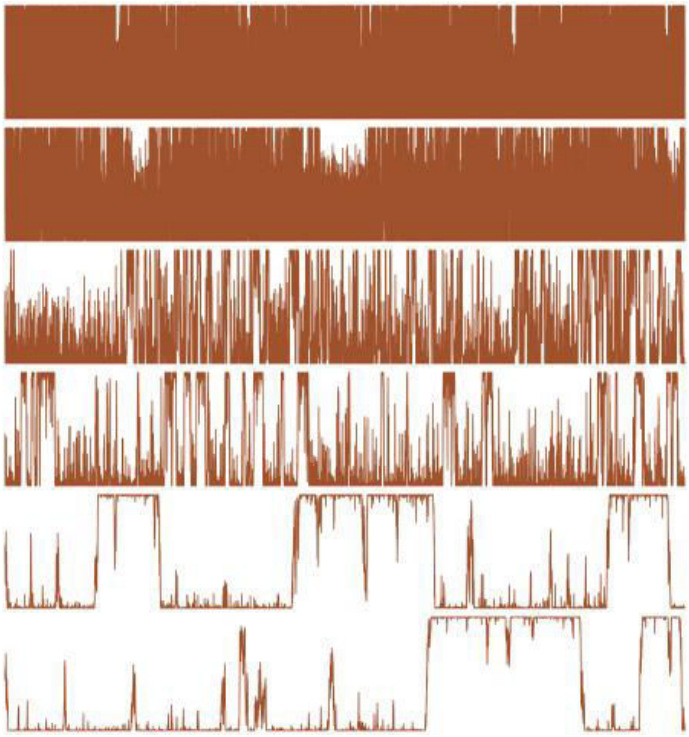


Figure 4.1: Gibbs sequences (α_t) on the first component weight for the mixture model $\alpha N(\mu, 1) + (1 - \alpha)N(0, 1)$ for a $N(0, 1)$ sample of size $N = 5, 10, 50, 100, 500, 10^3$ *(from top to bottom)* based on 10^5 simulations. The y-range range for all series is $(0, 1)$.

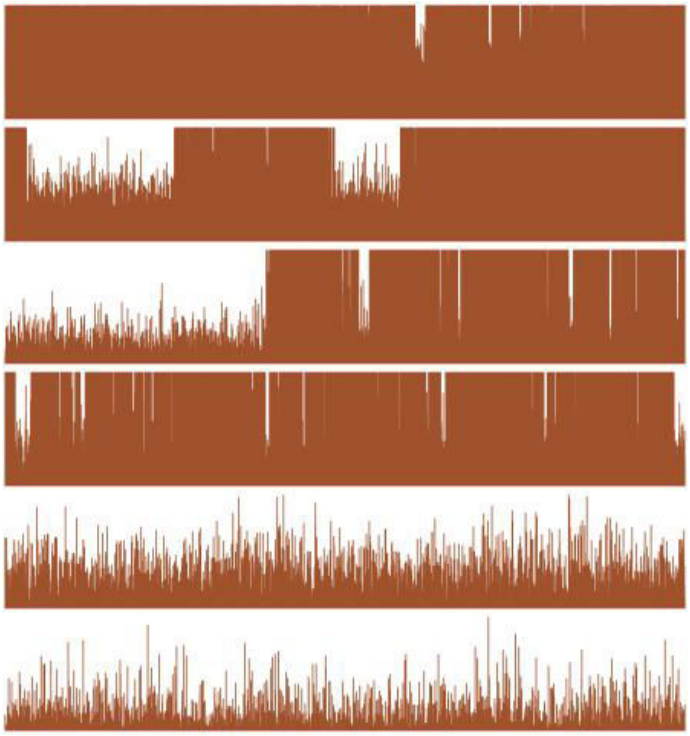


Figure 4.2: Metropolis–Hastings sequences (α_t) on the first component weight for the mixture model $\alpha N(\mu,1)+(1-\alpha)N(0,1)$ for a $N(0,1)$ sample of size $N = 5, 10, 50, 100, 500, 10^3$ *(from top to bottom)* based on 10^5 simulations.The y-range for all series is $(0,1)$.

4.3 Illustrations

In this Section, we proceed through a series of experiments in highly classical statistical settings in order to assess the performances of the mixture estimation approach for separating the models under comparison. As we will see throughout those examples, this experimentation brings a decisive confirmation of the consistency results obtained in Section 4.5. The first two examples are direct applications of Theorem 1 while the third is an application of Theorem 2.

Example 4.3.1 For a model choice test between a Poisson $\mathcal{P}(\lambda)$ and a Geometric $\mathcal{G}eo(p)$ (defined as a number of failures, hence also starting at zero) distribution, we can model the mixture (6.1) towards using the same parameter λ in the Poisson $\mathcal{P}(\lambda)$ and in the Geometric $\mathcal{G}eo(p)$ distribution if we set $p = 1/1+\lambda$. The resulting mixture, to be estimated, is then defined as

$$\mathfrak{M}_\alpha : \ \alpha\mathcal{P}(\lambda) + (1-\alpha)\mathcal{G}eo(1/1+\lambda)$$

This common parameterisation allows for the call to Jeffreys' (1939) improper prior $\pi(\lambda) = 1/\lambda$ since the resulting posterior is then proper. Indeed, in a Gibbs sampling implementation, the full posterior distribution on λ, conditional on the allocation vector ζ is given by

$$\pi(\lambda \mid \underline{x}, \zeta) \propto \exp(-n_1(\zeta)\lambda + \log\{\lambda\}\,(n\bar{x}_n - 1)\,(\lambda+1)^{-\{n_2(\zeta)+s_2(\zeta)\}}\,, \tag{4.3}$$

where $n_1(\zeta) = n - n_2(\zeta)$ is the number of observations allocated to the Poisson component, while $s_2(\zeta)$ is the sum of the observations that are allocated to the Geometric component. This conditional posterior is well-defined for every ζ when $n > 0$, which implies that the marginal posterior is similarly well-defined since ζ takes its values in a finite set. The distribution (4.3) can easily be simulated via a independent Metropolis-within-Gibbs step where the proposal distribution on λ is the Gamma distribution corresponding to the Poisson posterior. (The motivation for this choice is that, since both distributions share the same mean parameter, using the posterior distribution associated with either one of the components and all the observations should be realistic enough to produce high acceptance rates, even when the data is Geometric rather than Poisson. This is what happens in practice with acceptance rates higher than 75% in the Geometric case and close to 1 in the Poisson case. This strategy of relying on a model-based posterior as a proposal will be used throughout the examples. It obviously would not work in a regular mixture model.)

Under a $\mathcal{B}e(a_0, a_0)$ prior on α, the full conditional posterior density on α is a $\mathcal{B}e(n_1(\zeta)+a_0, n_2(\zeta)+a_0)$ distribution and the exact Bayes factor opposing the Poisson to the Geometric models is given by

$$\mathfrak{B}_{12} = n^{n\bar{x}_n} \prod_{i=1}^{n} x_i!\, \Gamma\left(n+2+\sum_{i=1}^{n} x_i\right) \Big/ \Gamma(n+2).$$

This Bayes factor is however undefined from a purely mathematical viewpoint, since it is associated with an improper prior on the parameter [Jeffreys 1939, DeGroot 1973, Berger 1998, Robert 2009b]. The posterior probability of the Poisson model is then derived as

$$\mathbb{P}(\mathfrak{M}_1|x) = \frac{\mathfrak{B}_{12}}{1+\mathfrak{B}_{12}}$$

when adopting (without much of a justification) identical prior weights on both models.

A first experiment in assessing our approach is based on 100 datasets simulated from a Poisson $\mathcal{P}(4)$ distribution. As shown in Figure 4.3, not only is the parameter λ properly estimated, but the estimation of α is very close to 1 for a sample size equal to $n = 1000$. In this case, the smaller the value of a_0, the better in terms of proximity to 1 of the posterior distribution on α. Note that the choice of a_0 does not significantly impact the posterior distribution of λ. Figure 4.4 gives an assessment of the convergence of the Metropolis-Hastings for λ and the mixture model weight α even if the sample size is very small (n=5).

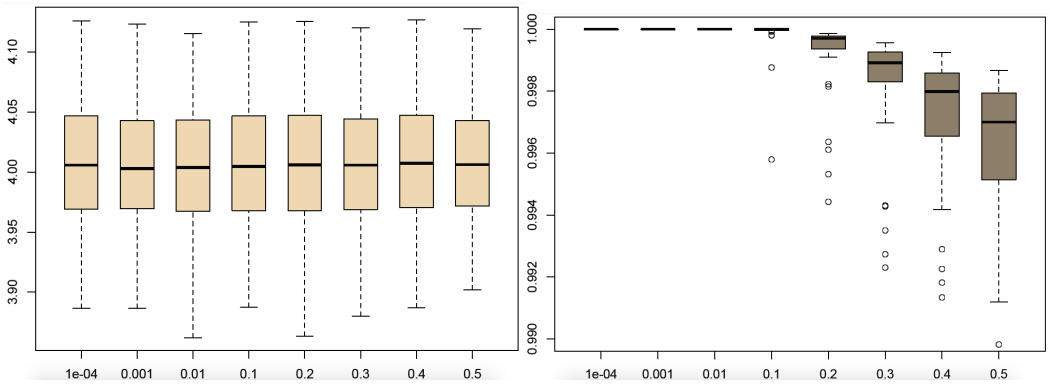


Figure 4.3: **Example 4.3.1:** Boxplots of the posterior means *(wheat)* of λ and the posterior medians *(dark wheat)* of α for 100 Poisson $\mathcal{P}(4)$ datasets of size $n = 1000$ for $a_0 = .0001, .001, .01, .1, .2, .3, .4, .5$. Each posterior approximation is based on 10^4 Metropolis-Hastings iterations.

Figure 4.5 highlights the convergence of the posterior means and posterior medians of α as the sample sizes n increase for the same Poisson $\mathcal{P}(4)$ simulated samples. The sensitivity of the posterior distribution of α on the hyperparameter a_0 is clearly expressed by that graph. While all posterior means and medians converge to 1 in this simulation, the impact of small values of a_0 on the estimates is such that we consider values $a_0 \leq .1$ as having too strong and too lengthy an influence on the posterior distribution to be acceptable.

We can also compare the outcome of a traditional (albeit invalid, since relying on improper priors) Bayesian analysis with our estimates of α. Figure 4.6 shows how the posterior probability of model $\mathfrak{M}_1$ and the posterior median of α relate as the sample size grows to 1000. The shaded areas indicate the range of all estimates of α, which varies between .2 and .8 for $a_0 = .5$ and between 0 and 1 for $a_0 \leq .1$. This

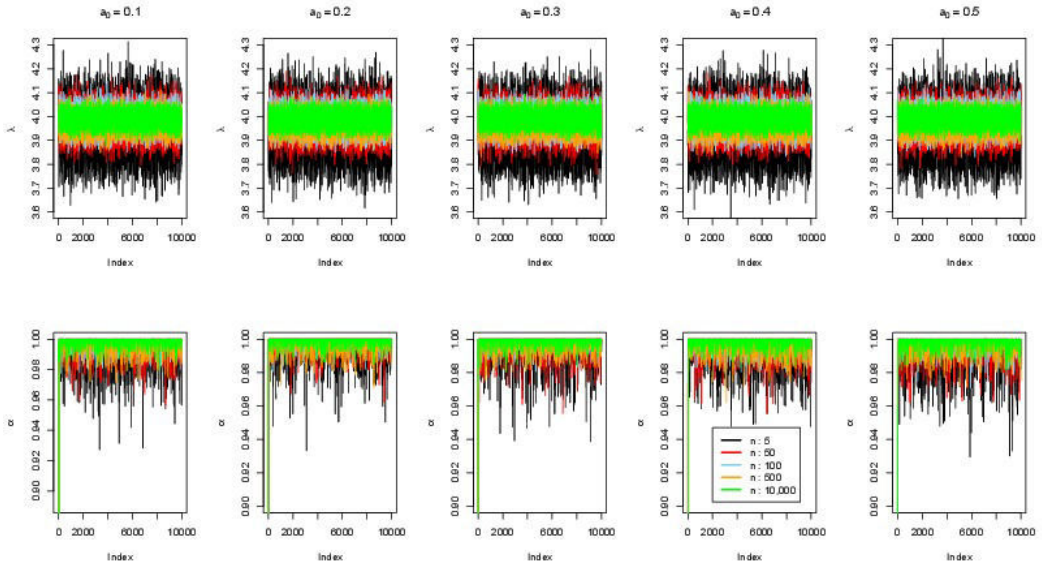


Figure 4.4: **Example 4.3.1:** Dataset from a Poisson distribution $\mathcal{P}(4)$: Estimations of *(Top)* λ and *(Bottom)* α via Metropolis-Hastings algorithm over 10^4 iterations for 5 samples of size $n = 5, 50, 100, 500, 10,000$.

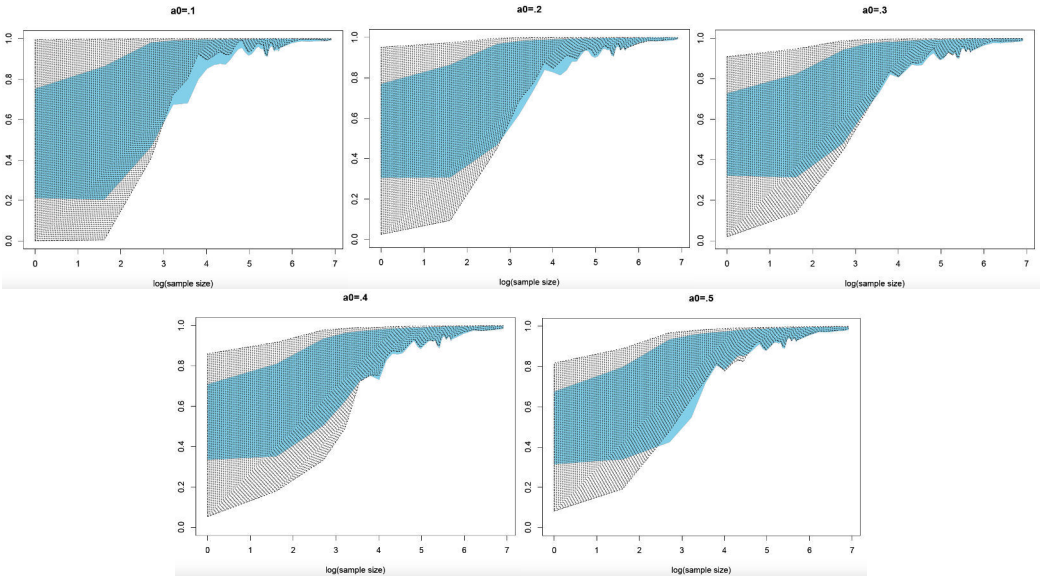
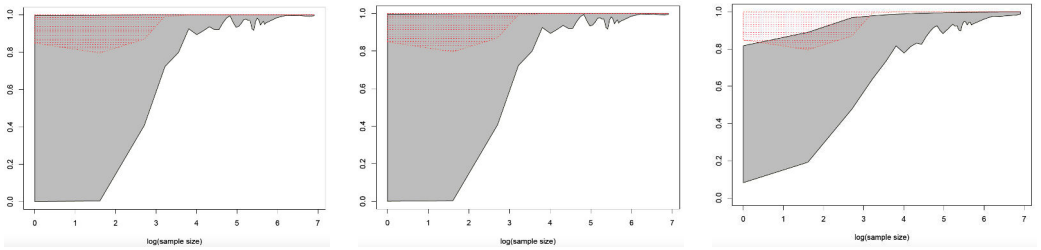


Figure 4.5: **Example 4.3.1:** Posterior means *(sky-blue)* and medians *(grey-dotted)* of the posterior distributions on α, displayed over 100 Poisson $\mathcal{P}(4)$ datasets for sample sizes from 1 to 1000. The shaded and dotted areas indicate the range of the estimates. Each plot corresponds to a Beta prior on α with parameter $a_0 = .1, .2, .3, .4, .5$ and each posterior approximation is based on 10^4 iterations.

difference reinforces our earlier recommendation that smaller values of a_0 should be avoided, as they overwhelm the information contained in the data for small sample



(a) a_0=.0001 (b) a_0=.1 (c) a_0=.5

Figure 4.6: **Example 4.3.1:** Comparison between the ranges of $\mathbb{P}(\mathfrak{M}_1|x)$ *(red dotted area)* and of the posterior medians of α for 100 Poisson $\mathcal{P}(4)$ datasets with sample sizes n ranging from 1 to 1000 and for several values of the hyperparameter a_0.

sizes.

A symmetric experiment is to study the behavior of the posterior distribution on α for data from the alternative model, i.e., a Geometric distribution. Based on 100 datasets from a Geometric $\mathcal{G}(0.1)$ distribution, Figure 4.7 displays the very quick convergence of the posterior median to 0 for all values of a_0 considered, even though the impact of this hyperprior is noticeable.

Example 4.3.2 For the model comparison of a normal $\mathcal{N}(\theta_1, 1)$ with a normal $\mathcal{N}(\theta_2, 2)$ distribution, we again model the mixture so that the same location parameter θ is used in both the normal $\mathcal{N}(\theta, 1)$ and the normal $\mathcal{N}(\theta, 2)$ distribution. Therefore, Jeffreys' (1939) noninformative prior $\pi(\theta) = 1$ can be used, in contrast with the corresponding Bayes factor. Indeed, when considering the mixture of normal models, $\alpha\mathcal{N}(\theta, 1) + (1-\alpha)\mathcal{N}(\theta, 2)$, and a Beta $\mathcal{B}(a_0, a_0)$ prior on α, considering the posterior distribution on (α, θ), conditional on the allocation vector ζ, leads to conditional independence between θ and α:

$$\theta|\mathbf{x}, \zeta \sim \mathcal{N}\left(\frac{n_1\bar{x}_1 + .5n_2\bar{x}_2}{n_1 + .5n_2}, \frac{1}{n_1 + .5n_2}\right), \quad \alpha|\zeta \sim \mathcal{B}e(a_0 + n_1, a_0 + n_2),$$

where n_i and $\bar{x}_i$ denote the number of observations and the empirical mean of the observations allocated to component i, respectively (with the convention that $n_i\bar{x}_i = 0$ when $n_i = 0$. Since this conditional posterior distribution is well-defined for every possible value of ζ and since the distribution ζ has a finite support, $\pi(\theta|x)$ is proper.[5]

For the same purpose of evaluating the convergence rates of the estimates of the mixture weights, we simulated 100 $\mathcal{N}(0, 1)$ datasets. Figure 4.8 displays the range of the posterior means and medians of α when either a_0 or n varies, showing the

[5]For this example, the conditional evidence $\pi(x|\zeta)$ can easily be derived in closed form, which means that a random walk on the allocation space $\{1, 2\}^n$ could be implemented. We did not follow that direction, as it seemed unlikely such a random walk would have been more efficient than a Metropolis–Hastings algorithm on the parameter space only.

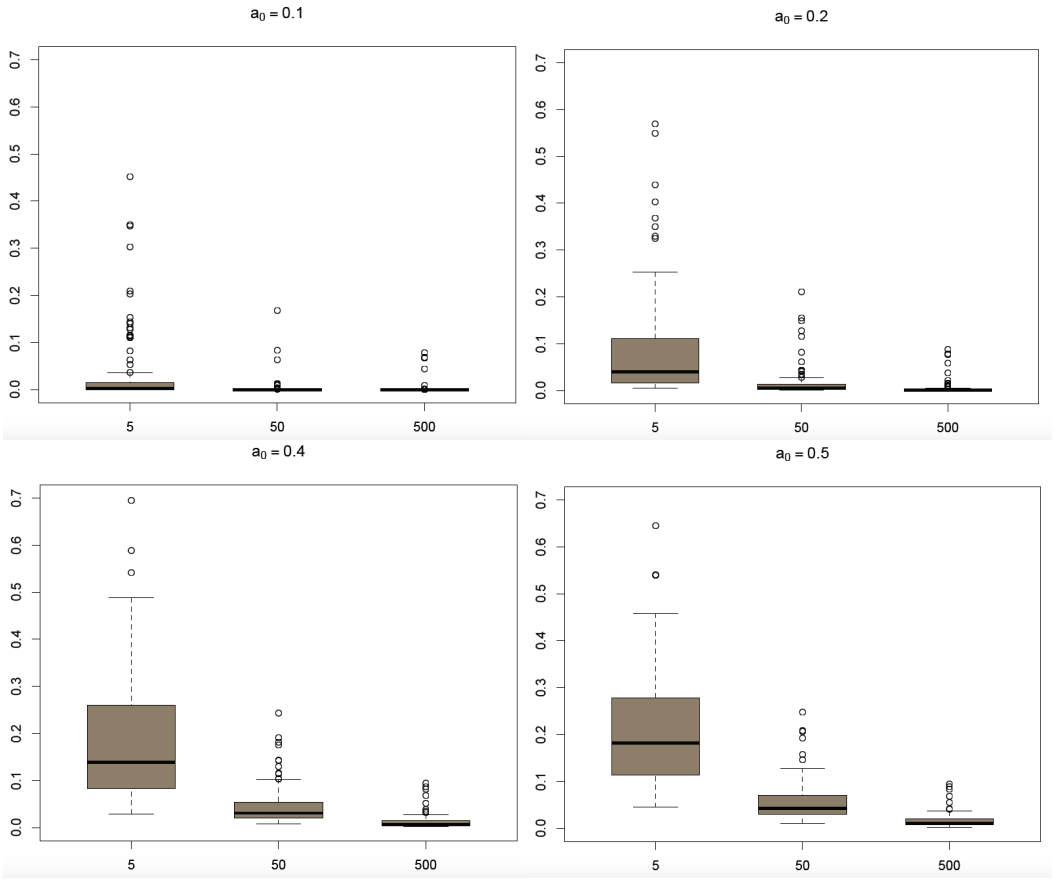


Figure 4.7: **Example 4.3.1:** Boxplots of the posterior medians of α for 100 geometric $\mathcal{G}eo(.1)$ datasets of size $n = 5, 50, 500$. Boxplots are plotted using four beta priors for α with $a0 = .1, .2, .4, .5$. Each posterior approximation is based on 10^4 iterations.

same concentration effect (if a lingering impact of a_o) when n increases. We also included the posterior probability of $\mathfrak{M}_1$ in the comparison, derived from the Bayes factor

$$\mathfrak{B}_{12} = 2^{n-1/2} \Big/ \exp 1/4 \sum_{i=1}^{n} (x_i - \bar{x})^2 \,,$$

with equal prior weights, even though it formally is not well-defined since based on an improper prior. The shrinkage of the posterior expectations towards 0.5 confirms our recommendation to use the posterior median instead. The same concentration phenomenon occurs for the $\mathcal{N}(0,2)$ case, as illustrated on Figure 4.10 for a single $\mathcal{N}(0,2)$ dataset.

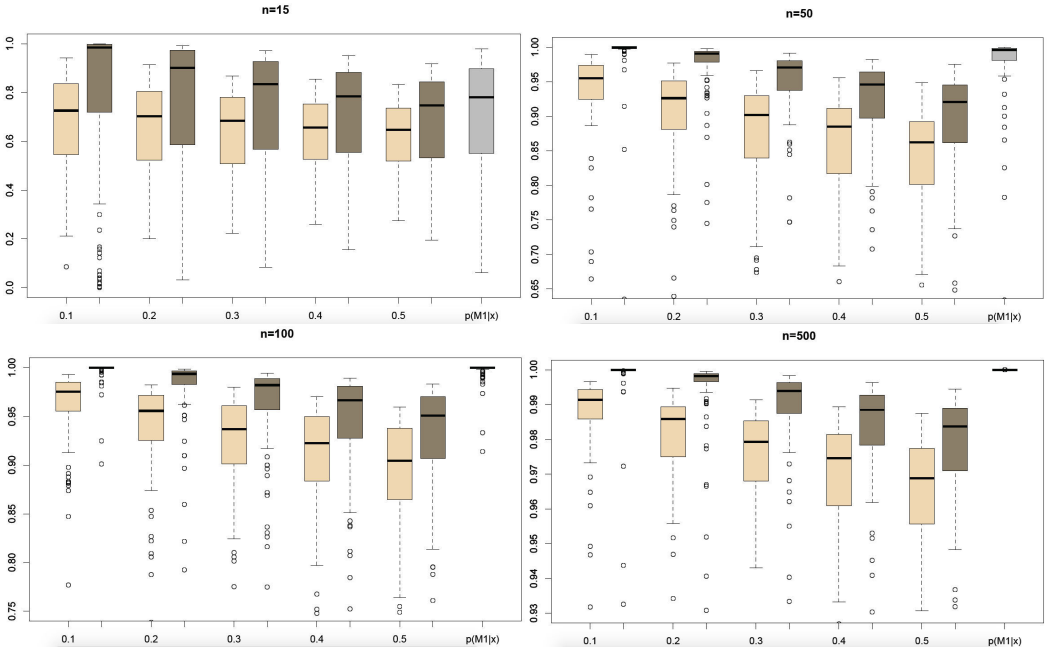


Figure 4.8: **Example 4.3.2:** Boxplots of the posterior means *(wheat)* and medians of α *(dark wheat)*, compared with a boxplot of the exact posterior probabilities of $\mathfrak{M}_0$ *(gray)* for a $\mathcal{N}(0,1)$ sample, derived from 100 datasets for sample sizes equal to $15, 50, 100, 500$. Each posterior approximation is based on 10^4 MCMC iterations.

In order to better understand the nature of the convergence of the posterior distribution of α towards the proper limiting value, we plotted in Figure 4.9 a zoomed version of this convergence, by comparing $\log(n)\ \log(1-\mathbb{E}[\alpha|\mathbf{x}])$ with $\log(1-p(\mathfrak{M}_1|\mathbf{x}))$ as the sample size n grows. Most interestingly, the variation range is of the same magnitude for both procedures, even though the choice of the hyperparameter a_0 impacts the variability of the mixture solution. This is due to the fact that the asymptotic regime is not quite reached for those sample sizes, as $1-\mathbb{P}(\mathfrak{M}_1|\mathbf{x}) \leq e^{-cn}$ for some positive c with high probability, while $\mathbb{E}[\alpha|\mathbf{x}] = O(n^{-1/2})$, leading to

$$\log(n)\ \log(1 - \mathbb{E}[\alpha|\mathbf{x}]) \asymp -(\log n)^2.$$

Furthermore, the alternative of considering the posterior probability of having the *entire sample* being generated from a single component is not relevant for the comparison as this estimate is always very close to zero. This means that, while α captures the model preferred by the data, the mixture modelling itself never favours a completely homogeneous sample that would come from one and only one component. By comparison, note that if we had instead called the algorithm of [van Havre 2014], we would have obtained mostly homogeneous samples for very small values of a_0. Their algorithm is a special type of tempering MCMC, where tempering is obtained by choosing successive values of a_0, ranging from large to very small.

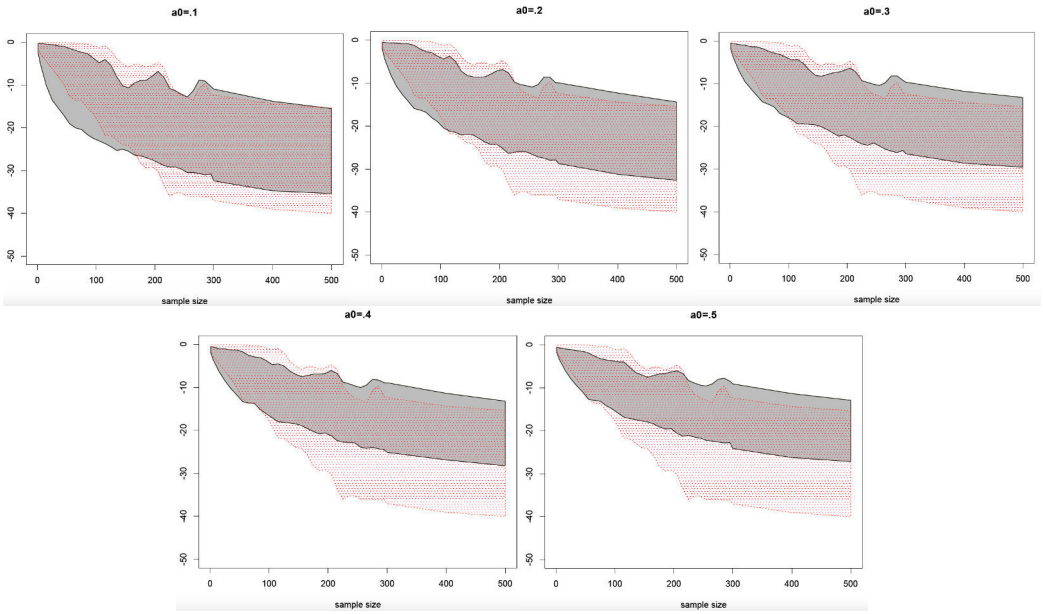


Figure 4.9: **Example 4.3.2:** Plots of ranges of $\log(n)\,\log(1-\mathbb{E}[\alpha|x])$ *(gray color)* and $\log(1-p(\mathfrak{M}_1|x))$ *(red dotted)* over 100 $\mathcal{N}(0,1)$ samples as sample size n grows from 1 to 500. and α is the weight of $\mathcal{N}(0,1)$ in the mixture model. The shaded areas indicate the range of the estimations and each plot is based on a Beta prior with $a_0 = .1, .2, .3, .4, .5, 1$ and each posterior approximation is based on 10^4 iterations.

Example 4.3.3 We now consider a setting where we oppose a $\mathcal{N}(0,1)$ model against a $\mathcal{N}(\mu,1)$ model, hence testing whether or not $\mu = 0$. This being an embedded case, we cannot use an improper prior on μ and thus settle for a $\mu \sim \mathcal{N}(0,1)$ prior. As discussed above in Section 6.2, Gibbs sampling applied to this mixture posterior model shows poor performances and should be replaced with a Metropolis–Hastings algorithm.

The resulting inference on the weight of the $\mathcal{N}(\mu,1)$ component, α, is unsurprisingly contrasted between the case when the data is distributed as $\mathcal{N}(0,1)$ and when it is not from this null distribution. In the former case, obtaining values of α close to one requires larger sample sizes than in the latter case. Figure 4.11 displays the

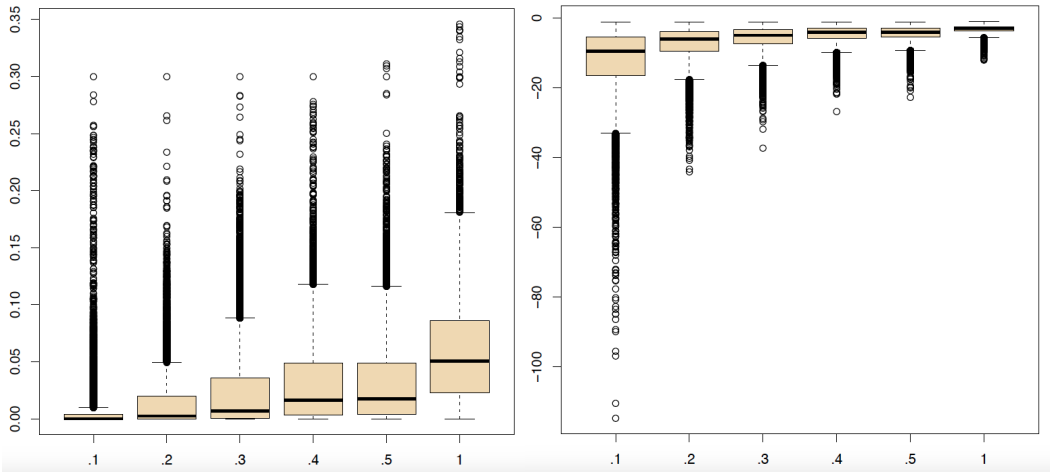


Figure 4.10: **Example 4.3.2:** *(left)* Posterior distributions of the mixture weight α and *(right)* of their logarithmic transform $\log\{\alpha\}$ under a Beta $\mathcal{B}(a_0, a_0)$ prior when $a_0 = .1, .2, .3, .4, .5, 1$ and for a normal $\mathcal{N}(0, 2)$ sample of 10^3 observations. The MCMC outcome is based on 10^4 iterations.

behavior of the posterior distribution of α when the sample comes from a normal distribution $\mathcal{N}(1, 1)$. For a sample of size 10^2, the accumulation of α on $(.8, 1)$ illustrates the strength of the support for the model $\mathcal{N}(\mu, 1)$ which is reduced with the increase of a_0. The impact of the small sample size on the posterior distributions of α is shown in the right side of Figure 4.11 for the case where $a_0 = .1$ such that for $n = 5$ we can not recognize which model is fitter to the data.

Example 4.3.4 Inspired from [Marin 2014], we oppose the normal $\mathcal{N}(\mu, 1)$ model to the double-exponential $\mathcal{L}(\mu, \sqrt{2})$ model. The scale $\sqrt{2}$ is intentionally chosen to make both distributions share the same variance. As in the normal case in Example 4.3.2, the location parameter μ can be shared by both models and allows for the use of the flat Jeffreys' prior. As in all previous examples, Beta distributions $\mathcal{B}(a_0, a_0)$ are compared wrt their hyperparameter a_0.

While, in those previous examples, we illustrated that the posterior distribution of the weight of the true model converged to 1, we now consider the setting of a dataset produce by another model than those in competition, using, e.g., $\mathcal{N}(0, .7^2)$ to simulate the data. In this specific case, both posterior means and medians of α fail to concentrate near 0 and 1 as the sample size increases, as shown on Figure 4.12. So in a majority of cases in this experiment, the outcome indicates that neither of both models is favored by the data. This example does not exactly follow the assumptions of Theorem 1 since the Laplace distribution is not differentiable everywhere, however it is almost surely differentiable and it is differentiable in quadratic mean and so we expect to see the same types of behavior as predicted by Theorem 1.

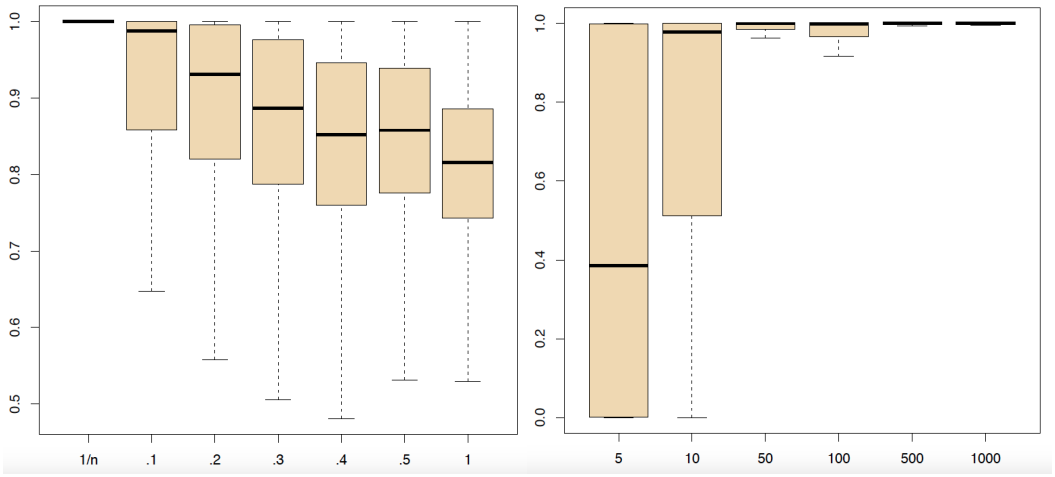


Figure 4.11: **Example 4.3.3** Posterior distributions of the $\mathcal{N}(\mu, 1)$ component weight α under a Beta $\mathcal{B}(a_0, a_0)$ prior *(left)* for $a_0 = 1/n, .1, .2, .3, .4, .5, 1$ with 10^2 $\mathcal{N}(1, 1)$ observations and *(right)* for $a_0 = .1$ with $n = 5, 10, 50, 100, 500, 10^3$ $\mathcal{N}(1, 1)$ observations. In both cases each posterior approximation is based on 10^5 MCMC iterations.

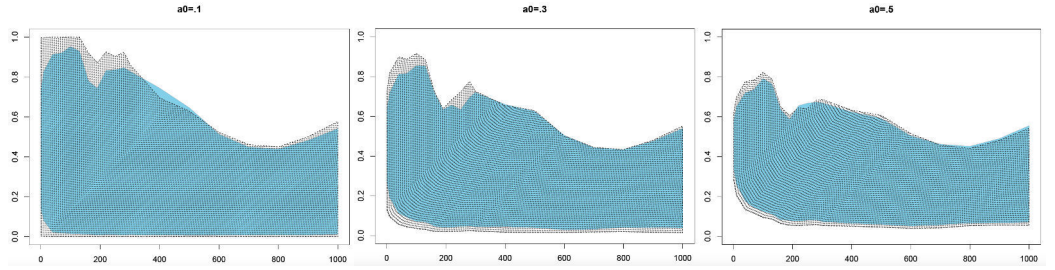


Figure 4.12: **Example 4.3.4:** Ranges of posterior means *(skyblue)* and medians *(dotted)* of the weight α of model $\mathcal{N}(\theta, 1)$ over 100 $\mathcal{N}(0, .7^2)$ datasets for sample sizes from 1 to 1000. Each estimate is based on a Beta prior with $a_0 = .1, .3, .5$ and 10^4 MCMC iterations.

In this example. the Bayes factor associated with Jeffreys' prior is defined as

$$\mathfrak{B}_{12} = \frac{\exp\left\{-\sum_{i=1}^n (x_i - \bar{x})^2/2\right\}}{(\sqrt{2\pi})^{n-1}\sqrt{n}} \Big/ \int_{-\infty}^{\infty} \frac{\exp\left\{-\sum_{i=1}^n |x_i - \mu|/\sqrt{2}\right\}}{(2\sqrt{2})^n} \mathrm{d}\mu$$

where the denominator is available in closed form (see 5.5). Since the prior is improper, it is formally undefined. Using nonetheless the above expression, we can compare Bayes estimators of α with the posterior probability of the model being a $\mathcal{N}(\mu, 1)$ distribution. Based on a Monte Carlo experiment involving 100 replicas of a $\mathcal{N}(0, .7^2)$ dataset, Figure 4.13 demonstrates how the mixture estimate mostly stay away from 0 and 1 while $\mathbb{P}(\mathfrak{M}_1|\mathbf{x})$ varies all over between 0 and 1 for all sample sizes considered here. While this is a weakly informative indication, the right hand side of Figure 4.13 shows that, on average, the posterior estimates of α converge toward a value between .1 and .4 for all a_0 while the posterior probabilities converge to .6. In that respect, both criteria offer a similar interpretation about the data because

neither α nor $P(\mathfrak{M}_1|x)$ confirm that either of the models is true.

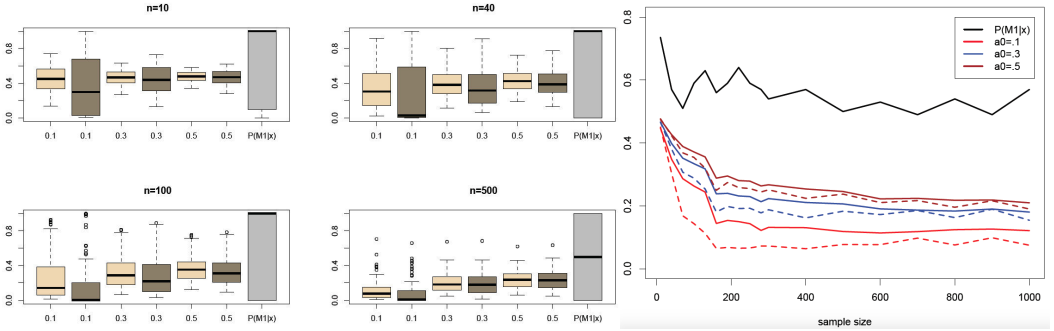


Figure 4.13: **Example 4.3.4:** *(left)* Boxplot of the posterior means *(wheat)* and medians *(dark wheat)* of α, and of the posterior probabilities of model $\mathcal{N}(\mu, 1)$ over 100 $\mathcal{N}(0, .7^2)$ datasets for sample sizes $n = 10, 40, 100, 500$; *(right)* averages of the posterior means and posterior medians of α against the posterior probabilities $\mathbb{P}(\mathfrak{M}_1|\mathbf{x})$ for sample sizes going from 1 to 1000. Each posterior approximation is based on 10^4 Metropolis-Hastings iterations.

In the following two examples we consider regression models. Although they do not strictly speaking follow the identically setup, the methodology can be extended to this case and so can the theory, asuming that the design is random. Hence example 4.3.5 is an application of Theorem 1 while example 4.3.6 is an application of Theorem 2.

Example 4.3.5 In this example, we apply our testing strategy to a binary dataset, using the R dataset about diabetes in Pima Indian women [R Development Core Team 2006] as a benchmark [Marin 2007]. This dataset contains a randomly selected table of 200 women tested for diabetes according to WHO criteria. The response variable y is "Yes" or "No", for presence or absence of diabetes and the explanatory variable $\mathbf{x}$ is restricted here to the *bmi*, body mass index weight in kg/(height in m)^2. For this binary dataset, either logistic or probit regression models could be suitable. We are thus comparing both fits via our method. If $\mathbf{y} = (y_1 \quad y_2 \ldots y_n)$ is the vector of binary responses and $X = [I_n \quad \mathbf{x}_1]$ is the $n \times 2$ matrix of corresponding explanatory variables, the models in competition can be defined as $(i = 1, \ldots, n)$

$$\begin{aligned} \mathfrak{M}_1 &: y_i \mid \mathbf{x}^i, \theta_1 \sim \mathcal{B}(1, p_i) \quad \text{where} \quad p_i = \frac{\exp(\mathbf{x}^i\theta_1)}{1+\exp(\mathbf{x}^i\theta_1)} \\ \mathfrak{M}_2 &: y_i \mid \mathbf{x}^i, \theta_2 \sim \mathcal{B}(1, q_i) \quad \text{where} \quad q_i = \Phi(\mathbf{x}^i\theta_2) \end{aligned} \tag{4.4}$$

where $\mathbf{x}^i = (1 \; x_{i1})$ is the vector of explanatory variables and where θ_j, $j = 1, 2$, is a 2×1 vector made of the intercept and of the regression coefficient under either $\mathfrak{M}_1$ or $\mathfrak{M}_2$. We once again consider the case where both models share the same parameter.

However, the model is a generalized linear model and there is no moment equation that relates θ_1 and θ_2. We thus adopt a local reparameterisation strategy by rescaling the parameters of the probit model $\mathfrak{M}_2$ so that the MLE's of both models coincide. This strategy follows from [Choudhury 2007] remark on the connection between the normal cdf and a logistic function

$$\Phi(\mathbf{x}^i\theta_2) \approx \frac{\exp(k\mathbf{x}^i\theta_2)}{1+\exp(k\mathbf{x}^i\theta_2)}$$

and we attempt to find the best estimate of k to bring both parameters into coherency. Given

$$(k_0, k_1) = (\widehat{\theta_{01}}/\widehat{\theta_{02}}, \widehat{\theta_{11}}/\widehat{\theta_{12}})\,,$$

ratios of the maximum likelihood estimates of the logistic model parameters to those for the probit model, we reparameterise $\mathfrak{M}_1$ and $\mathfrak{M}_2$ defined in (4.4) as

$$\begin{aligned} \mathfrak{M}_1 &: y_i \mid \mathbf{x}^i, \theta \sim \mathcal{B}(1, p_i) \quad \text{where} \quad p_i = \frac{\exp(\mathbf{x}^i\theta)}{1+\exp(\mathbf{x}^i\theta)} \\ \mathfrak{M}_2 &: y_i \mid \mathbf{x}^i, \theta \sim \mathcal{B}(1, q_i) \quad \text{where} \quad q_i = \Phi(\mathbf{x}^i(\kappa^{-1}\theta))\,, \end{aligned} \tag{4.5}$$

where $\kappa^{-1}\theta = (\theta_0/k_0, \theta_1/k_1)$.

Once the mixture model is thus parameterised, we set our now standard Beta $\mathcal{B}(a_0, a_0)$ on the weight of $\mathfrak{M}_1$, α, and choose the default g-prior on the regression parameter (see, e.g., Chapter 4.[Marin 2007]),

$$\theta \sim \mathcal{N}_2(0, n(X^{\mathrm{T}}X)^{-1})$$

In a Gibbs representation (not implemented here), the full conditional posterior distributions given the allocation vector ζ are that $\alpha \sim \mathcal{B}(a_0+n_1, a_0+n_2)$ and that

$$\begin{aligned} \pi(\theta \mid \mathbf{y}, X, \zeta) \propto & \frac{\exp\left\{\sum_i \mathbb{I}_{\zeta_i=1} y_i \mathbf{x}^i\theta\right\}}{\prod_{i;\zeta_i=1}[1+\exp(\mathbf{x}^i\theta)]} \exp\left\{-\theta^T(X^TX)\theta/2n\right\} \\ & \times \prod_{i;\zeta_i=2} \Phi(\mathbf{x}^i(\kappa^{-1}\theta))^{y_i}(1-\Phi(\mathbf{x}^i(\kappa^{-1}\theta)))^{(1-y_i)} \end{aligned} \tag{4.6}$$

where n_1 and n_2 are the number of observations allocated to the logistic and probit models, respectively. This conditional representation shows that the posterior distribution is then clearly defined, which is obvious when considering that for once the chosen prior is proper.

For the Pima dataset, the maximum likelihood estimates of the GLMs are $\hat{\theta_1} = (-4.11, 0.10)$ and $\hat{\theta_2} = (-2.54, 0.065)$, respectively, and so $k = (1.616, 1.617)$. We compare the outcomes of this Bayesian analysis when $a_0 = .1, .2, .3, .4, .5$ in Table 4.1. As clearly shown by the Table, the estimates of α are close to 0.5, no matter what the value of a_0 while the estimates of θ_0 and θ_1 are very stable (and quite similar to the MLEs). We note a slight increase of α towards 0.5 as a_0 increases, but do not want to over-interpret the phenomenon. This behavior leads us to conclude that (a) none or both of the models are appropriate for the Pima Indian data; (b) the sample

		Logistic model parameters		Probit model parameters	
a_0	α	θ_0	θ_1	$\frac{\theta_0}{k_0}$	$\frac{\theta_1}{k_1}$
.1	.352	-4.06	.103	-2.51	.064
.2	.427	-4.03	.103	-2.49	.064
.3	.440	-4.02	.102	-2.49	.063
.4	.456	-4.01	.102	-2.48	.063
.5	.449	-4.05	.103	-2.51	.064

Table 4.1: Dataset Pima.tr: Posterior medians of the mixture model parameters.

		$\mathfrak{M}_\alpha^1$					$\mathfrak{M}_\alpha^2$				
True model:		**logistic with** $\theta_1 = (5, 1.5)$					**probit with** $\theta_2 = (3.5, .8)$				
a_0		α	θ_0	θ_1	$\frac{\theta_0}{k_0}$	$\frac{\theta_1}{k_1}$	α	θ_0	θ_1	$\frac{\theta_0}{k_0}$	$\frac{\theta_1}{k_1}$
.1		.998	4.940	1.480	2.460	.640	.003	7.617	1.777	3.547	.786
.2		.972	4.935	1.490	2.459	.650	.039	7.606	1.778	3.542	.787
.3		.918	4.942	1.484	2.463	.646	.088	7.624	1.781	3.550	.788
.4		.872	4.945	1.485	2.464	.646	.141	7.616	1.791	3.547	.792
.5		.836	4.947	1.489	2.465	.648	.186	7.596	1.782	3.537	.788

Table 4.2: Simulated dataset: Posterior medians of the mixture model parameters.

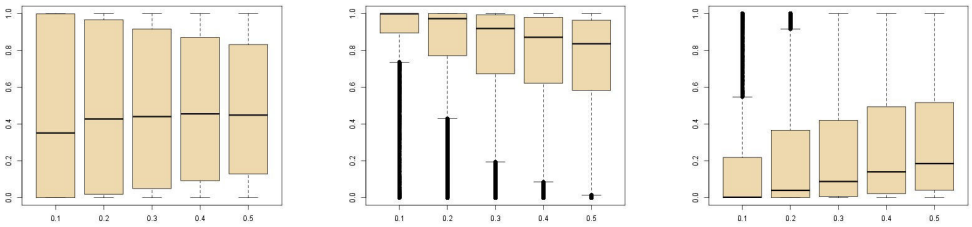
size may be insufficiently large for allowing a discrimination between the logit and the probit models.

Since the benchmark dataset is apparently too small to reach the asymptotic regime, we ran a second experiment with simulated logit and probit datasets and a larger sample size $n = 10,000$. For the logit model, we used the regression coefficients $(5, 1.5)$ and for the probit model the regression coefficients $(3.5, .8)$. The estimates of the parameters of both $\mathfrak{M}_{\alpha_1}$ and $\mathfrak{M}_{\alpha_2}$ and for both datasets are produced in Table 4.2. For every a_0, the estimates in the true model are quite close to the true values and the posterior estimates of α are either close to 1 in the logit case and to 0 in the probit case. For this large setting, there is thus consistency in the selection of the proper model. In addition, Figure 4.14 shows that when the sample size is large enough, the posterior distribution of α concentrates its mass near 1 and 0 when the data is simulated from a logit and a probit model, respectively.

Example 4.3.6 We now examine the classical issue of variable selection in a Gaussian linear regression model. Given a vector of outcomes $(y_1, y_2, \ldots, y_n)$ and the corresponding explanatory variables represented by the $n \times (k+1)$ matrix $X = [\mathbf{1}_n \quad X_1 \quad \ldots \quad X_k]$ (including $\mathbf{1}_n$, a first column of 1's), we assume that

$$y \mid X, \beta, \sigma^2 \sim \mathcal{N}_n(X\beta, \sigma^2 I_n) \tag{4.7}$$

where $\beta = (\beta_0, \beta_1, \ldots, \beta_k)^T$ is a $k+1$-vector of $k+1$ elements with β_0 the intercept. If we consider the generic case where any covariate could be removed from the model,



(a) Pima dataset (b) Data from logistic (c) Data from probit

Figure 4.14: **Example 4.3.5:** Posterior distributions of α in favor of the logistic model based on 10^4 Metropolis-Hastings iterations where $a_0 = .1, .2, .3, .4, .5$.

we are facing the comparison of $2^{k+1} - 1$ models, corresponding to every possible subset of explanatory variables. In our framework, this means evaluating a mixture model (6.1) with $\gamma = 2^{k+1} - 1$ components. For $j = 1, \ldots, \gamma$, $\mathfrak{M}_j$ will denote the corresponding model, υ_j the number of explanatory variables used in $\mathfrak{M}_j$, β^j the vector of the υ_j regression coefficients and X^j the sub-matrix of X derived from the covariate variables included in $\mathfrak{M}_j$.

The corresponding mixture model used for testing is therefore given by

$$\mathfrak{M}_\alpha : \; y \sim \sum_{j=1}^{\gamma} \alpha_j \mathcal{N}(X^j \beta^j, \sigma^2 I_n) \qquad \sum_{j=1}^{\gamma} \alpha_j = 1 \,. \tag{4.8}$$

When introducing a missing variable representation, each observation y_i is associated with a missing variable ζ_i taking values in $1, 2, \ldots, \gamma$. The weights of the mixture (4.8) are associated with a symmetric Dirichlet prior $(\alpha_1, \ldots, \alpha_\gamma) \sim \mathcal{D}_\gamma(a_0, \ldots, a_0)$.

Contrary to the previous examples of this section, we now consider two different settings, corresponding to the separate versus common parameterisations of the different models $\mathfrak{M}_j$.

Case 1. If $\mathfrak{M}_f$ denotes the full regression model, including all k explanatory variables, we impose that β^j is a subvector of β^f for all j's. Therefore the models $\mathfrak{M}_j$ and therefore the mixture model (4.8) all are parameterised in terms of *the same* β^f. To simplify the notation, we will denote this common parameter vector by $\beta = (\beta_0, \beta_1, \ldots, \beta_k)^T$. Therefore, conditional on $\zeta_i = j$, we have

$$y_i \sim \mathcal{N}\left(X(i) \cdot j_2 \beta, \sigma^2\right) ,$$

where $X(i)$ denotes the i-th row of X and j_2 is the binary (base 2) representation of the integer j, with the convention that $X(i) \cdot j_2$ means a term-by-term multiplication, i.e., that this vector contains zero entries for the components of j_2 that are equal to zero:

$$X(i) \cdot j_2 = (X(i)_1 j_2[1], \ldots, X(i)_k j_2[k]) \,.$$

Assuming $\upsilon_j > 0$ and gathering all observations such that $\zeta_i = j$ under the notation $y_{i;\zeta_i=j}$ and the corresponding covariates by $X_{i;\zeta_i=j}$, we then have

$$y_{i;\zeta_i=j} \sim \mathcal{N}_{\upsilon_j}\left(X_{i;\zeta_i=j} \cdot j_2\beta, \sigma^2 I_{\upsilon_j}\right) ,$$

with the same convention about the term-by-term multiplication. The overall model conditional on $\zeta = (\zeta_1, \ldots, \zeta_n)$, the conditional distribution of the dataset is therefore

$$\begin{pmatrix} y_{i;\zeta_i=1} \\ y_{i;\zeta_i=2} \\ \vdots \\ y_{i;\zeta_i=\gamma} \end{pmatrix}_{n\times 1} = \begin{pmatrix} \mathbf{1}_{i;\zeta_i=1} & 1_2[1][X_1]_{i;\zeta_i=1} & \ldots 1_2[k][X_k]_{i;\zeta_i=1} \\ \mathbf{1}_{i;\zeta_i=2} & 2_2[1][X_1]_{i;\zeta_i=2} & \ldots 2_2[k][X_k]_{i;\zeta_i=2} \\ \vdots & \vdots & \vdots \\ \mathbf{1}_{i;\zeta_i=\gamma} & \gamma_2[1][X_1]_{i;\zeta_i=\gamma} & \ldots \gamma_2[k][X_k]_{i;\zeta_i=\gamma} \end{pmatrix}_{n\times(k+1)} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix}_{(k+1)\times 1} + \varepsilon_{n\times 1}$$

where $\mathbf{1}_{i;\zeta_i=j}$ is a υ_j-dimensional vector of 1's. By convention, any value of j such that $\upsilon_j = 0$ does not appear in the above. If we summarize the above equation as $\mathbf{y}_\zeta = \mathbf{X}_\zeta\beta + \varepsilon$ and use a Zellner's [Zellner 1986] G-prior,

$$\beta|\sigma \sim \mathcal{N}_{k+1}\left(M_{k+1}, c\sigma^2(X^TX)^{-1}\right) , \quad \pi(\sigma^2) \propto 1/\sigma^2 ,$$

the full conditional posterior distribution on the parameters is defined as

$$(\alpha_1, \ldots, \alpha_\gamma)|\zeta \sim \mathcal{D}_\gamma(\upsilon_1+a_0, \ldots, \upsilon_\gamma+a_0) , \beta|y, \zeta, \sigma \sim \mathcal{N}_{k+1}(\bar{\beta}, \bar{\Sigma}) , \sigma^2|y, \beta \sim \mathcal{IG}(a, b) ,$$

where

$$\begin{aligned} \bar{\beta} &= \bar{\Sigma}\left\{X^TXM/c\sigma^2 + \mathbf{X}_\zeta{}^T\mathbf{y}_\zeta/\sigma^2\right\} \\ \bar{\Sigma} &= \left\{X^TX/c\sigma^2 + \mathbf{X}_\zeta{}^T\mathbf{X}_\zeta/\sigma^2\right\}^{-1} \\ a &= (n+k+1)/2 \\ b &= (\mathbf{y}_\zeta - \mathbf{X}_\zeta\beta)^T(\mathbf{y}_\zeta - \mathbf{X}_\zeta\beta)/2 + (\beta-M)^T(X^TX)(\beta-M)/2c . \end{aligned}$$

The MCMC implementation of this version of the model then leads to a straightforward Gibbs sampler.

Case 2. The alternative parameterisation of the mixture (6.1) is to consider all regression coefficients as independent between models. This means that, for $j = 1, \ldots, \gamma$, the regression model $\mathfrak{M}_j$ is written as $y = X^j\beta_{\mathfrak{M}_j} + \varepsilon$ and that the $\beta_{\mathfrak{M}_j}$'s are independent. We still assume σ is common to all components. In this representation, we allocate a Zellner's G-prior to each parameter vector,

$$\beta_{\mathfrak{M}_j} \sim \mathcal{N}_{\upsilon_j}(M_j, c\sigma^2(\{X^j\}^TX^j)^{-1})$$

and, conditional on the allocation vector ζ, the full conditional posterior distributions are easily derived:

$$(\alpha_1, \ldots, \alpha_\gamma)|\zeta \sim \mathcal{D}_\gamma(\upsilon_1+a_0, \ldots, \upsilon_\gamma+a_0) , \beta_{\mathfrak{M}_j}|y, \sigma, \zeta \sim \mathcal{N}_{\upsilon_j}(\eta_j, \varphi_j) , \sigma^2|y, \beta \sim \mathcal{IG}(a, b) ,$$

where

$$\begin{aligned}
\eta &= \varphi\left\{\{X^j\}^T X^j M_j/c\sigma^2 + X^j_{i;\zeta_i=j} y_{i;\zeta_i=j}/\sigma^2\right\}\\
\varphi &= \left\{\{X^j\}^T X^j/c\sigma^2 + \{X^j_{i;\zeta_i=j}\}^T X^j_{i;\zeta_i=j}/\sigma^2\right\}^{-1}\\
a &= (n+s)/2\\
b &= \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{\gamma} c(y_{i;\zeta_i=j} - X^j_{i;\zeta_i=j}\beta_{\mathfrak{M}_j})^T(y_{i;\zeta_i=j} - X^j_{i;\zeta_i=j}\beta_{\mathfrak{M}_j})\\
&\quad + c^{-1}(\beta_{\mathfrak{M}_j} - M_j)^T(\{X^j\}^T X^j)(\beta_{\mathfrak{M}_j} - M_j)
\end{aligned}$$

where s is the total number of the regression coefficients of all models under comparison and where the indexing conventions are the same as in Case 1.

The comparison of the performances of the mixture approach in both cases is conducted via simulated data with $k = 3$ covariates, meaning that $(1 \le i \le n)$

$$\mathbb{E}[y_i \mid \beta, X] = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3}\,.$$

This setting thus involves 15 models to be compared (since the model where the mean of the observations is zero is of no interest). The parameters used for the data simulation are $(\beta_0, \beta_1, \beta_2, \beta_3) = (2, -3, 0, 0)$, $\sigma = 1$, with X_1, X_2 and X_3 simulated from $\mathcal{N}(0,1), \mathcal{B}(1,.5)$ and $\mathcal{U}(10,11)$, respectively. We are seeking to identify the true regression model

$$\mathfrak{M}_2 : y_i = 2 - 3X_{i1} + \varepsilon_i\,,$$

by running (Gibbs) mixture estimations algorithms.

Based on a single simulated dataset, Figure 4.15 summarizes the results of those simulations by representing the convergence of the posterior medians of the true model weight in both cases as the sample size n increases. Comments that stem from these results are that

- ✓ all posterior medians of the true model weight α_2 converge to 1 when the sample size increases to $n = 10,000$, which means that the mixture procedure eventually supports $\mathfrak{M}_2$ against the other models;
- ✓ in those graphs, the impact of the prior modeling, i.e., of the value of a_0 is such that the convergence is faster when a_0 is smaller;
- ✓ even for small sample sizes, the posterior medians of α_2 are close to 1;
- ✓ the difference between both mixture parameterisations, i.e., Case 1 and Case 2, are negligible;
- ✓ for almost every sample size and prior hyperparameter, the method concludes that $\mathfrak{M}_2$ is likely to be more appropriate than the others.

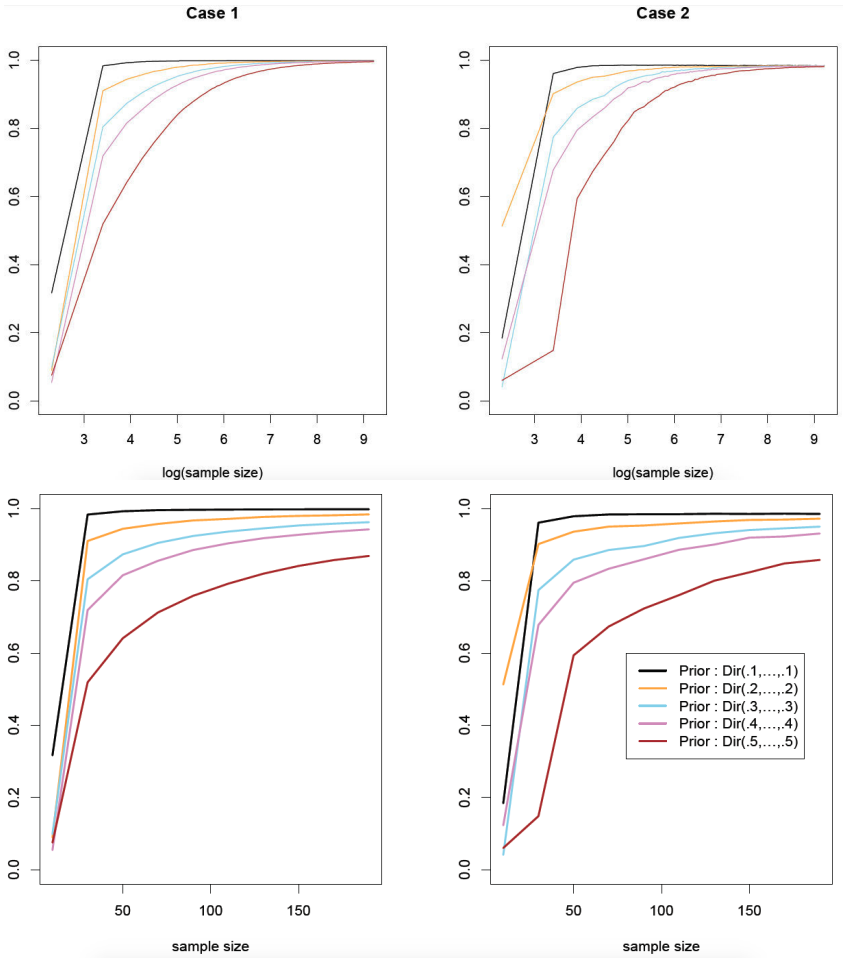


Figure 4.15: **Example 4.3.6:** *(top)* Posterior medians of the true model weight over 5 values of $a_0 = .1, .2, .3, .4, .5$ for sample sizes ranging from 1 to 10^4 and *(bottom)* from 1 to 200. Case 1 *(left)* and Case 2 *(right)* correspond to common and independent parameterisations of the mixture components. Each approximation is based on 10^4 Gibbs iterations.

The most interesting conclusion is therefore that using completely independent parameterisation between the components of the mixture does not induce a strong degradation in the performances of the method, although the convergence to 1 is slightly slower on the right hand side of Figure 4.15. Table 4.3 produces the posterior means of α_2 under different Dirichlet hyperparameters a_0, which shows a stronger difference only for $a_0 = 0.5$, which then appears as a less reliable upper bound.

In order to assess the difference with the classical Bayesian analysis of this model, we compare our posterior means of $1-\alpha_2$ with the posterior probability of $\mathfrak{M}_2$ computed using G-prior for the regression parameters in Figure 4.16. This picture shows that the convergence of $\log(1 - \mathbb{E}(\alpha_2|y, X))$ is faster than for $\log(1 - \mathbb{P}(\mathfrak{M}_2|y, X))$. It also exhibits a difference between Cases 1 and 2 for the larger sample size, with $\log(1 - \mathbb{E}(\alpha_2|y, X))$ concentrated between -6.5 and -5 in Case 1 method, about -4 in Case 2, and about -2 for $\log(1 - \mathbb{P}(\mathfrak{M}_2|y, X))$. Although those pictures are

a_0:	.1	.2	.3	.4	.5
Case 1:					
$\mathbb{E}[\alpha_2\|y, X]$	0.9836	0.9104	0.8043	0.7190	0.5190
Case 2:					
$\mathbb{E}[\alpha_2\|y, X]$	0.9611	0.9018	0.7743	0.6780	0.3905

Table 4.3: **Example 4.3.6:** Posterior means of α_2, weight of model $\mathfrak{M}_2$, when the sample size is n=30

based on a single dataset, they are conclusive about the performances of the mixture approach.

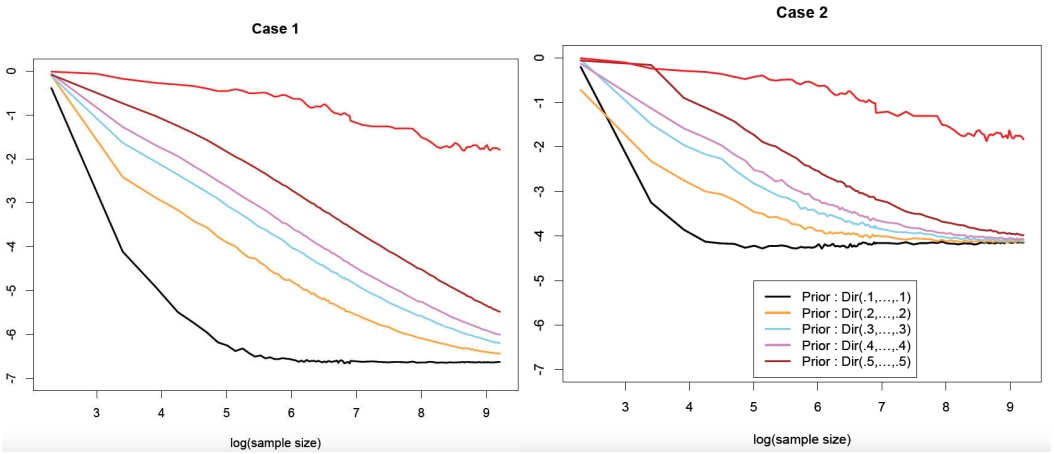


Figure 4.16: **Example 4.3.6:** $\log(1 - \mathbb{E}(\alpha_2|y, X))$ and $\log(1 - \mathbb{P}(\mathfrak{M}_2|y, X))$ *(red lines)* over logarithm of the sample size for $a_0 = .1, .2, .3, .4, .5$. Each posterior approximation is based on 10^4 iterations.

As a second check on the performances of the mixture approach for linear models, for the same set of three regressors, we simulated 50 datasets with 500 observations from each of the models $\mathfrak{M}_1, \ldots, \mathfrak{M}_{15}$ and looked at the respective averages of the Bayes estimates and of the posterior probabilities. In all cases reported in Table 4.4, the posterior means and medians support much more strongly the correct model than the posterior probability, which may sometimes get close to zero.

4.4 Case study : a survival analysis

Survival and reliability models are employed in a large number of disciplines ranging from engineering to health. An important modeling decision in these problems is the choice of the survival function. Among the many parametric alternatives, common choices include the Weibull, log-Normal, logistic, log-logistic, exponential, hypo- and hyper-exponential extensions, Gompertz, Birnhaum-Saunders, Erlang, Coxian, and Pareto distributions. The Weibull distribution is also a representative of the class

True model	$\mathbb{E}(\alpha_j\|y)$			$\text{med}(\alpha_j\|y)$			$\mathbb{P}(\mathfrak{M}_j\|y)$
a_0	.1	.3	.5	.1	.3	.5	
$\mathfrak{M}_1$	.952	.843	.791	1	1	.936	.465
$\mathfrak{M}_2$	.983	.962	.786	.989	.994	.915	.411
$\mathfrak{M}_3$	.976	.973	.821	1	1	.921	.494
$\mathfrak{M}_4$	.991	.867	.902	1	.987	.934	.503
$\mathfrak{M}_5$	.940	.952	.896	.978	.975	.909	.591
$\mathfrak{M}_6$	.974	.939	.898	1	1	.940	.617
$\mathfrak{M}_7$	.973	.899	.906	1	1	1	.888
$\mathfrak{M}_8$	.991	.918	.924	1	1	1	.938
$\mathfrak{M}_9$	.953	.940	.878	1	.993	.956	.505
$\mathfrak{M}_{10}$	.951	.967	.849	.988	.988	.947	.663
$\mathfrak{M}_{11}$	.958	.951	.820	1	.989	.971	.099
$\mathfrak{M}_{12}$	.969	.964	.951	.995	.967	.943	.196
$\mathfrak{M}_{13}$	.919	.951	.872	1	.962	.926	.547
$\mathfrak{M}_{14}$	.952	.964	.890	.998	.981	.911	.126
$\mathfrak{M}_{15}$	.991	.991	.955	1	.994	.908	.164

Table 4.4: **Example 4.3.6:** Comparison between posterior probabilities of the true linear models, posterior means and medians of the mixture model weights, averaged over 50 replicas of samples of size 500.

of models used for extreme value modelling. Other models in this class include the extreme value, Stable, Gumbel and Frechet.

We apply here our testing paradigm to choosing between three potential survival models. Given data $(x_1, ..., x_n)$ with corresponding censoring indicators $(c_1, .., c_n)$, we wish to test the hypothesis that the data are drawn from a log-Normal(ϕ, κ^2), a Weibull(α, λ), or a log-Logistic(γ, δ) distribution. The corresponding mixture is thus given by the density

$$\alpha_1 \exp\{-(\log x - \phi)^2/2\kappa^2\}/\sqrt{2\pi}x\kappa + \alpha_2 \frac{\alpha}{\lambda} \exp\{-(x/\lambda)^\alpha\}((x/\lambda)^{\alpha-1} + \alpha_3(\delta/\gamma)(x/\gamma)^{\delta-1}/(1+(x/\gamma)^\delta)^2$$

where $\alpha_3 = 1 - \alpha_1 - \alpha_2$. A more amenable version can be obtained by working on the scale $Y = -\log(X)$, which then provides a comparison between the N(θ, σ^2), Gumbel(μ, β), and Logistic(ξ, ζ) distributions. This gives rise to the mixture density

$$f_{\theta,\alpha}(y) = \alpha_1 \exp\{-(y-\phi)^2/2\sigma^2\}/(\sqrt{2\pi}\sigma) + \alpha_2/\beta \exp\{-(y-\mu)/\beta)\} \exp\left\{-e^{-(y-\mu)/\beta}\right\} + \alpha_3 \exp\{-(y-\xi)/\zeta\}/\{\zeta(1+\exp\{-(y-\xi)/\zeta\}))^2\}\,.$$

If we opt for a *common* parameterisation of those different models, we have the

following moments matching equations

$$\phi = \mu + \gamma\beta = \xi$$
$$\sigma^2 = \pi^2\beta^2/6 = \zeta^2\pi^2/3$$

where $\gamma \approx 0.5772$ is EulerâĂŞMascheroni constant. As above, this choice allows the use of a noninformative prior on the common location scale parameter, $\pi(\phi, \sigma^2) = 1/\sigma^2$. Once more, we use a Dirichlet prior $\mathcal{D}(a_0, a_0, a_0)$ on $(\alpha_1, \alpha_2, \alpha_3)$. Appendix 2 establishes that the corresponding posterior is proper provided the observations are not all equal.

A common feature in survival data is the presence of censoring. In this case, the mixture equation becomes

$$\begin{aligned} f_{\theta,\alpha}(y, c) = \alpha_1 &\left[e^{-(y-\phi)^2/2\sigma^2}/\sqrt{2\pi}\sigma\right]^c \Phi[(y-\phi)/\sigma]^{1-c} + \\ \alpha_2 &\left[1/\beta e^{-(y-\mu)/\beta} \exp\left\{-e^{-(y-\mu)/\beta}\right\}\right]^c \left[\exp\left\{-e^{-(y-\mu)/\beta}\right\}\right]^{1-c} + \\ \alpha_3 &\left[e^{-(y-\xi)/\zeta}/\left\{\zeta(1+e^{-(y-\xi)/\zeta})^2\right\}\right]^c \left[1/\left\{(1+e^{-(y-\xi)/\zeta})\right\}\right]^{1-c} \end{aligned}$$

Three experiments were performed. First, the performance of the model selection approach in distinguishing between the Weibull, lognormal and log-logistic distributions was assessed by simulating 1000 observations from a Normal$(0, 1)$ density (with no censoring), and testing a Normal versus Gumbel and Logistic distributions as described above. The experiment was then repeated using 1000 simulations from a Gumbel and then from a Logistic distribution. For illustration, the moment-matched Normal, Gumbel and Logistic densities are depicted Figure 4.17 by solid, dashed and dotted lines, respectively.

The Gibbs sampler was run for 10,000 iterations using a prior value of $a_0 = 1.0$ for the hyperparameter on the mixture weights. The resultant probabilities of selecting the various distributions are shown in Figure 4.17, left panel. It can be seen that in all cases, the correct model was overwhelmingly identified. As expected, the probabilities of a correct selection increase with the sample size; in an analogous experiment with $n = 10^5$, all probabilities were larger than 0.90 (figures not shown).

A second experiment was undertaken to assess the influence of the hyperparameter, a_0. The above Monte Carlo experiment was repeated with $n = 1000$ simulated observations, comparing the impact of four values of a_0, namely $a_0 = 0.01, 0.1, 1.0, 10.0$, for all three distributions and for each pair of distributions. As illustrated on Figure 4.18 and 4.19, the probabilities of selecting a (true) Normal or (true) Gumbel model, and in agreement with earlier comparisons, the value of a_0 impacts the probability of a correct model selection, although in all cases covered by this Figure, the correct model was overwhelmingly identified (as the most likely one). Note further that in this experiment the values of a_0 were higher than those we recommender above, namely $a + 0 \leq 0.5$. As before, increasing the sample size from $n = 1,000$ to $n = 10,000$ pushes the posterior probabilities toward the boundaries.

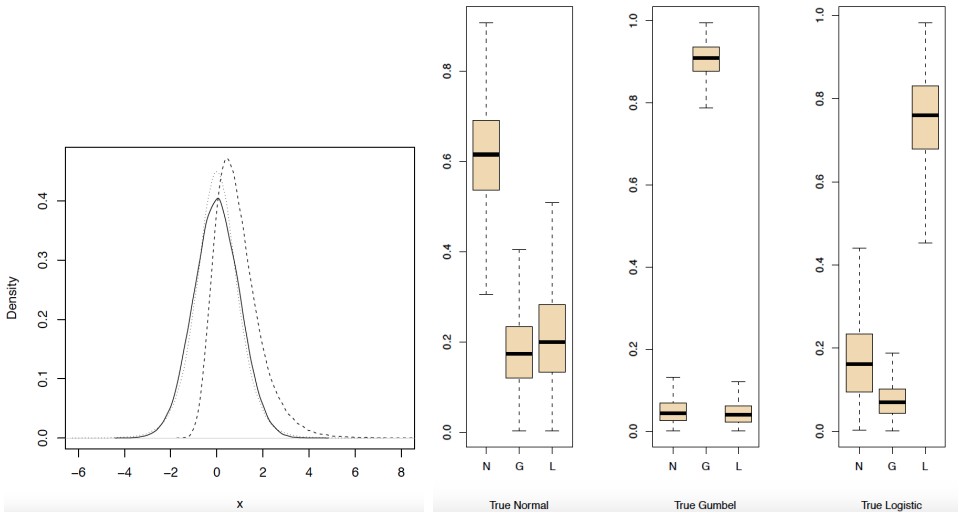


Figure 4.17: **Case study:** *(left)* Normal (solid), Gumbel (dashed) and Logistic (dotted) densities with $(0, 1)$ parameter; *(bottom)* Boxplots of the posterior distributions of the weights under the 3 scenario: truth = Normal *(left panel)*, truth = Gumbel *(middle panel)*, truth = logistic *(right panel)*.

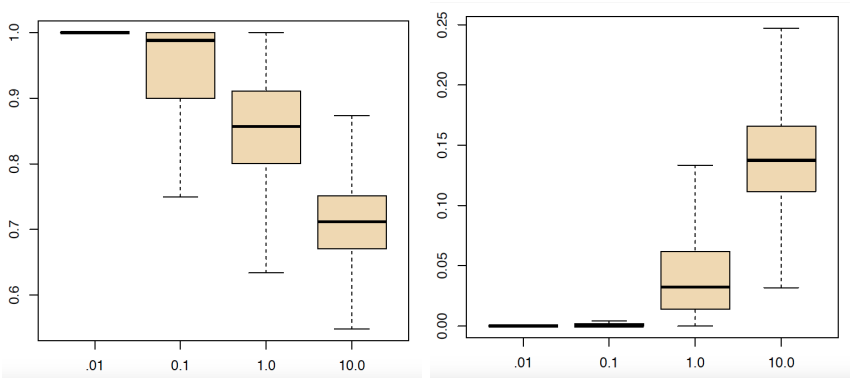


Figure 4.18: **Case study:** Boxplots of the posterior distributions of the Normal weight α_1 under the two scenarii: truth = Normal *(left panel)*, truth = Gumbel *(right panel)*, a_0=0.01, 0.1, 1.0, 10.0 *(from left to right in each panel)* and $n = 1,000$ observations.

In addition to this Monte Carlo evaluation of the mixture approach, we considered a real case study involving modeling survival times for breast cancer in Queensland, Australia. A sample of 25125 individuals with breast cancer was provided by Cancer Council Queensland. Among the subjects, 83.5% were recorded as censored and the remainder ($n = 4155$) were recorded as deaths from any cause. The median survival times were 4.35 and 2.02 years for each of these groups, respectively.

The response variable used in the following analyses is the hazard function, defined as the probability of death at $t+\Delta$ years given survival to t years, adjusted for age, sex and the expected mortality rate, that is, the age- and sex-adjusted background population risk of death. Of interest is whether or not this distribution is

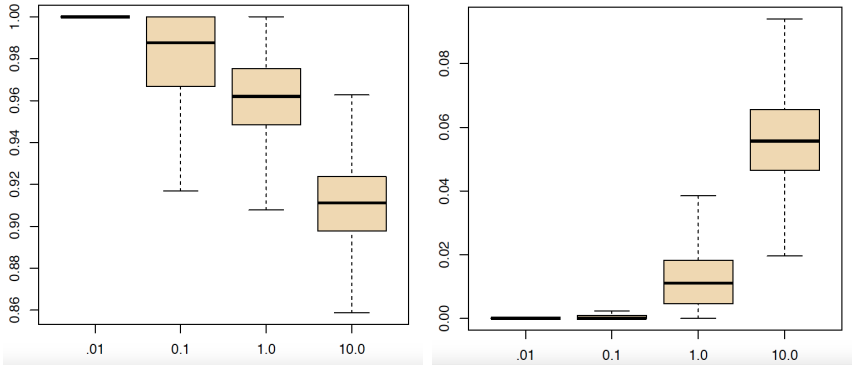


Figure 4.19: **Case study:** Boxplots of the posterior distributions of the Normal weight α_1 under the two scenarii: truth = Normal *(left panel)*, truth = Gumbel *(right panel)*, a_0=0.01, 0.1, 1.0, 10.0 *(from left to right in each panel)* and $n = 10,000$ simulated observations.

best fitted by a log Normal, Weibull, or log-Logistic distribution, or, equivalently, whether or not the log hazard is best fitted by a Normal, Gumbel, or Logistic distribution. Alternatively, it may be preferable to follow a model averaged approach, which is intrinsically part of the mixture model approach since the MCMC outcome provides in addition a posterior approximation of the overall mixture. (It could actually be argued that this approach is even better than standard model averaging as each observation in the sample selects the best fitted component of the mixture.) The choice of an appropriate model or of a combination of models is important for the prediction of survival for cancer patients, which then impacts on decisions about personalized management and treatment options.

Figure 4.20 provides histograms of both hazard and log hazard for all data and for deaths only (i.e., excluding censored observations). The corresponding q-q plots associated with fitting the three distributions, ignoring censoring, are also shown. The result of the modeling is that these distributions have different fit characteristics: whereas the Normal (and hence the log Normal) distribution fits the centre of the distribution more closely, the Weibull (and hence the Gumbel) distribution captures the tail behavior more accurately. The logistic distribution appears to have a similar fit to the Normal, but it accommodates slightly more diffuse tails. Based on a choice of hyperparameter $a_0 = 1.0$, the mixture test for the breast cancer data resulted in the choice of the logistic distribution with probability 0.996 (s.d. $1.4\,10^{-3}$), with the remaining probability mass almost equally split between the Normal and Gumbel distributions.

4.5 Asymptotic consistency

In this section we prove posterior consistency for our mixture testing procedure. More precisely we study the asymptotic behavior of the posterior distribution of α. We consider two different cases. In the first case, the two models, $\mathfrak{M}_1$ and $\mathfrak{M}_2$, are well separated while, in the second case, model $\mathfrak{M}_1$ is a submodel of $\mathfrak{M}_2$. We denote

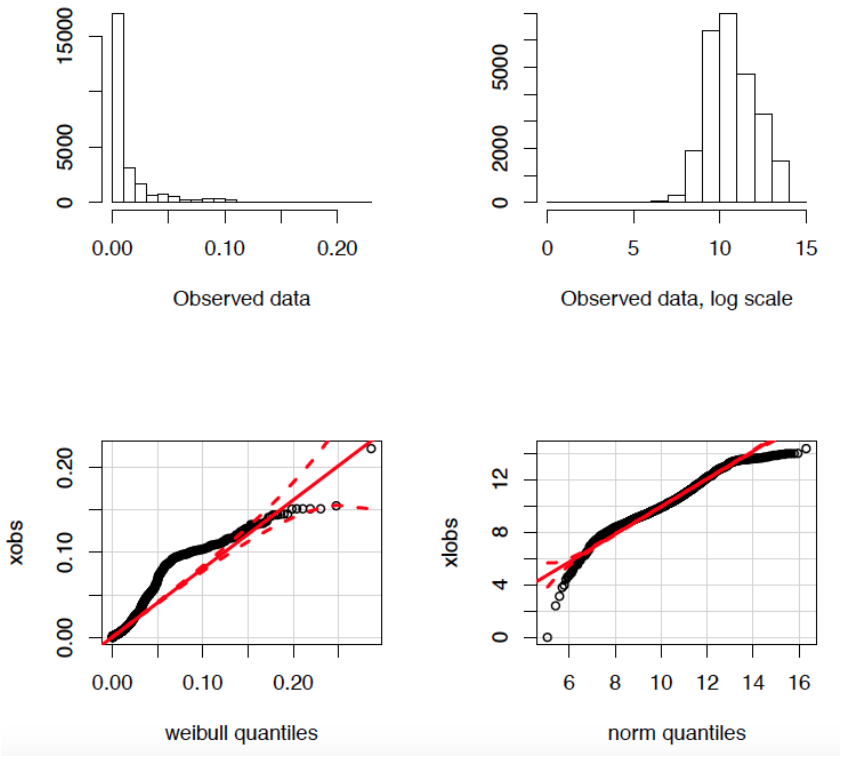


Figure 4.20: **Case study:** *(top panel)* Histograms of the hazard *(left)* and log hazard *(right)* for the non-censored data; *(bottom panel)* q-q plots for fitting the three distributions

by π the prior distribution on $(\alpha, \theta_1, \theta_2)$ and assume that $\theta_j \in \Theta_j \subset \mathbb{R}^{d_j}$. We first prove that, under weak regularity conditions on each model, we can obtain posterior concentration rates for the marginal density $f_{\theta,\alpha}(\cdot) = \alpha f_{1,\theta_1}(\cdot) + (1-\alpha) f_{2,\theta_2}(\cdot)$. Let $\mathbf{x}^n = (x_1, \cdots, x_n)$ a n sample with true density f^*.

Proposition 1 *Assume that, for all $C_1 > 0$, there exist Θ_n a subset of $\Theta_1 \times \Theta_2$ and $B > 0$ such that*

$$\pi\left[\Theta_n^c\right] \leq n^{-C_1}, \quad \Theta_n \subset \{\|\theta_1\| + \|\theta_2\| \leq n^B\} \tag{4.9}$$

and that there exist $H \geq 0$ and $L, \delta > 0$ such that, for $j = 1, 2$,

$$\begin{aligned} &\sup_{\theta,\theta' \in \Theta_n} \|f_{j,\theta_j} - f_{j,\theta_j'}\|_1 \leq L n^H \|\theta_j - \theta_j'\|, \quad \theta = (\theta_1, \theta_2),\ \theta' = (\theta_1', \theta_2')\,, \\ &\forall \|\theta_j - \theta_j^*\| \leq \delta; \quad KL(f_{j,\theta_j}, f_{j,\theta_j^*}) \lesssim \|\theta_j - \theta_j^*\|\,. \end{aligned} \tag{4.10}$$

We then have that, when $f^* = f_{\theta^*,\alpha^*}$, with $\alpha^* \in [0,1]$, there exists $M > 0$ such that

$$\pi\left[(\alpha,\theta); \|f_{\theta,\alpha} - f^*\|_1 > M\sqrt{\log n/n} | \mathbf{x}^n\right] = o_p(1)\,.$$

The proof of Proposition 1 is a direct consequence of Theorem 2.1 of [Ghosal 2000] and is omitted for the sake of conciseness. Condition (4.10) is a weak regularity condition on each of the candidate models. Combined with condition (4.9) it allows to consider noncompact parameter sets in the usual way, see for instance [Ghosal 2000]. It is satisfied in all examples considered in Section 4.3. We build on Proposition 1 to describe the asymptotic behavior of the posterior distribution on the parameters.

4.5.1 The case of separated models

Assume that both models are separated in the sense that there is identifiability:

$$\forall \alpha, \alpha' \in [0,1], \quad \forall \theta_j, \theta_j', \, j = 1,2 \quad P_{\theta,\alpha} = P_{\theta',\alpha'} \quad \Rightarrow \alpha = \alpha', \quad \theta = \theta', \tag{4.11}$$

where $P_{\theta,\alpha}$ denotes the distribution associated with $f_{\theta,\alpha}$. We assume that (4.11) also holds on the boundary of $\Theta_1 \times \Theta_2$. In other words, the following

$$\inf_{\theta_1 \in \Theta_1} \inf_{\theta_2 \in \Theta_2} \|f_{1,\theta_1} - f_{2,\theta_2}\|_1 > 0$$

holds. We also assume that, for all $\theta_j^* \in \Theta_j$, $j = 1,2$, if P_{θ_j} converges in the weak topology to $P_{\theta_j^*}$, then θ_j converges in the Euclidean topology to θ_j^*. The following result then holds:

Theorem 1 *Assume that (4.11) is satisfied, together with (4.9) and (4.10), then for all $\varepsilon > 0$*

$$\pi\left[|\alpha - \alpha^*| > \varepsilon | \mathbf{x}^n\right] = o_p(1).$$

In addition, assume that the mapping $\theta_j \to f_{j,\theta_j}$ is twice continuously differentiable in a neighborhood of θ_j^*, $j = 1,2$, and that

$$f_{1,\theta_1^*} - f_{2,\theta_2^*}, \nabla f_{1,\theta_1^*}, \nabla f_{2,\theta_2^*}$$

are linearly independent as functions of y and that there exists $\delta > 0$ such that

$$\nabla f_{1,\theta_1^*}, \, \nabla f_{2,\theta_2^*}, \, \sup_{|\theta_1 - \theta_1^*| < \delta} |D^2 f_{1,\theta_1}|, \, \sup_{|\theta_2 - \theta_2^*| < \delta} |D^2 f_{2,\theta_2}| \in L_1 \, .$$

Then

$$\pi\left[|\alpha - \alpha^*| > M\sqrt{\log n / n} | \mathbf{x}^n\right] = o_p(1). \tag{4.12}$$

Theorem 1 allows for the interpretation of the quantity α under the posterior distribution. In particular, if the data $\mathbf{x}^n$ is generated from model $\mathfrak{M}_1$ (resp. $\mathfrak{M}_2$), then the posterior distribution on α concentrates around $\alpha = 1$ (resp. around $\alpha = 0$), which establishes the consistency of our mixture approach.

Proof: Using Proposition 1, we have that

$$\pi\left(A_n|\mathbf{x}^n\right) = 1 + o_p(1)$$

with $A_n = \{(\alpha,\theta); \|f_{\theta,\alpha} - f_{\theta^*,\alpha^*}\|_1 \leq \delta_n\}$ and $\delta_n = M\sqrt{\log n/n}$. Consider a subsequence $\alpha_n, P_{1,\theta_{1n}}, P_{2,\theta_{2n}}$ which converges to α, μ_1, μ_2 where convergence holds in the sense that $\alpha_n \to \alpha$ and $P_{j,\theta_{jn}}$ converges weakly to μ_j. Note that $\mu_j(\mathcal{X}) \leq 1$ by precompacity of the unit ball under the weak topology. At the limit

$$\alpha\mu_1 + (1-\alpha)\mu_2 = \alpha^* P_{1,\theta_1^*} + (1-\alpha^*)P_{2,\theta_2^*}$$

The above equality implies that μ_1 and μ_2 are probabilities. Using (4.11), we obtain that

$$\alpha = \alpha^*, \quad \mu_j = P_{j,\theta_j^*},$$

which implies posterior consistency for α. The proof of (4.12) follows the same line as in [Rousseau 2011]. Consider first the case where $\alpha^* \in (0,1)$. Then the posterior distribution on θ concentrates around θ^*.

Writing

$$L' = (f_{1,\theta_1^*} - f_{2,\theta_2^*}, \alpha^*\nabla f_{1,\theta_1^*}, (1-\alpha^*)\nabla f_{2,\theta_2^*}) := (L_\alpha, L_1, L_2)$$

$$L^{"} = \mathrm{diag}(0, \alpha^* D^2 f_{1,\theta_1^*}, (1-\alpha^*)D^2 f_{2,\theta_2^*}) \quad \text{and} \quad \eta = (\alpha-\alpha^*, \theta_1-\theta_1^*, \theta_2-\theta_2^*), \quad \omega = \eta/|\eta|\,,$$

we then have

$$\|f_{\theta,\alpha} - f_{\theta^*,\alpha^*}\|_1 = |\eta|\left|\omega^{\mathrm{T}}L' + |\eta|/2\omega^{\mathrm{T}}L^{"}\omega + |\eta|\omega_1\left[\omega_2 L_2^{'} + \omega_3 L_3^{'}\right] + o(|\eta|)\right| \quad (4.13)$$

For all $(\alpha,\theta) \in A_n$, set $\eta = (\alpha-\alpha^*, \theta_1-\theta_1^*, \theta_2-\theta_2^*)$ goes to 0 and for n large enough there exists $\varepsilon > 0$ such that $|\alpha-\alpha^*| + |\theta-\theta^*| \leq \varepsilon$. We now prove that there exists $c > 0$ such that for all $(\alpha,\theta) \in A_n$

$$v(\omega) = \left|\omega^{\mathrm{T}}L' + \frac{|\eta|}{2}\omega^{\mathrm{T}}L^{"}\omega + |\eta|\omega_1\left[\omega_2^{\mathrm{T}}L_2^{'} + \omega_3^{\mathrm{T}}L_3^{'}\right] + o(|\eta|)\right| > c,$$

where ω is defined with respect to α,θ. Were it not the case, there would exist a sequence $(\alpha_n,\theta_n) \in A_n$ such that the associated $v(\omega_n) \leq c_n$ with $c_n = o(1)$. As ω_n belongs to a compact set we could find a subsequence converging to a point $\bar{\omega}$. At the limit we would obtain

$$\bar{\omega}^{\mathrm{T}}L^{'} = 0$$

and by linear independence $\bar{\omega} = 0$ which is not possible. Thus for all $(\alpha,\theta) \in A_n$

$$|\alpha-\alpha^*| + |\theta-\theta^*| \lesssim \delta_n.$$

Assume now instead that $\alpha^* = 0$. Then define $L^{'} = (L_\alpha, L_2)$ and

$$L^{"} = \mathrm{diag}(0, D^2 f_{2,\theta_2^*}) \quad \text{and} \quad \eta = (\alpha-\alpha^*, \theta_2-\theta_2^*), \quad \omega = \eta/|\eta|$$

and consider a Taylor expansion with θ_1 fixed, $\theta_1^* = \theta_1$ and $|\eta|$ going to 0. This leads to

$$\|f_{\theta,\alpha} - f_{\alpha^*,\theta^*}\|_1 = |\eta| \left| \omega^{\mathrm{T}} L' + \frac{|\eta|}{2} \omega^{\mathrm{T}} L" \omega + |\eta| \omega_1 \omega_3 L_3' \right| + o(|\eta|) \tag{4.14}$$

in place of (4.13) and the posterior concentration rate δ_n is obtained in the same way. □

We now consider the embedded case.

4.5.2 Embedded case

In this section we assume that $\mathfrak{M}_1$ is a submodel of $\mathfrak{M}_2$, in the sense that $\theta_2 = (\theta_1, \psi)$ with $\psi \in \mathcal{S} \subset \mathbb{R}^d$ and that $f_{2,\theta_2} \in \mathfrak{M}_1$ when $\theta_2 = (\theta_1, \psi_0)$ for some given value ψ_0, say $\psi_0 = 0$. Condition (4.11) is no longer verified for all α's: we assume however that it is verified for all $\alpha, \alpha^* \in (0, 1]$ and that $\theta_2^* = (\theta_1^*, \psi^*)$ satisfies $\psi^* \neq 0$. In this case, under the same conditions as in Theorem 1, we immediately obtain the posterior concentration rate $\sqrt{\log n/n}$ for estimating α when $\alpha^* \in (0, 1)$ and $\psi^* \neq 0$. We now treat the case where $\psi^* = 0$; in other words, f^* is in model $\mathfrak{M}_1$.

As in [Rousseau 2011], we consider both possible paths to approximate f^*: either α goes to 1 or ψ goes to $\psi_0 = 0$. In the first case, called path 1, $(\alpha^*, \theta^*) = (1, \theta_1^*, \theta_1^*, \psi)$ with $\psi \in \mathcal{S}$, in the second, called path 2, $(\alpha^*, \theta^*) = (\alpha, \theta_1^*, \theta_1^*, 0)$ with $\alpha \in [0, 1]$. In either case, we write P^* the distribution. We also denote $F^* g = \int f^*(x) g(x) d\mu(x)$ for any integrable function g. For sparsity reasons, we consider the following structure for the prior on (α, θ):

$$\pi(\alpha, \theta) = \pi_\alpha(\alpha) \pi_1(\theta_1) \pi_\psi(\psi), \quad \theta_2 = (\theta_1, \psi).$$

This means that the parameter θ_1 is common to both models, i.e., that θ_2 shares the parameter θ_1 with f_{1,θ_1}.

Condition (4.11) is replaced by

$$P_{\theta,\alpha} = P^* \quad \Rightarrow \alpha = 1, \quad \theta_1 = \theta_1^*, \quad \theta_2 = (\theta_1^*, \psi) \quad \text{or} \quad \alpha \leq 1, \quad \theta_1 = \theta_1^*, \quad \theta_2 = (\theta_1^*, 0) \tag{4.15}$$

Let Θ^* the above parameter set.

As in the case of separated models, the posterior distribution concentrates on Θ^*. We now describe more precisely the asymptotic behavior of the posterior distribution, using [Rousseau 2011]. We cannot apply directly Theorem 1 of [Rousseau 2011], hence the following result is an adaptation of it. We require the following assumptions with $f^* = f_{1,\theta_1^*}$. For the sake of simplicity, we assume that Θ_1 and $\mathcal{S}$ are compact. Extension to non compact sets can be handled similarly to [Rousseau 2011].

B1 *Regularity*: Assume that $\theta_1 \to f_{1,\theta_1}$ and $\theta_2 \to f_{2,\theta_2}$ are 3 times continuously differentiable and that

$$F^* \left(\frac{\bar{f}_{1,\theta_1^*}^3}{\underline{f}_{1,\theta_1^*}^3} \right) < +\infty, \quad \bar{f}_{1,\theta_1^*} = \sup_{|\theta_1 - \theta_1^*| < \delta} f_{1,\theta_1}, \quad \underline{f}_{1,\theta_1^*} = \inf_{|\theta_1 - \theta_1^*| < \delta} f_{1,\theta_1}$$

$$F^*\left(\frac{\sup_{|\theta_1-\theta_1^*|<\delta}|\nabla f_{1,\theta_1^*}|^3}{\underline{f}^3_{1,\theta_1^*}}\right)<+\infty,\quad F^*\left(\frac{|\nabla f_{1,\theta_1^*}|^4}{f^4_{1,\theta_1^*}}\right)<+\infty,$$
$$F^*\left(\frac{\sup_{|\theta_1-\theta_1^*|<\delta}|D^2 f_{1,\theta_1^*}|^2}{\underline{f}^2_{1,\theta_1^*}}\right)<+\infty,\quad F^*\left(\frac{\sup_{|\theta_1-\theta_1^*|<\delta}|D^3 f_{1,\theta_1^*}|}{\underline{f}_{1,\theta_1^*}}\right)<+\infty$$

B2 *Integrability*: There exists $\mathcal{S}_0 \subset \mathcal{S}\cap\{|\psi| > \delta_0\}$, for some positive δ_0 and satisfying $\mathrm{Leb}(\mathcal{S}_0)>0$, and such that for all $\psi\in\mathcal{S}_0$,

$$F^*\left(\frac{\sup_{|\theta_1-\theta_1^*|<\delta} f_{2,\theta_1,\psi}}{f^4_{1,\theta_1^*}}\right)<+\infty,\quad F^*\left(\frac{\sup_{|\theta_1-\theta_1^*|<\delta} f^3_{2,\theta_1,\psi}}{\underline{f}^3_{1,\theta 1^*}}\right)<+\infty,$$

B3 *Stronger identifiability* : Set

$$\nabla f_{2,\theta_1^*,\psi^*}(x)=\left(\nabla_{\theta_1} f_{2,\theta_1^*,\psi^*}(x)^{\mathrm{T}},\nabla_{\psi} f_{2,\theta_1^*,\psi^*}(x)^{\mathrm{T}}\right)^{\mathrm{T}}.$$

Then for all $\psi\in\mathcal{S}$ with $\psi\neq 0$, if $\eta_0\in\mathbb{R}$, $\eta_1\in\mathbb{R}^{d_1}$

$$\eta_0(f_{1,\theta_1^*}-f_{2,\theta_1^*,\psi})+\eta_1^{\mathrm{T}}[\nabla_{\theta_1} f_{1,\theta_1^*}-\nabla_{\theta_1} f_{2,\theta_1^*,\psi}(x)]=0 \quad \Leftrightarrow \eta_1=0,\,\eta_2=0 \tag{4.16}$$

We can now state the main theorem:

Theorem 2 *Given the model*

$$f_{\theta_1,\psi,\alpha}=\alpha f_{1,\theta_1}+(1-\alpha)f_{2,\theta_1,\psi},$$

assume that the data is made of the n sample $\mathbf{x}^n=(x_1,\cdots,x_n)$ *issued from* f_{1,θ_1^*} *for some* $\theta_1^*\in\Theta_1$, *that assumptions* $B1-B3$ *are satisfied, and that there exists* $M>0$ *such that*

$$\pi\left[(\alpha,\theta);\|f_{\theta,\alpha}-f^*\|_1>M\sqrt{\log n/n}|\mathbf{x}^n\right]=o_p(1).$$

If the prior π_α *on* α *is a Beta* $\mathcal{B}(a_1,a_2)$ *distribution, with* $a_2<d_2$, *and if the prior* $\pi_{\theta_1,\psi}$ *is absolutely continuous with positive and continuous density at* $(\theta_1^*,0)$, *then for all* M_n *going to infinity,*

$$\pi\left[|\alpha-\alpha^*|>M_n(\log n)^\gamma/\sqrt{n}|\mathbf{x}^n\right]=o_p(1),\quad \gamma=\max((d_1+a_2)/(d_2-a_2),1)/2, \tag{4.17}$$

Proof: We must find a precise lower bound on

$$D_n:=\int_\alpha\int_\Theta e^{l_n(f_{\theta,\alpha})-l_n(f^*)}\mathrm{d}\pi_\theta(\theta)\mathrm{d}\pi_\alpha(\alpha)$$

Consider the approximating set

$$S_n(\varepsilon)=\{(\theta,\alpha),\alpha>1-1/\sqrt{n},|\theta_1-\theta_1^*|\leq 1/\sqrt{n},|\psi-\bar{\psi}|\leq\varepsilon\}$$

with $|\bar{\psi}| > 2\varepsilon$ some fixed parameter in $\mathcal{S}$. Using the same computations as in [Rousseau 2011], it holds that for all $\delta > 0$ there exists $C_\delta > 0$ such that

$$P^*\left(D_n < e^{-C_\delta}\pi(S_n(\varepsilon))/2\right) < \delta. \tag{4.18}$$

So that with probability greater than $1-\delta$, $D_n \gtrsim n^{-(b+d_1)/2}$. Denote $B_n = \{(\theta,\alpha); \|f_{\theta,\alpha} - f^*\|_1 \leq M\sqrt{\log n/n}\}$ and $A_n = \{(\theta,\alpha) \in B_n; 1-\alpha > z_n/\sqrt{n}\}$ with $z_n = M_n(\log n)^\gamma/\sqrt{n}$ and M_n a sequence increasing to infinity. We split B_n into

$$B_{n,1}(\varepsilon) = B_n \cap \{(\theta,\alpha), \theta = (\theta_1,\psi); |\psi| < \varepsilon\}, \quad B_{n,2}(\varepsilon) = B_n \cap B_{n,1}(\varepsilon)^c.$$

To prove Theorem 2 it is enough to verify that

$$\pi(A_n) = o(n^{-(a_2+d_1)/2}).$$

To simplify notations we also write $\delta_n = M\sqrt{\log n/n}$. First we prove that for all $\varepsilon > 0$, $A_n \cap B_{n,2}(\varepsilon) = \emptyset$, when n is large enough. Let $\varepsilon > 0$, then for any $(\theta,\alpha) \in A_n \cap B_{n,2}(\varepsilon)$, We thus have $|\psi| \neq o(1)$, $\alpha = 1 + o(1)$ and $|\theta_1 - \theta_1^*| = o(1)$. Consider a Taylor expansion of $f_{\theta,\alpha}$ around $\alpha = 1$ and $\theta_1 = \theta_1^*$, with ψ fixed. This leads to

$$\begin{aligned}
f_{\theta,\alpha} - f^* &= (\alpha-1)[f_{1,\theta_1^*} - f_{2,\theta_1^*,\psi}] + (\theta_1 - \theta_1^*)[\nabla_{\theta_1} f_{1,\theta_1^*} - \nabla_{\theta_1} f_{2,\theta_1^*,\psi}(x)] \\
&\quad + \frac{1}{2}(\theta_1 - \theta_1^*)^{\mathrm{T}}\left(\bar{\alpha} D^2_{\theta_1} f_{1,\bar{\theta}_1} + (1-\bar{\alpha}) D^2_{\theta_1} f_{2,\bar{\theta}_1,\psi}\right)(\theta_1 - \theta_1^*) \\
&\quad + (\alpha-1)(\theta_1 - \theta_1^*)^{\mathrm{T}}[\nabla_{\theta_1} f_{1,\bar{\theta}_1} - \nabla_{\theta_1} f_{2,\bar{\theta}_1,\psi}] \\
&= (\alpha-1)[f_{1,\theta_1^*} - f_{2,\theta_1^*,\psi}] + (\theta_1 - \theta_1^*)\nabla_{\theta_1} f_{1,\theta_1^*} + o(|\alpha-1| + |\theta_1 - \theta_1^*|)
\end{aligned}$$

with $\bar{\alpha} \in (0,1)$ and $\bar{\theta}_1 \in (\theta_1, \theta_1^*)$ and the $o(1)$ is uniform over $A_n \cap B_{n,2}(\varepsilon)$. Set $\eta = (\alpha - 1, \theta_1 - \theta_1^*)$ and $x = \eta/|\eta|$ if $|\eta| > 0$. Then

$$\|f_{\theta,\alpha} - f^*\|_1 = |\eta|\left(x^{\mathrm{T}} L_1(\psi) + o(1)\right), \quad L_1 = (f_{1,\theta_1^*} - f_{2,\theta_1^*,\psi}, \nabla_{\theta_1} f_{1,\theta_1^*})$$

We now prove that on $A_n \cap B_{n,2}(\varepsilon)$, $\|f_{\theta,\alpha} - f^*\|_1 \gtrsim |\eta|$. Assume that it is not the case then there exist $c_n > 0$ going to 0 and a sequence (θ_n, α_n) such that along that subsequence $|x_n^{\mathrm{T}} L_1(\psi_n) + o(1)| \leq c_n$ with $x_n = \eta_n/|\eta_n|$. Since it belongs to a compact, together with ψ_n, any converging subsequence satisfies at the limit $(\bar{x}, \bar{\psi})$,

$$\bar{x}^{\mathrm{T}} L_1(\bar{\psi}) = 0\,,$$

which is not possible. Hence $|\alpha - 1| \lesssim M\sqrt{\log n}/\sqrt{n} = o(M_n(\log n)^\gamma/\sqrt{n})$, which is not possible so that $A_n \cap B_{n,2}(\varepsilon) = \emptyset$ when n is large enough. We now bound $\pi(A_n \cap B_{n,1}(\varepsilon))$ for $\varepsilon > 0$ small enough but fixed. We consider a Taylor expansion around $\theta^* = (\theta_1^*, 0)$, leaving α fixed. Note that $\nabla_{\theta_1} f_{2,\theta^*} = \nabla_{\theta_1} f_{1,\theta_1^*}$. We have

$$f_{\theta,\alpha} - f^* = (\theta_1 - \theta_1^*)^{\mathrm{T}} \nabla_{\theta_1} f_{2,\theta^*} + (1-\alpha)\psi^{\mathrm{T}} \nabla_\psi f_{2,\theta^*} \frac{1}{2}(\theta - \theta^*)^{\mathrm{T}} H_{\alpha,\bar{\theta}}(\theta - \theta^*)$$

where $H_{\alpha,\bar{\theta}}$ is the bloc matrix

$$H_{\alpha,\bar{\theta}} = \begin{pmatrix} \alpha D^2_{\theta_1} f_{1,\bar{\theta}_1} + (1-\alpha) D^2_{\theta_1,\theta_1} f_{2,\bar{\theta}} & (1-\alpha) D^2_{\theta_1,\psi} f_{2,\bar{\theta}} \\ (1-\alpha) D^2_{\psi,\theta_1} f_{2,\bar{\theta}} & (1-\alpha) D^2_{\psi,\psi} f_{2,\bar{\theta}} \end{pmatrix}$$

Since $H_{\alpha,\bar{\theta}}$ is bounded in L_1 (in the sense that each of its components is bounded as functions in L_1), uniformly in neighborhoods of θ^*, we have writing $\eta = (\theta_1 - \theta_1^*, (1-\alpha)\psi)$ and $x = \eta/|\eta|$, that $|\eta| = o(1)$ on $A_n \cap B_{n,1}(\varepsilon)$ and

$$\|f_{\theta,\alpha} - f^*\|_1 \gtrsim |\eta| \left(x^{\mathrm{T}} \nabla f_{2,\theta^*} + o(1) \right),$$

if ε is small enough. Using a similar argument to before, this leads to $|\eta| \lesssim \delta_n$ on $A_n \cap B_{n,1}(\varepsilon)$, so that

$$\pi\left(A_n \cap B_{n,1}(\varepsilon)\right) \lesssim \delta_n^{d_1} \int_{z_n/\sqrt{n}}^{1} (\delta_n/u)^{d_2} u^{b-1} du \lesssim \delta_n^{d_1+b} z_n^{b-d_2} \lesssim n^{-(d_1+a_2)/2} M_n^{a_2-d_2},$$

which terminates the proof. □

4.6 Conclusion

Bayesian inference has been used in a very wide range over the past twenty years, mostly thanks to enhanced computing abilities, and many of those applications of the Bayesian paradigm have concentrated on the comparison of scientific theories and on testing of null hypotheses. Due to the ever increasing complexity of the statistical models handled in such applications, the natural and understandable tendency of practitioners has been to rely on the default solution of the posterior probability (or equivalently of the Bayes factor) without ever questioning its validity. It is only in rare cases that warnings were heeded [Robert 2011] about the poorly understood sensitivity of such tools to both prior modeling and posterior calibration. In this area, objective Bayes solutions remain tentative and do not meet with consensus.

We thus believe Bayesian analysis has reached the time for a paradigm shift in the matter of hypothesis testing and model selection, albeit the solution does not have to be found outside the Bayesian paradigm, as for instance the frequentist priors of [Johnson 2013b, Johnson 2013a] and the integrated likelihood setting of [Aitkin 2010]. The novel paradigm we proposed here for Bayesian testing of hypotheses and Bayesian model comparison offers many incentives while answering some of the classical attacks against posterior probabilities and Bayes factors. Our alternative to the construction of traditional posterior probabilities that a given hypothesis is true or that the data originates from a specific model is therefore to rely on the encompassing mixture model. Not only do we replace the original testing problem with a better controlled estimation target that focus on the frequency of a given model within the mixture model, but we also allow for posterior variability over this frequency as opposed to the deterministic characteristics of the standard

Bayesian approach. The posterior distribution on the weights of both components in the mixture offers a setting for deciding about which model is most favored by the data that is at least as intuitive as the sole number corresponding to either the posterior probability or the Bayes factor. The range of acceptance, rejection and indecision conclusions can easily be calibrated by simulation under both models, as well as by deciding on the values of the weights that are extreme enough in favor of one model. The examples provided in this paper have showed that the posterior medians of such weights are very quickly settling near the boundary values of 0 and 1, depending on which model is right. Even though we do not advocate such practice, it is even possible to derive a Bayesian p-value by looking at the posterior area under the tail of the distribution of the weight.

Besides decision making, another issue of potential concern about this new approach is the impact of the prior modeling. We demonstrated through all our examples that a partly common parameterisation is always feasible and hence allows for reference priors, at least on the common parameters. This proposal thus allows for a removal of the absolute prohibition of using improper priors in hypothesis testing [DeGroot 1973], a problem which has plagued the objective Bayes literature for decades. Concerning the prior on the weight parameter, we analyzed the sensitivity on the resulting posterior distribution of various prior Beta modelings on those weights. While the sensitivity is clearly present, it naturally vanishes as the sample size increases, in agreement with our consistency results, and remains of a moderate magnitude, which leads us to suggest the default value of $a_0 = 0.5$ in the Beta prior, in connection with both the earlier result of [Rousseau 2011] and Jeffreys' prior in the simplest mixture setting.

A last point about our proposal is that it does not induce additional computational strain on the analysis. Provided algorithmic solutions exist for both models under comparison, such solutions can be recycled towards estimating the encompassing mixture model. As demonstrated through the various examples in the paper, the setting is actually easier than with a standard mixture estimation problem [Diebolt 1994, Marin 2005] because of the existence of common parameters that allow for the original MCMC samplers to be turned into proposals. Gibbs sampling completions are useful for assessing the potential outliers in a model but altogether not essential to achieve a conclusion about the overall problem.

Chapter 5

Supplementary material: Testing hypotheses as a mixture estimation model

In Chapter 4, we expressed how Bayesian model choice via posterior probabilities of models can be replaced by an estimation based on the probability weight of a model within a mixture model and the reasonable performance of this transformation illustrated by several examples. This chapter deals with some more Bayesian inferences, MCMC algorithms, the behavior of resulting Markov chains and also statistical tools in more details.

5.1 Mixture weight distribution

Bayesian inference of the mixture model weights is based on a beta distribution as a conjugate prior probability distribution with shape parameters that take values $a_0 = .1, .2, .3, .4, .5$ while $a_0 = .5$ yields Jeffreys prior. The identical shape parameters lead a symmetric density function about .5 that looks like a basin in a one-dimensional curve and tends to infinity in the boundaries of unit interval as shown in Figure 5.1. If n_1 denotes the number of observations associated with the mixture component $\mathfrak{M}_1$, the evolution of the posterior probability over n_1 based on the Jeffreys prior, which is $\mathcal{B}eta(.5 + n_1, .5 + n_2)$, is shown on the right side of Figure 5.1. This evolution implies that the smaller the value of n_1 is the more the posterior density of $\mathfrak{M}_1$ tightens up near zero. This means that only in the case where a very large number of the observations is allocated to the model $\mathfrak{M}_1$, the Bayesian estimate of corresponding mixture weight is very close to 1. This behavior plays a fundamental role in replacing the posterior probabilities of the models by the posterior estimations of the mixture model weights.

5.2 Poisson versus geometric

In the first example of Chapter 4, Poisson distribution is compared with Geometric distribution under the assumption of using the same parameter in both models. This allows us to consider a non-informative prior for the common parameter λ. Here, we firstly study this case with more details in section 5.2.1 and then in section 5.2.2, we proceed with comparing these distributions when they have deferent parameters in

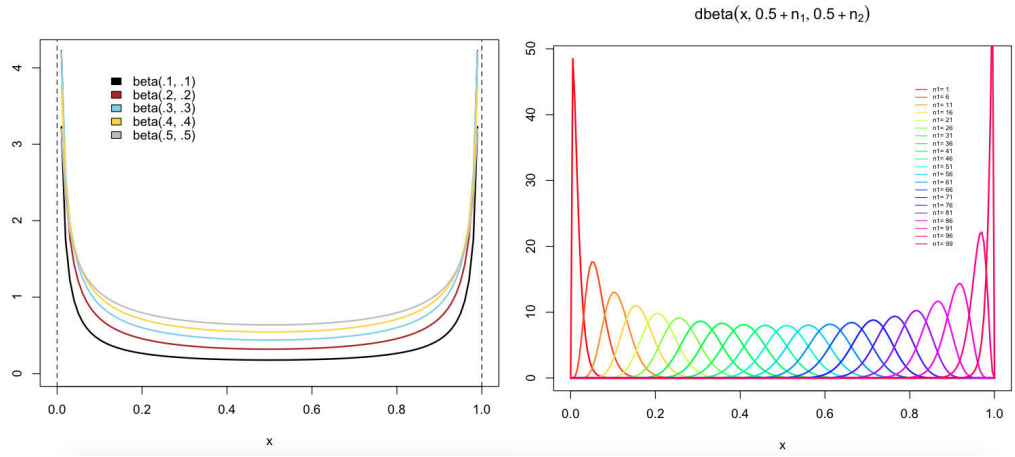


Figure 5.1: **Mixture weights distribution** 5.1: *(Left)* Probability density function of beta prior for the weight of the mixture model. *(right)* Posterior probability of the mixture weights as a function of n_1 when $a_0 = .5$ and $n_1 + n_2 = 100$.

order to assess the behavior of the mixture model weights in the case of informative prior modeling.

5.2.1 Non-informative prior modeling

Let us to consider an i.i.d sample $x_1, \ldots, x_n$ from model $\mathfrak{M}_\alpha$. With the assumption of the same parameter in both components of the mixture model, the likelihood is such that

$$\ell(\lambda, \alpha|x) = \prod_{i=1}^{n} \alpha \exp(-\lambda)\lambda^{x_i}/x_i! + (1-\alpha)\lambda^{x_i}/(1+\lambda)^{x_i+1}$$

and under the condition of the missing variable ζ_i associated with each x_i, we will have

$$\ell(\lambda, \alpha|x, \zeta) = \alpha^{n_1}(1-\alpha)^{n_2} \prod_{i;\zeta_i=1} \exp(-\lambda)\lambda^{x_i}/x_i! \prod_{i;\zeta_i=2} (1/1+\lambda)(1-1/1+\lambda)^{x_i}.$$

When λ and α are independent and $\pi(\lambda) = 1/\lambda$, the posterior distribution of λ is given by

$$\pi(\lambda|x, \zeta) \propto \exp(-n_1\lambda)\lambda^{\sum_i \mathbb{I}_{\zeta_i=1} x_i + \sum_i \mathbb{I}_{\zeta_i=2} x_i - 1}/(1+\lambda)^{n_2 + \sum_i \mathbb{I}_{\zeta_i=2} x_i}$$

In order to simulate parameter λ from this non-standard posterior distribution, we apply the Metropolis-within-Gibbs algorithm. The related **R** code is available in [Kamary 2016b] in which α is simulated from the conditional posterior density $\mathcal{B}eta(a_0 + n_1, a_0 + n_2)$ and λ from gamma proposal with parameter $(\sum_{i=1}^{n} x_i/2, n/2)$. The proposal distribution of λ results in an unbiased estimate for λ because the

expected value is $\sum_{i=1}^{n} x_i/n$ which converges to $\mathbb{E}(\sum_{i=1}^{n} x_i/n) = \lambda$. This algorithm works well in terms of accurately estimating the parameters λ and θ by comparing them with the true values and the resulting distribution of α supports the true model when the sample size is high enough. However, in the case where the sample size is small, this algorithm results in poor estimation of α due to the problem of label switching as shown in Figure 5.2.

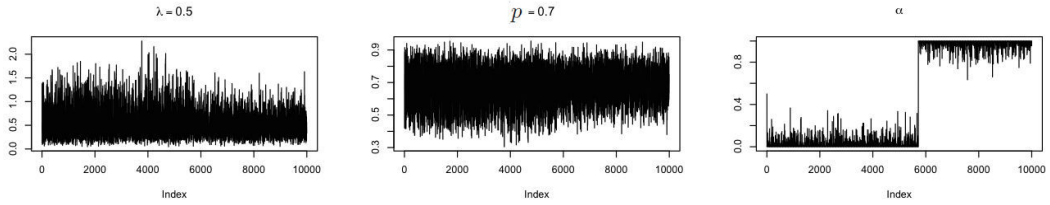


Figure 5.2: **Poisson versus geometric** 5.2.1: Sequence of $\lambda[t], p[t]$ and $\alpha[t]$ simulated by `GSmix` function with $10,000$ iterations for a dataset of size 30 from $\mathcal{P}(.5)$ when $a_0 = .1$.

Another **R** code can be written by considering the acceptance ratio of the Metropolis-Hastings step based on the target distributions without associating any indicator variable ζ to the observations. In other words, we can sample from the following posterior densities

$$\pi(\lambda|x) = \left(\prod_{i=1}^{n} \alpha \exp(-\lambda)\lambda^{x_i}/x_i! + (1-\alpha)\lambda^{x_i}/(1+\lambda)^{x_i+1}\right) 1/\lambda$$
$$\pi(\alpha|x) \propto \left(\prod_{i=1}^{n} \alpha \exp(-\lambda)\lambda^{x_i}/x_i! + (1-\alpha)\lambda^{x_i}/(1+\lambda)^{x_i+1}\right) \alpha^{a_0-1}(1-\alpha)^{a_0-1}$$

which are both non-standard and require MCMC algorithm to sample. The advantages of the implementation of corresponding algorithm in **R** is that it is almost twice faster than the Gibbs sampler method and the label switching does not happen anymore in the output of α even for small sample sizes. The **R** code is available in [Kamary 2016b]. Our first check on convergence of Markov chains provided by this algorithm is to consider four samples of size $50, 1000, 400, 600$ simulated from $\mathcal{P}(.64)$, $\mathcal{P}(10)$, $\mathcal{G}eo(.4)$ and $\mathcal{P}(.59)$, respectively, and to plot histories of $\lambda[t], p[t], \alpha[t]$, as shown in Figures 5.3 and 5.4. The trace plots indicate that the Markov chains have stabilized and appear constant over the graphs. Moreover, the chains have good mixing and are dense in the sense that they quickly traverse the support of the distribution. They are also able to explore both the tails and the mode areas efficiently. The autocorrelation plots show a very small degree of autocorrelation among the posterior samples, and the histograms estimate the posterior marginal distributions for the parameters. Note that we observe the same behavior for the simulated samples when $a_0 = .2, .3, .4, .5$.

In order to evaluate the accuracy of the parameter estimations, we simulate 50 datasets of sizes from 10 to 1000 once from Poisson distribution with parameter

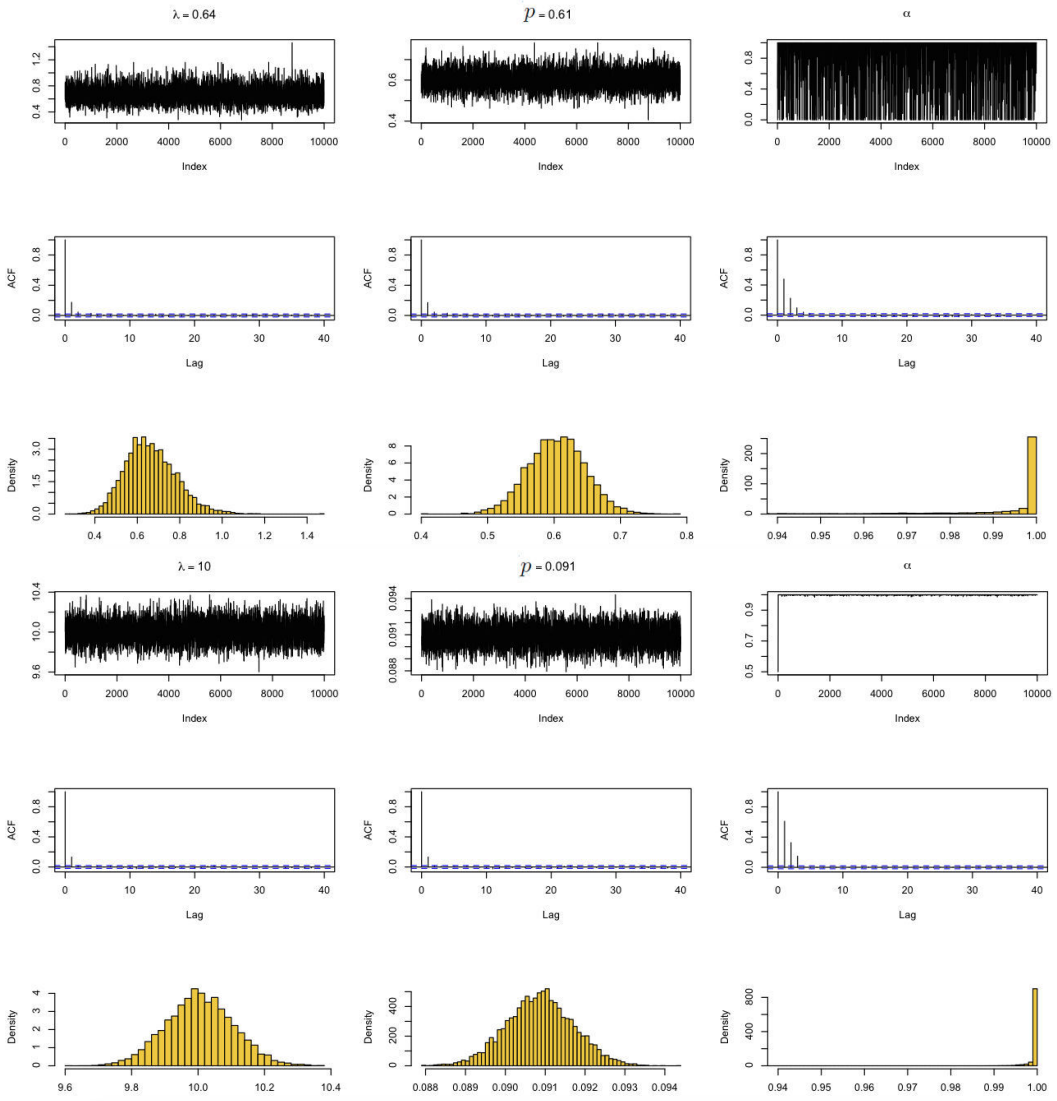


Figure 5.3: **Poisson versus geometric** 5.2.1: Sequence *(Top)*, empirical autocorrelations using `acf` function in **R** *(Middle)* and histograms *(Bottom)* of $\lambda[t], p[t]$ and $\alpha[t]$ with $10,000$ MCMC iterations for datasets of sizes $50, 1000$ from $\mathcal{P}(.64)$ and $\mathcal{P}(10)$ when $a_0 = .1$.

λ varied from .5 to 10 and another from geometric distribution with parameter p diverse from .06 to 1. By running MCMC algorithm for all 100 datasets, the mean absolute errors of the resulting estimates of λ and p based on the median of the simulated samples are summarized in Table 5.1 . Very small values of MAE lead us to conclude that the parameters of both models are accurately estimated in both cases. In addition to that, the evolution of the corresponding posterior estimations of α over the sample size shown in Figure 5.5 indicates the degree of the support of α toward the true model.

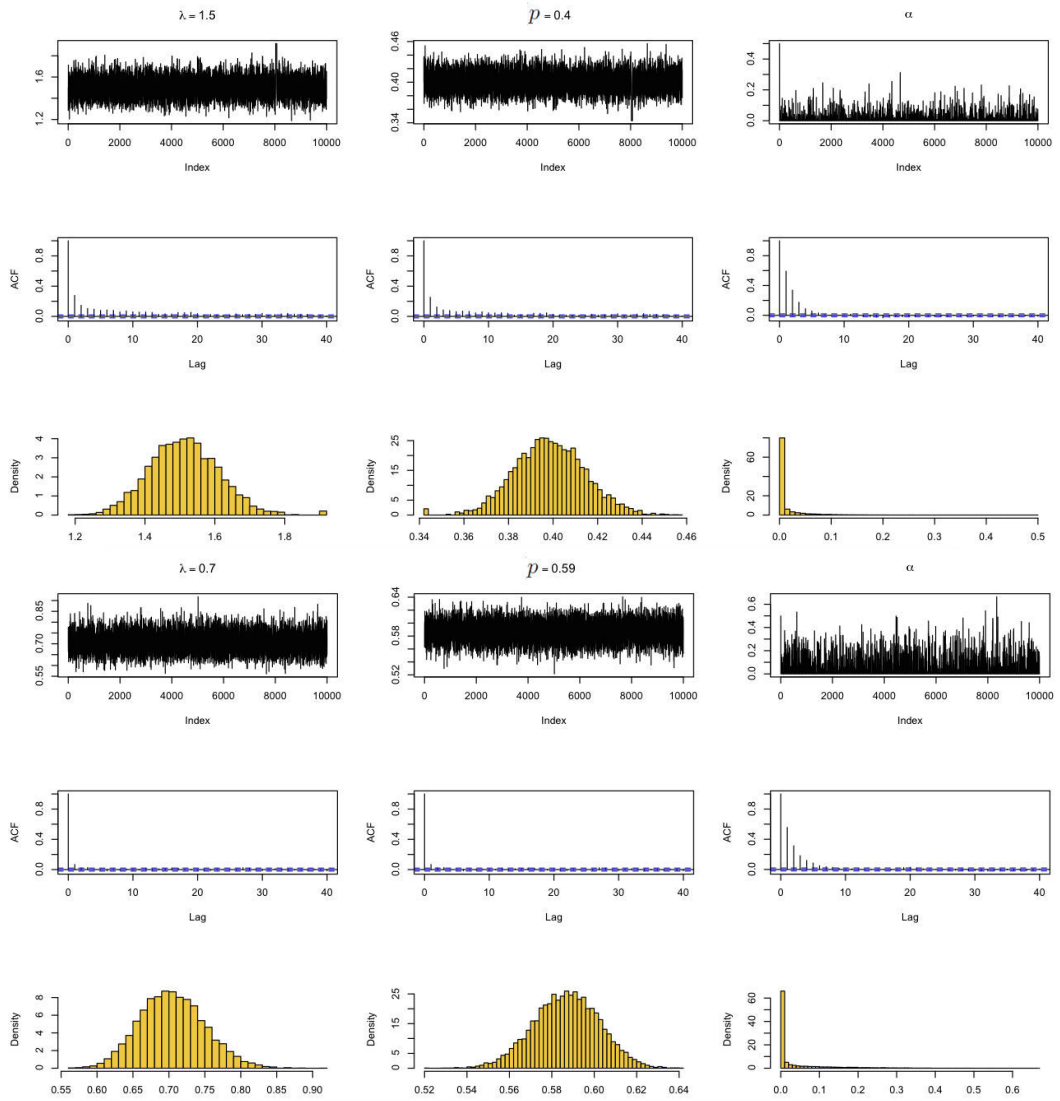


Figure 5.4: **Poisson versus geometric** 5.2.1: Sequence *(Top)*, empirical autocorrelations using `acf` function in **R** *(Middle)* and histograms *(Bottom)* of $\lambda[t], p[t]$ and $\alpha[t]$ with $10,000$ MCMC iterations for datasets of sizes $400, 600$ from $\mathcal{G}eo(.4)$ and $\mathcal{P}(.59)$ when $a_0 = .1$.

5.2.2 Informative prior modeling

Suppose that $\mathcal{P}(\lambda)$ is tested against $\mathcal{G}eo(p)$. We consider the following conjugate priors for the parameters λ and p,

$$\lambda \sim \mathcal{G}(\beta_1, \beta_2); \quad p \sim \mathcal{B}eta(\delta_1, \delta_2).$$

The joint posterior distribution of λ and p is therefore given by

$$\pi(\lambda, p|x) \propto \left(\prod_{i=1}^{n} \alpha^{\exp(-\lambda)\lambda^{x_i}/x_i!} + (1-\alpha)(1-p)^{x_i} p \right) \lambda^{\beta_1 - 1} \exp(-\beta_2 \lambda) p^{\delta_1 - 1} (1-p)^{\delta_2 - 1}$$

Datasets from Poisson distribution					
a_0	0.1	0.2	0.3	0.4	0.5
$\tilde{\lambda}$	.0064	.0077	.0085	.0082	.0101
$\tilde{\theta}$	.0016	.0018	.0019	.0017	.0019
Datasets from Geometric distribution					
a_0	0.1	0.2	0.3	0.4	0.5
$\tilde{\lambda}$	.0531	.0602	.0580	.0648	.0701
$\tilde{\theta}$	.0542	.0619	.0010	.0009	.0011

Table 5.1: **Poisson versus Geometric** 5.2.1: Mean absolute error.

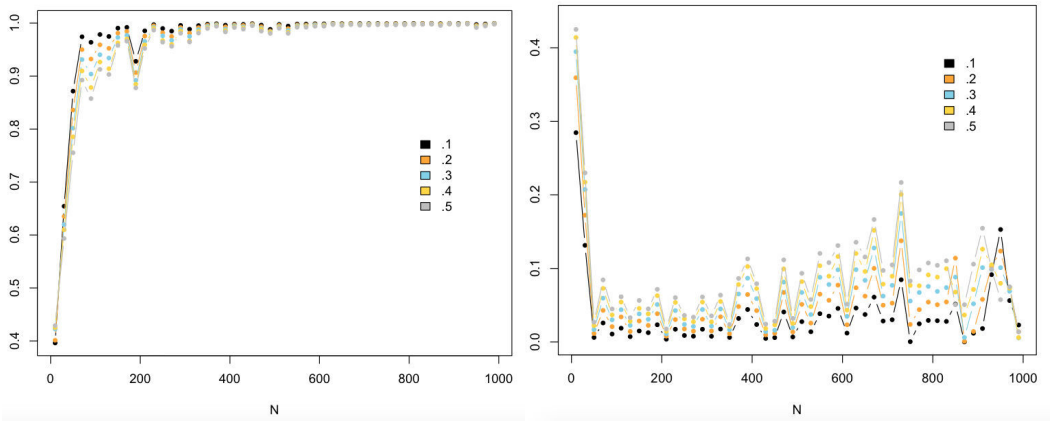


Figure 5.5: **Poisson versus geometric** 5.2.1: Evolution of α over the sample size when data is simulated from *(left)* Poisson distribution with parameter λ taking values from .5 to 10; *(right)* Geometric distribution with parameter p taking values from .06 to 1. Each estimation is based on $10,000$ MCMC iterations.

The posterior samples from this non-standard posterior density are produced by implementing the Metropolis-within-Gibbs in **R** [Kamary 2016b]. The candidate values for λ and p are independently proposed by gamma and beta distributions. Once again, we simulate two datasets of sizes $50, 500$ from $\mathcal{P}(.64)$ and $\mathcal{G}eo(.49)$, respectively, and for each dataset, we execute the **R** code 50 times by choosing different initial values for the chains in order to assess the convergence of the simulated samples. For both datasets, the marginal posterior distributions of λ and p are stable under 50 repetitions of the Metropolis-within-Gibbs algorithm with $10,000$ iterations, as shown in Figure 5.6. They are also identical and centered on the true values whatever the value of a_0 is. The estimated densities for α shown on the right side of Figure 5.6 have the same behavior as the case where both models under comparison share the same parameter. This means that the distributions are concentrated over 1 for the true model and the smaller the value of a_0 is, the more the densities tighten up over the boundaries of unit interval. In addition to this, for each value of a_0, the generated samples of α resulted by 50 repetitions of MCMC algorithm have the same behavior and this guarantees the convergence of the chains

toward stationarity in this case.

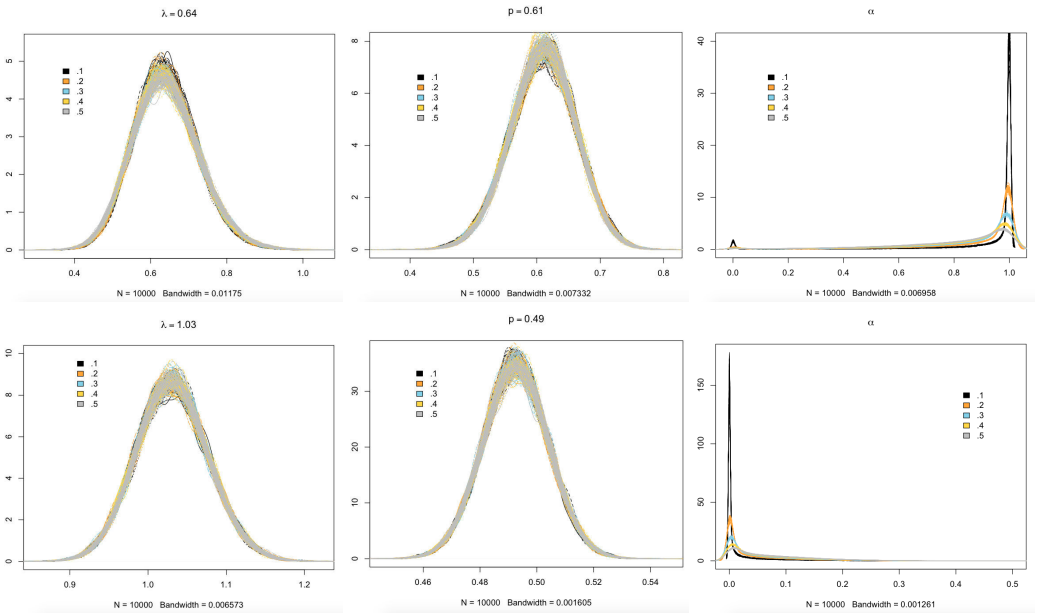


Figure 5.6: **Poisson versus geometric** 5.2.2: Marginal posterior distributions of λ, p and α obtained by running MCMC algorithm for 50 times with $a_0 = 0.1, 0.2, 0.3, 0.4, 0.5$ for *(Top)* a sample of size 50 simulated from $\mathcal{P}(.64)$ when the true value of $(\beta_1, \beta_2, \delta_1, \delta_2)$ is $(32, 50, 50, 32)$; *(Bottom)* a sample of size 500 simulated from $\mathcal{G}eo(.49)$ when the true value of $(\beta_1, \beta_2, \delta_1, \delta_2)$ is $(505, 490, 490, 505)$. Each density is based on $10,000$ MCMC iterations.

5.3 $\mathcal{N}(\theta, 1)$ versus $\mathcal{N}(\theta, 2)$

When comparing two normal distributions with the same location parameter θ, both Gibbs sampler based on allocating missing variable ζ to the observations and the Metropolis-Hastings algorithm based on following posterior

$$\pi(\theta|x, \alpha) \propto \ell(\theta, \alpha|x)\pi(\theta) \tag{5.1}$$

$$= \prod_{i=1}^{n} \left(\alpha \exp(-(x_i-\theta)^2/2) + (1-\alpha) \exp(-(x_i-\theta)^2/4)/\sqrt{2} \right) \tag{5.2}$$

yield the same Bayesian inference for the mixture model. The label switching does not happen using Gibbs sampler in this case, as shown at the bottom of Figures 5.7 and 5.8. The implementations of both algorithms using programming language **R** are given in [Kamary 2016b].

In the Metropolis-Hastings algorithm, the proposals of θ are drawn from a normal distribution centered on the empirical mean of the observations and its standard deviation can be calibrated by an argument in the input of the corresponding function. By simulating 3 datasets of sizes $10, 510, 1000$ once from $\mathcal{N}(\theta, 1)$ and another

from $\mathcal{N}(\theta, 2)$, we analyze the output of both algorithms. Comparison between the estimated marginal posterior distributions of θ derived from the outputs of both Metropolis-Hastings and Gibbs sampler is shown in Figures 5.7 and 5.8. Both methods accurately estimate the common location parameter of the normal distribution for any value of a_0 while the resulting posterior densities of α are identical and in favor of the true model in both cases even for samples of size 10.

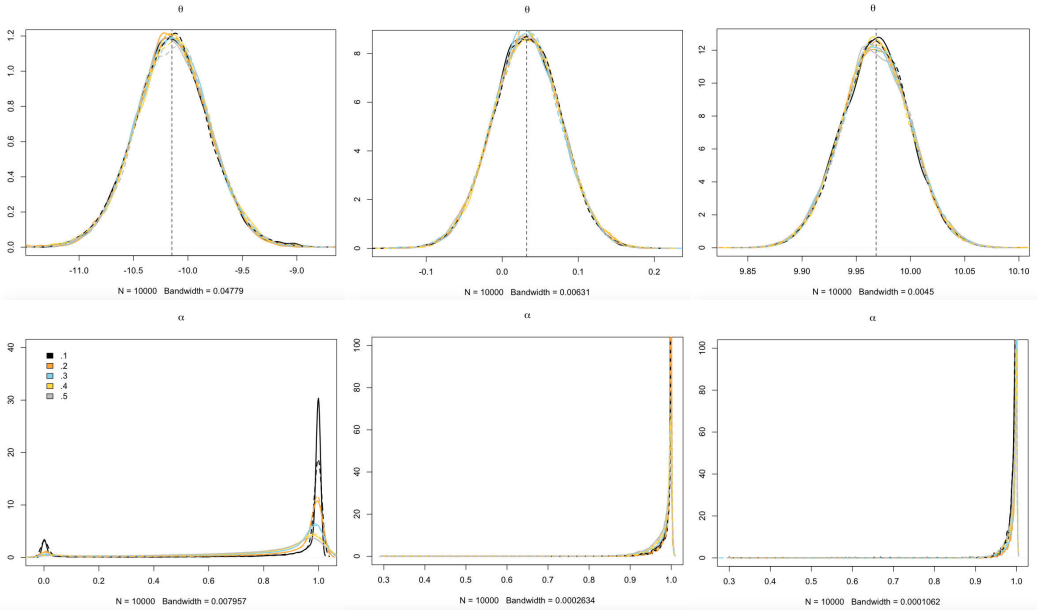


Figure 5.7: $\mathcal{N}(\theta, 1)$ **versus** $\mathcal{N}(\theta, 2)$ 5.3: Marginal posterior densities of θ *(Top)* and α *(Bottom)* obtained by running *(dashed lines)* Gibbs sampling algorithm and *(solid lines)* Metropolis-Hastings algorithm with different values of a_0 for samples of sizes $10, 510$ and 1000 simulated from $\mathcal{N}(-10.15, 1), \mathcal{N}(-0.03, 1)$ and $\mathcal{N}(9.97, 1)$, respectively. Each density is based on $10,000$ MCMC iterations and the vertical dotted line in θ plots corresponds to the true value.

Another data analysis is summarized in Tables 5.2 and 5.3. The tables report a series of posterior summaries related to 11 datasets simulated from normal distributions, $\mathcal{N}(\theta, 1)$ and $\mathcal{N}(\theta, 2)$ such that the posterior median and standard deviation of posterior draws for θ and α. The number of the observations N and the true value of θ are also listed. Included in the tables is also a convergence test using [Gelman 1992]'s criterion that is done by `gelman.diag` function in **R**, named `gd.`. This test is based on four MCMC chains produced in parallel starting from an arbitrary position for each parameter. Tables 5.2 and 5.3 display that for all datasets, the parameter θ is accurately estimated by the median of the posterior draws with a standard deviation less than .5 while the point estimate of α is always very close to 1 for the true model. The results of `gd.` test shows a clear stabilization around the target value 1 from 10,000 iterations which indicates that the four chains have converged on the same region, resulting in a perfect fit to the target.

		D.1	D.2	D.3	D.4	D.5	D.6	D.7	D.8	D.9	D.10	D.11
N		10	30	40	50	70	90	110	310	510	810	1000
θ		-10.6	-6.4	-3.9	-2.1	2.2	6	10	-4.1	0	6	10
$a_0 = 0.1$	sd.	0.33	0.18	0.16	0.15	0.12	0.11	0.095	0.06	0.04	0.04	0.03
$\hat{\theta}$	md.	-10.65	-6.38	-3.92	-2.09	2.17	6.01	10.02	-4.07	0.03	5.97	9.97
	gd.	1	1	1	1.02	1.01	1	1	1	1	1	1.01
	sd.	0.26	0.09	0.07	0.14	0.09	0.07	0.04	0.02	0.02	0.01	0.02
$\hat{\alpha}$	md.	0.99	0.99	0.99	0.99	0.99	0.99	0.99	1	1	1	0.99
	gd.	1	1.01	1	1	1.001	1	1	1.01	1	1	1
$a_0 = 0.2$	sd.	0.33	0.19	0.16	0.15	0.12	0.11	0.09	0.06	0.05	0.04	0.03
$\hat{\theta}$	md.	-10.64	-6.38	-3.93	-2.1	2.18	6.02	10.01	-4.06	-0.02	5.96	10.01
	gd.	1.002	1	1	1	1.02	1	1	1	1	1	1
	sd.	0.27	0.11	0.11	0.17	0.11	0.08	0.06	0.03	0.02	0.014	0.02
$\hat{\alpha}$	md.	0.98	0.99	0.99	0.98	0.99	0.99	0.99	0.99	0.99	0.99	0.99
	gd.	1.01	1	1	1	1	1	1.02	1	1	1.01	1
$a_0 = 0.3$	sd.	0.34	0.19	0.16	0.15	0.12	0.11	0.098	0.06	0.04	0.04	0.03
$\hat{\theta}$	md.	-10.54	-6.41	-3.91	-2.01	2.13	6.00	9.98	-4.09	0.01	6.01	9.97
	gd.	1	1	1	1	1.01	1	1	1	1	1	1
	sd.	0.28	0.13	0.12	0.17	0.12	0.10	0.07	0.03	0.03	0.02	0.02
$\hat{\alpha}$	md.	0.94	0.98	0.98	0.95	0.97	0.98	0.98	0.99	0.99	0.99	0.99
	gd.	1	1	1	1	1	1	1	1.02	1	1	1.01
$a_0 = 0.4$	sd.	0.34	0.19	0.16	0.15	0.12	0.11	0.10	0.06	0.05	0.03	0.03
$\hat{\theta}$	md.	-10.63	-6.37	-3.90	-2.13	2.17	6.01	10.01	-4.1	0.03	5.97	9.98
	gd.	1	1	1	1	1	1.01	1	1	1	1	1
	sd.	0.27	0.14	0.12	0.18	0.13	0.11	0.08	0.03	0.03	0.02	0.03
$\hat{\alpha}$	md.	0.92	0.96	0.97	0.92	0.95	0.96	0.98	0.99	0.99	0.99	0.99
	gd.	1	1	1	1	1	1	1	1	1.02	1.01	1
$a_0 = 0.5$	sd.	0.34	0.19	0.16	0.15	0.12	0.11	0.10	0.06	0.04	0.04	0.03
$\hat{\theta}$	md.	-10.65	-6.42	-3.89	-2.11	2.18	5.98	10	-4.07	-0.03	6	10.03
	gd.	1	1.01	1	1	1	1	1	1	1	1	1
	sd.	0.27	0.15	0.14	0.19	0.14	0.12	0.08	0.04	0.03	0.02	0.03
$\hat{\alpha}$	md.	0.88	0.95	0.96	0.89	0.93	0.95	0.97	0.99	0.99	0.99	0.98
	gd.	1	1	1	1	1	1	1	1.02	1	1.01	1.01

Table 5.2: $\mathcal{N}(\theta, 1)$ **versus** $\mathcal{N}(\theta, 2)$ 5.3: Posterior summaries; Datasets (D.1, ..., D.11) are simulated from $\mathcal{N}(\theta, 1)$ and each point estimator is based on $10,000$ iterations of the Metropolis-Hastings algorithm.

		D.1	D.2	D.3	D.4	D.5	D.6	D.7	D.8	D.9	D.10	D.11
N		10	30	40	50	70	90	110	310	510	810	1000
θ		-9.3	-6.2	-3.9	-1.6	2	6.1	10.1	-3.9	0	6	10
$a_0 = 0.1$	sd.	0.44	0.26	0.24	0.2	0.17	0.15	0.13	0.08	0.06	0.05	0.04
$\hat{\theta}$	md.	-9.31	-6.18	-3.87	-1.62	1.97	6.12	10.12	-3.87	0.02	6.01	10.01
	gd.	1	1	1.01	1	1	1	1	1	1	1	1
	sd.	0.13	0.11	0.08	0.12	0.06	0.16	0.09	0.09	0.02	0.02	0.03
$\hat{\alpha}$	md.	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	6e-5	4e-5	2e-5
	gd.	1	1	1	1	1	1	1.01	1	1	1.01	1
$a_0 = 0.2$	sd.	0.44	0.24	0.22	0.21	0.17	0.14	0.13	0.08	0.06	0.05	0.04
$\hat{\theta}$	md.	-9.33	-6.35	-3.88	-1.62	1.97	6.12	10.13	-3.88	0.01	6.01	10.01
	gd.	1.01	1	1	1.02	1	1	1.01	1	1	1	1
	sd.	0.16	0.14	0.11	0.14	0.08	0.18	0.12	0.10	0.03	0.02	0.03
$\hat{\alpha}$	md.	0.01	0.01	0.01	0.02	0.01	0.05	0.01	0.04	1e-3	1e-3	0.00
	gd.	1	1	1	1	1	1	1	1	1	1.01	1
$a_0 = 0.3$	sd.	0.45	0.24	0.24	0.18	0.17	0.14	0.14	0.08	0.06	0.05	0.05
$\hat{\theta}$	md.	-9.36	-6.18	-3.88	-1.62	1.98	6.11	10.12	-3.87	0.02	6.02	10.01
	gd.	1	1	1.01	1	1.01	1	1.01	1	1	1	1
	sd.	0.18	0.15	0.11	0.16	0.1	0.19	0.13	0.10	0.04	0.03	0.04
$\hat{\alpha}$	md.	0.04	0.04	0.02	0.05	0.02	0.12	0.05	0.07	0.01	0.01	0.01
	gd.	1	1	1	1	1	1	1	1	1.01	1	1
$a_0 = 0.4$	sd.	0.45	0.24	0.22	0.2	0.16	0.14	0.14	0.08	0.06	0.05	0.04
$\hat{\theta}$	md.	-9.37	-6.18	-3.88	-1.62	1.97	6.12	10.12	-3.87	0.02	6.00	10.01
	gd.	1.01	1	1	1	1	1	1.01	1	1.01	1	1.01
	sd.	0.18	0.17	0.13	0.17	0.11	0.19	0.13	0.11	0.04	0.03	0.02
$\hat{\alpha}$	md.	0.06	0.06	0.05	0.08	0.04	0.17	0.07	0.09	0.01	0.01	0.02
	gd.	1	1	1	1	1	1	1.01	1	1	1	1
$a_0 = 0.5$	sd.	0.45	0.24	0.21	0.19	0.16	0.14	0.13	0.08	0.06	0.05	0.04
$\hat{\theta}$	md.	-9.35	-6.18	-3.87	-1.62	1.98	6.12	10.12	-3.79	-0.02	6.01	9.98
	gd.	1.01	1	1	1	1	1	1.01	1	1	1	1
	sd.	0.19	0.17	0.14	0.18	0.11	0.19	0.14	0.11	0.04	0.03	0.04
$\hat{\alpha}$	md.	0.09	0.08	0.07	0.11	0.05	0.18	0.09	0.12	0.02	0.01	0.02
	gd.	1	1	1	1	1	1	1	1	1	1	1.02

Table 5.3: $\mathcal{N}(\theta, 1)$ **versus** $\mathcal{N}(\theta, 2)$ 5.3: Posterior summaries; Datasets (D.1, ..., D.11) are simulated from $\mathcal{N}(\theta, 2)$ and each point estimator is based on $10,000$ iterations of the Metropolis-Hastings algorithm.

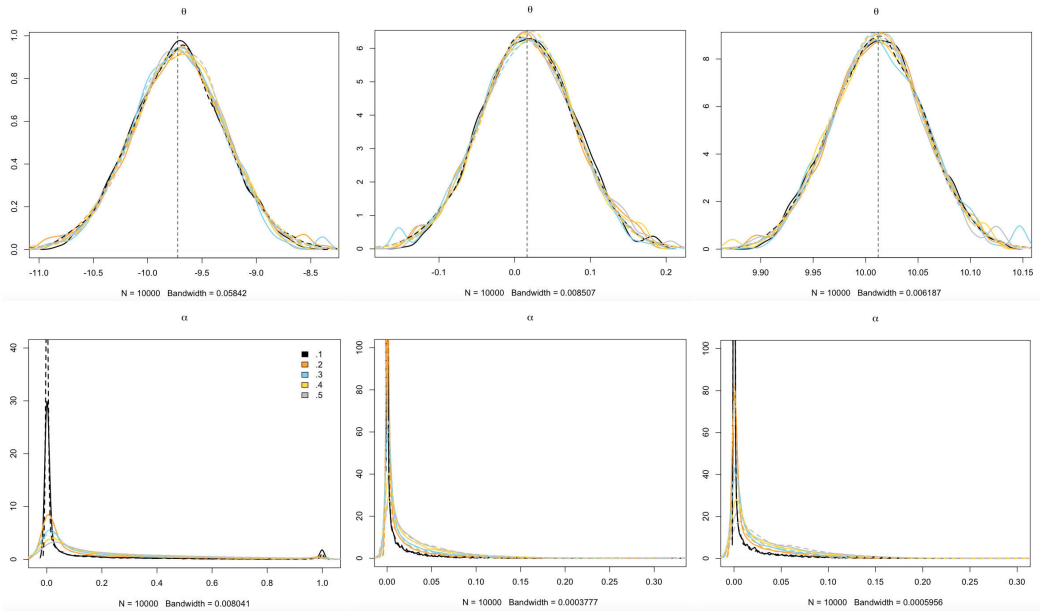


Figure 5.8: $\mathcal{N}(\theta, 1)$ **versus** $\mathcal{N}(\theta, 2)$ 5.3: Marginal posterior densities of θ *(Top)* and α *(Bottom)* obtained by running *(dashed lines)* the Gibbs sampling algorithm and *(solid lines)* the Metropolis-Hastings algorithm with different values of a_0 for samples of sizes $10, 510$ and 1000 simulated from $\mathcal{N}(-9.72, 2), \mathcal{N}(0.02, 2)$ and $\mathcal{N}(10.01, 2)$, respectively. Each density is based on $10,000$ MCMC iterations and the vertical dotted line in θ plots corresponds to the true value.

5.4 Standard normal distribution versus $\mathcal{N}(\mu, 1)$

The mixture of the standard normal distribution and $\mathcal{N}(\mu, 1)$ is defined by $\mathfrak{M}_\alpha$: $\alpha\mathcal{N}(0,1) + (1-\alpha)\mathcal{N}(\mu, 1)$ and the conditional posterior distributions of μ and α are given by

$$\pi(\mu|x) \propto \left(\prod_{i=1}^{n} \alpha \exp(-x_i^2/2) + (1-\alpha)\exp(-(x_i-\mu)^2/2)\right) \exp(-\mu^2/2)$$

$$\pi(\alpha|x) \propto \left(\prod_{i=1}^{n} \alpha \exp(-x_i^2/2) + (1-\alpha)\exp(-(x_i-\mu)^2/2)\right) \alpha^{a_0-1}(1-\alpha)^{a_0-1}.$$

The Metropolis-within-Gibbs algorithm will be applied to simulate from the conditional posteriors above and the corresponding **R** code can be seen in [Kamary 2016b]. Convergence verification of the chains produced by this algorithm is done by plotting the resulting posterior draws for some datasets simulated from both competing models as shown in Figure 5.9. Small autocorrelations indicate very low degree of correlation between the draws. The trace plots illustrate good mixing of the chains which are moving around the parameter space. The marginal posterior distribution of each parameter is also shown by the histograms in Figure 5.9. The plots are related to $a_0 = .1$ and we get the same results for the cases where $a_0 = .2, .3, .4, .5$.

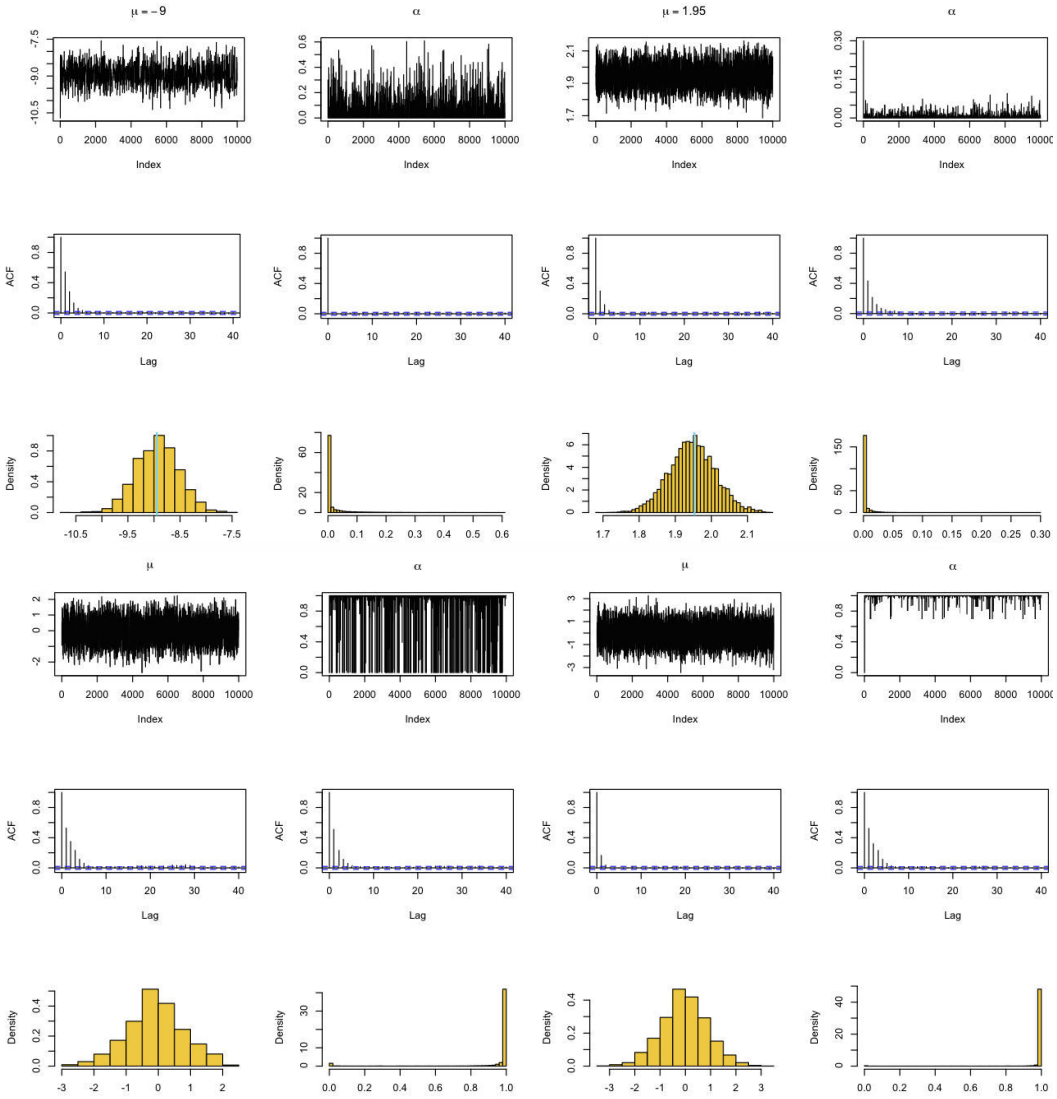


Figure 5.9: $\mathcal{N}(0,1)$ **versus** $\mathcal{N}(\mu,1)$ 5.4: Sequence, empirical autocorrelations using `acf` function in **R** and histograms of $\mu[t]$ and $\alpha[t]$ simulated by the Metropolis-within-Gibbs algorithm with $10,000$ iterations for: *(Top)* two datasets of sizes 5 *(left)* and 250 *(right)* from $\mathcal{N}(-9,1)$ and $\mathcal{N}(1.95,1)$, respectively; *(Bottom)* two datasets of sizes 850 *(left)* and $50,000$ *(right)* simulated from standard normal distribution. $a_0 = .1$.

Another experiment is to run the algorithm several times for different datasets of different sizes simulated from each model under comparison. Table 5.4 lists the information about datasets such as the number of observations and the true value of the parameter μ in the case of simulating the data points from $\mathcal{N}(\mu,1)$. The table reports also the posterior mean and standard deviation of μ which are accurately estimated for all datasets.

Figure 5.10 displays that when $\mathcal{N}(\mu,1)$ is the model from which the dataset is simulated, the posterior estimate of α strongly supports this model for the data

Table (a)

	Data:	**D.1**	**D.2**	**D.3**	**D.4**	**D.5**	**D.6**	**D.7**	**D.8**
	N	5	15	25	35	55	65	75	95
$\hat{\mu}$	Mean	-0.17	0.16	0.2	0.09	-0.03	0.02	0.00	0.02
	Sd.	0.83	0.82	0.9	0.84	0.86	0.89	0.86	0.85
	Data:	**D.9**	**D.10**	**D.11**	**D.12**	**D.13**	**D.14**	**D.15**	**D.16**
	N	250	450	650	750	850	950	1000	5e+4
$\hat{\mu}$	Mean	-0.02	-0.02	-0.01	-0.02	-0.08	0.05	-0.10	-0.03
	Sd.	0.86	0.81	0.83	0.85	0.88	0.89	0.80	0.90

Table (b)

	Data:	**D.1**	**D.2**	**D.3**	**D.4**	**D.5**	**D.6**	**D.7**	**D.8**
	N	5	15	25	35	45	55	65	250
	μ	-9	-8.3	-7.8	-6.8	-5.8	-5	-4.1	2
$\hat{\mu}$	Mean	-8.93	-8.38	-7.75	-6.89	-5.75	-4.96	-4.06	1.94
	Sd.	0.41	0.24	0.19	0.17	0.15	0.13	0.12	0.06
	Data:	**D.9**	**D.10**	**D.11**	**D.12**	**D.13**	**D.14**	**D.15**	**D.16**
	N	350	450	550	650	750	850	950	1000
	μ	3	4	4.9	6	6.9	8	8.9	10
$\hat{\mu}$	Mean	3.01	4.00	4.98	6.00	6.93	7.98	8.98	10.04
	Sd.	0.05	0.05	0.04	0.04	0.04	0.03	0.03	0.03

Table 5.4: $\mathcal{N}(0, 1)$ **versus** $\mathcal{N}(\mu, 1)$ 5.4: Observation information and posterior summaries; *Table (a)* Datasets (D.1, ..., D.16) are simulated from $\mathcal{N}(0, 1)$; *Table (b)* Datasets (D.1, ..., D.16) are simulated from $\mathcal{N}(\mu, 1)$. a_0 is supposed to be .1 and each point estimator is based on $10,000$ MCMC iterations.

whatever the sample size is. By comparing the posterior distributions of α related to the datasets **D.1, D.2, D.3** with $N = 5, 15, 25$ in the case where the datasets are from standard normal distribution, we can see that the variation of the posterior draws is high, especially for the sample of small size.

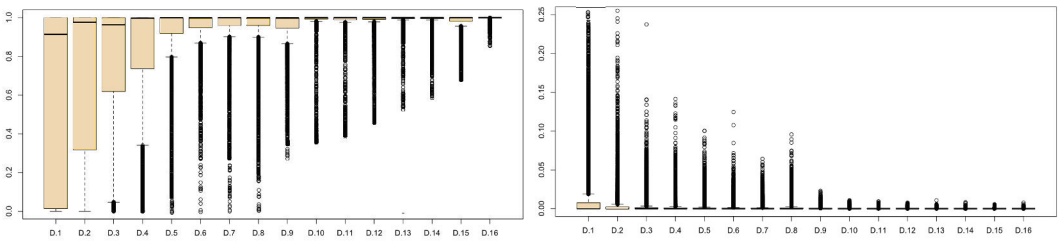


Figure 5.10: $\mathcal{N}(0,1)$ **versus** $\mathcal{N}(\mu,1)$ 5.4: Posterior distribution of $\alpha[t]$, the weight of the standard normal in the mixture model, under a beta prior with parameter $a_0 = 0.1$ for *(left)* 16 standard normal datasets and *(right)* 16 datasets simulated from $\mathcal{N}(\mu,1)$ when each posterior approximation is based on 10^4 MCMC iterations. Table 5.4 lists more details about the datasets.

5.5 Normal versus double-exponential distribution

The mixture of a normal and a double-exponential distributions can be defined as $\alpha\mathcal{N}(\mu,1) + (1-\alpha)\mathcal{L}(\mu,\sqrt{2})$ and the conditional posterior distributions of μ and α are therefore given by

$$\pi(\mu|x,\alpha) \propto \left(\prod_{i=1}^{n} \alpha\exp(-(x_i-\mu)^2/2) + (1-\alpha)\exp(-|x_i-\mu|/\sqrt{2})\right)$$
$$\pi(\alpha|x,\mu) \propto \left(\prod_{i=1}^{n} \alpha\exp(-(x_i-\mu)^2/2) + (1-\alpha)\exp(-|x_i-\mu|/\sqrt{2})\right)(\alpha(1-\alpha))^{a_0-1}.$$

An implementation of the Metropolis-within-Gibbs can be used to simulate from these non-standard posteriors in which the parameter μ is simulated from a normal distribution centered on the empirical mean of the dataset while the standard deviation is calibrated by the user. The code in **R** is available in [Kamary 2016b]. In Chapter 4, we illustrated that the posterior estimates of the mixture model weights fail to concentrate near 0 or 1 even for high sample sizes when the analyzed dataset is produced by another model than those in competition. Here, we proceed by analyzing the behavior of α for the datasets simulated from one of the models under comparison. To do so, 21 samples of sizes from 5 to 1000 are simulated once from standard normal distribution another time from $\mathcal{L}(0,\sqrt{2})$. The trace plots of μ are shown in Figure 5.11 and indicate the stabilization of the Markov chains over the true value 0 for any sample sizes. Figure 5.11 also shows that the higher the sample size is the more the concentration of the Markov chain is over the true value while

the marginal posterior densities of α are always in favor of the true model even for the samples of size 5. In other words, the marginal posterior densities of α tightenup near 1 when the data comes from $\mathcal{N}(0,1)$. We emphasize here that these results are preliminary for the convergence check.

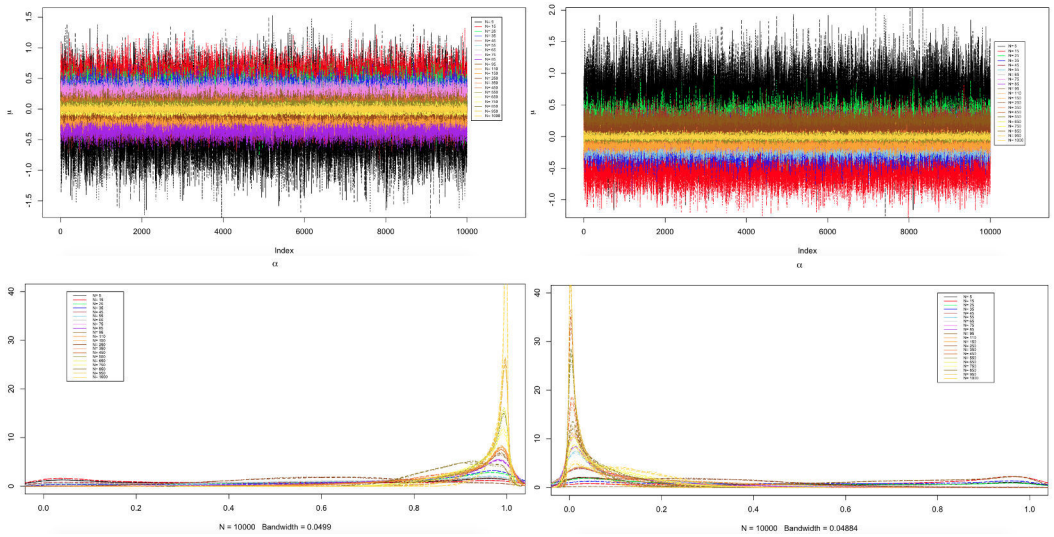


Figure 5.11: $\mathcal{N}(\mu,1)$ **versus** $\mathcal{L}(\mu,\sqrt{2})$ 5.5: *(Top)* Sequence of $\mu[t]$ and *(Bottom)* empirical densities of the posterior draws of α for 21 datasets from $\mathcal{N}(0,1)$ *(Left)* and from $\mathcal{L}(0,\sqrt{2})$ *(Right)*, based on 10^4 iterations of MCMC algorithm when $a_0 = 0.5$.

In order to check the convergence of the Markov chains, [Gelman 2003] suggests to compute the statistic `gelman.diag` for each scalar estimate of interest, and to continue running the chains until the statistics are all less than 1.1. We therefore compute this criterion based on four chains produced for μ and for α and list the values in Table 5.5. This table also displays a 97.5% upper limit of this diagnostic for both parameters α and μ. The values of [Gelman 2003]'s statistics shown in this table are all less than 1.1, that illustrate satisfactory convergence has been achieved.

When we test a normal $\mathfrak{M}_1 = \mathcal{N}(\mu,1)$ against a double-exponential distribution $\mathfrak{M}_2 = \mathcal{L}(\mu,\sqrt{2})$ under the flat prior by using the Bayes factor, we need to compute the marginal likelihood under both models. The marginal distribution under $\mathfrak{M}_1$ is

$$\begin{aligned}\pi_1(x) &= \int_{-\infty}^{\infty} (2\pi)^{-n/2} \exp\left(-\textstyle\sum_{i=1}^n (x_i-\mu)^2/2\right) \mathrm{d}\mu \\ &= \exp\left(-\textstyle\sum_{i=1}^n (x_i-\bar{x})^2/2\right)\Big/(2\pi)^{-n/2} \int_{-\infty}^{\infty} \exp(-n(\mu-\bar{x})^2/2)\mathrm{d}\mu \\ &= \exp\left(-\textstyle\sum_{i=1}^n (x_i-\bar{x})^2/2\right)\Big/(2\pi)^{-(n-1)/2}\sqrt{n}\end{aligned}$$

Potential scale reduction factors:

		μ		α	
Data	N	Point est.	97.5% quantile	Point est.	97.5% quantile
D.1	5	1	1	1	1
D.2	15	1	1	1	1
D.3	25	1	1.01	1	1
D.4	35	1	1	1	1
D.5	45	1	1	1	1.02
D.6	55	1	1.01	1	1
D.7	65	1	1	1	1
D.8	75	1	1	1	1
D.9	85	1	1.01	1	1.01
D.10	95	1	1	1	1
D.11	110	1	1	1	1
D.12	150	1	1	1	1
D.13	250	1	1	1	1
D.14	350	1	1.01	1	1.01
D.15	450	1	1	1	1
D.16	550	1	1.01	1	1.01
D.17	650	1	1	1	1
D.18	750	1	1	1	1
D.19	850	1	1	1	1
D.20	950	1	1	1	1
D.21	1000	1	1	1	1

Table 5.5: $\mathcal{N}(\mu, 1)$ **versus** $\mathcal{L}(\mu, \sqrt{2})$ 5.5: Datasets (D.1, ..., D.21) are simulated from $\mathcal{L}(0, \sqrt{2})$ and a_0 is .5.

and under $\mathfrak{M}_2$, we have

$$\begin{aligned}\pi_2(x) &= \int_{-\infty}^{\infty} (2\sqrt{2})^{-n} \exp\left(-\textstyle\sum_{i=1}^{n} |x_i-\mu|/\sqrt{2}\right) \mathrm{d}\mu \\ &= (2\sqrt{2})^{-n} \int_{-\infty}^{x_{(1)}} \exp\left(-\textstyle\sum_{i=1}^{n} |x_i-\mu|/\sqrt{2}\right) \mathrm{d}\mu \\ &+ (2\sqrt{2})^{-n} \sum_{i=1}^{n-1} \int_{x_{(i)}}^{x_{(i+1)}} \exp\left(-\textstyle\sum_{j=1}^{n} |x_j-\mu|/\sqrt{2}\right) \mathrm{d}\mu \\ &+ (2\sqrt{2})^{-n} \int_{x_{(n)}}^{\infty} \exp\left(-\textstyle\sum_{i=1}^{n} |x_i-\mu|/\sqrt{2}\right) \mathrm{d}\mu\end{aligned}$$

where $x_{(1)} < \ldots < x_{(n)}$. From $\mu < x_{(1)}$, we obtain $|x_i - \mu| = x_i - \mu$ for $i = 1, \ldots, n$ and we can rewrite the first integral as following

$$\begin{aligned}\int_{-\infty}^{x_{(1)}} \exp\left(-\textstyle\sum_{i=1}^{n} |x_i-\mu|/\sqrt{2}\right) \mathrm{d}\mu &= \exp\left(-\sum_{i=1}^{n} x_i/\sqrt{2}\right) \int_{-\infty}^{x_{(1)}} \exp\left(n\mu/\sqrt{2}\right) \mathrm{d}\mu \\ &= \sqrt{2}/n \exp\left(-\sum_{i=1}^{n} x_i/\sqrt{2} + nx_{(1)}/\sqrt{2}\right)\end{aligned}$$

Since $\mu > x_{(n)}$, $|x_i - \mu| = \mu - x_i$, the third integral can be rewritten as

$$\begin{aligned}\int_{x_{(n)}}^{\infty} \exp\left(-\textstyle\sum_{i=1}^{n} |x_i-\mu|/\sqrt{2}\right) \mathrm{d}\mu &= \exp\left(\sum_{i=1}^{n} x_i/\sqrt{2}\right) \int_{x_{(n)}}^{\infty} \exp\left(-\sum_{i=1}^{n} \mu/\sqrt{2}\right) \mathrm{d}\mu \\ &= \sqrt{2}/n \exp\left(\sum_{i=1}^{n} x_i/\sqrt{2} - nx_{(n)}/\sqrt{2}\right)\end{aligned}$$

For $i = 1, \ldots, n-1$, we also have $x_{(i)} < \mu < x_{(i+1)}$ from which we deduce

$$|x_{(j)} - \mu| = \begin{cases} \mu - x_{(j)} & \text{for} \quad j < i+1 \\ x_{(j)} - \mu & \text{for} \quad j \geq i+1 \end{cases}$$

and we will therefore have

$$\begin{aligned}\int_{x_{(i)}}^{x_{(i+1)}} \exp\left(-\textstyle\sum_{j=1}^{n} |x_i-\mu|/\sqrt{2}\right) \mathrm{d}\mu &= \int_{x_{(i)}}^{x_{(i+1)}} \exp\left(-\sum_{j=1}^{i} \mu - x_{(j)}/\sqrt{2} - \sum_{j=i+1}^{n} x_{(j)} - \mu/\sqrt{2}\right) \mathrm{d}\mu \\ &= \exp\left(\textstyle\sum_{j=1}^{i} x_{(j)} - \sum_{j=i+1}^{n} x_{(j)}/\sqrt{2}\right) \\ &\int_{x_{(i)}}^{x_{(i+1)}} \exp\left(-\sum_{j=1}^{i} \mu/\sqrt{2} + \sum_{j=i+1}^{n} \mu/\sqrt{2}\right) \mathrm{d}\mu \\ &= \exp\left(\textstyle\sum_{j=1}^{i} x_{(j)} - \sum_{j=i+1}^{n} x_{(j)}/\sqrt{2}\right) \int_{x_{(i)}}^{x_{(i+1)}} \exp\left((n-2i)\mu/\sqrt{2}\right) \mathrm{d}\mu\end{aligned}$$

The last integral is equal to $x_{n/2+1} - x_{n/2}$ when $i = n/2$ and we can therefore write

$$\begin{aligned}
\sum_{i=1}^{n-1} \int_{x_{(i)}}^{x_{(i+1)}} \exp\left(-\textstyle\sum_{j=1}^{n} |x_j-\mu|/\sqrt{2}\right) \mathrm{d}\mu = & \sum_{i=1;i\neq n/2}^{n-1} \sqrt{2}/(n-2i) \exp\left(\textstyle\sum_{j=1}^{i} x_{(j)} - \sum_{j=i+1}^{n} x_{(j)}/\sqrt{2}\right) \\
& \exp\left((n-2i)x_{(i+1)}/\sqrt{2}\right) - \exp\left((n-2i)x_{(i)}/\sqrt{2}\right) \\
& + \exp\left(\sum_{j=1}^{n/2} x_{(j)}/\sqrt{2} - \sum_{j=n/2+1}^{n} x_{(j)}/\sqrt{2}\right) \left(x_{(n/2+1)} - x_{(n/2)}\right) \\
= & \sum_{i=1;i\neq n/2}^{n-1} \sqrt{2}/n-2i \exp\left(-\textstyle\sum_{j=i+1}^{n} x_{(j)} - \sum_{j=1}^{i} x_{(j)} - (n-2i)x_{(i+1)}/\sqrt{2}\right) \\
& - \exp\left(-\textstyle\sum_{j=i+1}^{n} x_{(j)} - \sum_{j=1}^{i-1} x_{(j)} - (n-2i+1)x_{(i)}/\sqrt{2}\right) \\
& + \exp\left(\sum_{j=1}^{n/2} x_{(j)}/\sqrt{2} - \sum_{j=n/2+1}^{n} x_{(j)}/\sqrt{2}\right) \left(x_{(n/2+1)} - x_{(n/2)}\right)
\end{aligned}$$

The Bayes factor in Example 4.3.4 can be derivate from the marginal likelihood of the double-exponential $\mathcal{L}(\mu, \sqrt{2})$ model under a flat prior, that is:

$$\begin{aligned}
\int_{-\infty}^{\infty} \exp\left\{-1/\sqrt{2} \sum_{i=1}^{n} |x_i - \mu|\right\} \mathrm{d}\mu = & \sqrt{2}/n \exp\left\{-1/\sqrt{2}\left(\sum_{j=1}^{n} x_{(j)} - nx_{(1)}\right)\right\} \\
& + \sqrt{2} \sum_{\substack{i=1 \\ i\neq n/2}}^{n-1} 1/n-2i \exp\left\{-1/\sqrt{2}\left(\sum_{j=i+1}^{n} x_{(j)} - \sum_{j=1}^{i} x_{(j)} - (n-2i)x_{(i+1)}\right)\right\} \\
& - \sqrt{2} \sum_{\substack{i=1 \\ i\neq n/2}}^{n-1} 1/n-2i \exp\left\{-1/\sqrt{2}\left(\sum_{j=i+1}^{n} x_{(j)} - \sum_{j=1}^{i-1} x_{(j)} - (n-2i+1)x_{(i)}\right)\right\} \\
& + (x_{n/2+1} - x_{n/2}) \exp\left\{-1/\sqrt{2}\left(\sum_{j=\frac{n}{2}+1}^{n} x_{(j)} - \sum_{j=1}^{\frac{n}{2}} x_{(j)}\right)\right\} \\
& + \sqrt{2}/n \exp\left\{-1/\sqrt{2}\left(nx_{(n)} - \sum_{j=1}^{n} x_{(j)}\right)\right\}.
\end{aligned}$$

We computed the Bayes factor in Chapter 4 with the intention of comparing it with the results of our approach. However, the use of the improper prior avoid considering the Bayes factor as a validate criterion in this case.

5.6 Logistic versus probit regression model

We return to the problem of testing a logistic against a probit model for the binary outcomes. In chapter 4, we illustrated that when estimating a mixture of these regression models, the posterior estimate of α strongly supports the true model. This means that we can easily distinguish the true model when the sample size is large enough. The analyses are based on the mixture model defined as

$$\mathfrak{M}_\alpha : \alpha^{(\exp(y_i \mathbf{x}^i\theta)/(1+\exp(\mathbf{x}^i\theta)))} + (1-\alpha)\Phi(\mathbf{x}^i(\kappa^{-1}\theta))^{y_i}(1-\Phi(\mathbf{x}^i(\kappa^{-1}\theta)))^{1-y_i}$$

The likelihood and the conditional posterior distributions of θ and α derived from the g-prior and beta prior without considering missing variable ζ can be written as

$$\ell(\theta,\alpha|y,X) = \left(\prod_{i=1}^{n} \alpha^{(\exp(y_i \mathbf{x}^i\theta)/(1+\exp(\mathbf{x}^i\theta)))} + (1-\alpha)\Phi(\mathbf{x}^i(\kappa^{-1}\theta))^{y_i}(1-\Phi(\mathbf{x}^i(\kappa^{-1}\theta)))^{1-y_i}\right)$$

$$\begin{aligned}\pi(\theta|y,X,\alpha) &\propto \prod_{i=1}^{n} \alpha^{(\exp(y_i \mathbf{x}^i\theta)/(1+\exp(\mathbf{x}^i\theta)))} \\ &\quad + (1-\alpha)\Phi(\mathbf{x}^i(\kappa^{-1}\theta))^{y_i}(1-\Phi(\mathbf{x}^i(\kappa^{-1}\theta)))^{1-y_i}\exp(-\theta^T(X^TX)\theta/2n) \\ \pi(\alpha|y,X,\theta) &\propto \prod_{i=1}^{n} \alpha^{(\exp(y_i \mathbf{x}^i\theta)/(1+\exp(\mathbf{x}^i\theta)))} \\ &\quad + (1-\alpha)\Phi(\mathbf{x}^i(\kappa^{-1}\theta))^{y_i}(1-\Phi(\mathbf{x}^i(\kappa^{-1}\theta)))^{1-y_i}(\alpha(1-\alpha))^{a_0-1}\end{aligned}$$

We can use the Metropolis-within-Gibbs algorithm to simulate from the conditional posteriors above in which the parameter θ is simulated according to a random walk multivariate normal distribution $\mathcal{N}(\theta[t], \tau\hat{\Sigma})$ starting from the maximum likelihood estimates. $\hat{\Sigma}$ is the asymptotic covariance matrix of the maximum likelihood estimates of the coefficients. For the **R** code, see [Kamary 2016b]. In chapter 4, two models were analyzed with an explanatory variable. Here, we test the regression models with intercept once over 3 another over 6 explanatory variables. The datasets are simulated once from logistic another from probit model. The explanatory variables are simulated from $\mathcal{N}(0,1)$, $\mathcal{U}(0,1)$,$\mathcal{U}(-1,1)$, $\mathcal{N}(1,4)$, $\mathcal{U}(2,3)$ and $\mathcal{U}(-1,1)$. The approximate Bayes estimates of θ are obtained by running the Metropolis-within-Gibbs with scale $\tau = 1$ over 10^4 iterations. The results summarized in Table 5.6 are slightly close to the true values.

Figure 5.12 gives an assessment of the convergence of the chains obtained for 10^4 data points simulated from the logistic model which is summarized in Table 5.6 in the case where $a_0 = .5$. The row sequences and the autocorrelation graphs illustrate the good mixing behavior of the chains. The posterior distributions of α displayed in Figure 5.13 are related to the four datasets used to estimate the regression coefficients in Table 5.6. Once again, the accumulation of the posterior draws of α is over 1 for the true model even if the number of the explanatory variables is more than 2.

Sample of size $1e4$

Data simulated from logistic model					**Data simulated from probit model**			
θ:	5	1.5	-0.5	2	-4.2	0.9	1.5	1.3
a_0	$\hat{\theta}_0$	$\hat{\theta}_1$	$\hat{\theta}_2$	$\hat{\theta}_3$	$\hat{\theta}_0$	$\hat{\theta}_1$	$\hat{\theta}_2$	$\hat{\theta}_3$
0.1	4.86	1.46	-.49	1.95	-4.01	.92	1.55	-1.38
0.2	4.85	1.47	-.48	1.94	-4.00	.91	1.54	-1.38
0.3	4.86	1.46	-.49	1.96	-4.01	.92	1.54	-1.39
0.4	4.86	1.47	-.48	1.96	-4.01	.91	1.55	-1.38
0.5	4.86	1.47	-.49	1.95	-4.01	.91	1.55	-1.39

Sample of size $2e4$

Data simulated from logistic model								**Data simulated from probit model**						
θ:	3	1.5	-0.5	2	-0.3	1.1	-0.8	-3.5	0.9	1.8	-1.2	0.7	2.6	-5.5
a_0	$\hat{\theta}_0$	$\hat{\theta}_1$	$\hat{\theta}_2$	$\hat{\theta}_3$	$\hat{\theta}_4$	$\hat{\theta}_5$	$\hat{\theta}_6$	$\hat{\theta}_0$	$\hat{\theta}_1$	$\hat{\theta}_2$	$\hat{\theta}_3$	$\hat{\theta}_4$	$\hat{\theta}_5$	$\hat{\theta}_6$
0.1	3.13	1.51	-.57	1.96	-.33	1.07	-.64	-3.7	.88	1.73	-1.2	.69	2.2	-5.3
0.2	3.15	1.51	-.56	1.97	-.33	1.07	-.63	-3.6	.88	1.73	-1.2	.70	2.3	-5.2
0.3	3.16	1.52	-.58	1.96	-.33	1.06	-.63	-3.7	.88	1.73	-1.2	.69	2.3	-5.3
0.4	3.15	1.51	-.56	1.97	-.32	1.07	-.64	-3.6	.89	1.72	-1.2	.69	2.3	-5.2
0.5	3.18	1.52	-.57	1.97	-.33	1.06	-.63	-3.7	.89	1.73	-1.2	.70	2.3	-5.2

Table 5.6: **Logistic versus probit regression** 5.6: Observation information and Bayesian estimate of the regression coefficients, $\hat{\theta}$; Each point estimator is based on $10,000$ MCMC iterations.

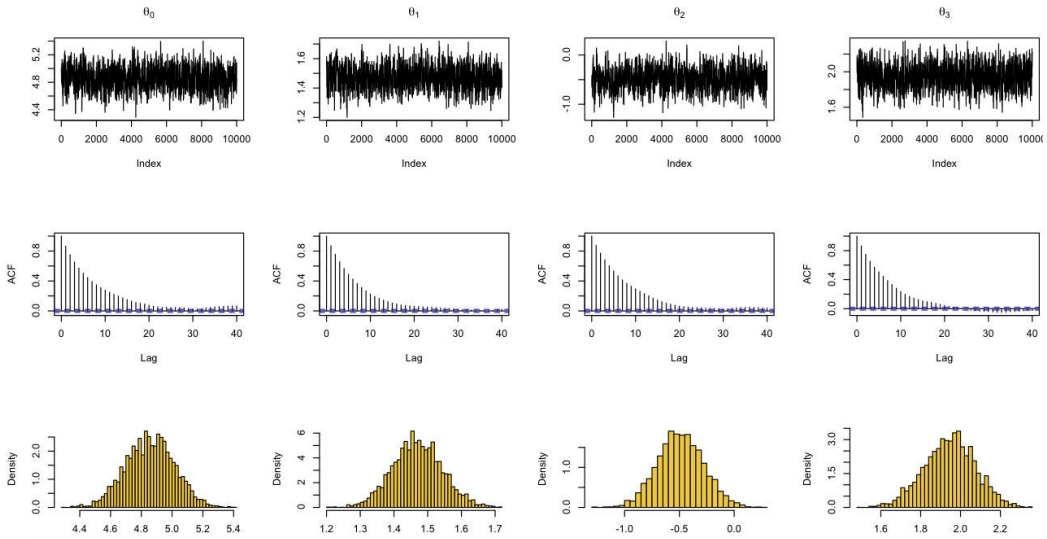


Figure 5.12: **Logistic versus probit** 5.6: *(Top)* Sequence of $\theta[t]$; *(Center)* Autocorrelation over 10^4 iterations; Histogram over the last 9000 iterations. $a_0 = .5$.

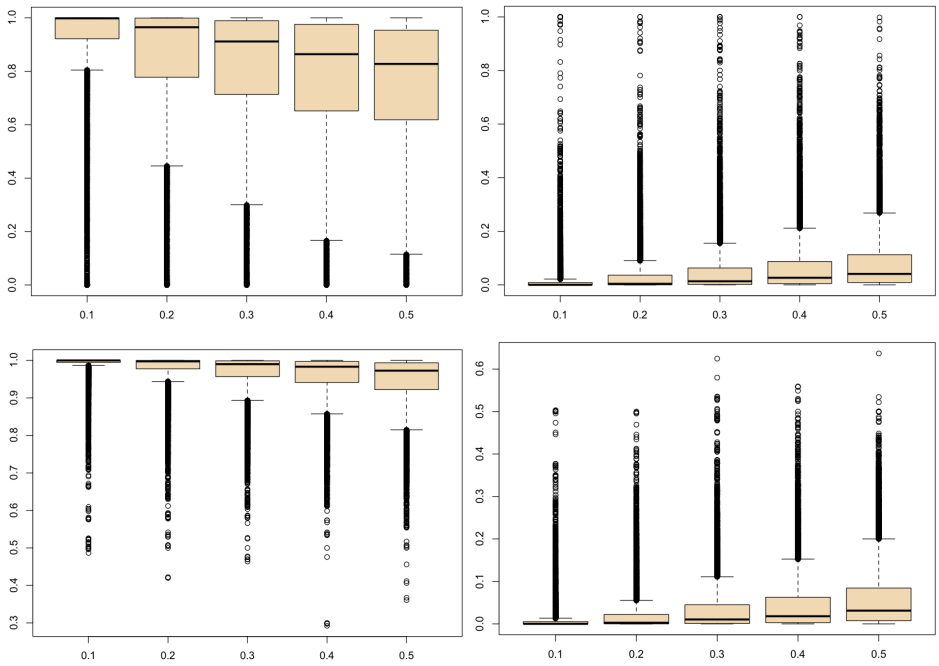


Figure 5.13: **Logistic versus probit** 5.6: Posterior distributions of α in favor of the logistic model based on 10^4 MCMC iterations where $a_0 = .1, .2, .3, .4, .5$. *(Left)* Two datasets of sizes $10^4, 20,000$ simulated from logistic model; *(Right)* Two datasets of sizes $10^4, 20,000$ simulated from probit model that are related to the analyses shown in Table 5.6.

5.7 Variable selection

The use of the mixture model for the variable selection in a Gaussian regression model is considered as a decision problem in that all potential models have to be considered as the mixture components against the mixture weights that ranks them in this context. If k is the number of predictor variables to explain the output y, every subset of explanatory variables can constitute a proper set of explanatory variables for the regression of y and the related model should be considered as a mixture model component. As an example, when $k = 3$, all possible models for y are shown in Table 5.7.

In the case where the mixture model is parametrized in terms of the same potential parameter β, the regression model is denoted by $\mathbf{y}_\zeta = \mathbf{X}_\zeta\beta + \varepsilon$. The likelihood function is

$$\ell(\beta, \sigma^2, \alpha|\mathbf{y}_\zeta, \mathbf{X}_\zeta, \zeta) = \prod_{j=1}^{\gamma} \alpha_j^{\upsilon_j} (\sqrt{2\pi}\sigma)^{-n} \exp\left(-(\mathbf{y}_\zeta - \mathbf{X}_\zeta\beta)^T(\mathbf{y}_\zeta - \mathbf{X}_\zeta\beta)/2\sigma^2\right)$$

where $\upsilon_j = \sum_{i=1}^{n} \mathbb{I}_{\zeta_i=j}; j = 1, \ldots, \gamma$. The joint prior for β, σ^2 and α is the improper prior

$\mathfrak{M}_1 : y_i = \beta_0 + \varepsilon_i$
$\mathfrak{M}_2 : y_i = \beta_0 + \beta_1 X_{i1} + \varepsilon_i$
$\mathfrak{M}_3 : y_i = \beta_0 + \beta_2 X_{i2} + \varepsilon_i$
$\mathfrak{M}_4 : y_i = \beta_0 + \beta_3 X_{i3} + \varepsilon_i$
$\mathfrak{M}_5 : y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i$
$\mathfrak{M}_6 : y_i = \beta_0 + \beta_1 X_{i1} + \beta_3 X_{i3} + \varepsilon_i$
$\mathfrak{M}_7 : y_i = \beta_0 + \beta_2 X_{i2} + \beta_3 X_{i3} + \varepsilon_i$
$\mathfrak{M}_8 : y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \varepsilon_i$
$\mathfrak{M}_9 : y_i = \beta_1 X_{i1} + \varepsilon_i$
$\mathfrak{M}_{10} : y_i = \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i$
$\mathfrak{M}_{11} : y_i = \beta_1 X_{i1} + \beta_3 X_{i3} + \varepsilon_i$
$\mathfrak{M}_{12} : y_i = \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \varepsilon_i$
$\mathfrak{M}_{13} : y_i = \beta_2 X_{i2} + \varepsilon_i$
$\mathfrak{M}_{14} : y_i = \beta_2 X_{i2} + \beta_3 X_{i3} + \varepsilon_i$
$\mathfrak{M}_{15} : y_i = \beta_3 X_{i3} + \varepsilon_i$

Table 5.7: **Variable selection** 5.7: All potential models for regression y when $k = 3$ and $i = 1, \ldots, n$.

$$\pi(\beta, \sigma^2, \alpha) \propto (\sigma^2)^{-k+1/2-1} \exp\left(-(\beta - M_{k+1})^T (X^T X)(\beta - M_{k+1}) / 2c\sigma^2\right) \prod_{j=1}^{\gamma} \alpha_j^{a_0 - 1}$$

where $\alpha = (\alpha_1, \ldots, \alpha_k)$ is the vector of γ component weights. The conditional posterior distributions of β, σ^2 and α given latent variable ζ can be computed as follows

$$\begin{aligned}
\pi(\beta|\sigma^2, \alpha, \mathbf{y}_\zeta, \mathbf{X}_\zeta, \zeta) &\propto \exp\left(-(\mathbf{y}_\zeta - \mathbf{X}_\zeta\beta)^T(\mathbf{y}_\zeta - \mathbf{X}_\zeta\beta)/2\sigma^2 - (\beta - M_{k+1})^T(X^TX)(\beta - M_{k+1})/2c\sigma^2\right) \\
&\propto \exp\left(-\binom{\mathbf{X}_\zeta\beta - \mathbf{y}_\zeta}{\beta - M_{k+1}}^T \begin{pmatrix} \sigma^2 I_n & 0 \\ 0 & c\sigma^2(X^TX)^{-1} \end{pmatrix}^{-1} \binom{\mathbf{X}_\zeta\beta - \mathbf{y}_\zeta}{\beta - M_{k+1}} / 2\right) \\
&\propto \exp\left(-\{\binom{\mathbf{X}_\zeta}{I_{k+1}}\beta - \binom{\mathbf{y}_\zeta}{M_{k+1}}\}^T \begin{pmatrix} \sigma^2 I_n & 0 \\ 0 & c\sigma^2(X^TX)^{-1} \end{pmatrix}^{-1} \{\binom{\mathbf{X}_\zeta}{I_{k+1}}\beta - \binom{\mathbf{y}_\zeta}{M_{k+1}}\}/2\right)
\end{aligned}$$

The last term of the equation above yields the following posterior density for β

$$\begin{aligned}
\pi(\beta|\sigma^2, \alpha, \mathbf{y}_\zeta, \mathbf{X}_\zeta, \zeta) &\propto \exp\left(-(\beta - \bar{\beta})^T(\sigma^{-2}{\mathbf{X}_\zeta}^T\mathbf{X}_\zeta + (X^TX))(\beta - \bar{\beta})/2\right) \\
\bar{\beta} &= \{\sigma^{-2}{\mathbf{X}_\zeta}^T\mathbf{X}_\zeta + (c\sigma^2)^{-1}X^TX\}^{-1}\{\sigma^{-2}{\mathbf{X}_\zeta}^T\mathbf{y}_\zeta + (c\sigma^2)^{-1}X^TXM_{k+1}\}
\end{aligned}$$

which implies the multivariate Gaussian distribution obtained in Chapter 4. By dropping the term that does not involve σ^2 from the multiplication of likelihood and the joint prior of β, σ^2 and α, we will obtain

$$\pi(\sigma^2|\beta, \alpha, \mathbf{y}_\zeta, \mathbf{X}_\zeta, \zeta) \propto (\sigma^2)^{-(n+k+1)/2-1} \exp\left(-(\mathbf{y}_\zeta - \mathbf{X}_\zeta\beta)^T(\mathbf{y}_\zeta - \mathbf{X}_\zeta\beta)/2 - (\beta - M_{k+1})^T(X^TX)(\beta - M_{k+1})/2c/\sigma^2\right)$$

from which we can easily deduce inverse-gamma distribution for σ^2. For mixture weights, we have

$$\pi(\alpha|\zeta) = \prod_{j=1}^{\gamma} \alpha_j^{v_j + a_0 - 1}$$

which results in a Dirichlet distribution with the concentration parameter $\upsilon_j + a_0; j = 1, \ldots, \gamma$. In Chapter 4, three stages Gibbs sampler algorithm is applied in order to obtain samples from the conditional posterior distributions above and the corresponding **R** code is available in [Kamary 2016b].

The example 4.3.6 illustrates the efficient performance of the Gibbs sampler in this context because the weight of the potential model from which the outputs are simulated converges to 1 with the sample size. However, in the case where the sample size is small, Gibbs sampler needs a high number of MCMC iterations to get the convergence of the chains and that extremely increases the system time. As an example, we analyze **caterpillar** dataset extracted from a 1973 study on pine processionary caterpillars [Marin 2007] when the response variable is the logarithmic transform of the average number of nests of caterpillars per tree. Three explanatory variables are considered for the regression model, which are supposed to be x_1: the altitude, x_2: the slope and x_3: the number of the pines in the area. According to the classical analysis, the coefficient β_3 is not significant. This means that the appropriate model to y would have the form as $\mathfrak{M}_5$ in Table 5.7 in this case and the maximum likelihood estimate of the components, $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2$ is $4.94, -0.002, -0.035$.

After running the Gibbs sampler algorithm by considering c equal to the sample size and $M_{k+1} = 0_4$, the convergence of the chains is achieved when the number of the iterations is 10^5 as shown in Figure 5.15. In this case, the posterior distributions of $\beta_0, \beta_1, \beta_2, \beta_3$ are centered on the same values obtained by maximum likelihood method. α_5 is concentrated over 1 which indicates that the posterior draws support the model $\mathfrak{M}_5$ for the output y. The result that is in agreement with the classical conclusion.

We obtain the same results when we implement the Metropolis-within-Gibbs algorithm to sample from the posterior distributions of the mixture parameters, except that we do not need to consider a large number of iterations and the convergence is achieved by producing 10^4 MCMC iterations. The code in **R** is also shown in [Kamary 2016b] in which the parameters β, σ^2 and α are simulated from multivariate normal, inverse-gamma and Dirichlet proposal distributions when the acceptance probabilities are based on the following posteriors

$$\pi(\beta|\sigma^2, \alpha, y, X) \propto \sum_{j=1}^{\gamma} \alpha_j \exp\left(-(y-X^j\beta^j)^T(y-X^j\beta^j)/2\sigma^2\right) \exp\left(-(\beta-M_{k+1})^T(X^TX)(\beta-M_{k+1})/2c\sigma^2\right)$$

$$\pi(\sigma^2|\beta, \alpha, y, X) \propto (\sigma^2)^{-n+k+1/2-1} \sum_{j=1}^{\gamma} \alpha_j \exp\left(-(y-X^j\beta^j)^T(y-X^j\beta^j)/2\sigma^2\right)$$
$$\exp\left(-(\beta-M_{k+1})^T(X^TX)(\beta-M_{k+1})/2c\sigma^2\right)$$

$$\pi(\alpha|\beta, \sigma^2, y, X) \propto \sum_{j=1}^{\gamma} \alpha_j \exp\left(-(y-X^j\beta^j)^T(y-X^j\beta^j)/2\sigma^2\right) \prod_{j=1}^{\gamma} \alpha_j^{a_0-1}.$$

The second case studied in variable selection problem is related to the condition that the possible regression models are independent in the sense that each model

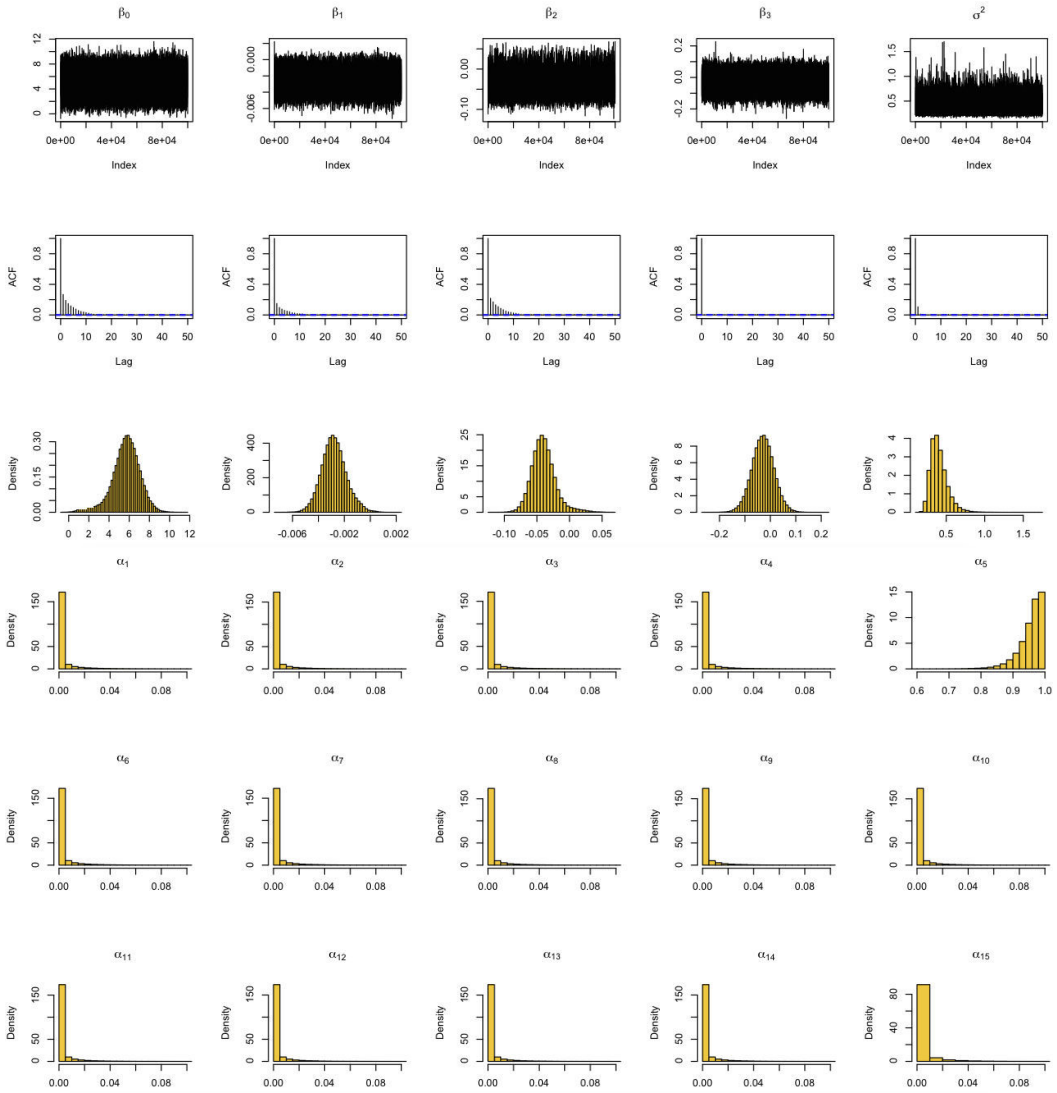


Figure 5.14: **Case 1. Caterpillar dataset** 5.7: *(Top)* Sequences of $1e5$ Gibbs sampler iterations; Empirical autocorrelations using `acf` **R** function; Histograms of the last $90,000$ iterations. *(Bottom)* Histograms of the posterior distributions of $\alpha_1, \ldots, \alpha_{15}$ based on 10^5 MCMC iterations when $a_0 = .1$.

has the regression coefficients that should be independently estimated from those of the other models. In this case, the number of the parameters rises very quickly by increasing the number of the explanatory variables. It means that using a large number of explanatory variables requires a huge number of parameters to be estimated. Consequently, the time system of the MCMC programs increases a lot. When we have three explanatory variables for the response y, 32 regression coefficients should be estimated for 15 potential models shown in Table 5.8.

The regression model is defined as $y = X^j \beta_{\mathfrak{M}_j} + \varepsilon$ for each model, $\mathfrak{M}_j$ and the likelihood conditional on missing variable ζ can therefore be written by

$\mathfrak{M}_1 : y_i = \beta_0^1 + \varepsilon_i$
$\mathfrak{M}_2 : y_i = \beta_0^2 + \beta_1^2 X_{i1} + \varepsilon_i$
$\mathfrak{M}_3 : y_i = \beta_0^3 + \beta_2^3 X_{i2} + \varepsilon_i$
$\mathfrak{M}_4 : y_i = \beta_0^4 + \beta_3^4 X_{i3} + \varepsilon_i$
$\mathfrak{M}_5 : y_i = \beta_0^5 + \beta_1^5 X_{i1} + \beta_2^5 X_{i2} + \varepsilon_i$
$\mathfrak{M}_6 : y_i = \beta_0^6 + \beta_1^6 X_{i1} + \beta_3^6 X_{i3} + \varepsilon_i$
$\mathfrak{M}_7 : y_i = \beta_0^7 + \beta_2^7 X_{i2} + \beta_3^7 X_{i3} + \varepsilon_i$
$\mathfrak{M}_8 : y_i = \beta_0^8 + \beta_1^8 X_{i1} + \beta_2^8 X_{i2} + \beta_3^8 X_{i3} + \varepsilon_i$
$\mathfrak{M}_9 : y_i = \beta_1^9 X_{i1} + \varepsilon_i$
$\mathfrak{M}_{10} : y_i = \beta_1^{10} X_{i1} + \beta_2^{10} X_{i2} + \varepsilon_i$
$\mathfrak{M}_{11} : y_i = \beta_1^{11} X_{i1} + \beta_3^{11} X_{i3} + \varepsilon_i$
$\mathfrak{M}_{12} : y_i = \beta_1^{12} X_{i1} + \beta_2^{12} X_{i2} + \beta_3^{12} X_{i3} + \varepsilon_i$
$\mathfrak{M}_{13} : y_i = \beta_2^{13} X_{i2} + \varepsilon_i$
$\mathfrak{M}_{14} : y_i = \beta_2^{14} X_{i2} + \beta_3^{14} X_{i3} + \varepsilon_i$
$\mathfrak{M}_{15} : y_i = \beta_3^{15} X_{i3} + \varepsilon_i$

Table 5.8: **Variable selection** 5.7: All potential models for regression y when $k = 3$ and $i = 1, \ldots, n$.

$$\ell(\beta, \sigma^2, \alpha | y, X, \zeta) = (2\pi\sigma^2)^{-n/2} \prod_{j=1}^{\gamma} \alpha_j^{\upsilon_j} \exp\left(-\sum_{i=1}^{n} (y_{i;\zeta_i=j} - X^j_{i;\zeta_i=j}\beta_{\mathfrak{M}_j})^T (y_{i;\zeta_i=j} - X^j_{i;\zeta_i=j}\beta_{\mathfrak{M}_j}) / 2\sigma^2\right)$$

where $\sum_{j=1}^{\gamma} \upsilon_j = n$ and the joint prior of β, α and σ^2 is

$$\pi(\beta, \sigma^2, \alpha) \propto (\sigma^2)^{-s/2-1} \exp\left(-\sum_{j=1}^{\gamma} (\beta_{\mathfrak{M}_j} - M_j)^T (\{X^j\}^T X^j)(\beta_{\mathfrak{M}_j} - M_j) / 2c\sigma^2\right) \prod_{j=1}^{\gamma} \alpha_j^{a_0 - 1}.$$

The conditional posterior distribution of $\beta_{\mathfrak{M}_j}; j = 1, \ldots, \gamma$ is given by

$$\begin{aligned}
\pi(\beta_{\mathfrak{M}_j} | \sigma^2, \alpha, y, X, \zeta) &\propto \exp\left(-\sum_{i=1}^{n} (y_{i;\zeta_i=j} - X^j_{i;\zeta_i=j}\beta_{\mathfrak{M}_j})^T (y_{i;\zeta_i=j} - X^j_{i;\zeta_i=j}\beta_{\mathfrak{M}_j}) / 2\sigma^2\right) \\
&\quad \exp\left(-(\beta_{\mathfrak{M}_j} - M_j)^T (\{X^j\}^T X^j)(\beta_{\mathfrak{M}_j} - M_j) / 2c\sigma^2\right) \\
&\propto \exp\left(-(\beta_{\mathfrak{M}_j} - \eta)^T (\{X^j\}^T X^j / c\sigma^2 + \{X^j_{i;\zeta_i=j}\}^T X_{i;\zeta_i=j} / \sigma^2)(\beta_{\mathfrak{M}_j} - \eta) / 2\right) \\
\eta &= \left(\{X^j\}^T X^j / c\sigma^2 + \{X^j_{i;\zeta_i=j}\}^T X_{i;\zeta_i=j} / \sigma^2\right)^{-1} \left(\{X^j\}^T X^j M_j / c\sigma^2 + X^j_{i;\zeta_i=j} y_{i;\zeta_i=j} / \sigma^2\right)
\end{aligned}$$

which leads us to deduce the multivariate Gaussian distribution for the regression model coefficients as defined in the variable selection section 4.3.6. For σ^2, we can write

$$\begin{aligned}
\pi(\sigma^2 | \beta, \alpha, y, X, \zeta) \propto (\sigma^2)^{-(n+s)/2} &\exp\left(-\sum_{j=1}^{\gamma} \sum_{i=1}^{n} (y_{i;\zeta_i=j} - X^j_{i;\zeta_i=j}\beta_{\mathfrak{M}_j})^T (y_{i;\zeta_i=j} - X^j_{i;\zeta_i=j}\beta_{\mathfrak{M}_j}) / 2\sigma^2\right) \\
&\exp\left(-\sum_{j=1}^{\gamma} (\beta_{\mathfrak{M}_j} - M_j)^T (\{X^j\}^T X^j)(\beta_{\mathfrak{M}_j} - M_j) / 2c\sigma^2\right)
\end{aligned}$$

that results in inverse-gamma distribution with the parameters a, b as pointed out in 4.3.6 and the posterior density of α is the same as the **Case 1.**. The **R** program related to the Gibbs sampler algorithm that samples from the conditional

	$\mathbb{E}^{\pi}(\beta_0\|y,X)$	$\mathbb{E}^{\pi}(\beta_1\|y,X)$	$\mathbb{E}^{\pi}(\beta_2\|y,X)$	$\mathbb{E}^{\pi}(\sigma^2\|y,X)$	$\mathbb{E}^{\pi}(\alpha_5\|y,X)$
Case 1.	5.18	-0.003	-0.039	0.54	0.96
Case 2.	5.21	-0.003	-0.051	0.56	0.77
MLE.	4.94	-0.002	-0.035	0.65	

Table 5.9: **Variable selection** 5.7: Point estimate of the regression coefficients of the model $\mathfrak{M}_5$, σ^2 and α_5 based on $10,000$ MCMC iterations when $a_0 = 0.5$.

posteriors of β, σ^2 and α is indicated in [Kamary 2016b]. The analyses spoken of in the **Case 2.** part of the variable selection section in Chapter 4 are based on this Gibbs sampler algorithm. However, running this program is time consuming even for a dataset with small sample sizes and with 10^4 MCMC iterations. For example, for the **Caterpillar** dataset, the convergence of the chains is achieved when we produce a large number of MCMC iterations. However, time system is much more than the one for the **Case 1.**. We can also use the Metropolis-within-Gibbs algorithm in this case where the acceptance probability of the proposal distributions of α, σ^2 and $\beta_{\mathfrak{M}_j}; j = 1, \ldots, \gamma$ are based on the following posterior distributions

$$\begin{aligned}
\pi(\beta_{\mathfrak{M}_j}|\sigma^2,\alpha,y,X) &\propto \sum_{j=1}^{\gamma} \alpha_j \exp\left(-(y-X^j\beta_{\mathfrak{M}_j})^T(y-X^j\beta_{\mathfrak{M}_j})/2\sigma^2\right) \\
&\quad \exp\left(-(\beta_{\mathfrak{M}_j}-M_j)^T(\{X^j\}^TX^j)(\beta_{\mathfrak{M}_j}-M_j)/2c\sigma^2\right) \\
\pi(\sigma^2|\beta,\alpha,y,X) &\propto (\sigma^2)^{-(n+s)/2} \sum_{j=1}^{\gamma} \alpha_j \exp\left(-(y-X^j\beta_{\mathfrak{M}_j})^T(y-X^j\beta_{\mathfrak{M}_j})/2\sigma^2\right) \\
&\quad \exp\left(-\sum_{j=1}^{\gamma}(\beta_{\mathfrak{M}_j}-M_j)^T(\{X^j\}^TX^j)(\beta_{\mathfrak{M}_j}-M_j)/2c\sigma^2\right) \\
\pi(\alpha|\beta,\sigma^2,y,X) &\propto \sum_{j=1}^{\gamma} \alpha_j \exp\left(-(y-X^j\beta_{\mathfrak{M}_j})^T(y-X^j\beta_{\mathfrak{M}_j})/2\sigma^2\right) \prod_{j=1}^{\gamma} \alpha_j^{a_0-1}.
\end{aligned}$$

The related code in **R** is pointed out in [Kamary 2016b]. After running the algorithm for the **Caterpillar** dataset, a graphical convergence check for the posterior draws of the regression coefficients is shown in Figures 5.15 and 5.16 that illustrate the Markov chains have stabilized and mixed very well in this case. The Bayes estimates of the regression components of the model $\mathfrak{M}_5$ and the maximum likelihood estimates are displayed in Table 5.9. The Bayesian outputs of both cases are very close to the maximum likelihood estimates while the posterior estimate of α_5 is also strongly in favor of $\mathfrak{M}_5$ in both cases. Note that in **Case 1.**, the point estimate of α_5 is closer to 1 than in **Case 2.**.

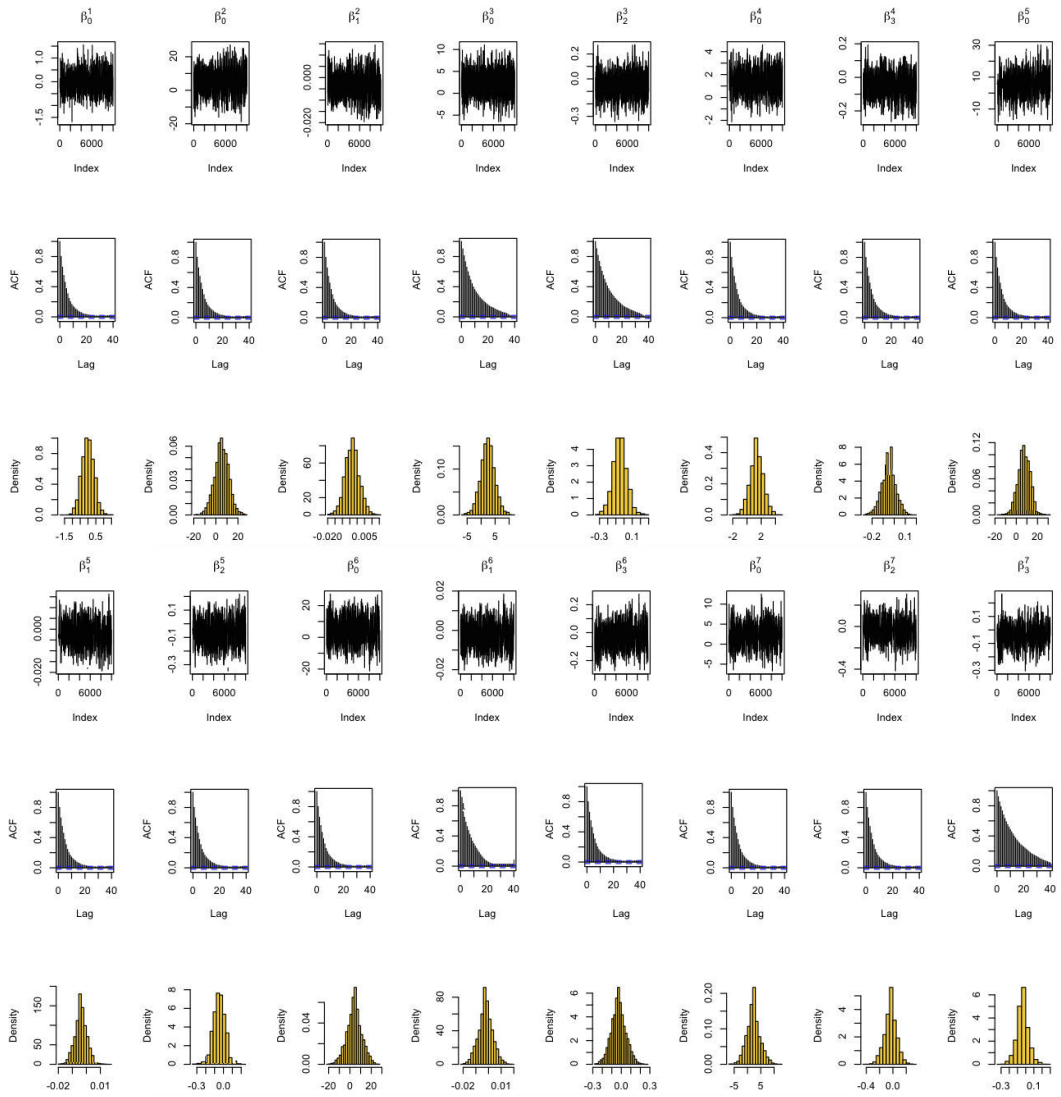


Figure 5.15: **Case 2. Caterpillar dataset** 5.7: Sequences, autocorrelations and histograms of 10^4 Metropolis-within-Gibbs iterations for the regression coefficients β^j of the model $\mathfrak{M}_j; j = 1, \ldots, 7$ shown in Table 5.8 when $a_0 = .5$.

5.8 Propriety of the posterior in the case study of Section 4

To prove the propriety of the posterior it is enough to prove the propriety of the subposterior distribution associated to each component since the parameter (θ, σ) is shared between the components. It is known that in the case of a Gaussian model $\mathcal{N}(\theta, \sigma)$ the posterior associated to the prior $\pi(\theta, \sigma) = 1/\sigma$ is proper as soon as $n \geq 2$ and at least 2 observations are distinct. We now show that this results extends to the case of a Gumbel(θ, σ^2) and of a Logistic(θ, σ). Let I denote the

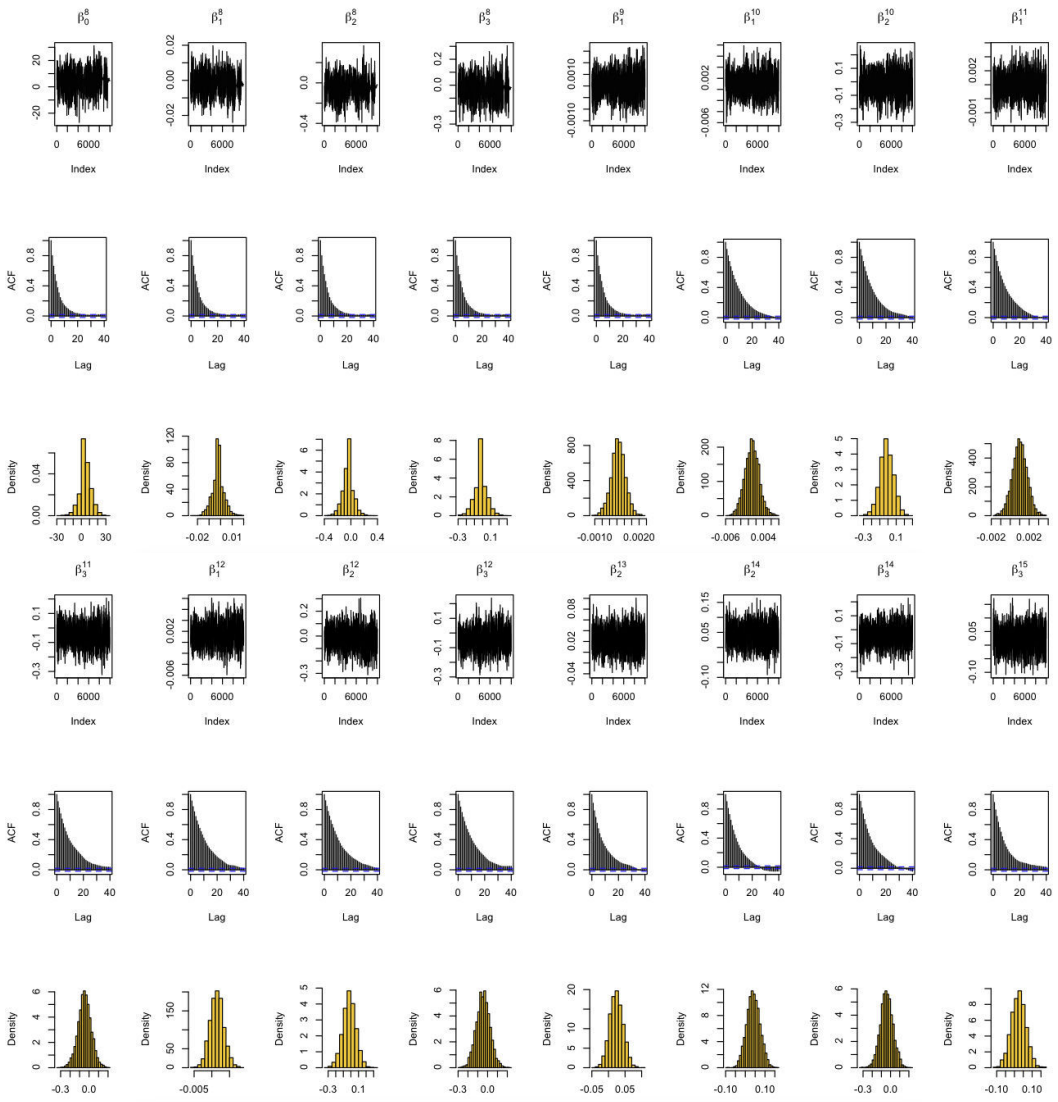


Figure 5.16: **Case 2. Caterpillar dataset** 5.7: Sequences, autocorrelations and histograms of 10^4 Metropolis-within-Gibbs sampler iterations for the regression coefficients β^j of the model $\mathfrak{M}_j; j = 8, \ldots, 15$ shown in Table 5.8 when $a_0 = .5$.

marginal likelihood, in the Gumbel case.

$$I = \int_{\mathbb{R}\times\mathbb{R}^+} \frac{1}{\sigma^{n+1}} \exp\left\{-\sum_{i=1}^{n}(Y_i-\theta)/\sigma\right\} \exp\left\{-\sum_{i=1}^{n} e^{-(Y_i-\theta)/\sigma}\right\} \mathrm{d}\theta\mathrm{d}\sigma$$

We set $\gamma_n = \sum_{i=1}^{n} e^{-Y_i/\sigma}$, then

$$I(\sigma) = e^{-n\bar{Y}_n/\sigma} \int_{\mathbb{R}} \exp(n\theta/\sigma) \exp(-\gamma_n e^{\theta/\sigma})\mathrm{d}\theta \propto e^{-n\bar{Y}_n/\sigma}\sigma \int_{\mathbb{R}} u^{n-1}\exp(-\gamma_n u)\mathrm{d}u$$

$$\propto e^{-n\bar{Y}_n/\sigma}\sigma\gamma_n^{-n} \propto \exp\left(-\frac{n\bar{Y}_n}{\sigma}\right)\left(\sum_{i=1}^{n} e^{-\frac{Y_i}{\sigma}}\right)^{-n}.$$

So

$$I \propto \int_{\mathbb{R}^+} \sigma^{-n} \exp\left(-\frac{n\bar{Y}_n}{\sigma}\right) \left(\sum_{i=1}^{n} e^{-\frac{Y_i}{\sigma}}\right)^{-n} \mathrm{d}\sigma = \int_{\mathbb{R}^+} \sigma^{-n} \left(\sum_{i=1}^{n} e^{-\frac{1}{\sigma}\left(Y_i - \bar{Y}_n\right)}\right)^{-n} \mathrm{d}\sigma < +\infty$$

if only if $\min_i \left(Y_i - \bar{Y}_n\right) < 0$. This is almost surely true when $n \geq 2$. We now study the Logistic case, using similar computations, so that

$$\begin{aligned}
I &\propto \int_{\mathbb{R}\times\mathbb{R}^+} \frac{e^{-n\frac{\bar{Y}_n}{\sigma}}}{\sigma^{(n+1)}} \frac{e^{\theta n/\sigma}}{\prod_i (1 + e^{-Y_i/\sigma} e^{\theta/\sigma})^2} \mathrm{d}\theta \mathrm{d}\sigma \propto \int_{\mathbb{R}^+} \frac{e^{-n\frac{\bar{Y}_n}{\sigma}}}{\sigma^n} \int_{\mathbb{R}} \frac{u^{n-1}}{\prod_i (1 + e^{-Y_i/\sigma} u)^2} \mathrm{d}u \mathrm{d}\sigma \\
&\leq \int_{\mathbb{R}^+} \frac{e^{-n\frac{\bar{Y}_n}{\sigma}}}{\sigma^n} \int_{\mathbb{R}} \frac{u^{n-1}}{\left(1 + u^{n-1} e^{-(n-1)\bar{Y}_n/\sigma} \max_i e^{-(Y_i - \bar{Y}_n)/\sigma}\right)^2} \mathrm{d}u \\
&\propto \int_{\mathbb{R}^+} \frac{1}{\sigma^n} \int_{\mathbb{R}} \frac{u^{n-1}}{\left(1 + u^{n-1} \max_i e^{-(Y_i - \bar{Y}_n)/\sigma}\right)^2} \mathrm{d}u \\
&\propto \int_{\mathbb{R}^+} \frac{1}{\sigma^n} e^{2n \min_i (Y_i - \bar{Y}_n)/\sigma} \mathrm{d}\sigma < +\infty
\end{aligned}$$

if and only if $\min_i \left(Y_i - \bar{Y}_n\right) < 0$. Thus means the observations cannot be all equal.

Chapter 6

Non-informative reparameterisations for location-scale mixtures

Joint work with Kate Lee and Christian P. Robert

Abstract

While mixtures of Gaussian distributions have been studied for more than a century (Pearson, 1894), the construction of a reference Bayesian analysis of those models still remains unsolved, with a general prohibition of the usage of improper priors [Frühwirth-Schnatter 2006] due to the ill-posed nature of such statistical objects. This difficulty is usually bypassed by an empirical Bayes resolution [Richardson 1997]. By creating a new parameterisation centered on the mean and variance of the mixture distribution itself, we are able to develop here a genuine non-informative prior for Gaussian mixtures with an arbitrary number of components. We demonstrate that the posterior distribution associated with this prior is almost surely proper and provide MCMC implementations that exhibit the expected exchangeability. While we only study here the Gaussian case, extension to other classes of location-scale mixtures is straightforward.

Keywords: Noninformative prior, improper prior, Mixture of distributions, Bayesian analysis, Dirichlet prior, exchangeability, plane-sphere intersection, polar coordinates

6.1 Introduction

A mixture density is traditionally represented as a weighted average of densities from standard families, i.e.,

$$f(x|\boldsymbol{\theta}, \mathbf{p}) = \sum_{i=1}^{k} p_i f(x|\theta_i) \qquad \sum_{i=1}^{k} p_i = 1\,. \tag{6.1}$$

Each component of the mixture is characterized by a component-wise parameter θ_i and the weights p_i of those components translate the importance of each of those components in the model.

This particular representation gives a separate meaning to each component through its parameter θ_i, even though there is a well-known lack of identifiability in

such models, due to the invariance of the sum by permutation of the indices. This issue relates to the equally well-known "label switching" phenomenon in the Bayesian approach to the model, which pertains both to inference and to simulation of the corresponding posterior [Celeux 2000, Stephens 2000, Frühwirth-Schnatter 2001, Frühwirth-Schnatter 2004, Jasra 2005]. From this Bayesian viewpoint, the choice of the prior distribution on the component parameters is quite open, the only constraint being that the corresponding posterior is proper [Diebolt 1994, Frühwirth-Schnatter 2004]. [Diebolt 1994] and [Wasserman 1999] discussed the alternative approach of *imposing* proper posteriors on improper priors by banning almost empty components from the likelihood function. While consistent, this approach induces dependence between the observations, higher computational costs and is not handling overfitting very well. It has therefore seen little following.

The prior distribution on the weights p_i is equally open for choice, but a standard version is a Dirichlet distribution with common hyperparameter a, $\mathrm{Dir}(a,\ldots,a)$. Recently, [Rousseau 2011] demonstrated that the choice of this hyperparameter a relates to the inference on the total number of components, namely that a small enough value of a manages to handle over-fitted mixtures in a convergent manner. In a Bayesian non-parametric modeling, [Griffin 2010] showed that the prior on the weights may have a higher impact when inferring about the number of components, relative to the prior on the component-specific parameters. As indicated above, the prior distribution on the θ_i's has received less attention and conjugate choices are most standard, since they facilitate simulation via Gibbs samplers [Diebolt 1990, Escobar 1995, Richardson 1997] if not estimation, since posterior moments remain unavailable in closed form. In addition, [Richardson 1997] among others proposed data-based priors that derive some hyperparameters as functions of the data, towards an automatic scaling of such priors. An R package, bayesm [Rossi 2010] incorporates some of those ideas. In the case when $\theta_i = (\mu_i, \sigma_i)$ is a location-scale parameter, [Mengersen 1996] proposed a reparameterisation of (6.1) that express each component as a local perturbation of the previous one, namely ($i > 1$)

$$\mu_i = \mu_{i-1} + \sigma_{i-1}\delta_i\,, \quad \sigma_i = \tau_i\sigma_{i-1}\,,\ \tau_i < 1\,,$$

with μ_1 and σ_1 being the reference values. Based on this reparameterisation, [Robert 1998] established that a particular improper prior on (μ_1, σ_1) still leads to a proper prior. We propose here to modify further this reparameterisation towards using the global mean and global variance of the mixture distribution as reference location and scale, respectively. This modification has foundational consequences in terms of using improper and non-informative priors over mixtures, in sharp contrast with the existing literature (see, e.g. [Diebolt 1993, Diebolt 1994, O'Hagan 1994, Wasserman 1999]).

Bayesian computing for mixtures covers a wide variety of proposals, starting with the introduction of the Gibbs sampler [Diebolt 1990, Gelman 1990, Escobar 1995], some concerned with approximations [Roeder 1990, Wasserman 1999] and MCMC features [Richardson 1997, Celeux 2000, Casella 2002], and others with asymptotic

justifications, in particular when over-fitting mixtures [Rousseau 2011, Kamary 2014], but most attempting to overcome the methodological hurdles in estimating mixture model [Chib 1995, Neal 1999, Berkhof 2003, Marin 2005, Frühwirth-Schnatter 2006, Lee 2009, Mengersen 2011].

In this paper, we introduce and study the global mean-variance reparameterisation (Section 6.2), which main consequence is to constrain all other parameters to a compact space. We study several possible parameterisations of that kind and demonstrate that the improper Jeffreys-like prior associated with them is proper. In Section 6.3, we propose some MCMC implementation to estimate the parameters of the mixture, discussing label switching (Section 6.3.2) and its resolution by tempering. Extensions to non-Gaussian mixtures are briefly discussed in Section 6.6.

6.2 Mixture representation

6.2.1 Mean-variance reparameterisation

Let us first recall how both mean and variance of a mixture distribution can be represented in terms of the mean and variance parameters of the component of the mixture:

Lemma 1 *If μ_i and σ_i^2 denote the mean and variance of the distribution with density $f(\cdot|\theta_i)$, respectively, the mean of the mixture distribution (6.1) is given by*

$$\mathbb{E}_{\boldsymbol{\theta},\mathbf{p}}[X] = \sum_{i=1}^{k} p_i \mu_i$$

and its variance by

$$\mathrm{var}_{\boldsymbol{\theta},\mathbf{p}}(X) = \sum_{i=1}^{k} p_i \sigma_i^2 + \sum_{i=1}^{k} p_i (\mu_i^2 - \mathbb{E}_{\boldsymbol{\theta},\mathbf{p}}[X]^2)$$

Proof: The population mean given by

$$\mathbb{E}_{\boldsymbol{\theta},\mathbf{p}}[X] = \sum_{i=1}^{k} p_i \mathbb{E}_{f(\cdot|\theta_i)}[X] = \sum_{i=1}^{k} p_i \mu_i$$

where $\mathbb{E}_{f(\cdot|\theta_i)}[X]$ is the expected value component i. Similarly, the population variance is given by

$$\mathrm{var}_{\boldsymbol{\theta},\mathbf{p}}(X) = \sum_{i=1}^{k} p_i \mathbb{E}_{f(\cdot|\theta_i)}[X^2] - \mathbb{E}_{\boldsymbol{\theta},\mathbf{p}}[X]^2 = \sum_{i=1}^{k} p_i (\sigma_i^2 + \mu_i^2) - \mathbb{E}_{\boldsymbol{\theta},\mathbf{p}}[X]^2 \,,$$

which concludes the proof □

For any location-scale mixture, we then propose a reparameterisation of the mixture model that starts by scaling all parameters in terms of its global mean μ and global variance σ^2. For instance, we can switch to the representation

$$\mu_i = \mu + \sigma\alpha_i \quad \text{and} \quad \sigma_i = \sigma\tau_i \tag{6.2}$$

of the component-wise parameters, where $\tau_i > 0$ and $\alpha_i \in \mathbb{R}$. This is formally equivalent to the reparameterisation of [Mengersen 1996], except that they put no special meaning on the global mean and variance parameters. Once the global mean and variance are set, this imposes natural constraints on the other parameters of the model. For instance, setting the global variance to σ^2 implies that $(\mu_1, \ldots, \mu_k, \sigma_1, \ldots, \sigma_k)$ belongs to a specific ellipse conditional on the weights and σ^2, by virtue of Lemma 1.

Considering the α_i's and the τ_i's in (6.2) as the new parameters of the components, the following result states that the global mean and variance parameters are the sole freely varying parameters. In other words, once both the global mean and variance are set, there exists a parameterisation such that all remaining parameters of a mixture distribution are restricted to a compact set, which is most helpful in selecting a non-informative prior distribution.

Lemma 2 *The parameters α_i and τ_i in (6.2) are constrained by*

$$\sum_{i=1}^{k} p_i\alpha_i = 0 \quad \textit{and} \quad \sum_{i=1}^{k} p_i\tau_i^2 + \sum_{i=1}^{k} p_i\alpha_i^2 = 1\,.$$

Proof: The result is a trivial consequence of Lemma 1. The population mean is

$$\mathbb{E}_{\boldsymbol{\theta},\mathbf{p}}[X] = \sum_{i=1}^{k} p_i\mu_i = \sum_{i=1}^{k} p_i(\mu + \sigma\alpha_i) = \mu + \sum_{i=1}^{k} p_i\alpha_i = \mu$$

and the first constraint follows. The population variance is

$$\begin{aligned}
\text{var}_{\boldsymbol{\theta},\mathbf{p}}(X) &= \sum_{i=1}^{k} p_i\sigma_i^2 + \sum_{i=1}^{k} p_i(\mu_i^2 - \mathbb{E}_{\boldsymbol{\theta},\mathbf{p}}[X]^2) \\
&= \sum_{i=1}^{k} p_i\sigma^2\tau_i^2 + \sum_{i=1}^{k} p_i p_i(\mu^2 + 2\sigma\mu\alpha_i + \sigma^2\alpha_i^2 - \mu^2) \\
&= \sum_{i=1}^{k} p_i\sigma^2\tau_i^2 + \sum_{i=1}^{k} p_i\sigma^2\alpha_i^2 = \sigma^2
\end{aligned}$$

The last equation simplifies to the second constraint above. □

6.2.2 Reference priors

The constraints in Lemma 2 define a set of values of $(p_1, \ldots, p_k, \alpha, \ldots, \alpha, \tau, \ldots, \tau)$ that is obviously compact. From a Bayesian perspective, this allows for the call to

uniform and other non-informative proper priors, conditional on (μ, σ). Furthermore, since (μ, σ) is a location-scale parameter, we may invoke [Jeffreys 1939] to use the Jeffreys prior $\pi(\mu, \sigma) = 1/\sigma$ on this parameter, even though this is not the genuine Jeffreys prior for the mixture model [Grazian 2015]. In the same spirit as [Robert 1998] who established properness of the posterior distribution derived by [Mengersen 1996], we now establish that this choice of prior produces a proper posterior distribution for a minimal sample size of two.

Theorem 3 *The posterior distribution associated with the prior $\pi(\mu, \sigma) = 1/\sigma$ and with the likelihood derived from (6.1) is proper when the components $f(\cdot|\mu, \sigma)$ are Gaussian densities, provided (a) proper distributions are used on the other parameters and (b) there are at least two observations in the sample.*

Proof: When $n = 1$, it is easy to show that the Jeffreys posterior is not proper. The marginal likelihood is then

$$\begin{aligned} M_k(x_1) &= \sum_{i=1}^{k} \int p_i f(x_1|\mu + \sigma\alpha_i, \sigma^2\tau_i^2)\pi(\mu, \sigma, \mathbf{p}, \boldsymbol{\alpha}, \boldsymbol{\tau})\, \mathrm{d}(\mu, \sigma, \mathbf{p}, \boldsymbol{\alpha}, \boldsymbol{\tau}) \\ &= \sum_{i=1}^{k} \int \left\{ \int \frac{p_i}{\sqrt{2\pi}\sigma^2\tau_i} \exp\left(\frac{-(x_1 - \mu - \sigma\alpha_i)^2}{2\tau_i^2\sigma^2} \right) \mathrm{d}(\mu, \sigma) \right\} \pi(\mathbf{p}, \boldsymbol{\alpha}, \boldsymbol{\tau})\, \mathrm{d}(\mathbf{p}, \boldsymbol{\alpha}, \boldsymbol{\tau}) \\ &= \sum_{i=1}^{k} \int \left\{ \int_0^{\infty} \frac{p_i}{\sigma} \mathrm{d}\sigma \right\} \pi(\mathbf{p}, \boldsymbol{\alpha}, \boldsymbol{\tau})\, \mathrm{d}(\mathbf{p}, \boldsymbol{\alpha}, \boldsymbol{\tau}) \end{aligned}$$

The integral against σ is then not defined.

For two data-points, $x_1, x_2 \sim \sum_{i=1}^{k} p_i f(\mu + \sigma\alpha_i, \sigma^2\tau_i^2)$, the associated marginal likelihood is

$$\begin{aligned} M_k(x_1, x_2) &= \int \prod_{j=1}^{2} \left\{ \sum_{i=1}^{k} p_i f(x_j|\mu + \sigma\alpha_i, \sigma^2\tau_i^2) \right\} \pi(\mu, \sigma, \mathbf{p}, \boldsymbol{\alpha}, \boldsymbol{\tau})\, \mathrm{d}(\mu, \sigma, \mathbf{p}, \boldsymbol{\alpha}, \boldsymbol{\tau}) \\ &= \sum_{i=1}^{k} \sum_{j=1}^{k} \int p_i p_j f(x_1|\mu + \sigma\alpha_i, \sigma^2\tau_i^2) f(x_2|\mu + \sigma\alpha_j, \sigma^2\tau_j^2) \pi(\mu, \sigma, \mathbf{p}, \boldsymbol{\alpha}, \boldsymbol{\tau})\, \mathrm{d}(\mu, \sigma, \mathbf{p}, \boldsymbol{\alpha}, \boldsymbol{\tau}) \,. \end{aligned}$$

If all those k^2 integrals are proper, the Jeffrey posterior distribution is proper. An arbitrary integral $(1 \leq i, j \leq k)$ in this sum leads to

$$\begin{aligned} & \int p_i p_j f(x_1|\mu + \sigma\alpha_i, \sigma^2\tau_i^2) f(x_2|\mu + \sigma\alpha_j, \sigma^2\tau_j^2) \pi(\mu, \sigma, \mathbf{p}, \boldsymbol{\alpha}, \boldsymbol{\tau})\, \mathrm{d}(\mu, \sigma, \mathbf{p}, \boldsymbol{\alpha}, \boldsymbol{\tau}) \\ = & \int \left\{ \int \frac{p_i p_j}{2\pi\sigma^3\tau_i\tau_j} \exp\left[\frac{-(x_1 - \mu - \sigma\alpha_i)^2}{2\tau_i^2\sigma^2} + \frac{-(x_2 - \mu - \sigma\alpha_j)^2}{2\tau_j^2\sigma^2} \right] \mathrm{d}(\mu, \sigma) \right\} \pi(\mathbf{p}, \boldsymbol{\alpha}, \boldsymbol{\tau})\, \mathrm{d}(\mathbf{p}, \boldsymbol{\alpha}, \boldsymbol{\tau}) \\ = & \int \left\{ \int_0^{\infty} \frac{p_i p_j}{\sqrt{2\pi}\sigma^2\sqrt{\tau_i^2 + \tau_j^2}} \exp\left[\frac{-1}{2(\tau_i^2 + \tau_j^2)} \left(\frac{1}{\sigma^2}(x_1 - x_2)^2 + \frac{2}{\sigma}(x_1 - x_2)(\alpha_i - \alpha_j) \right.\right.\right. \\ & \left.\left.\left. + (\alpha_i - \alpha_j)^2 \right) \right] \mathrm{d}\sigma \right\} \pi(\mathbf{p}, \boldsymbol{\alpha}, \boldsymbol{\tau})\, \mathrm{d}(\sigma, \mathbf{p}, \boldsymbol{\alpha}, \boldsymbol{\tau}) \,. \end{aligned}$$

Substituting $\sigma = 1/z$, the above is integrated with respect to z, leading to

$$\begin{aligned}
\int & \left\{ \int_0^\infty \frac{p_i p_j}{\sqrt{2\pi}\sqrt{\tau_i^2+\tau_j^2}} \exp\left(\frac{-1}{2(\tau_i^2+\tau_j^2)} \left(z^2(x_1-x_2)^2 + 2z(x_1-x_2)(\alpha_i-\alpha_j) \right.\right.\right. \\
& \qquad \left.\left.\left. +(\alpha_i-\alpha_j)^2 \right)\right) \mathrm{d}z \right\} \pi(\mathbf{p},\boldsymbol{\alpha},\boldsymbol{\tau})\, \mathrm{d}(\mathbf{p},\boldsymbol{\alpha},\boldsymbol{\tau}) \\
= & \int \left\{ \int_0^\infty \frac{p_i p_j}{\sqrt{2\pi}\sqrt{\tau_i^2+\tau_j^2}} \exp\left(\frac{-(x_1-x_2)^2}{2(\tau_i^2+\tau_j^2)} \Big(z + \frac{\alpha_i-\alpha_j}{x_1-x_2} \Big)^2 \right) \mathrm{d}z \right\} \pi(\mathbf{p},\boldsymbol{\alpha},\boldsymbol{\tau})\, \mathrm{d}(\mathbf{p},\boldsymbol{\alpha},\boldsymbol{\tau}) \\
= & \int \frac{p_i p_j}{|x_1-x_2|} \Phi\left(-\frac{\alpha_i-\alpha_j}{x_1-x_2} \frac{|x_1-x_2|}{\sqrt{\tau_i^2+\tau_j^2}} \right) \pi(\mathbf{p},\boldsymbol{\alpha},\boldsymbol{\tau})\; \mathrm{d}(\mathbf{p},\boldsymbol{\alpha},\boldsymbol{\tau})\,,
\end{aligned}$$

where Φ is the cumulative distribution function of the standardized Normal distribution. Given that the prior is proper on all remaining parameters of the mixture and that the integrand is bounded by $1/|x_1-x_2|$, it integrates against the remaining components of $\boldsymbol{\theta}$.

Let us now consider the case $n \geq 3$. Since the posterior $\pi(\boldsymbol{\theta}|x_1,x_2)$ is proper, it constitutes a proper prior when considering only the observations $x_3,\ldots,x_n$. Therefore, the posterior is almost everywhere proper. □

6.2.3 Further reparameterisations

Before proposing relevant priors, let us note that the constraints in Lemma 2 suggest a new reparameterisation (among many possible ones): this reparameterisations uses the weights p_i in the definition of the component parameters, as to achieve a more generic constraint. The component location and scale parameters in (6.2) can indeed be reparameterised as

$$\alpha_i = \sigma\gamma_i/\sqrt{p_i} \quad \text{and} \quad \tau_i = \sigma\eta_i/\sqrt{p_i}\,,$$

leading to the mixture representation

$$f(x|\boldsymbol{\theta},\mathbf{p}) = \sum_{i=1}^k p_i f(x|\mu+\sigma\gamma_i/\sqrt{p_i}, \sigma\eta_i/\sqrt{p_i})\,, \quad \eta_i > 0\,, \tag{6.3}$$

Given $(p_1,\cdots,p_k)$, these new parameters are constrained by

$$\sum_{i=1}^k \sqrt{p_i}\gamma_i = 0 \quad \text{and} \quad \sum_{i=1}^k (\eta_i^2+\gamma_i^2) = 1\,,$$

which means that $(\gamma_1,\ldots,\eta_k)$ belongs to an hypersphere of $\mathbb{R}^{2k}$ intersected with an hyperplane of this space.

Given these constraints, further simplifications via new reparameterisations can be contemplated, as for instance separating mean and variance parameters in (6.3)

by introducing a radius φ such that

$$\sum_{i=1}^{k} \gamma_i^2 = \varphi^2 \quad \text{and} \quad \sum_{i=1}^{k} \eta_i^2 = 1 - \varphi^2 \,.$$

This choice naturally leads to a hierarchical prior where, e.g., φ^2 and $(p_1, \ldots, p_k)$ are distributed from a $\mathcal{B}e(a_1, a_2)$ and a $\mathcal{D}ir(\alpha_0, \ldots, \alpha_0)$ distributions, respectively, while the vectors $(\gamma_1, \ldots, \gamma_k)$ and $(\eta_1, \ldots, \eta_k)$ are uniformly distributed on the spheres of radius φ and $\sqrt{1-\varphi^2}$, respectively, under the additional linear constraint $\sum_{i=1}^{k} \sqrt{p_i}\gamma_i = 0$.

We now describe how this reparameterisation leads to a practical construction of the constrained parameter space, for an arbitrary number of components k.

6.2.3.1 Spherical coordinate representation of the γ's.

The vector $(\gamma_1, \ldots, \gamma_k)$ belongs both to the hypersphere of radius φ and to the hyperplane orthogonal to $(\sqrt{p_1}, \ldots, \sqrt{p_k})$. Therefore, $(\gamma_1, \ldots, \gamma_k)$ can be expressed in terms of spherical coordinates within that hyperplane. Namely, if $(\mathit{F}_1, \ldots, \mathit{F}_{k-1})$ denotes an orthonormal basis of the hyperplane, $(\gamma_1, \ldots, \gamma_k)$ can be written as

$$(\gamma_1, \ldots, \gamma_k) = \varphi \cos(\varpi_1)\mathit{F}_1 + \varphi \sin(\varpi_1)\cos(\varpi_2)\mathit{F}_2 + \ldots + \varphi \sin(\varpi_1)\cdots\sin(\varpi_{k-2})\mathit{F}_{k-1}$$

with the angles $\varpi_1, \ldots, \varpi_{k-3}$ in $[0, \pi]$ and ϖ_{k-2} in $[0, 2\pi]$. The s-th orthonormal base F_s can be derived from the k-dimensional orthogonal vectors $\widetilde{\mathit{F}}_s$ where

$$\widetilde{\mathit{F}}_{1,j} = \begin{cases} -\sqrt{p_2}, & j = 1 \\ \sqrt{p_1}, & j = 2 \\ 0, & j > 2 \end{cases}$$

and the s-th vector is given by

$$\widetilde{\mathit{F}}_{s,j} = \begin{cases} -(p_j p_{s+1})^{1/2} \Big/ \left(\sum_{l=1}^{s} p_l\right)^{1/2}, & s > 1,\ j \le s \\ \left(\sum_{l=1}^{s} p_l\right)^{1/2}, & s > 1,\ j = s+1 \\ 0, & s > 1,\ j > s+1 \end{cases}$$

Note the special case of $k = 2$ since the angle ϖ_1 is then missing. In this special case, the mixture location parameter is defined by $(\gamma_1, \gamma_2) = \varphi \mathit{F}_1$ and φ takes both positive and negative values. In the general setting, the parameter vector $(\gamma_1 \cdots, \gamma_k)$ is a transform of $(\varphi^2, p_1, \cdots, p_k, \varpi_1, \cdots, \varpi_{k-2})$. A natural reference prior for ϖ is made of uniforms, $\varpi_1, \cdots, \varpi_{k-3} \sim \mathcal{U}[0, \pi]$ and $\varpi_{k-2} \sim \mathcal{U}[0, 2\pi]$, although other choices are obviously possible and should be explored to test the sensitivity to the prior.

6.2.3.2 Dual representation of the η_i's.

For the component variance parameters, the vector $(\eta_1, \cdots, \eta_k)$ belongs to the k-dimension sphere of radius $\sqrt{1-\varphi^2}$. A natural prior is then a Dirichlet distribution with common hyperparameter a,

$$\pi(\eta_1^2, \cdots, \eta_k^2, \varphi^2) = \text{Dir}(\alpha, \cdots, \alpha)$$

If k is small enough, $(\eta_1, \cdots, \eta_k)$ can then be simulated from the corresponding posterior with no computational challenge. However, as k increases, sampling may become more delicate and benefits from a similar spherical reparameterisation. In this approach, the vector $(\eta_1, \cdots, \eta_k)$ is rewritten through spherical coordinates with angle components $(\xi_1, \cdots, \xi_{k-1})$,

$$\eta_i = \begin{cases} \sqrt{1-\varphi^2}\cos(\xi_i)\,, & i = 1 \\ \sqrt{1-\varphi^2}\displaystyle\prod_{j=1}^{i-1}\sin(\xi_j)\cos(\xi_i)\,, & 1 < i < k \\ \sqrt{1-\varphi^2}\displaystyle\prod_{j=1}^{i-1}\sin(\xi_j)\,, & i = k \end{cases}$$

Unlike ϖ, the support for all angles $\xi_1, \cdots, \xi_{k-1}$ is limited to $[0, \pi/2]$, due to the positivity requirement on the η_i's. In this case, a reference prior on the angles is

$$(\xi_1, \cdots, \xi_{k-1}) \sim \mathcal{U}([0, \pi/2]^{k-1})\,,$$

while again other choices are possible.

6.3 MCMC implications

6.3.1 The Metropolis-within-Gibbs sampler

Given the reparameterisations introduced in Section 6.2, different MCMC implementations are possible and we investigate in this section some of these. To this effect, we distinguish between two cases: (i) only $(\mu_1, \cdots, \mu_k)$ is expressed in spherical coordinates; and (ii) both the μ_i's and the σ_i's are associated with spherical coordinates.

Although the target density is similar to the target explored by early Gibbs samplers in [Diebolt 1990] and [Gelman 1990], simulating directly the new parameters implies managing constrained parameter spaces. The hierarchical nature of the parameterisation also leads us to consider a block Gibbs sampler that coincides with this hierarchy. Since the corresponding full conditional posteriors are not in closed form, a Metropolis-within-Gibbs sampler is implemented here with random walk proposals. In this approach, the scales of the proposal distributions are automatically calibrated towards optimal acceptance rates [Roberts 1997, Roberts 2001, Roberts 2009, Rosenthal 2011]. Convergence of a simulated chain is assessed based

on the rudimentary convergence monitoring technique of [Gelman 1992]. The description of the algorithm is provided by the pseudo-code version in Figure 6.1. Note that the Metropolis-within-Gibbs version does not rely on latent variables and complete likelihood as in [Tanner 1987] and [Diebolt 1990]. Following the adaptive MCMC method in Section 3 of [Roberts 2009], we derive the optimal scales associated with proposal densities, based on 10 batches with size 50. The scales ε are identified by a subscript with the corresponding parameter.

For the reparameterisation (i), all steps are the same except that steps 2.5 and 2.7 are combined together and that $((\varphi^2)^{(t)}, (\eta_1^2)^{(t)}, \ldots, (\eta_k^2)^{(t)})$ is updated in the same manner. One potential proposal density is a Dirichlet distribution,

$$((\varphi^2)', (\eta_1^2)', \ldots, (\eta_k^2)') \sim \mathrm{Dir}((\varphi^2)^{(t-1)}\varepsilon, (\eta_1^2)^{(t-1)}\varepsilon, \ldots, (\eta_k^2)^{(t-1)}\varepsilon)\,.$$

Alternative proposal densities will be discussed later along with simulation studies in Section 4.

6.3.2 Removing and detecting label switching

The standard parameterisation of mixture models contains weights $\{p_i\}_{i=1}^k$ and component-wise parameters $\{\theta_i\}_{i=1}^k$ as shown in (6.1). The likelihood function is invariant under permutations of the component indices. If an exchangeable prior is chosen on weights and component-wise parameters, the posterior density reproduces the likelihood invariance and component labels are not identifiable. This phenomenon is called *label switching* and is well-studied in the literature [Celeux 2000, Stephens 2000, Frühwirth-Schnatter 2001, Frühwirth-Schnatter 2004, Jasra 2005]. This means that the posterior distribution consists of $k!$ symmetric modes and a Markov chain with such target distribution is expected to explore all of them. However, a chain often fails and rather ends up exploring a particular mode.

In our reparameterisation of Gaussian mixture models, each component mean and variance are functions of angular and radius parameters with weights. The mapping between both parameterisations is a one-to-one map conditional on the weights. In other words, there are unique component-wise means and variances given particular values for angular and radius parameters and weights. Although the new parameterisation is not exchangeable, due to the choice of the orthogonal basis, adopting an exchangeable prior on the weights (e.g., a Dirichlet distribution with a common parameter) and uniform priors on all angular parameters leads to an exchangeable posterior on the natural parameters of the mixture. Therefore, label switching should also occur with this prior modeling.

When an MCMC chain manages to jump between modes, the inference on each of the mixture components becomes harder [Geweke 2007]. To get component-specific inference and to give a meaning to each component, various relabelling methods have been proposed in the literature (see, e.g., [Frühwirth-Schnatter 2004]). A first available alternative is to reorder labels so that the mixture weights are in increasing order [Frühwirth-Schnatter 2001]. A second alternative method proposed by, e.g., [Lee 2009] is that labels are reordered towards producing the shortest distance

Metropolis-within-Gibbs algorithm for reparameterised mixture model

1 Generate initial values $(\mu^{(0)}, \sigma^{(0)}, \mathbf{p}^{(0)}, \varphi^{(0)}, \xi_1^{(0)}, \ldots, \xi_{k-1}^{(0)}, \varpi_1^{(0)}, \ldots, \varpi_{k-2}^{(0)})$.

2 For $t = 1, \ldots, T$, the update of $(\mu^{(t)}, \sigma^{(t)}, \mathbf{p}^{(t)}, \varphi^{(t)}, \xi_1^{(t)}, \ldots, \xi_{k-1}^{(t)}, \varpi_1^{(t)}, \ldots, \varpi_{k-2}^{(t)})$ follows;

2.1 Generate a proposal $\mu' \sim \mathcal{N}(\mu^{(t-1)}, \varepsilon_\mu)$ and update $\mu^{(t)}$ against $\pi(\cdot|\mathbf{x}, \sigma^{(t-1)}, \mathbf{p}^{(t-1)}, \varphi^{(t-1)}, \boldsymbol{\xi}^{(t-1)}, \boldsymbol{\varpi}^{(t-1)})$.

2.2 Generate a proposal $\log(\sigma)' \sim \mathcal{N}(\log(\sigma^{(t-1)}), \varepsilon_\sigma)$ and update $\sigma^{(t)}$ against $\pi(\cdot|\mathbf{x}, \mu^{(t)}, \mathbf{p}^{(t-1)}, \varphi^{(t-1)}, \boldsymbol{\xi}^{(t-1)}, \boldsymbol{\varpi}^{(t-1)})$.

2.3 Generate proposals $\xi_i' \sim \mathcal{U}[0, \pi/2]$, $i = 1, \cdots, k-1$, and update $(\xi_1^{(t)}, \ldots, \xi_{k-1}^{(t)})$ against $\pi(\cdot|\mathbf{x}, \mu^{(t)}, \sigma^{(t)}, \mathbf{p}^{(t-1)}, \varphi^{(t-1)}, \boldsymbol{\varpi}^{(t-1)})$.

2.4 Generate proposals $\varpi_i' \sim \mathcal{U}[0, \pi]$, $i = 1, \cdots, k-3$, and $\varpi_{k-2}' \sim \mathcal{U}[0, 2\pi]$. Update $(\varpi_1^{(t)}, \ldots, \varpi_{k-2}^{(t)})$ against $\pi(\cdot|\mathbf{x}, \mu^{(t)}, \sigma^{(t)}, \mathbf{p}^{(t-1)}, \varphi^{(t-1)}, \boldsymbol{\xi}^{(t)})$.

2.5 Generate a proposal $(\varphi^2)' \sim \mathcal{B}eta((\varphi^2)^{(t)}\varepsilon_\varphi + 1, (1 - (\varphi^2)^{(t)})\varepsilon_\varphi + 1)$ and update $\varphi^{(t)}$ against $\pi(\cdot|\mathbf{x}, \mu^{(t)}, \sigma^{(t)}, \mathbf{p}^{(t-1)}, \boldsymbol{\xi}^{(t)}, \boldsymbol{\varpi}^{(t)})$.

2.6 Generate a proposal $\mathbf{p}' \sim \mathrm{Dir}(p_1^{(t-1)}\varepsilon_p + 1, \ldots, p_k^{(t-1)}\varepsilon_p + 1)$, and update $\mathbf{p}^{(t)}$ against $\pi(\cdot|\mathbf{x}, \mu^{(t)}, \sigma^{(t)}, \varphi^{(t)}, \boldsymbol{\xi}^{(t)}, \boldsymbol{\varpi}^{(t)})$.

2.7 Generate proposals $\xi_i' \sim U[\xi_i^{(t)} - \varepsilon_\xi, \xi_i^{(t)} + \varepsilon_\xi]$, $i = 1, \cdots, k-1$, and update $(\xi_1^{(t)}, \ldots, \xi_{k-1}^{(t)})$ against $\pi(\cdot|\mathbf{x}, \mu^{(t)}, \sigma^{(t)}, \mathbf{p}^{(t)}, \varphi^{(t)}, \boldsymbol{\varpi}^{(t)})$.

2.8 Generate proposals $\varpi_i' \sim U[\varpi_i^{(t)} - \varepsilon_\varpi, \varpi_i^{(t)} + \varepsilon_\varpi]$, $i = 1, \cdots, k-2$, and update $(\varpi_1^{(t)}, \ldots, \varpi_{k-2}^{(t)})$ against $\pi(\cdot|\mathbf{x}, \mu^{(t)}, \sigma^{(t)}, \mathbf{p}^{(t)}, \varphi^{(t)}, \boldsymbol{\xi}^{(t)})$.

Figure 6.1: Pseudo-code representation of the Metropolis-within-Gibbs algorithm used in this paper for the reparameterisation (ii) based on two sets of spherical coordinates. For simplicity's sake, we denote $\mathbf{p}^{(t)} = (p_1^{(t)}, \ldots, p_k^{(t)})$, $\mathbf{x} = (x_1, \ldots, x_n)$, $\boldsymbol{\xi}^{(t)} = (\xi_1^{(t)}, \ldots, \xi_{k-1}^{(t)})$ and $\boldsymbol{\varpi}^{(t)} = (\varpi_1^{(t)}, \ldots, \varpi_{k-2}^{(t)})$.

between the current posterior sample and the (or a) maximum posterior probability (MAP) estimate.

Let us denote by h the map from our reparameterisation to the standard parameterisation of (6.1), i.e.,

$$(\mu_1, \ldots, \mu_k, \sigma_1, \ldots, \sigma_k, \mathbf{p}) = h(\mathbf{p}, \boldsymbol{\theta}),$$

with its inverse h^{-1} available as well. We also denote by $\mathfrak{S}_k$ the set of permutations of $\{1, \ldots, k\}$. Then, given an MCMC sample $\{\mathbf{p}^{(t)}, \boldsymbol{\theta}^{(t)}\}_{t=1}^{T}$, the above relabelling technique procedure follows;

1. Reparameterise the MCMC sample $\{\mathbf{p}^{(t)}, \boldsymbol{\theta}^{(t)}\}_{t=1}^{T}$ into component-wise means and standard deviations via the function h, resulting in $\{\mu_1^{(t)}, \ldots, \mu_k^{(t)}, \sigma_1^{(t)}, \ldots, \sigma_k^{(t)}, \mathbf{p}^{(t)}\}_{t=1}^{T}$.

2. Find the MAP estimate by computing the posterior values of the sample; denote the solution as $(\mu_1^*, \ldots, \mu_k^*, \sigma_1^*, \ldots, \sigma_k^*, \mathbf{p}^*)$.

3. Reorder $(\mu_1^{(t)}, \ldots, \mu_k^{(t)}, \sigma_1^{(t)}, \ldots, \sigma_k^{(t)}, \mathbf{p}^{(t)})$ as

$$(\widetilde{\mu}_1^{(t)}, \ldots, \widetilde{\mu}_k^{(t)}, \widetilde{\sigma}_1^{(t)}, \ldots, \widetilde{\sigma}_k^{(t)}, \widetilde{\mathbf{p}}^{(t)}) = \delta_j(\mu_1^{(t)}, \ldots, \mu_k^{(t)}, \sigma_1^{(t)}, \ldots, \sigma_k^{(t)}, \mathbf{p}^{(t)})$$

where $\delta_j = \arg\min_{\delta \in \mathfrak{S}_k} \|\delta(\mu_1^{(t)}, \ldots, \mu_k^{(t)}, \sigma_1^{(t)}, \ldots, \sigma_k^{(t)}, \mathbf{p}^{(t)}) - (\mu_1^*, \ldots, \mu_k^*, \sigma_1^*, \ldots, \sigma_k^*, \mathbf{p}^*)\|$.

The resulting permutation is then denoted $\lambda^{(t)} \in \mathfrak{S}_k$. Label switching occurrences in an MCMC sequence can be monitored via the changes in the sequence $\lambda^{(1)}, \ldots, \lambda^{(T)}$. If the chain fails to switch modes, the sequence is likely to remain at the same permutation. On the opposite, if a chain moves between some of the $k!$ symmetric posterior modes, the $\lambda^{(t)}$'s are expected to vary.

We proceed here by a simulation studies section and all algorithms used in this section are publicly available within the R package Ultimixt [Kamary 2015]. The package Ultimixt contains functions that implement adaptive determination of optimal scales and convergence monitoring based on [Gelman 1992] criterion. In addition, Ultimixt includes functions that summarize the simulations and compute point estimates of each parameter, such as posterior mean and median. It also produces an estimated mixture density in numerical and graphical formats. The output further includes graphical representations of the generated parameter samples. For the potentially unimodal parameters μ, σ and φ, averaging and calculating the median over the generated chains directly returns valid point estimators, as those parameters are not subjected to label switching. For the other parameters (component weights, means and variances), since label switching is a possible issue, we need to postprocess the MCMC draws as discussed earlier, by first relabelling these simulations. We then derive point estimates by clustering over the parameter space, using k-mean clustering [Hastie 2001].

6.4 Simulation studies

In this section, we examine the performances of the above Metropolis-within-Gibbs algorithm, when applied to both reparameterisations defined above. We also consider the special case $k = 2$ in Section 6.4.1. All simulations were conducted using the package Ultimixt [Kamary 2015].

6.4.1 The case $k = 2$

In this specific case, we do not have to simulate any angle. Two straightforward proposals are compared over simulation experiments. One is based on Beta and Dirichlet proposals:

$$p^* \sim \mathcal{B}eta(p^{(t)}\varepsilon_p, (1-p^{(t)})\varepsilon_p)\,, \quad (\varphi^{2^*}, \eta_1^{2^*}, \eta_2^{2^*}) \sim \mathcal{D}ir(\varphi^{2^{(t)}}\varepsilon, \eta_1^{2^{(t)}}\varepsilon, \eta_2^{2^{(t)}}\varepsilon)$$

(this will be called Proposal 1) and another one is based on Gaussian random walks:

$$\begin{aligned}
\log(p^*/(1-p^*)) &\sim \mathcal{N}(\log(p^{(t)}/(1-p^{(t)})), \varepsilon_p) \\
(\vartheta_1^*, \vartheta_2^*)^T &\sim \mathcal{N}(\chi_2^{(t)}, \varepsilon_\vartheta I_2) \quad \text{with} \\
(\varphi^{2^*}, \eta_1^{2^*}, \eta_2^{2^*}) &= (\exp(\vartheta_1^*)/\bar{\vartheta}^*, \exp(\vartheta_2^*)/\bar{\vartheta}^*, 1/\bar{\vartheta}^*)\,, \\
\chi_2^{(t)} &= (\log(\varphi^{2^{(t)}}/\eta_2^{2^{(t)}}), \log(\eta_1^{2^{(t)}}/\eta_2^{2^{(t)}}) \\
\text{and} \quad \bar{\vartheta}^* &= 1 + \exp(\vartheta_1^*) + \exp(\vartheta_2^*)
\end{aligned}$$

(which will be called Proposal 2). The global parameters are proposed using Normal and Inverse-Gamma proposals

$$\mu^* \sim \mathcal{N}(\bar{x}, \varepsilon_\mu) \quad \text{and} \quad \sigma^{2^*} \sim \mathcal{IG}((n+1)/2, (n-1)\bar{\sigma}^2/2)$$

where $\bar{x}$ and $\bar{\sigma}^2$ are sample mean and variance respectively. We present below some analyses and also explain how MCMC methods can be used to fit the reparameterised mixture distribution.

Example 6.4.1 In this experiment, a dataset of size 50 is simulated from the mixture $0.65\mathcal{N}(-8, 2) + 0.35\mathcal{N}(-0.5, 1)$, which implies that while the true value of $(\varphi, \eta_1, \eta_2)$ is $(0.91, 0.16, 0.38)$. Figure 6.2 illustrates the performances of a Metropolis-within-Gibbs algorithm based on Proposal 1. It shows the outcomes of 10 parallel chains, each started randomly from different starting values. The estimated densities are almost indistinguishable among the different chains and they all converge to a neighborhood of the true values. The chains are well-mixed and the sampler output covers the entire sample space in this case.

We also run the Metropolis-within-Gibbs algorithm based on Proposal 2 using the same simulated dataset for comparison purposes. As shown in Figure 6.3, the outputs for both proposals are quite similar but Proposal 1 produces more symmetric chains on $p, \varphi, \eta_1, \eta_2$, thus suggesting higher mixing abilities.

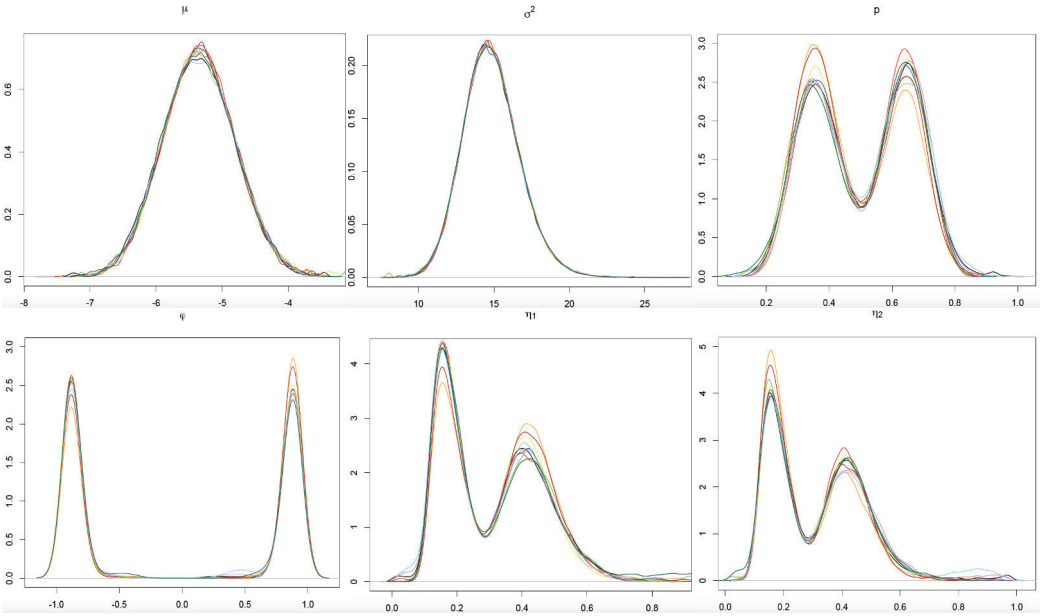


Figure 6.2: **Example 6.4.1:** Kernel estimates of the posterior densities of the parameters μ, σ, p, φ, η_i, based on 10 parallel MCMC chains for Proposal 1 and $2\,10^5$ iterations, based on a single simulated sample of size 50. The true value of $(\varphi, \eta_1, \eta_2)$ is $(0.91, 0.16, 0.38)$.

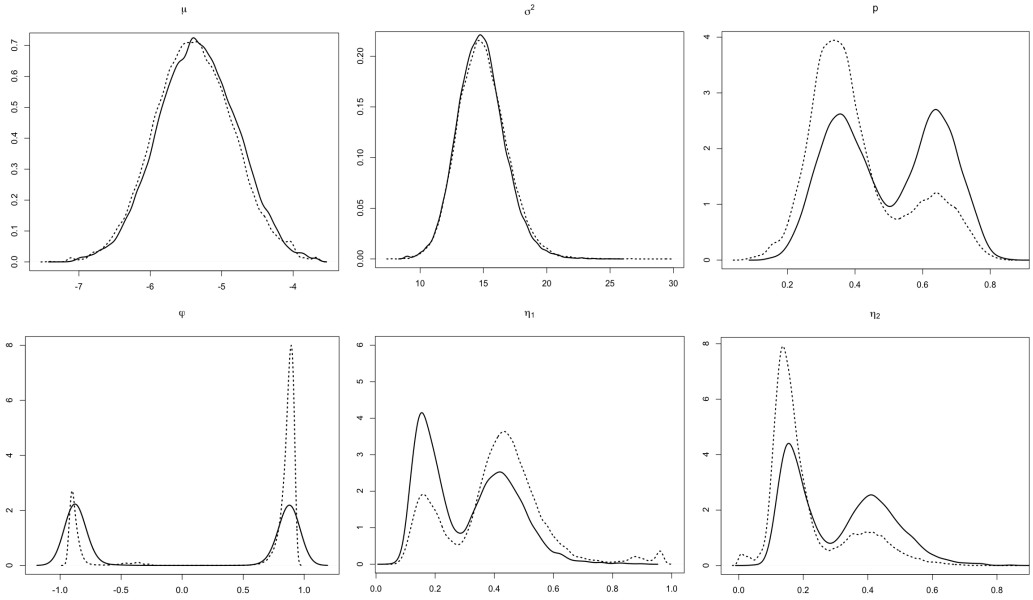


Figure 6.3: **Example 6.4.1:** Comparison between MCMC samples from our Metropolis-within-Gibbs algorithm using Proposal 1 (*solid line*) or Proposal 2 (*dashed line*), with $90,000$ iterations and the same sample as in Figure 6.2. The true value of $(\varphi, \eta_1, \eta_2)$ is $(0.91, 0.16, 0.38)$.

Proposal 1	ar_μ	ar_σ	ar_p	$ar_{\varphi,\eta}$	ε_μ	ε_p	ε
	0.40	0.47	0.45	0.24	0.56	77.06	99.94
Proposal 2	ar_μ	ar_σ	ar_p	$ar_{\varphi,\eta}$	ε_μ	ε_p	ε_ϑ
	0.38	0.46	0.45	0.27	0.55	0.29	0.35

Table 6.1: **Example 6.4.1**: Acceptance rate (ar) and corresponding proposal scale (ε) when the adaptive Metropolis-within-Gibbs sampler is used.

The scales of the various proposals are determined by aiming at [Roberts 1997] goal of an average acceptance rate of either 0.44 or 0.234 depending on the dimension of the simulated parameter. As shown in Table 6.1, an adaptive Metropolis-within-Gibbs strategy manages to recover acceptance rates close to optimal values. ◀

Having exposed how our sampler behaves we now discuss a second example, in which we briefly outline how this method may behave for a benchmark dataset with a slightly larger sample size.

Example 6.4.2 We now analyze the benchmark Old Faithful dataset, available from R, using the 272 observations of eruption times and a mixture with two components. The empirical mean and variance of the observations are $(3.49, 1.30)$.

When using Proposal 1, the optimal scales $\varepsilon_\mu, \varepsilon_p, \varepsilon$ after $50,000$ burn-in iterations are 0.07, 501.1, 802.19, respectively. The posterior distributions of the generated samples shown in Figure 6.4 demonstrate a strong concentration of (μ, σ^2) near the empirical mean and variance. Trace plots for the other parameters indicate a high dependence between successive iterations.There is a strong indication that the chain gets trapped into a single mode of the posterior density. In Section 6.5, we reanalyse this dataset when using parallel tempering. ◀

6.4.2 The general case

We now consider the general case of estimating a reparameterised mixture for any k when the variance vector $(\eta_1^2, \ldots, \eta_k^2)$ also has the spherical coordinate system as represented in Section 6.2.3.

Example 6.4.3 We simulated 50 data points from the mixture

$$0.27\mathcal{N}(-4.5, 1) + 0.4\mathcal{N}(10, 1) + 0.33\mathcal{N}(3, 1)\,.$$

Running our adaptive Metropolis-within-Gibbs algorithm shows that the simulated samples are quite close to the true values. However, the sampler has apparently visited only one of the posterior modes. This lack of label switching helps us in producing point estimates directly from this MCMC output [Geweke 2007] but this also shows an incomplete convergence of the MCMC sampler [Celeux 2000]. When considering the new parameters of this mixture, the single ϖ plays a significant role in the lack of label switching since transforming ϖ to $\pi - \varpi$ swaps first and second components.

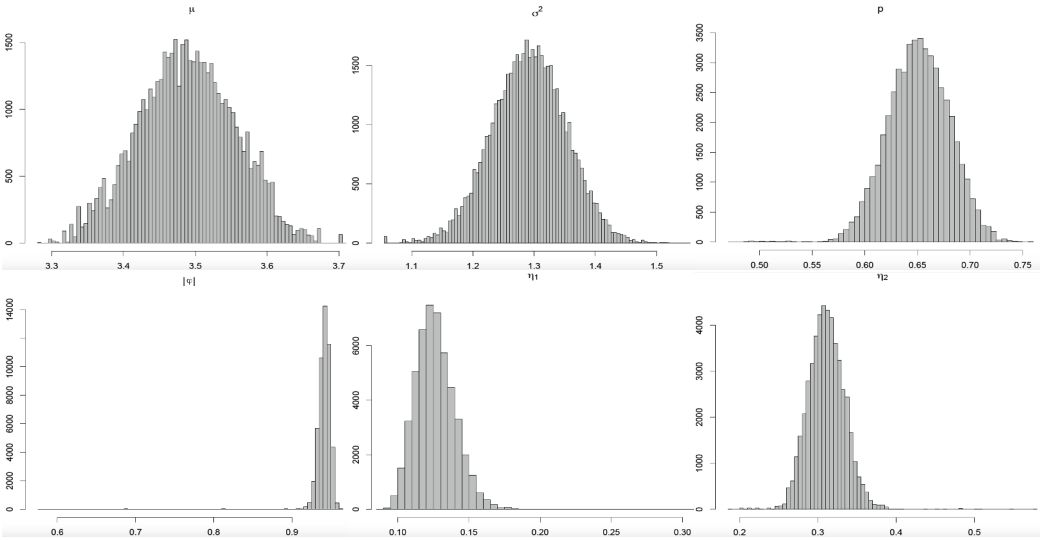


Figure 6.4: **Old Faithful dataset (Example 6.4.2):** Posterior distributions of the parameters of a two-component mixture distribution based on 50,000 MCMC iterations.

If we restrict the proposal on ϖ to step 2.4 of the Metropolis-within-Gibbs algorithm, namely using only a uniform $\mathcal{U}(0, 2\pi)$ distribution, Figure 6.5 shows that the MCMC chains of the p_i's are both well-mixed and exhibiting strong exchangeability. However, the corresponding acceptance rate is quite low at 0.051.

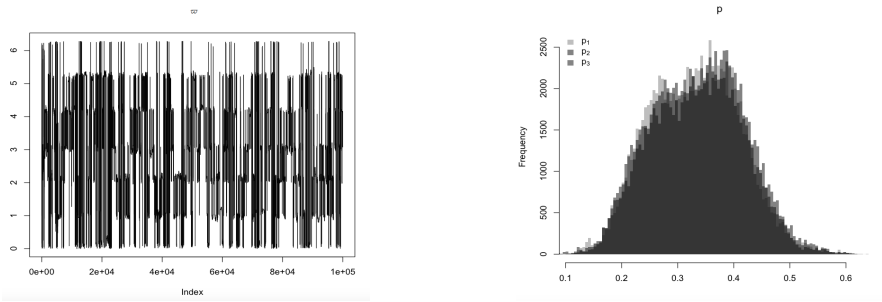


Figure 6.5: **Example 6.4.3:** *(Left)* Evolution of the sequence $(\varpi^{(t)})$ and *(Right)* histograms of the simulated weights based on 10^5 iterations of an adaptive Metropolis-within-Gibbs algorithm with independent proposal on ϖ.

If we consider in addition the random walk proposal of Step 2.8 on ϖ, namely a $\mathcal{U}(\varpi^{(t)} - \varepsilon_\varpi, \varpi^{(t)} + \varepsilon_\varpi)$ distribution, this step clearly improves performances, as illustrated in Figure 6.6, with acceptance rates all close to 0.234 and 0.44. Almost perfect label switching occurs in this case.

The marginal posterior distributions of the means and standard deviations are shown in Figure 6.7. They are almost indistinguishable due to label switching. Point estimates are once more produced by relabelling and k-mean clustering, to be com-

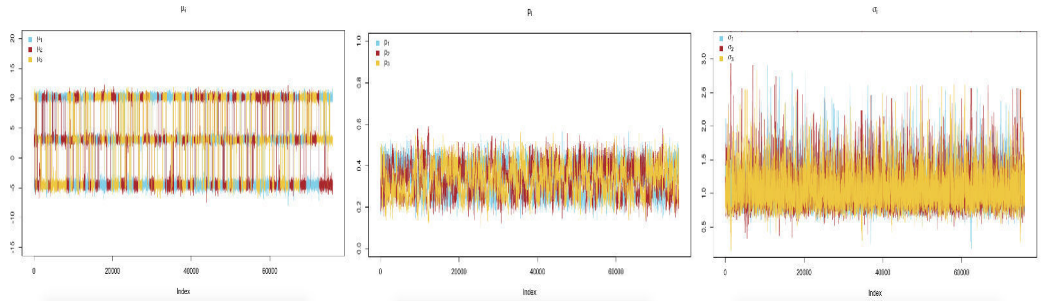


Figure 6.6: **Example 6.4.3:** Traces of the last $70,000$ simulations from the posterior distributions of the component means, standard deviations and weights, involving an additional random walk proposal on ϖ, based on 10^5 iterations.

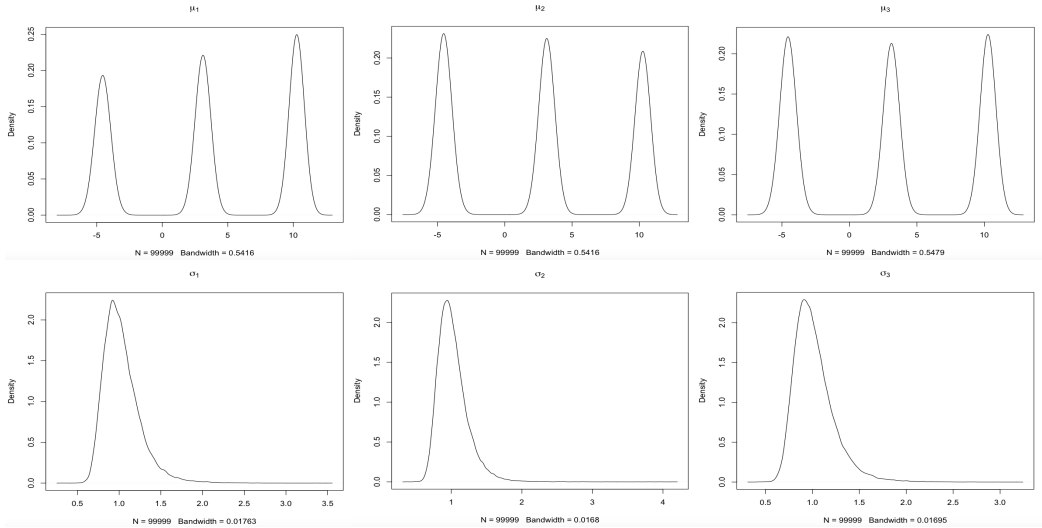


Figure 6.7: **Example 6.4.3:** Estimated marginal posterior densities of component means and standard deviations, based on 10^5 MCMC iterations.

pared with the MAP estimates automatically deduced from the simulation output. Those estimate are shown on the left and right sides of Table 6.2, respectively. Estimates computed by both methods are almost identical and all parameters are close to the true values.

However, Bayesian inference for parameters related to individual components of the mixture using averaging over posterior draws is not possible in this case since the posterior means of the component specific parameters such as $p, \mu_i, \sigma_i; i = 1, 2, 3$ are the same for all components. We therefore revert to both methods of k-means clustering algorithm presented at the beginning of this section and removing label switching based on the distance between posterior sample and MAP estimate which are shown in left and right sides of Table 6.2, respectively. Bayesian estimations computed by both methods are almost identical and all parameters of the mixture distributions are accurately estimated in comparison with those of the true model

	Angular & component-wise parameters					
	k-means clustering			MAP estimate		
	ϖ	ξ_1	ξ_2	ϖ	ξ_1	ξ_2
Median	3.54	0.97	0.73	3.32	0.94	0.83
Mean	3.53	0.98	0.72	3.45	0.94	0.82
	p_1	p_2	p_3	p_1	p_2	p_3
Median	0.40	0.27	0.33	0.41	0.27	0.33
Mean	0.41	0.27	0.33	0.41	0.27	0.33
	μ_1	μ_2	μ_3	μ_1	μ_2	μ_3
Median	10.27	-4.55	3.11	10.27	-4.55	3.11
Mean	10.27	-4.54	3.12	10.26	-4.45	3.11
	σ_1	σ_2	σ_3	σ_1	σ_2	σ_3
Median	0.93	1.04	1.01	0.93	1.04	1.03
Mean	0.95	1.08	1.05	0.95	1.07	1.05

	Global parameters		
	μ	σ	φ
Median	3.98	6.03	0.98
Mean	3.98	6.02	0.99

Proposal scales					
ε_μ	ε_σ	ε_p	ε_φ	ε_ϖ	ε_ξ
0.33	0.06	190	160	0.09	0.39

Acceptance rates					
ar_μ	ar_σ	ar_p	ar_φ	ar_ϖ	ar_ξ
0.22	0.34	0.23	0.43	0.42	0.22

Table 6.2: **Example** 6.4.3: Point estimators of the parameters of a mixture of 3 components, proposal scales and corresponding acceptance rates.

with the acceptance rates of the proposal distributions of the Metropolis-within-Gibbs very close to the optimal ones.

Example 6.4.4 We now consider an 8 component mixture,

$$0.08\mathcal{N}(0, 0.8) + 0.12\mathcal{N}(1.5, 1.1) + 0.2\mathcal{N}(3, 0.9) + 0.1\mathcal{N}(5, 1.2) + 0.15\mathcal{N}(7.5, 2) + 0.1\mathcal{N}(9, 1.3) + 0.13\mathcal{N}(10.2, 0.7) + 0.12\mathcal{N}(11.5, 1.1)\,,$$

from which we simulated 20 samples of size 250. Calibration of the random walks is achieved after 10^4 for almost all samples.

When computing point estimates of the natural parameters of the components, we obtain the maximum errors of 0.08 and 0.11 for μ and σ, respectively. The average absolute error over the 20 samples is quite low. Furthermore, when comparing the true and estimated mixtures, we can resort to the Kullback-Leibler divergence. For the 20 simulated samples, the maximum value is 0.02, which means an information loss of at most 2%. If we consider the upper bound introduced by [Sayyareh 2011] on Kullback-Leibler divergence, the obtained values indicates a good similarity between P_{true} and $P_{estimated}$ and illustrates the consistency of the estimates resulting from our Metropolis-within-Gibbs algorithm. ◀

Example 6.4.5 When an MCMC chain converges to a very small value for at least one component weight p_i, this may lead to an extremely large mean or large variance in the corresponding component. This happens partly because there is hardly any information from the data for this component and partly because the new parameters are functions of $1/\sqrt{p_i}$. We may thus face extreme points in the simplex parameter spaces. This phenomenon is illustrated with the *Galaxy dataset*, a constant benchmark for mixture estimation [Roeder 1990, Richardson 1997], when we impose $k = 6$ components. The MCMC sample is again summarized by k-means clustering and MAP estimates, as presented in Section 6.3.2. The resulting means, medians and 95% credible intervals of the parameters of the mixture components

are displayed in Table 6.3. Unsurprisingly, global mean and standard deviation are quite similar to the empirical estimates. Table 6.3 also displays estimates based on the Gibbs sampler of bayesm [Rossi 2010] and on the EM algorithms of mixtools [Benaglia 2009], with our approach being produced by Ultimixt [Kamary 2015].

Obtaining very close estimations for two component means μ_i, as $\mu_1 = 19.59$ and $\mu_5 = 19.93$, and $\mu_2 = 21.97$ and $\mu_6 = 22$ and $\mu_4 = 22.21$ for bayesm, and $\mu_1 = 24.27$ and $\mu_6 = 24.26$ for mixtools, signals that overfitting occurs: there are more components than supported by the data. With our analysis, overfitting is handled in a different way: the mean of one or more component weights is close to zero. For instance, we obtained estimates of p_1 very close to zero, inducing estimates for μ_1 and σ_1 of 61.59 and 32.23 for μ_1 and σ_1 (obtained by k-means clustering) and of 67.26 and 20.53 (using MAP estimates), as shown in Table 6.3. If we examine the MCMC sequences in detail, the minimum simulated value for the first component weight and the corresponding first component mean and standard deviation are $1.045\,10^{-6}$, 449.25 and 284.34, respectively. Such extreme values are produced because of the extremely small weight. However, such large values have no impact on the resulting estimate of the mixture itself. This is clearly exhibited in Figure 6.8 for the *Galaxy dataset*, which shows that extreme values have no effect on the predictive density plots due to the small weights. Using our modeling, the resulting density estimate is remarkably smooth when considering that the number of observations is 82 and a number of components equal to 6.

If we repeat running the algorithm on the *Galaxy dataset* for $50,000$ iterations and a smaller number of components, for instance $k = 4$, summary and model fit statistics are provided in Table 6.3. In this case, extreme values do not occur and the predictive density plots show that a four component model fits the data equally well as displayed in Figure 6.9. The posterior estimates of the component parameters computed by three methods (k-means clustering, MAP, and EM estimates) are almost similar, while the Gibbs sampler results from bayesm yield two very close estimates of component means, $\mu_2 = 21.05$ and $\mu_4 = 20.90$ in this case.

The common priors for the standard parameters are

$$\mu_i \sim N(\bar{\mu}, 10\sigma_R)\,, \quad \sigma_R^2 \sim \text{IW}(\nu, 3) \quad \text{and} \quad (p_1, \ldots, p_k) \sim \mathcal{D}ir(\alpha_0, \ldots, \alpha_0)$$

where $\text{IW}(\nu, 3)$ is the Inverse-Wishart distribution with the scale parameter of 3 and the degrees of freedom of ν. Unknown hyperparameters $\bar{\mu}$, σ_R, α_0 and ν are given by bayesm from the empirical estimation of data and, the comparison of the proposed priors and the prior obtained from bayesm are graphically presented in Figure 6.10.

	6 components, $k = 6$						**4 components,** $k = 4$			
	k-means clustering						**k-means clustering**			
	p_1	p_2	p_3	p_4	p_5	p_6	p_1	p_2	p_3	p_4
Median	0.01	0.08	0.13	0.43	0.05	0.24	0.56	0.27	0.06	0.10
Mean	0.02	0.06	0.14	0.46	0.05	0.24	0.58	0.25	0.06	0.11
	μ_1	μ_2	μ_3	μ_4	μ_5	μ_6	μ_1	μ_2	μ_3	μ_4
Median	25.95	9.72	22.06	19.83	32.71	22.87	20.19	21.52	32.79	9.72
Mean	61.59	9.725	22.09	19.84	32.70	22.93	20.27	21.48	33.29	9.73
	σ_1	σ_2	σ_3	σ_4	σ_5	σ_6	σ_1	σ_2	σ_3	σ_4
Median	4.53	4.91	1.91	0.52	2.86	0.65	0.52	1.62	3.00	1.05
Mean	32.23	4.61	2.41	0.58	4.23	1.10	0.57	2.08	3.66	3.44
	MAP estimate						**MAP estimate**			
	p_1	p_2	p_3	p_4	p_5	p_6	p_1	p_2	p_3	p_4
Median	0.04	0.09	0.13	0.37	0.10	0.15	0.32	0.46	0.08	0.08
Mean	0.05	0.09	0.10	0.39	0.14	0.22	0.34	0.43	0.13	0.09
2.5%	$< 10^{-5}$	< 0.01	< 0.01	< 0.01	$< 10^{-3}$	< 0.01	0.04	0.02	0.01	0.04
97.5%	0.2.1	0.13	0.69	0.39	0.56	0.68	0.87	0.82	0.51	0.15
	μ_1	μ_2	μ_3	μ_4	μ_5	μ_6	μ_1	μ_2	μ_3	μ_4
Median	30.96	9.70	21.75	19.73	20.61	23.12	19.84	22.17	28.23	9.71
Mean	67.26	8.18	21.58	18.73	20.84	24.33	19.83	22.34	29.03	9.50
2.5%	22.87	-9.28	19.60	9.68	12.83	21.29	17.59	20.14	22.27	9.17
97.5%	606.16	10.21	23.44	20.47	25.69	33.07	21.47	26.87	36.20	10.21
	σ_1	σ_2	σ_3	σ_4	σ_5	σ_6	σ_1	σ_2	σ_3	σ_4
Median	4.82	0.54	1.76	0.60	3.41	1.73	0.69	2.22	3.22	0.53
Mean	20.53	2.05	2.06	0.73	15.59	2.34	0.96	3.23	4.15	0.91
2.5%	0.79	0.30	0.31	0.19	0.41	0.17	0.29	0.87	0.68	0.29
97.5%	198.23	17.28	7.63	2.13	35.95	7.62	2.44	9.62	10.57	1.34
	Gibbs sampler (bayesm)						**Gibbs sampler** (bayesm)			
	p_1	p_2	p_3	p_4	p_5	p_6	p_1	p_2	p_3	p_4
	0.17	0.09	0.14	0.23	0.19	0.19	0.33	0.31	0.18	0.18
	μ_1	μ_2	μ_3	μ_4	μ_5	μ_6	μ_1	μ_2	μ_3	μ_4
	19.59	21.97	20.83	22.21	19.93	22.00	20.53	21.05	21.75	20.90
	σ_1	σ_2	σ_3	σ_4	σ_5	σ_6	σ_1	σ_2	σ_3	σ_4
	0.35	0.23	0.22	0.24	0.26	0.31	0.22	0.19	0.21	0.27
	EM estimate (mixtools)						**EM estimate** (mixtools)			
	p_1	p_2	p_3	p_4	p_5	p_6	p_1	p_2	p_3	p_4
	0.04	0.08	0.17	0.41	0.09	0.20	0.52	0.33	0.04	0.11
	μ_1	μ_2	μ_3	μ_4	μ_5	μ_6	μ_1	μ_2	μ_3	μ_4
	24.27	9.71	22.33	19.88	33.04	24.26	19.72	22.72	33.04	10.14
	σ_1	σ_2	σ_3	σ_4	σ_5	σ_6	σ_1	σ_2	σ_3	σ_4
	0.08	0.42	0.44	.70	0.19	8.33	0.62	1.77	0.92	2.73

Table 6.3: **Galaxy dataset**: Estimates of the parameters of a mixture of 6 and 4 components.

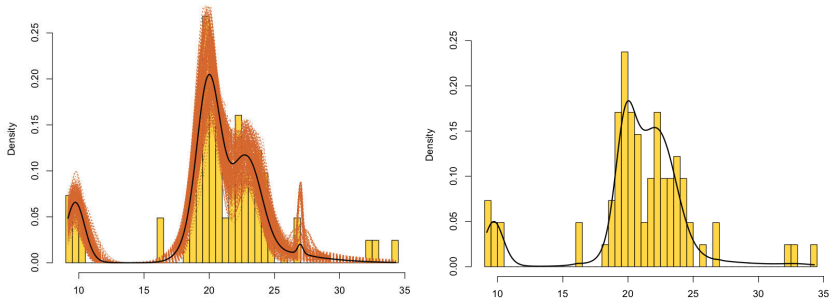


Figure 6.8: **Galaxy dataset:** (*Left*) Representation of 500 MCMC iterations as mixture distributions with the overlaid average curve for $k = 6$ components (*dark line*); (*Right*) mixture density estimate based on $15,000$ MCMC iterations for $k = 6$ components.

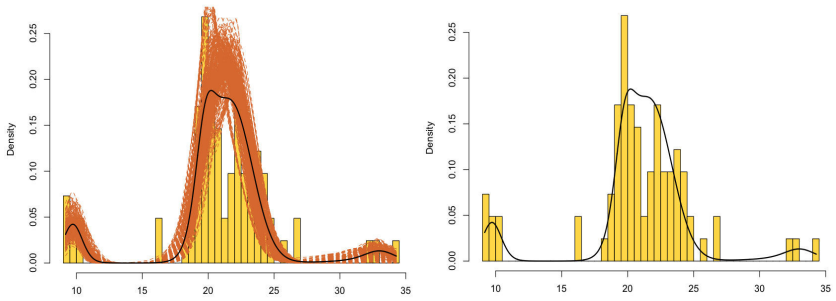


Figure 6.9: **Galaxy dataset:** (*left*) Representation of 500 Metropolis-within-Gibbs iterations for the mixture estimation and the overlay curve (*dark line*) obtained by averaging over iterations; (*right*) The mixture density estimate to histogram of dataset computed by averaging over $50,000$ MCMC iterations.

It is seen that the proposed prior is more dispersed for μ_1 and p_1 and is very skewed toward 0 for σ_1 with long tail. When $k = 6$, bayesm yields a more concentrated prior for p to accommodate all components and the proposed prior becomes dispersed to give flexible support on component-wise location and scale.

6.5 Parallel tempering

In Example 6.4.2 we have seen that for the Old Faithful dataset, the multimodality of the mixture model is not reproduced in the MCMC output, which means the adaptive Metropolis-within-Gibbs sampler cannot escape one of the modes. In this case, parallel tempering may be used [Marinari 1992, Neal 1996]. This method allows for better mixing in multimodal target distributions, when using straightforward Metropolis-Hastings algorithms fail [Miasojedow 2013]. It is indeed designed

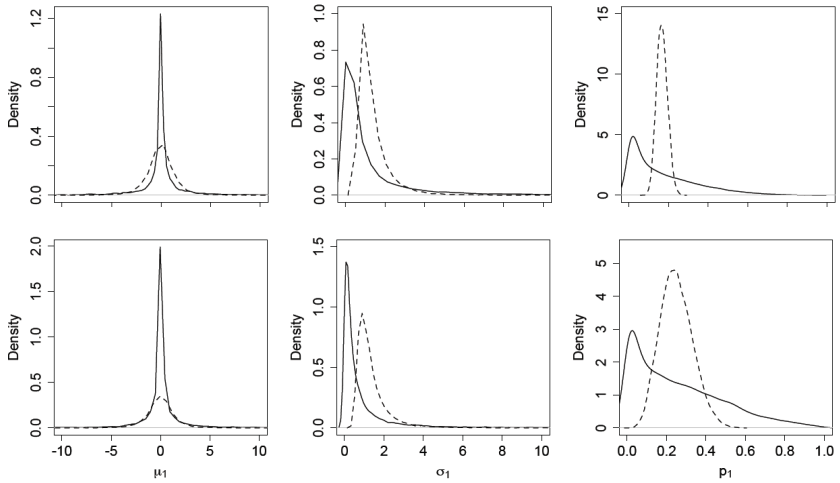


Figure 6.10: **Galaxy dataset:** Empirical prior densities based on 10^4 samples for μ_1, σ_1 and p_1 when (*Top*) $k = 6$ and (*Bottom*) $k = 4$. For the proposed prior (*solid lines*), the priors induced are $\pi(\mu_1) \propto \pi(\sigma\gamma_1/\sqrt{p_1})$ and $\pi(\sigma_1) \propto \pi(\sigma\eta_1/\sqrt{p_1})$. For the prior by bayesm (*dashed lines*), hyperparameters are $\alpha_0 = 5$ for $k = 4$ and $\alpha_0 = 25$ for $k = 6$ while $\bar{\mu} = 0$ and $\nu = 3$.

to overcome low probability regions between modal areas. Given the posterior density $f(\theta|x)$, we define tempered versions $f_\beta(\theta|x) \propto f(\theta|x)^\beta$, where $0 \leq \beta \leq 1$ is the inverse temperature and $\beta = 1$ corresponds to the original target distribution [Geyer 1991]. The tempered MCMC algorithm then runs a basic MCMC algorithm on a range of tempered distribution and, at each iteration, the current samples are considered for potential exchanges between adjacent temperatures, with a Metropolis–Hastings acceptance probability

$$\alpha_h = \min\left(1, \frac{f_{\beta_{h-1}}(\theta_h^{(t)}) f_{\beta_h}(\theta_{h-1}^{(t)})}{f_{\beta_{h-1}}(\theta_{h-1}^{(t)}) f_{\beta_h}(\theta_h^{(t)})}\right),$$

as the chances of accepting a swap are higher for nearby temperatures. Proposal scales are calibrated by adaptive MCMC method and is used for all tempered versions of the target. Temperatures are chosen of the form 2^j $(j = 1, \ldots)$ and the sequence is determined according to the degree of symmetry in the distribution of the p_i's or when the minimum acceptance rate for swaps between adjacent temperatures is larger than a default threshold.

Example 6.5.1 Considering again the Old Faithful benchmark, we set this symmetry threshold to .1 and this acceptance threshold to 0.3. Using the same proposals as in Example 6.4.2 and $N_{sim} = 50,000$, the algorithm selects 4 temperatures, thus equal to $1, 2, 4, 8$. Figure 6.11 demonstrates that the parallel tempering sampler visits all modes in the posterior distribution and that the mixing of the chains is greatly improved. ◀

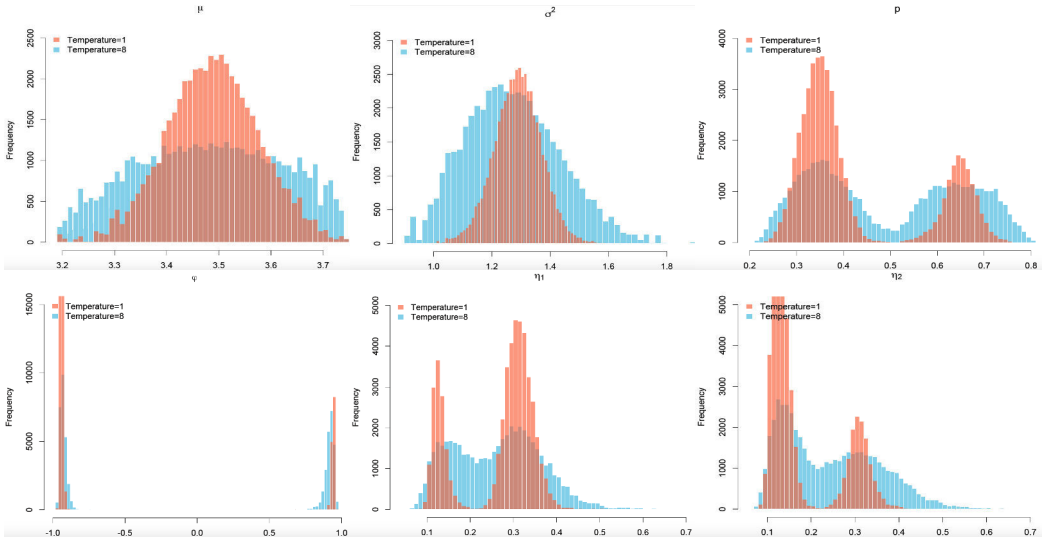


Figure 6.11: **Faithful dataset 6.4.2:** Posterior distribution of the mixture distribution parameters and comparison between the lowest and highest temperatures (target distribution and $f(x|\theta)^{1/8}$) of parallel tempering outputs based on $50,000$ iterations.

Example 6.5.2 We now implement parallel tempering for a mixture of $k = 3$ components applied to a benchmark dataset from [Marin 2007]. This dataset is derived from an image of a car license plate, and made of 2625 observations. In [Marin 2007], a lack of label switching is observed when using a Gibbs sampler. Once again, this means each component can be estimated by its mean and standard deviation. The sample size is larger here and more likely to mixing problems. This is clearly exhibited in the six top plots of Figure 6.12 where the estimates provided for the three components are quite distinct. When implementing parallel tempering, the temperature increase stops when when all acceptance rates of swaps are above .4, meaning for this dataset 7 temperatures ranging from 1 to 64.

The six bottom plots of Figure 6.12 show that parallel tempering immensely improves the swaps between the posterior modes. The sample of ϖ's produced by parallel tempering visits a much larger region in $(0, 2\pi)$, when compared with the highly peaked output of the original MCMC output.

The histograms in Figure 6.12 show that the posterior on p and η are now close to identical for each component. Two-dimensional plots also highlight this correct label switching behavior, which demonstrates better mixing and convergence of the produced chain. ◀

6.6 Conclusion

This paper has introduced a new parametrisation for mixtures of location-scale models. By constraining the parameters in terms of the global mean and global variance

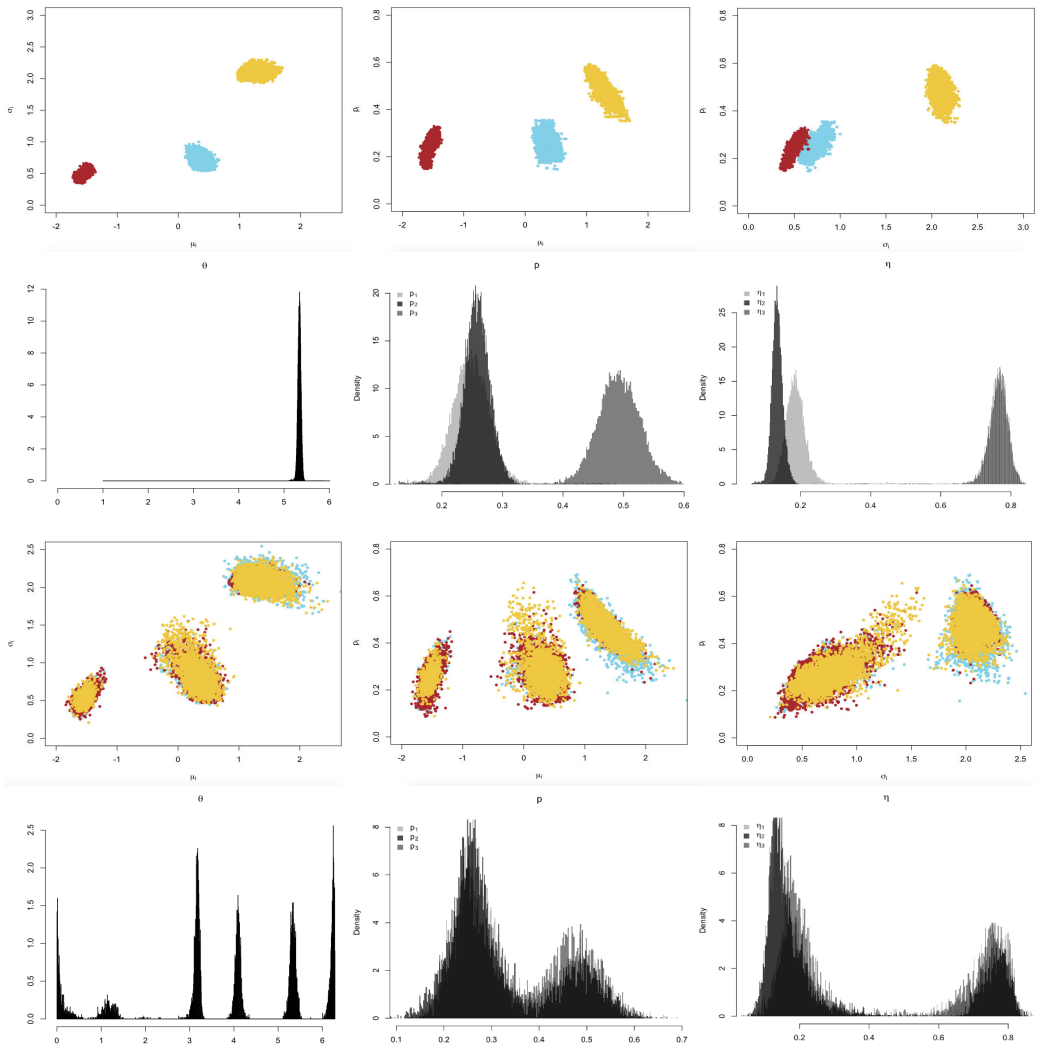


Figure 6.12: **Licence dataset (Example 6.5.2):** Comparison between Metropolis-within-Gibbs and parallel tempering outputs: The distributions of the samples of 10^4 last points and corresponding 2×2 plots.

of the mixture, i.e., by recognizing the location-scale nature of such mixtures, it has been shown that the remaining parameters can be expressed as varying within a compact set. Therefore, it is possible to use a well-defined uniform prior on these parameters (as well as any proper prior) and we established that an improper prior of Jeffreys' type on the global mean and global variance returns a proper posterior distribution when handling at least two observations from the mixture. While the notion of *non-informative* or *objective* prior is open to interpretations and sometimes controversies, we believe we have defined in this paper what can be considered as the first reference prior for mixture models.

We have demonstrated that relatively standard simulation algorithms are able

to handle this new parametrisation and that they can manage the computing issues connected with label switching. In case of poor switching, we also established that parallel tempering can be easily implemented. As exhibited in the Ultimixt package, relabelling techniques are readily available.

While the extension to non-Gaussian cases with location-scale parameterisation (and beyond) is conceptually straightforward, considering this parameterisation in higher dimensions is delicate in terms of the covariance matrix. Indeed, even though we can easily set the global variance of the mixture as a parameter, reparameterising the component variances against this reference matrix remains an open question that we have not yet explored.

Chapter 7

Supplementary material: Non-informative reparameterisations for location-scale mixtures

Chapter 6 focuses on the reparametrisation of a mixture of Gaussian distributions with the purpose of using non-informative prior distributions on the parameters. We proceed here by some theoretical aspects, more data analyses and discuss the results.

The model of interest is a convex combination of the univariate Gaussian distributions defined by

$$f(x|\mu_i, \sigma_i) = \sum_{i=1}^{k} p_i \mathcal{N}(x|\mu_i, \sigma_i)$$

A feature of this combination of densities is that it allows to produce a probability density function because of preserving the properties of non negativity and integrating to 1. Multimodality of the produced density is another property which causes the "label switching" issue in the Bayesian analysis of the model as mentioned before while conditions for the number of modes have been explored by [Robertson 1969] and [Behboodian 1970]. As Bayesian methods enable the uncertainty in the model parameters to be directly quantified by examining the posterior distribution, they are useful for fitting these models to data. Despite that the mixture distributions have a range of applications, making an objective choice of prior for the component parameters is difficult in the case where no information is available to determine a subjective prior. Basically, assigning independent improper non-informative priors to the parameters of the mixture components results in improper posterior distribution as shown by [Marin 2006] which unable these prior to be used for the mixtures. This difficulty motivated the idea of shifting the parameters of ith mixture component to two variability parameters α_i and τ_i by defining linear functions $\mu_i = \mu + \sigma\alpha_i, \sigma_i = \sigma\tau_i$ based on μ and σ acting as the intercept and slope of these linear equations. This change leads to a proper posterior derived from a Jeffreys prior for the global parameters of the mixtures, as demonstrated in Chapter 6. Since σ is positive, both μ_i and σ_i are increasing with respect to the values of α_i and τ_i. In a special case where $(\alpha_i, \tau_i); i = 1, \ldots, k$ converges toward $(0, 1)$, the mixture is

transformed to a simple normal distribution with mean and standard deviation μ and σ, respectively.

7.1 Spherical coordinate concept

The modifications $\alpha_i = {}^{\sigma\gamma_i}/_{\sqrt{p_i}}$ and $\tau_i = {}^{\sigma\eta_i}/_{\sqrt{p_i}}$ bring about a hypersphere and hyperplane equations due to the constraints obtained from the mean and variance of the population that drive the resulting mixture model 6.3 more compact in terms of specifying the prior distribution for the resulting variability parameters. In addition, from the intersection of hypersphere $\sum_{i=1}^{k} \gamma_i^2 = \varphi^2$ centered at the origin and the hyperplane $\sum_{i=1}^{k} \sqrt{p_i}\gamma_i = 0$ which also passes through the origin, we deduce that γ_is belong to a circle of radius φ centered at the origin, as mentioned before. For example if we consider $k = 3$, $\varphi = 0.5$ and $p = (0.35, 0.25, 0.4)$, we will have

$$HP_3 : 0.59\gamma_1 + 0.5\gamma_2 + 0.63\gamma_3 = 0; \quad HS_3 : \gamma_1^2 + \gamma_2^2 + \gamma_3^2 = 0.25,$$

and a 3-dimensional graphical representation of $HP_3 \cap HS_3$ is shown in Figure 7.1 which illustrates that the intersection is precisely a set of points in hyperplane at a distance of φ from the origin.

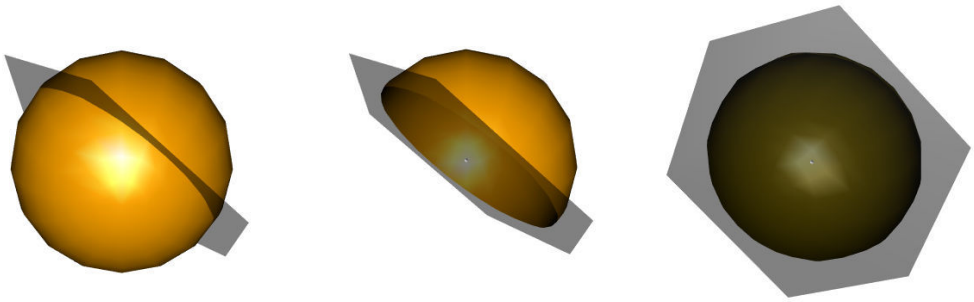


Figure 7.1: **Intersection between 3-dimensional hyperplane and hypersphere**.

For any k, here we represent the spherical coordinate of γ's obtained in Chapter 6 in more details. Suppose that

$$\begin{aligned} HP_k : &\quad \gamma_1\sqrt{p_1} + \gamma_2\sqrt{p_2} + \ldots + \gamma_k\sqrt{p_k} = 0 \\ HS_k : &\quad \gamma_1^2 + \gamma_2^2 + \ldots + \gamma_k^2 = \varphi^2 \end{aligned}$$

Where $\sum_{i=1}^{n} p_i = 1$. From HP_k, two vectors

$$(\sqrt{p_1}, \ldots, \sqrt{p_k}); \quad \Gamma = (\gamma_1, \ldots, \gamma_k)$$

are orthogonal. Let

$$\begin{aligned} \boldsymbol{\Gamma}_k &= (\sqrt{p_1}, \ldots, \sqrt{p_k}) \Big/ \sqrt{\sum_{i=1}^{k} p_i} \\ &= (\sqrt{p_1}, \ldots, \sqrt{p_k}) \end{aligned} \tag{7.1}$$

be a unit-length vector. Since $\boldsymbol{E}$ denotes any Euclidean space of finite dimension k, the hyperplane HP_k has dimension $k-1$ and we can find an orthonormal basis for HP_k as $(\boldsymbol{\Gamma}_1, \ldots, \boldsymbol{\Gamma}_{k-1})$ orthogonal to $\boldsymbol{\Gamma}_k$, [Gallier 2011].
An orthonormal basis including $\boldsymbol{\Gamma}_k$ based on the orthogonal basis $\{\tilde{\boldsymbol{\Gamma}}_1, \ldots, \tilde{\boldsymbol{\Gamma}}_k\}$ is given by

$$\begin{aligned}
\boldsymbol{\Gamma}_1 &= (-\sqrt{p_2}, \sqrt{p_1}, 0, \ldots, 0, 0) \Big/ \sqrt{p_1+p_2} \\
\boldsymbol{\Gamma}_2 &= (-\sqrt{p_1 p_3} \Big/ \sqrt{p_1+p_2}, -\sqrt{p_2 p_3} \Big/ \sqrt{p_1+p_2}, \sqrt{p_1+p_2}, 0, \ldots, 0) \Big/ \sqrt{p_1+p_2+p_3} \\
\boldsymbol{\Gamma}_3 &= (-\sqrt{p_1 p_4} \Big/ \sqrt{p_1+p_2+p_3}, -\sqrt{p_2 p_4} \Big/ \sqrt{p_1+p_2+p_3}, -\sqrt{p_3 p_4} \Big/ \sqrt{p_1+p_2+p_3}, \sqrt{p_1+p_2+p_3}, 0, \ldots, 0) \Big/ \sqrt{p_1+p_2+p_3+p_4} \\
&\ldots \\
\boldsymbol{\Gamma}_{k-1} &= (-\sqrt{p_1 p_k} \Big/ \sqrt{\textstyle\sum_{i=1}^{k-1} p_i}, -\sqrt{p_2 p_k} \Big/ \sqrt{\textstyle\sum_{i=1}^{k-1} p_i}, \ldots, -\sqrt{p_{k-1} p_k} \Big/ \sqrt{\textstyle\sum_{i=1}^{k-1} p_i}, \sqrt{\textstyle\sum_{i=1}^{k-1} p_i}) \Big/ \sqrt{\textstyle\sum_{i=1}^{k} p_i} \\
\boldsymbol{\Gamma}_k &= (\sqrt{p_1}, \ldots, \sqrt{p_k})
\end{aligned}$$

We can easily show that for all $i \in 1, \ldots, k$; $\boldsymbol{\Gamma}_i$ has unit-length and for all $i \neq j$, dot product of $\boldsymbol{\Gamma}_i$ and $\boldsymbol{\Gamma}_j$ is zero. Using this orthonormal basis, any point on the hyperplane can therefore be expressed as

$$\Gamma = b_1 \boldsymbol{\Gamma}_1 + b_2 \boldsymbol{\Gamma}_2 + \ldots + b_k \boldsymbol{\Gamma}_k$$

which gives

$$\begin{aligned}
\gamma_1 &= -b_1 \sqrt{p_2} \Big/ \sqrt{p_1+p_2} - b_2 \sqrt{p_1 p_3} \Big/ \sqrt{p_1+p_2} \sqrt{\textstyle\sum_{i=1}^{3} p_i} - \ldots - b_{k-1} \sqrt{p_1 p_k} \Big/ \sqrt{\textstyle\sum_{i=1}^{k-1} p_i} \sqrt{\textstyle\sum_{i=1}^{k} p_i} + b_k \sqrt{p_1} \\
\gamma_2 &= b_1 \sqrt{p_1} \Big/ \sqrt{p_1+p_2} - b_2 \sqrt{p_2 p_3} \Big/ \sqrt{p_1+p_2} \sqrt{\textstyle\sum_{i=1}^{3} p_i} - \ldots - b_{k-1} \sqrt{p_2 p_k} \Big/ \sqrt{\textstyle\sum_{i=1}^{k-1} p_i} \sqrt{\textstyle\sum_{i=1}^{k} p_i} + b_k \sqrt{p_2} \\
\gamma_3 &= b_2 \sqrt{p_1+p_2} \Big/ \sqrt{p_1+p_2+p_3} - b_3 \sqrt{p_3 p_4} \Big/ \sqrt{\textstyle\sum_{i=1}^{3} p_i} \sqrt{\textstyle\sum_{i=1}^{4} p_i} - \ldots - b_{k-1} \sqrt{p_3 p_k} \Big/ \sqrt{\textstyle\sum_{i=1}^{k-1} p_i} \sqrt{\textstyle\sum_{i=1}^{k} p_i} + b_k \sqrt{p_3} \\
&\ldots \\
\gamma_{k-1} &= b_{k-2} \sqrt{\textstyle\sum_{i=1}^{k-2} p_i} \Big/ \sqrt{\textstyle\sum_{i=1}^{k-1} p_i} - b_{k-1} \sqrt{p_{k-1} p_k} \Big/ \sqrt{\textstyle\sum_{i=1}^{k-1} p_i} \sqrt{\textstyle\sum_{i=1}^{k} p_i} + b_k \sqrt{p_{k-1}} \\
\gamma_k &= b_{k-1} \sqrt{\textstyle\sum_{i=1}^{k-1} p_i} \Big/ \sqrt{\textstyle\sum_{i=1}^{k} p_i} + b_k \sqrt{p_k}.
\end{aligned}$$

$(\gamma_1, \ldots, \gamma_k)$ belongs to both HP_k and HS_k and thus replacing it in hyperplane results in

$$\begin{aligned}\sqrt{p_1}\gamma_1 + \sqrt{p_2}\gamma_2 + \ldots + \sqrt{p_k}\gamma_k = &-b_1\sqrt{p_1p_2}/\sqrt{p_1+p_2} + \ldots + b_kp_1\\ &+ b_1\sqrt{p_1p_2}/\sqrt{p_1+p_2} + \ldots + b_kp_2\\ &+ \ldots\\ &- b_{k-1}p_{k-1}\sqrt{p_k}/\sqrt{\textstyle\sum_{i=1}^{k-1} p_i}\sqrt{\textstyle\sum_{i=1}^{k} p_i} + b_kp_{k-1}\\ &+ b_{k-1}\sqrt{p_k}\sqrt{\textstyle\sum_{i=1}^{k-1} p_i}/\sqrt{\textstyle\sum_{i=1}^{k} p_i} + b_kp_k\end{aligned}$$

and by canceling positive and negative similar terms we obtain $\sum_{i=1}^{k} b_kp_i = 0$ that ends up with $b_k = 0$. In this case, replacing γ_i's in HS_k leads to a hypersphere of radius φ in $k-1$-dimensional Euclidean space

$$b_1^2 + b_2^2 + \ldots + b_{k-1}^2 = \varphi^2$$

and thus any reparametrization of this object such as spherical coordinate system may be considered.

In the special case of $k = 2$, the orthonormal basis is defined by two following vectors

$$F_1 = (-\sqrt{p_2}/\sqrt{p_1+p_2}, \sqrt{p_1}/\sqrt{p_1+p_2}); \quad F_2 = (\sqrt{p_1}, \sqrt{p_2})$$

where $p_1 = p; p_2 = 1-p$ and we can therefore write

$$\begin{aligned}(\gamma_1, \gamma_2) &= b_1F_1 + b_2F_2\\ &= (-b_1\sqrt{1-p} + b_2\sqrt{p}, b_1\sqrt{p} + b_2\sqrt{1-p})\end{aligned}$$

Replacing γ_1 and γ_2 above in both

$$HP_2 : \sqrt{p}\gamma_1 + \sqrt{1-p}\gamma_2 = 0; \quad HS_2 : \gamma_1^2 + \gamma_2^2 = \varphi^2$$

yields

$$\begin{aligned}-b_1\sqrt{p(1-p)} + b_2p + b_1\sqrt{p(1-p)} + b_2(1-p) = 0; \quad b_2 = 0\\ b_1^2(1-p) + b_1^2p = \varphi^2; \quad b_1^2 = \varphi^2\end{aligned}$$

$b_1^2 = \varphi^2$ can be considered as a sphere in 1-dimensional Euclidean space that represents a pair of points $\{-\varphi, +\varphi\}$ which is the boundary of a line segment (a part of a line that is bounded by two distinct end points, and contains every point on the line between its endpoints). We can therefore rewrite γ_i's as

$$(\gamma_1, \gamma_2) = (\pm\varphi\sqrt{1-p}, \pm\varphi\sqrt{p}).$$

Since $\eta_1^2 + \eta_2^2 = 1 - \varphi^2$, spherical coordinate representation of η_1 and η_2 will be

$$\eta_1 = \sqrt{1-\varphi^2}\cos(\xi); \quad \eta_2 = \sqrt{1-\varphi^2}\sin(\xi)$$

where $\xi \in [0, \pi/2]$. In example 6.4.1, the analyses related to 50 data points simulated from the mixture $0.65\mathcal{N}(-8, 2) + 0.35\mathcal{N}(-0.5, 1)$ are based on the case where a Dirichlet prior $\mathcal{D}ir(0.5, 0.5, 0.5)$ is assigned to $(\varphi^2, \eta_1^2, \eta_2^2)$. It means that only the component means are expressed in spherical coordinates. Here, we reanalyze this model by considering the spherical coordinates above for η_i's. In this case, we place a beta prior on the parameter φ^2 with the same hyper parameters $0.5, 0.5$. Figure 7.2 shows that the estimates of the marginal posterior distributions of means and standard deviations are symmetric and each μ_i (σ_i) is very similar to one another due to the label switching phenomenon. As the component parameters are not identifiable marginally, estimating them on the basis of these MCMC output is not straightforward and we thus revert to the k-means clustering algorithm. The procedure is implemented by using the package `Ultimixt` in Chapter 8. The estimations of the component parameters displayed in Table 7.1 are quite similar to those of the true model while the calibration of the proposal scale results in the acceptance rates close to the optimal.

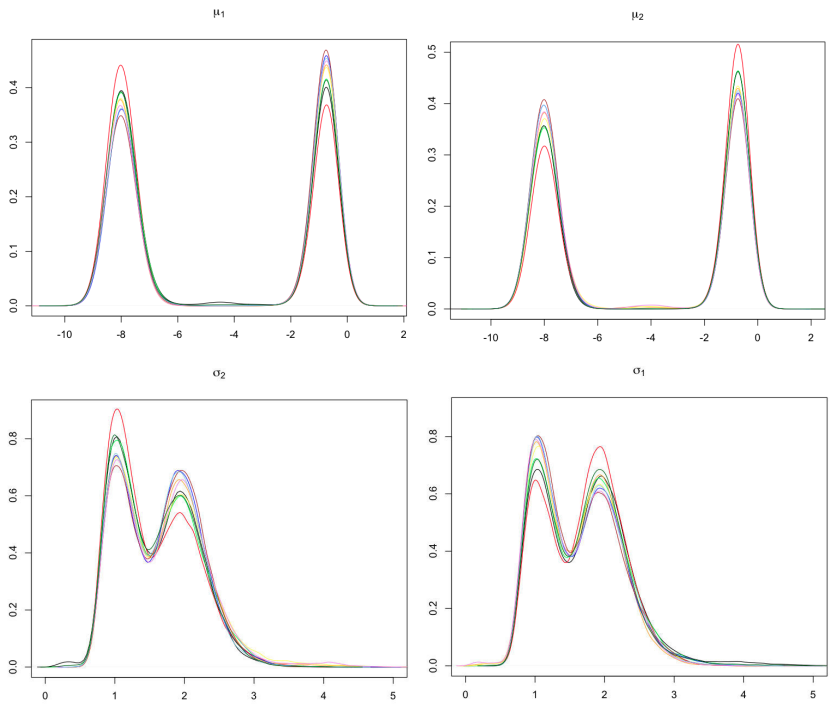


Figure 7.2: **Mixture of two normal distributions** 7.1: Estimated marginal posterior densities of component means μ_1, μ_2 and standard deviations σ_1, σ_2 based on 10^5 MCMC iterations.

On the other hand, since the estimate of the predictive density based on the MCMC output does not depend on the labelling of the components, Figure 7.3 shows that the estimated mixture is very smooth and unaffected by the label switching. Note that the black line is the estimate of the density that is obtained by averaging the simulated densities over the last 500 iterations.

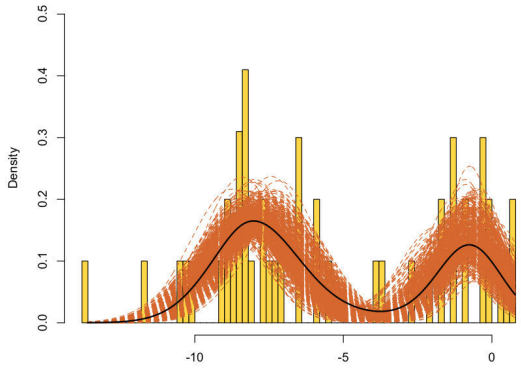


Figure 7.3: **Mixture of two normal distributions** 7.1: Representation of 500 MCMC iterations as mixture distributions with the overlaid average curve (dark line).

	k-means clustering					
	μ	σ	φ	ξ		
Med.	-5.35	3.89	0.89	.747		
Mean	-5.34	3.89	0.87	.810		
	p_1	p_2	μ_1	μ_2	σ_1	σ_2
Med.	0.64	0.36	-7.99	-0.75	1.12	2.05
Mean	0.64	0.36	-7.95	-0.79	1.18	2.09
	Proposal scales					
	s_μ	s_σ	ε_p	ε_φ	ε_ξ	
	0.56	0.11	65	540	0.29	
	Acceptance rates					
	ar_μ	ar_σ	ar_p	ar_φ	ar_ξ	
	0.38	0.48	0.43	0.43	0.43	

Table 7.1: **Mixture of two normal distributions** 7.1: Point estimates, proposal scales and acceptance rates.

The results of the example above illustrates that when $k = 2$, the Bayes estimates of the mixture parameters based on the spherical coordinate of η_i's are identical to the case where $(\varphi^2, \eta_1^2, \eta_2^2)$ is supposed to be from Dirichlet distribution. In both cases, the parameters are accurately estimated as long as the convergence towards the stationary distribution is achieved as shown in Figure 7.2.

In the following, we apply the reparametrisation of the mixture distribution based on the spherical coordinate of the component parameters for some other datasets and summarize the resulting Bayesian analyses by implementing the functions of `Ultimixt` package, Chapter 8.

7.2 Data analyses

Two datasets *Acidity dataset* and *Enzyme dataset* were initially used by [Richardson 1997] while *Fishery dataset* and *Darwin's dataset* are taken from [Frühwirth-Schnatter 2006]. For all datasets, we run the MCMC algorithm with 10^4 iterations and in Figures 7.4, 7.5, 7.6 and 7.7, the predictive mixture densities are computed once for the last 500 iterations another by averaging over 10^4 iterations (dark line in the figures).

7.2.1 Acidity data

The *Acidity dataset* is related to an acidity index measured in a sample of 155 lakes in the Northeastern United States. A histogram of the data points is shown in 7.4 and so a mixture of 2 components is well suited to model the data. In Figure 7.4 the histogram of the data is overlaid with the predictive density estimate obtained by applying `Ultimixt` package 8 that indicates the model represents the data well, with no need for any more components.

	Angular & component-wise parameters					
	Acidity data		Enzyme data		Darwin's data	
	ξ		ξ		ξ	
Med	0.71		1.39		0.13	
Mean	0.75		1.4		0.12	
	p_1	p_2	p_3	p_1	p_2	p_3
Med	0.40	0.60	0.41	0.59	0.85	0.15
Mean	0.41	0.59	0.40	0.60	0.84	0.16
	μ_1	μ_2	μ_1	μ_2	μ_1	μ_2
Med	6.24	4.32	1.25	0.19	21.68	12.0
Mean	6.23	4.33	1.24	0.19	21.63	12.2
	σ_1	σ_2	σ_1	σ_2	σ_1	σ_2
Med	0.53	0.38	0.52	0.08	0.33	1.5
Mean	0.54	0.37	0.53	0.08	0.47	1.4

		Global parameters		
Acidity	Obs.	5.1	1.04	
		μ	σ	φ
	Med	5.1	1.04	0.90
	Mean	5.1	1.04	0.89
Enzyme	Obs.	0.62	0.62	
		μ	σ	φ
	Med	0.62	0.63	0.84
	Mean	0.62	0.62	0.83
Darwin	Obs.	20.2	3.6	
		μ	σ	φ
	Med	20.2	3.7	0.92
	Mean	20.2	3.6	0.92

Table 7.2: **Data analyses 7.2**: Point estimators of the parameters of a mixture of two components. Each estimate is obtained based on 10^4 MCMC iterations. Med indicates the estimate based on the median of draws and two values behind Obs. are the mean and standard deviation of the dataset $\bar{x} = \sum_{j=1}^{n} {}^{x_j}/_n; s = (\sum_{j=1}^{n} {}^{(x_j-\bar{x})}/_{(n-1)})^{1/2}$, respectively.

7.2.2 Enzyme data

Enzyme dataset concerns the distribution of enzymatic activity in the blood, for an enzyme involved in the metabolism of carcinogenic substances, among of group of 245 unrelated individuals with the purpose of identifying subgroups of slow or fast metabolizers as a marker of genetic polymorphism in the general population. This dataset has been reanalyzed by [van Havre 2014] who compared the probabilities that this data can be modeled by two or three components. [van Havre 2014] showed that fitting a mixture model with two components to the data is more likely than three components. We therefore analyze *Enzyme dataset* with our reparametrized mixture distribution by considering $k = 2$. The predictive density estimates shown in Figure 7.5 illustrate a good fitting of the mixture with two components to the data.

7.2.3 Darwin's data

Darwin's data consists of 15 observations of differences in heights between pairs of self-fertilized and cross-fertilized plants grown under the same condition. The histogram of the data overlaid with the predictive density estimates based on fitting a mixture of two normal components to the data is shown in Figure 7.7. Despite that the sample size is small, the mixture estimate fits very well the data.

Table 7.2 shows the point estimates of the component parameters including mean and median for the datasets *Acidity, Enzyme* and Darwin. The table reveals very negligible difference between the estimates based on mean and median of the posterior draws and the errors between the point estimates of μ and $\bar{x}$ are zero for all datasets while for σ the error between the median of the posterior draws and the standard deviations of datasets are negligible (the error of posterior estimates based on the average of draws is also zero for all datasets).

	Global parameters; $k = 4$		
	μ	σ	φ
Med	6.1	1.90	0.81
Mean	6.1	1.89	0.80

Angular & component parameters				
	ϖ_1	ϖ_2		
Med	2.5	5.3		
Mean	2.4	5.4		
	ξ_1	ξ_2	ξ_3	
Med	0.27	0.49	0.73	
Mean	0.28	0.48	0.73	
	p_1	p_2	p_3	p_4
Med	0.58	0.25	0.14	0.03
Mean	0.56	0.26	0.14	0.04
	μ_1	μ_2	μ_3	μ_4
Med	5.2	7.7	3.5	3.2
Mean	5.2	7.9	3.6	3.2
	σ_1	σ_2	σ_3	σ_4
Med	0.24	0.68	1.02	0.74
Mean	0.25	0.71	1.23	0.74

	Global parameters; $K = 3$		
Obs.	6.1	1.9	
	μ	σ	φ
Med	6.1	1.90	0.71
Mean	6.1	1.90	0.70

Angular & component parameters			
	ϖ	ξ_1	ξ_2
Med	3.9	0.34	0.22
Mean	3.7	0.34	0.21
	p_1	p_2	p_3
Med	0.38	0.53	0.09
Mean	0.38	0.54	0.08
	μ_1	μ_2	μ_3
Med	7.3	5.2	3.2
Mean	7.3	5.2	3.3
	σ_1	σ_2	σ_3
Med	0.53	0.29	1.8
Mean	0.52	0.30	1.8

Table 7.3: **Fishery data** 7.2: Point estimators of the parameters of a mixture of two components. Each estimate is obtained based on 10^4 MCMC iterations. Med indicates the estimate based on the median of draws and two value behind Obs. are the mean and standard deviation of the dataset $\bar{x} = \sum_{j=1}^{n} x_j/n; s = (\sum_{j=1}^{n} (x_j - \bar{x})/(n-1))^{1/2}$, respectively.

7.2.4 Fishery data

We fit a Gaussian mixture model on the *Fishery data* which consists of data on the lengths of 256 snappers. The heterogeneity in the data comes from the different age groups a fish might belong to depending if it comes from the current year's spawning or the previous, and so on.

Overfitting the *Fishery dataset* produces two possible alternate configurations with four or three components and we therefore analyze this data once by considering $k = 4$ another by modeling a mixture of $k = 3$ components. Table 7.3 displays the posterior parameters describing each configuration. When $k = 4$, we obtain two almost similar component means $\mu_3 = 3.6$ and $\mu_4 = 3.2$ and since one of them has an estimated weight close to zero this case can be considered as overfitted model. The Bayes estimates of the parameters of a mixture with three components exhibit a better performance as shown on the right of Table 7.3. We note here that the point estimates of global parameters μ and σ in the case where $k = 4$ are almost identical to the ones for $k = 3$. This indicates that the change of the component number of the mixture model does not impact the Bayesian inference of these two parameters. The radius φ decreases when k diminishes from 4 to 3 and since σ_i's are expressed by spherical coordinate with radius $(1 - \varphi^2)^{1/2}$, the component standard deviations naturally increase in comparison with the case of $k = 4$.

Figure 7.6 shows the posterior predictive densities based on MCMC output for the case where $k = 4$ that is almost identical to the one for $k = 3$. This illustrates that the estimated mixtures are unaffected by the overfitting phenomenon as mentioned before for *Galaxy data* in Chapter 6.

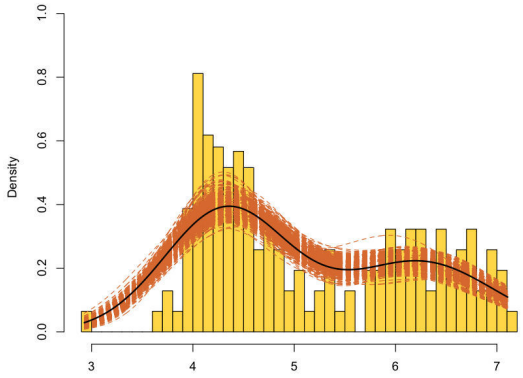


Figure 7.4: ***Acidity data*** 7.2.1.

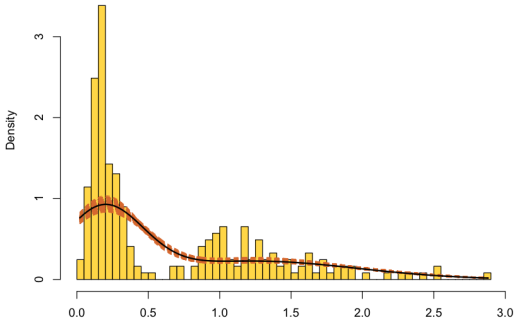


Figure 7.5: ***Enzyme data*** 7.2.2.

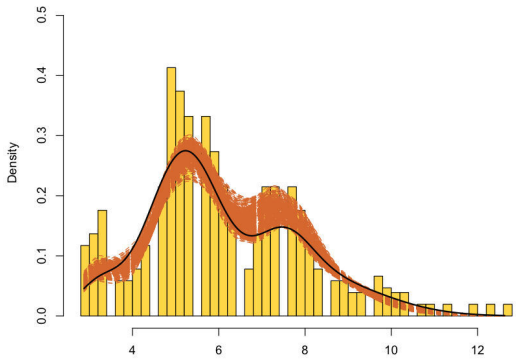


Figure 7.6: ***Fishery data*** 7.2.4.

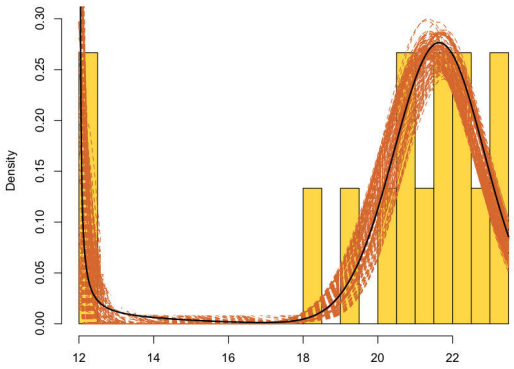


Figure 7.7: ***Darwin's data*** 7.2.3.

Note that in both Tables 7.2 and 7.3, for the global parameters, the means and medians are computed based the obtained MCMC draws. These summary statistics are considered as the posterior estimations while for the angular and component-wise parameters the estimates are obtained by applying k-means clustering algorithm on the related posterior samples.

7.3 Parallel tempering algorithm

Last part of Chapter 6 deals with the analyses based on parallel tempering method. We showed that this method greatly improves mixing of MCMC chains. In some examples, we have seen that for a sample of size large enough, multimodality of the mixture model causes mixing problem or eventuates a good fit of one of the components to the extent that it is difficult for the Metropolis-within-Gibbs sampler to escape. In this case, using parallel tempering or replica exchange is suggested [Gill 2004] which can be tracked down in a paper written by [Swendsen 1986]. The applications of this method has used not only for problems in statistical physics but in chemistry, biology, engineering and materials science [Earl 2005]. The important

feature is preventing the Markov chain from sticking in minor modal areas for long periods of time. Parallel tempering MCMC as a method for generating candidate samples from all over a distribution is designed to overcome low probability regions between areas of importance. This means that a temperature parameter could be used to flatten out the target distribution such that the more temperature raised the more distribution flattens out and this makes the random chain more likely to mix quickly, for that temperature [Lewandowski 2014, Gill 2004, Earl 2005, Neal 1996, Wang 2011, Li 2009, Swendsen 1986].

The method is that the target distribution is transformed to the Boltzmann distribution for a given temperature which is called "replicas". In order to simulate parameter θ from a posterior distribution non-standard $f(\theta|x)$, in temperature $1/\beta$, the replicas is defined as $f_\beta(\theta|x) = f(\theta|x)^\beta$ where $0 \leq \beta \leq 1$. This means that the chains can be constructed from "tempered" versions of the target of interest by raising it to a power between 0 and 1, with 0 corresponding to a complete flattening of the distribution, and 1 corresponding to the desired target [Altekar 2004, Geyer 2011, Geyer 1991]. [Miasojedow 2013] argue that for a target distribution $f()$, tempering of $f()$ often provides better mixing within modes of the target distribution. However, this method is often more effective than the non-tempered approach [Hamze 2010].

Having this expression for the target distribution, we run the basic MCMC algorithm on each distribution and in each iteration, the current samples are considered probabilistically for exchanges between different temperature levels with probability α_h as shown in parallel tempering algorithm below. Note that here only pairs between neighboring temperatures are considered for swapping because the chances of accepting a trade are more likely to be higher. Before performance of parallel tempering MCMC becomes optimal, we should tune the number of replicas h and their temperatures which are not actually evident and besides this, the simulation of multiple chains does increase the computation time. Several suggestions for the number of replicas and temperature of the replicas have been offered. An example is a geometric progression of temperatures [Miasojedow 2013, Earl 2005]. [Miasojedow 2013] propose an adaptive algorithm in order to tune the temperature schedule and the parameters of the random-walk Metropolis kernel. Here, we consider that the proposal scale is automatically calibrated using adaptive Metropolis-within-Gibbs algorithm for the target distribution according to the optimal acceptance rates as explained before and it is simultaneously used for the proposal distribution of all tempered versions of the target distribution. The number of temperatures is automatically chosen according to the degree of the symmetry of the generated samples of p to 0.5 in the case where the number of mixture components is $k = 2$. The number of temperatures is also determined according to the acceptance rate of the swaps between the neighboring temperatures, ar_{swap}.

Let $\theta = (\mu, \sigma, p, \varphi, \xi, \varpi)$, ε_θ be the scale of the proposal distribution in Metropolis-within-Gibbs step. We suppose that δ_1 is the symmetry threshold and δ_2 acceptance rate of swaps threshold. We therefore define parallel tempering algorithm in 7.8. Note that if R_1 and R_2 indicate the ratios of the simulated samples for p, greater

Parallel tempering algorithm for reparameterised mixture when $k = 2$:

1 Generate initial values $\theta^{(0)}$;

2 Compute ε_θ by implementing the adaptive Metropolis-within-Gibbs algorithm 9.1.

3 Input $N_\beta = 2$ and choose $\beta = (\beta_1, \beta_2)$ such that $2^{-N_\beta} < \beta_2 < \beta_1$.

4 For $t = 1, \ldots, T$, the update of $\theta^{(t)}$ follows

4.1 For $h = 1, \ldots, N_\beta$

- ✓ Generate $\theta_h^{(t)}$ from MCMC$(\theta_h^{(t)}, \pi_{\beta_h})$.
- ✓ Accept $\theta_{h-1}^{[t]} = \theta_h^{[t]}$ with probability

$$\alpha_h = \min\left(1, f_{\beta_{h-1}}(\theta_h^{[t]}) f_{\beta_h}(\theta_{h-1}^{[t]}) / f_{\beta_{h-1}}(\theta_{h-1}^{[t]}) f_{\beta_h}(\theta_h^{[t]})\right).$$

5 Stop if $d(R_1, R_2) < \delta_1$ or $\min(ar_{swap}) > \delta_2$.

6 Otherwise compute $\beta = (\beta, 2^{-N_\beta})$, $N_\beta = N_\beta + 1$ and go to line 4.

Figure 7.8: MCMC(θ, π) denotes a Metropolis-within-Gibbs step defined in 6.1 with starting point θ and target distribution π and N_β is the number of temperatures.

than and less than 0.5 to the total number of iterations, respectively, the degree of symmetry of draws of p to 0.5 is computed by $|R_1 - R_2|$.

The output of the algorithm 7.8 consists of the chain of samples for θ, the number of replicas and the temperature set, $1/\beta$. Both examples 6.4.2 and 6.5.2 illustrate the good performance of this algorithm in terms of improving the mixing of the MCMC chains.

Chapter 8

Ultimixt package

8.1 Ultimixt

Type Package
Title Bayesian Analysis of a Non-Informative Parametrization for Gaussian Mixture Distributions
Version 1.0
Date 2015-12-10
Author Kaniav Kamary, Kate Lee
Maintainer Kaniav Kamary <kamary@ceremade.dauphine.fr>
Depends coda, gtools, graphics, grDevices, stats
Description A generic reference Bayesian analysis of unidimensional mixtures of Gaussian distributions obtained by a location-scale parameterisation of the model is implemented. Included functions can be applied to produce a Bayesian analysis of Gaussian mixtures with an arbitrary number of components, with no need to define the prior distribution.
License GPL (>=2.0)

R topics documented:

Contents

Index

Description

Despite a comprehensive literature on estimating mixtures of Gaussian distributions,

there does not exist a well-accepted reference Bayesian approach to such models. One reason for the difficulty is the general prohibition against using improper priors (Fruhwirth-Schnatter, 2006) due to the ill-posed nature of such statistical objects. Kamary, Lee and Robert (2015) took advantage of a mean-variance reparametrisation of a Gaussian mixture model to propose improper but valid reference priors in this setting. This R package implements the proposal and computes posterior estimates of the parameters of a Gaussian mixture distribution. The approach applies with an arbitrary number of components. The Ultimixt R package contains an MCMC algorithm function and further functions for summarizing and plotting posterior estimates of the model parameters for any number of components.

Details

Package:	Ultimixt
Type:	Package
Version:	1.0
Date:	2015-10-30
License:	GPL (>=2.0)

Beyond simulating MCMC samples from the posterior distribution of the Gaussian mixture model, this package also produces summaries of the MCMC outputs through numerical and graphical methods.

Note: The proposed parameterisation of the Gaussian mixture distribution is given by

$$f(x|\mu,\sigma,\mathbf{p},\varphi,\varpi,\xi) = \sum_{i=1}^{k} p_i f\left(x|\mu+\sigma\gamma_i/\sqrt{p_i},\sigma\eta_i/\sqrt{p_i}\right)$$

under the non-informative prior $\pi(\mu,\sigma) = 1/\sigma$. Here, the vector of the $\gamma_i = \varphi\Psi_i\Big(\varpi,\mathbf{p}\Big)_i$'s belongs to an hypersphere of radius φ intersecting with an hyperplane. It is thus expressed in terms of spherical coordinates within that hyperplane that depend on $k-2$ angular coordinates ϖ_i. Similarly, the vector of $\eta_i = \sqrt{1-\varphi^2}\Psi_i\Big(\xi\Big)_i$'s can be turned into a spherical coordinate in a k-dimensional Euclidean space, involving a radial coordinate $\sqrt{1-\varphi^2}$ and $k-1$ angular coordinates ξ_i. A natural prior for ϖ is made of uniforms, $\varpi_1,\ldots,\varpi_{k-3} \sim U[0,\pi]$ and $\varpi_{k-2} \sim U[0,2\pi]$, and for φ, we consider a beta prior $Beta(\alpha,\alpha)$. A reference prior on the angles ξ is $(\xi_1,\ldots,\xi_{k-1}) \sim U[0,\pi/2]^{k-1}$ and a Dirichlet prior $Dir(\alpha_0,\ldots,\alpha_0)$ is assigned to the weights $p_1,\ldots,p_k$

Author(s)

Kaniav Kamary
Maintainer: <kamary@ceremade.dauphine.fr>

References

Fruhwirth-Schnatter, S. (2006). Finite Mixture and Markov Switching Models. Springer-Verlag, New York, New York.
Kamary, K., Lee, J.Y., and Robert, C.P. (2015) Non-informative reparameterisation

of location-scale mixtures. arXiv.

See Also

Ultimixt

Examples

```
data(faithful)
xobs=faithful[,1]
estimate=K.MixReparametrized(xobs, k=2, alpha0=.5, alpha=.5, Nsim=1e4)
```

8.2 K.MixReparametrized function

`K.MixReparametrized`

Sample from a Gaussian mixture posterior associated with a noninformative prior and obtained by Metropolis-within-Gibbs sampling

---

Description

This function returns a sample simulated from the posterior distribution of the parameters of a Gaussian mixture under a non-informative prior. This prior is derived from a mean-variance reparameterisation of the mixture distribution, as proposed by Kamary et al. (2015). The algorithm is a Metropolis-within-Gibbs scheme with an adaptive calibration of the proposal distribution scales. Adaptation is driven by the formally optimal acceptance rates of 0.44 and 0.234 in one and larger dimensions, respectively (Roberts et al.,1997). This algorithm monitors the convergence of the MCMC sequences via Gelman's and Rubin's (1992) criterion.

Usage

```
K.MixReparametrized(xobs, k, alpha0, alpha, Nsim)
```

Arguments

`xobs`	vector of the observations or dataset
`k`	number of components in the mixture model
`alpha0`	hyperparameter of Dirichlet prior distribution of the mixture model weights which is .5 by default
`alpha`	hyperparameter of beta prior distribution of the radial coordinate which is .5 by default
`Nsim`	number of MCMC iterations after calibration step of proposal scales

Details

The output of this function contains a simulated sample for each parameter of the mixture distribution, the evolution of the proposal scales and acceptance rates over the number of iterations during the calibration stage, and their final values after calibration.

Value

The output of this function is a list of the following variables, where the dimension of the vectors is the number of simulations:

`mean global`	vector of simulated draws from the conditional posterior of the mixture model mean
`sigma global`	vector of simulated draws from the conditional posterior of the mixture model standard deviation
`weights`	matrix of simulated draws from the conditional posterior of the mixture model weights with a number of columns equal to the number of components k
`angles xi`	matrix of simulated draws from the conditional posterior of the angular coordinates of the component standard deviations with a number of columns equal to $k-1$
`phi`	vector of simulated draws from the conditional posterior of the radian coordinate
`angles varpi`	matrix of simulated draws from the conditional posterior of the angular coordinates defined for component means with a number of columns equal to $k-2$
`accept rat`	vector of resulting acceptance rates of the proposal distributions without calibration step of the proposal scales
`optimal para`	vector of resulting proposal scales after optimization obtained by adaptive MCMC
`adapt rat`	list of acceptance rates of batch of 50 iterations obtained when calibrating the proposal scales by adaptive MCMC. The number of columns depends on the number of proposal distributions.
`adapt scale`	list of proposal scales calibrated by adaptive MCMC for each batch of 50 iterations with respect to the optimal acceptance rate. The number of columns depends on the number of proposal distribution scales.
`component means`	matrix of MCMC samples of the component means of the mixture model with a number of columns equal to k
`component sigmas`	matrix of MCMC samples of the component standard deviations of the mixture model with a number of columns equal to k

Note: The number of the iterations in this algorithm is automatically determined depending on the convergence of the generated samples for the means and standard deviations of the components.

Author(s)

Kaniav Kamary

References

Kamary, K., Lee, J.Y., and Robert, C.P. (2015) Non-informative reparameterisation

of location-scale mixtures. arXiv.
Robert, C. and Casella, G. (2009). Introducing Monte Carlo Methods with R. Springer-Verlag.
Roberts, G. O., Gelman, A. and Gilks, W. R. (1997). Weak convergence and optimal scaling of random walk Metropolis algorithms. *Ann. Applied Probability*, 7, 110–120.
Gelman, A. and Rubin, D. (1992). Inference from iterative simulation using multiple sequences (with discussion). *Statistical Science*, 457–472.

See Also
Ultimixt

Examples
```
data(faithful)
xobs=faithful[,1]
estimate=K.MixReparametrized(xobs, k=2, alpha0=.5, alpha=.5, Nsim=10000)
```

8.3 Plot.MixReparametrized function

`Plot.MixReparametrized` *plot of the MCMC output produced by K.MixReparametrized*

Description
This is a generic function for a graphical rendering of theMCMCsamples produced by K.MixReparametrized function. The function draws boxplots for unimodal variables and for multimodal arguments after clustering them by applying a k-means algorithm. It also plots line charts for other variables..

Usage
```
plot.MixReparametrized(xobs, estimate)
```

Arguments

`xobs`	vector of the observations
`estimate`	output of the K. MixReparametrized function

Details
Boxplots are produced using the boxplot.default method.

Value
The output of this function consists of

`boxplot`	three boxplots for the radial coordinates, the mean and the standard deviation of the mixture distribution, k boxplots for each of the mixture model weights, component means and component standard deviations.
`histogram`	an histogram of the observations against an overlaid curve of the density estimate, obtained by averaging over all mixtures corresponding to the MCMC draws,
`line chart`	line charts that report the evolution of the proposal scales and of the acceptance rates over the number of batch of 50 iterations.

Note: The mixture density estimate is based on the draws simulated of the parameters obtained by K.MixReparametrized function.

Author(s)

Kaniav Kamary

References

Kamary, K., Lee, J.Y., and Robert, C.P. (2015) Non-informative reparameterisation of location-scale mixtures. arXiv.

See Also

K.MixReparametrized

Examples

```
data(faithful)
xobs=faithful[,1]
estimate=K.MixReparametrized(xobs, k=2, alpha0=.5, alpha=.5, Nsim=20000)
plo=Plot.MixReparametrized(xobs, estimate)
```

8.4 SM.MAP.MixReparametrized function

`SM.MixReparametrized` *summary of the output produced by K.MixReparametrized*

Description

Label switching in a simulated Markov chain produced by K.MixReparametrized is removed by the technique of Marin et al. (2004). Namely, component labels are reordered by the shortest Euclidian distance between a posterior sample and the maximum a posteriori (MAP) estimate. Let θ_i be the i-th vector of computed component means, standard deviations and weights. The MAP estimate is derived from the MCMC sequence and denoted by θ_{MAP}. For a permutation $\tau \in \Im_k$ the labelling of θ_i is reordered by

$$\tilde{\theta}_i = \tau_i(\theta_i)$$

where $\tau_i = \arg\min_{\tau \in \Im_k} || \tau(\theta_i) - \theta_{MAP} ||$.

Angular parameters $\xi_1^{(i)}, \ldots, \xi_{k-1}^{(i)}$ and $\varpi_1^{(i)}, \ldots, \varpi_{k-2}^{(i)}$s are derived from $\tilde{\theta}_i$. There exists an unique solution in $\varpi_1^{(i)}, \ldots, \varpi_{k-2}^{(i)}$ while there are multiple solutions in $\xi^{(i)}$ due to the symmetry of $|\cos(\xi)|$ and $|\sin(\xi)|$. The output of $\xi_1^{(i)}, \ldots, \xi_{k-1}^{(i)}$ only includes angles on $[-\pi, \pi]$.

The label of components of θ_i (before the above transform) is defined by

$$\tau_i^* = \arg \min_{\tau \in \Im_k} || \theta_i - \tau(\theta_{MAP}) || .$$

The number of label switching occurrences is defined by the number of changes in τ^*.

Usage

```
SM.MAP.MixReparametrized(estimate, xobs, alpha0, alpha)
```

Arguments

`estimate`	output of K.MixReparametrized
`xobs`	Data set
`alpha0`	Hyperparameter of Dirichlet prior distribution of the mixture model weights
`alpha`	Hyperparameter of beta prior distribution of the radial coordinate

Details

Details.

Value

`MU`	Matrix of MCMC samples of the component means of the mixture model
`SIGMA`	Matrix of MCMC samples of the component standard deviations of the mixture model
`P`	Matrix of MCMC samples of the component weights of the mixture model
`Ang-SIGMA`	Matrix of computed ξ's corresponding to SIGMA
`Ang-MU`	Matrix of computed ϖ's corresponding to MU. This output only appears when $k > 2$
`Global-mean`	Mean, median and 95% credible interval for the global mean parameter
`Global-Std`	Mean, median and 95% credible interval for the global standard deviation parameter
`Phi`	Mean, median and 95% credible interval for the radius parameter
`component-mu`	Mean, median and 95% credible interval of MU
`component-sigma`	Mean, median and 95% credible interval of SIGMA
`component-p`	Mean, median and 95% credible interval of P
`l-stay`	Number of MCMC iterations between changes in labelling
`n-switch`	Number of label switching occurrences

Note:

Note.

Author(s)
Kate Lee

References
Marin, J.-M., Mengersen, K. and Robert, C. P. (2004) Bayesian Modelling and Inference on Mixtures of Distributions, Handbook of Statistics, Elsevier, Volume 25, Pages 459–507.

See Also
K.MixReparametrized

Examples
```
data(faithful)
xobs=faithful[,1]
estimate=K.MixReparametrized(xobs, k=2, alpha0=.5, alpha=.5, Nsim=20000)
result=SM.MAP.MixReparametrized(estimate,xobs,alpha0=0.5,alpha=0.5)
```

8.5 SM.MixReparametrized function

`SM.MixReparametrized` *summary of the output produced by K.MixReparametrized*

Description
This is a generic function that summarizes the MCMC samples produced by K.MixReparametrized. The function invokes several estimation methods which choice depends on the unimodality or multimodality of the argument.

Usage
```
SM.MixReparametrized(xobs, estimate)
```

Arguments

`xobs`	vector of the observations
`estimate`	output of K.MixReparametrized

Details
This function outputs posterior point estimates for all parameters of the mixture model. They mostly differ from the generaly useless posterior means. The output summarizes unimodal MCMC samples by computing measures of centrality, including mean and median, while multimodal outputs require a pre-processing, due to the label switching phenomenon (Jasra et al., 2005). The summary measures are then computed after performing a multi-dimensional k-means clustering (Hartigan and Wong, 1979) following the suggestion of Fruhwirth-Schnatter (2006).

Value

`Mean`	vector of mean and median of simulated draws from the conditional posterior of the mixture model mean
`Sd`	vector of mean and median of simulated draws from the conditional posterior of the mixture model standard deviation
`Phi`	vector of mean and median of simulated draws from the conditional posterior of the radial coordinate
`Angles. 1.`	vector of means of the angular coordinates used for the component means in the mixture distribution
`Angles. 2.`	vector of means of the angular coordinates used for the component standard deviations in the mixture distribution
`weight.i`	vector of mean and median of simulated draws from the conditional posterior of the component weights of the mixture distribution; $i = 1, \ldots, k$
`mean.i`	vector of mean and median of simulated draws from the conditional posterior of the component means of the mixture distribution; $i = 1, \ldots, k$
`sd.i`	vector of mean and median of simulated draws from the conditional posterior of the component standard deviations of the mixture distribution; $i = 1, \ldots, k$
`Acc rat`	vector of final acceptance rate of the proposal distributions of the algorithm with no calibration stage for the proposal scales
`Opt scale`	vector of optimal proposal scales obtained by calibration stage

Note: For multimodal outputs such as the mixture model weights, component means, and component variances, for each MCMC draw, first the labels of the weights $p_i, i = 1, \ldots, k$ and corresponding component means and standard deviations are permuted in such a way that $p_1 \leq \ldots \leq p_k$. Then the component means and standard deviations are jointly partitioned into k clusters by applying a standard k-means algorithm with k clusters to a sample of size Tk (where T is the number of iterations), following Fruhwirth-Schnatter (2006) method. For each group, cluster centers are considered as parameter estimates.

Author(s)

Kaniav Kamary

References

Jasra, A., Holmes, C. and Stephens, D. (2005). Markov Chain Monte Carlo methods and the label switching problem in Bayesian mixture modeling. *Statistical Science*, 20, 50–67.

Hartigan, J. A. and Wong, M. A. (1979). A K-means clustering algorithm. *Applied Statistics* 28, 100–108.

Fruhwirth-Schnatter, S. (2006). Finite mixture and Markov switching models. Springer-Verlag.

See Also

K.MixReparametrized

Examples

```
data(faithful)
xobs=faithful[,1]
estimate=K.MixReparametrized(xobs, k=2, alpha0=.5, alpha=.5, Nsim=20000)
summari=SM.MixReparametrized(estimate)
```

Chapter 9

Supplementary material: Ultimixt package

Mixture models as a popular method for modeling unobserved heterogeneity, find application in a very wide number of applied fields. As mentioned before, they can be used to model a statistical population with subpopulations with the densities on the subpopulations as the mixture components and the proportions of each subpopulation in the overall population as the weights. For maximum likelihood estimation of the mixture model parameters, the Expectation-Maximization (EM) algorithm is most frequently used by the classical community which is provided in **R** by packages such as mclust [Fraley 2002], flexmix [Gruen 2008] and mixtools [Benaglia 2009]. Bayesian estimation has become feasible with the advent of Markov chain Monte Carlo (MCMC) simulation and **R** packages such as BayesMix [Gruen 2015] and bayesm [Rossi 2010] have been made for estimating univariate Gaussian finite mixtures with MCMC methods. However, the model class that is implemented in these packages only allows informative structure of the prior distributions.

This chapter involves the algorithms of `Ultimixt` package described in Chapter 8.

9.1 Description of implementation

The first function of the package, `K.MixReparametrized`, provides the functionality for estimating univariate Gaussian mixture models regarding the non-informative parameterization expressed in Chapter 6 with MCMC methods. Within a given model class users can modify the prior specification of the mixture distribution weights and the number of the components for developing a suitable model for the dataset. `K.MixReparametrized` function shares the following features:

Calibration step which consists of determining the scales of the proposal distributions by applying adaptive Metropolis-within-Gibbs algorithm;

The motivation of this step is to avoid using the method of trial and error in order to obtain proposal scales. In some special cases, the method of trial and error can be useful to make some ideas about which value could work well in the sense that the proposal distribution results in good mixing MCMC chains. However, trial and error method is time consuming and sometimes fails to attain a satisfactory value to get the convergence with the movements in the support of the target distribution. Adaptive MCMC

algorithm determines the scales such a way that the chain does not move neither too slowly (most of the proposals are accepted) nor very quickly (says most of the proposals will usually be rejected). This method asks the computer to automatically learn better parameter values while an algorithm runs. Under some conditions this method updates the scale of each proposal (different for each parameter) at each iteration with the intention of finding the best value. In other words, the scale of the proposal distribution is automatically calibrated according to the optimal acceptance rate (the fraction of the proposed moves which are accepted) and with a factor that decreases to 0 in such a way that the convergence conditions are held [Robert 2009a, Rosenthal 2011, Roberts 2001, Roberts 2009, Roberts 1997]. So when the acceptance rate on batches of 50 iterations (by default) is too high, the proposal variance is automatically increased whereas in the case of the acceptance rate too small, it will decrease by adding or subtracting $\min(0.01, 1/\sqrt{t})$ after the t^{th} iterations. [Roberts 1997] show that for a random walk Metropolis-Hastings with a d-dimensional target distribution which consists of i.i.d components, when the number of the parameters to simulate tends to ∞ the optimal acceptance rate is 0.234 and for a one dimensional problem, the optimal acceptance rate is approximately 0.44. With this description, the adaptive Metropolis-within-Gibbs algorithm can be summarized as follows:

For simplicity's sake, let $\theta = (\mu, \sigma, p, \varphi, \xi, \varpi)$ be the mixture parameters and $\varepsilon_\theta = (\varepsilon_\mu, \varepsilon_\sigma, \varepsilon_p, \varepsilon_\varphi, \varepsilon_\xi, \varepsilon_\varpi)$ be the scale of the proposal distribution $q_\theta()$ in Metropolis-within-Gibbs step from which the proposals of θ are generated. We also create an associated variable $\log(\varepsilon_\theta)$ giving the logarithm of the standard deviation to be used when simulating a proposal increment to parameter θ. For a total number of iterations T, suppose that ar_{opt} denotes the optimal acceptance rate of the proposals and after the j^{th} batch of 50 and l^{th} batch of 500 iterations, n_j and N_l are supposed to be the number of times that the proposals have been accepted while $ar_j = {}^{n_j}/_{50}$ and $AR_l = {}^{N_l}/_{500}$ are the related acceptance rate, respectively, for $j = 1, \ldots, {}^{T}/_{50}$ and $l = 1, \ldots, {}^{T}/_{500}$. If T_t denotes the number of the batch of 500 iterations from which we start testing AR_{T_t} towards optimal acceptance rate with threshold values δ, the adaptive Metropolis-within-Gibbs algorithm will be defined in Figure 9.1.

In the programs of `Ultimixt`, $T_t = 30$ by default and so after $15,000$ iterations, the algorithm starts comparing the acceptance rate of the last 500 iterations with the optimal acceptance rate. The calibration step is terminated when the resulting acceptance rate is located in a small neighborhood of the optimal acceptance rate. Note that the total number of MCMC iterations at most is $T = 30,000$ by default in adaptive Metropolis-within-Gibbs step of `K.MixReparametrized` function.

Convergence monitoring of the chain provided for all the parameters of the mixture distribution by applying [Gelman 1992] criterion;

Adaptive Metropolis-within-Gibbs algorithm for reparameterised mixture:

1 Initialize $\theta^{(0)}$, $\varepsilon_\theta^{(0)}$; $d_1 = 0; d_2 = 0; j = 1; l = 1$.

2 For $t = 1, \ldots, T$, the update of $\theta^{(t)}$ and $\varepsilon_\theta^{(t)}$ follows

2.1 Generate a proposal $\theta' \sim q(\cdot|\theta^{(t-1)})$ and update $\theta^{(t)}$ against $\pi(\cdot|\mathbf{x})$.

2.2 If $d_1 = 50$ compute ar_j

- ✓ If $ar_j < ar_{opt}$ do $\log(\varepsilon_\theta) - \min(0.01, 1/\sqrt{t})$, $d_1 = 0$ and $j = j + 1$;
- ✓ If $ar_j > ar_{opt}$ do $\log(\varepsilon_\theta) + \min(0.01, 1/\sqrt{t})$, $d_1 = 0$ and $j = j + 1$;
- ✓ Otherwise $d_1 = d_1 + 1$.

2.3 If $t \geq T_t$ and $d_2 = 500$ compute AR_l

- ✓ Compute $d_l(AR_l, ar_{opt}) = |AR_l - ar_{opt}|$;
- ✓ Stop the algorithm if $d_l(AR_l, ar_{opt}) < \delta$;
- ✓ Otherwise $d_2 = d_2 + 1$ and $l = l + 1$. Go to line 2.1.

Figure 9.1: Pseudo-code representation of the Adaptive Metropolis-within-Gibbs algorithm used in `K.MixReparametrized` function of `Ultimixt`. Note that for each t, line [2.1] is done according to a Metropolis-within-Gibbs algorithm step described in Figure 6.1 of Chapter 6.

This step is started after calibrating the proposal scales and in each 1000 additional iterations, [Gelman 1992] criterion is automatically computed for the posterior draws of the component parameters, each one based on 4 chains produced in parallel. The simulation is stopped when this criterion is close to 1 for all produced chains. [Gelman 1992] criterion is a convergence monitoring diagnostic and this method allows us not to continue simulating the parameters after achieving the convergence.

Thus the `Ultimixt` package calls function `gelman.diag()` from package `coda` for the convergence monitoring step.

`SM.MixReparametrized` and `SM.MAP.MixReparametrized` are able to analyze the output of the MCMC simulations by numerical methods. For the unimodal terms such as the mean μ, the variance σ and the radius coordinate φ, the draws are regarded as posterior draws and averaging and calculating median over these draws will be considered as the point estimator.

This method is not satisfactory in the case where the label switching problem occurs in the posterior draws because of the multimodality of the posterior distribution as discussed before. The label switching occurs for the parameters such as component weights, means and variances p_i, μ_i and σ_i; $i = 1, \ldots, k$. In this case, the function `SM.MixReparametrized` eliminates the resulting unidentifiability using a method of post processing the MCMC draws by imposing a restriction on the

ordering of the mixture component parameters, as, e.g., the constraint that the weights of the component distributions are ascending. After that, a clustering procedure is applied to permute the MCMC draws of component means and standard deviations by applying a two dimensional k-means type algorithm with $k!$ clusters. This method is a combination of model identification methods such as identifiability constrain and unsupervised clustering suggested in [Frühwirth-Schnatter 2006]. The simulated draws of the angles are also summarized by applying two one-line k-means algorithm with $k-2$ and $k-1$ clusters to a sample of size $T(k-2)$ for $\varpi_i; i = 1, \ldots, k-2$ and a sample of size $T(k-1)$ for $\xi_i; i = 1, \ldots, k-1$, respectively. Note that in order to avoid the problem caused by unidentifiability, the number of the mixture components should initially be properly chosen.

Using `Plot.MixReparametrized` function, the MCMC results can be visually analyzed by using trace plots and by plotting the estimated densities for the mixture distribution over the draws. With the functions of `Ultimixt` package the user can comfortably reproduce some of the results presented in simulation study section 6.4 of Chapter 6.

9.2 Application

In the following we illustrate the use of `Ultimixt` on an example for a simulated dataset with 50 data points for which the histogram is shown in Figure 9.2. The sample mean and standard deviation are $4.06, 3.38$, respectively. We therefore fit a Gaussian mixture model on the data by choosing $k = 5$. In order to sample from the posterior distribution of the mixture model parameters, we apply the function `K.MixReparametrized`. To do so, we have to specify the hyper parameters α_0, α and the total number of MCMC iterations. We therefore choose $\alpha_0 = 0.5$, $\alpha = 0.5$ and 10^4 iterations. The output is firstly summarized by `Plot.MixReparametrized` function. The distribution of the draws of μ and σ shown in Figures 9.3 are centered on the empirical mean and standard deviation of the dataset. The posterior distribution of each component parameter is evaluated by plotting boxplot in Figure 9.3 that helps us to quickly examine the output of `K.MixReparametrized` from a graphical way. Figure 9.2 also shows the related estimate of the predictive density that is efficiently fitted the data.

The evolution of the scales of the proposal distributions in the Metropolis-within-Gibbs algorithm is one of the outputs of `Plot.MixReparametrized` function that is displayed in Figure 9.4. This figure shows that the update of the proposal scales is terminated for 300 batch of 50 iterations of the adaptive Metropolis-within-Gibbs step. This means that the acceptance rates attain the stability over the optimal one after 15000 MCMC iterations.

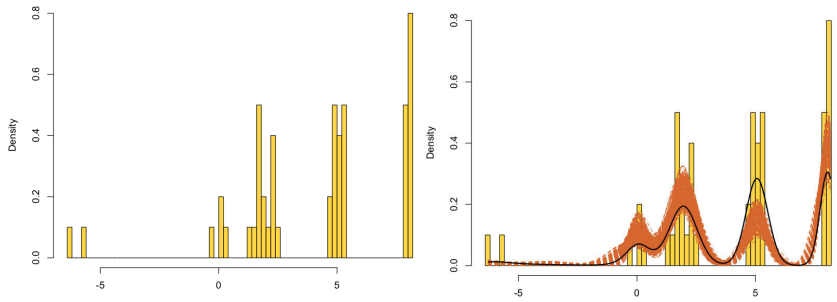


Figure 9.2: **Mixture of 5 normal distributions** 9.2: *(Left)* Histogram of a sample of size 50; *(Right)* Representation of 500 MCMC iterations as mixture distributions with the overlaid average curve (dark line) over 10^4 MCMC iterations obtained by applying `Plot.MixReparametrized` function.

Finally, calling two functions `SM.MixReparametrized` and `SM.MAP.MixReparametrized` leads to the following results for the component specific parameters:

```
> SM.MixReparametrized(xobs, estimate)
       Mean: Mean of mixture distribution
Median                                3.996
Mean                                  4.007
 ################################
       Sd: Sd of mixture distribution
Median                             3.382
Mean                               3.381
 ################################
         Phi
Median 0.9915
Mean   0.9829
 ################################
$`Angles. 1.`
[1] 0.8438176 0.9496362 3.8469520

 ################################
$`Angles. 2.`
[1] 1.3154771 0.3632422 1.0788443 0.8003213

 ################################
Component means, standard deviations and weights:

     weight  weight.1  weight.2  weight.3  weight.4
Median 0.23842  0.3437  0.2859  0.04079  0.09119
mean   0.23283  0.3473  0.2860  0.04080  0.09307
```

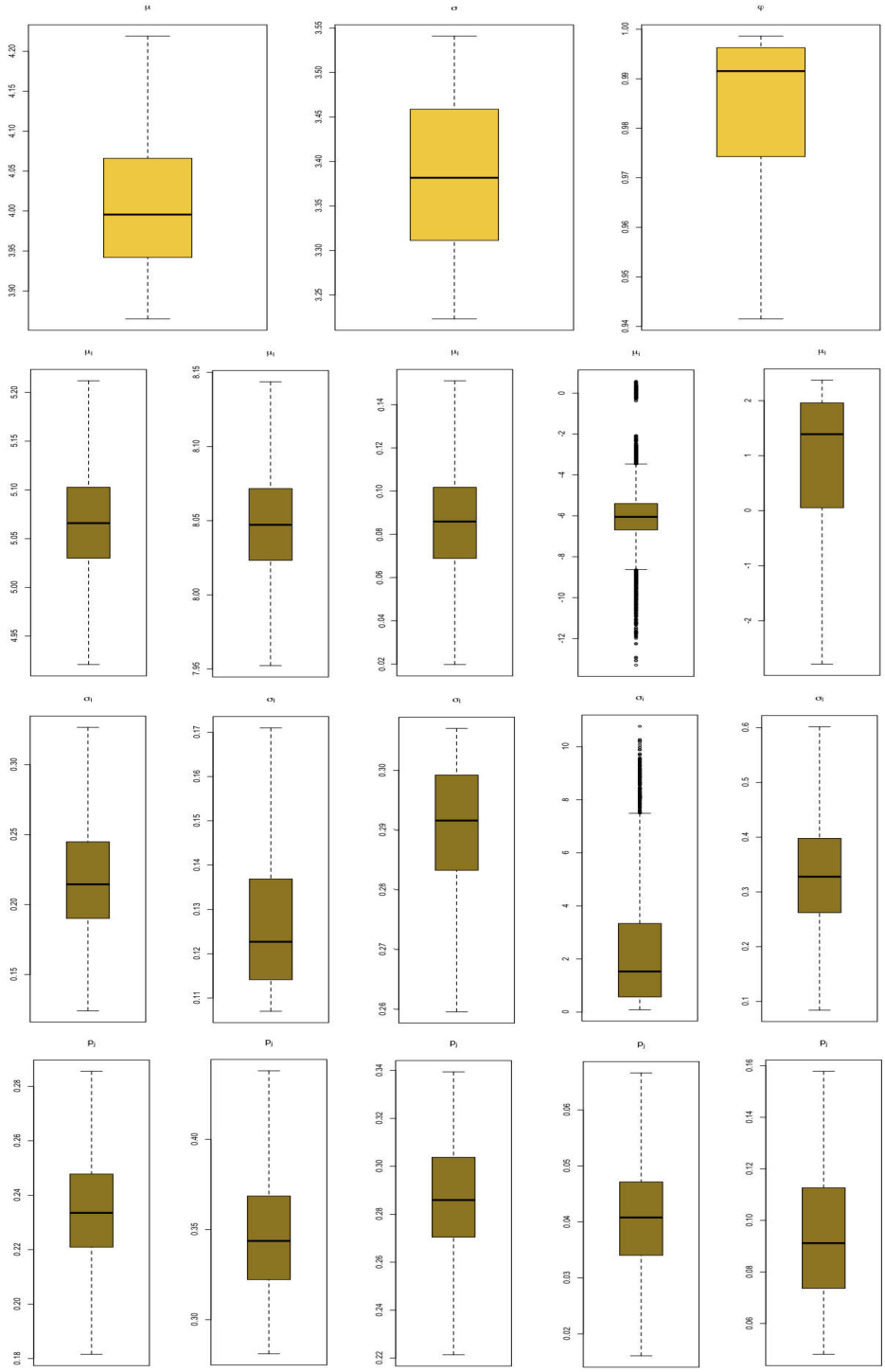


Figure 9.3: **Mixture of 5 normal distributions** 9.2: Posterior distributions of the global and component wise parameters of the mixture model by applying `Plot.MixReparametrized` function on the MCMC output.

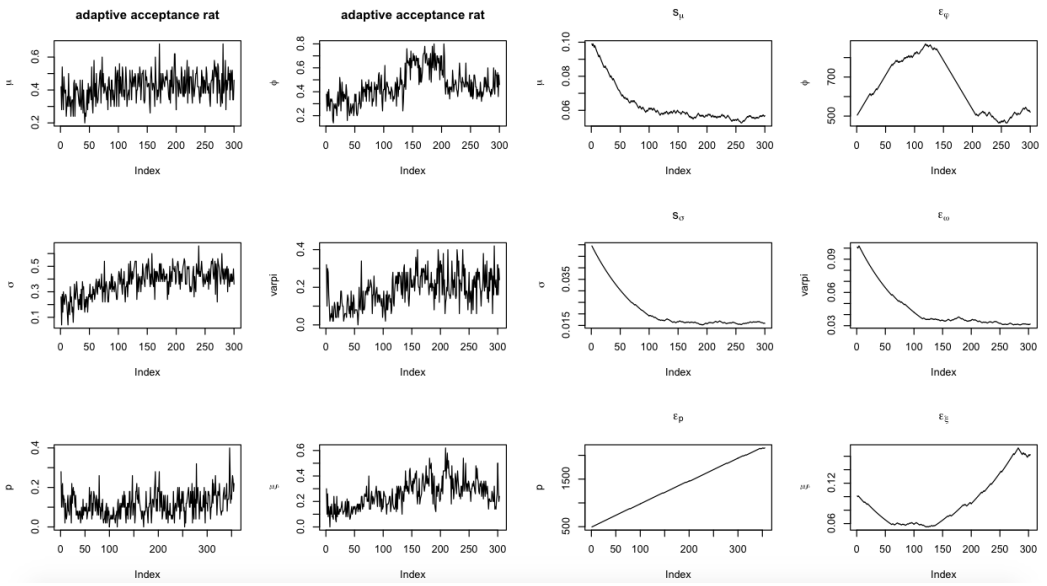


Figure 9.4: **Mixture of** 5 **normal distributions** 9.2: Evolution of the proposal scales and related acceptance rates over the number of the batch of size 50.

```
      mean  mean.1  mean.2  mean.3  mean.4
Median 5.066  8.046  0.071  -6.051  1.9103
mean   5.066  8.047  0.080  -5.817  1.8954

        sd   sd.1   sd.2   sd.3   sd.4
Median 0.248  0.103  0.292  1.529  0.328
mean   0.257  0.107  0.290  2.115  0.402
 ################################
$‘Acceptance rate of proposals‘
     mu    sigma    p    phi    theta   xi
0.2116   0.2322  0.1713  0.5156  0.2285 0.2250

 ################################
$‘Optimal proposal scales‘
s_mu     s_sigma  eps_p    eps_phi   eps_theta  eps_xi
5.7e-02  1.6e-02  2.2e+03  5.2e+02  3.1e-02  1.6e-01

> SM.MAP.MixReparametrized(estimate, xobs, .5, .5)
 ################################
Global mean
    Mean   Median     2.5%    97.5%
4.007435 3.995590 3.878364 4.189164
 ################################
Global standard deviation
    Mean   Median     2.5%    97.5%
```

```
3.381317 3.381742 3.226926 3.521489
 ##############################
Radius(phi)
     Mean    Median      2.5%     97.5%
0.9829364 0.9915425 0.9302170 0.9978621
 ##############################
Component means, standard deviations and weights:

    weight  weight.1  weight.2  weight.3  weight.4
Mean   0.0413  0.3335  0.2472  0.0926  0.2855
Median 0.0408  0.3399  0.2422  0.0910  0.2866
2.5%   0.0196  0.1680  0.1984  0.0489  0.2155
97.5%  0.0669  0.4380  0.3255  0.1410  0.3533

      mean  mean.1  mean.2  mean.3  mean.4
Mean   -6.067  5.066  8.047  0.082  1.959
Median -6.056  5.066  8.046  0.062  1.961
2.5%   -9.509  4.956  7.988  -0.270  1.780
97.5%  -3.002  5.180  8.106  0.598  2.135

      sd   sd.1   sd.2   sd.3   sd.4
Mean   2.233  0.221  0.107  0.336  0.351
Median 1.750  0.214  0.103  0.275  0.343
2.5%   0.232  0.156  0.071  0.126  0.248
97.5%  7.450  0.321  0.167  0.853  0.503
 ##############################
Angle components associated with component means
 and standard deviations:

      angle_sigma angle_sigma.1 angle_sigma.2 angle_sigma.3
Mean   0.717  1.048  1.323  1.099
Median 0.646  1.057  1.333  1.134
2.5%   0.217  0.789  1.157  0.613
97.5%  1.389  1.260  1.438  1.390

        angle_mu angle_mu.1 angle_mu.2
Mean   0.8918079  0.8981641   2.436233
Median 0.8900492  0.8961931   2.446848
2.5%   0.7619400  0.7793419   2.245740
97.5%  1.0952951  1.0577988   2.586439
$‘Acceptance rate of proposals‘
     mu   sigma   p   phi   theta   xi
0.2116   0.2322  0.1713  0.5156  0.2285 0.2250
```

```
 ##############################
$`Optimal proposal scales`
s_mu     s_sigma  eps_p    eps_phi   eps_theta  eps_xi
5.7e-02  1.6e-02  2.2e+03  5.2e+02  3.1e-02  1.6e-01
```

The output of the last two functions helps us to easily make inference for the parameter of the mixture model. If we compare the results of the functions `SM.MixReparametrized` and `SM.MAP.MixReparametrized`, we can see that both methods of k-means clustering algorithm and the one based on MAP estimate result in the same Bayesian inference about the mixture parameters.

Bibliography

[Adams 1987] M. Adams. William ockham. University of Notre Dame Press, Notre Dame, Indiana, 1987. (Cited on page 36.)

[Aitkin 1991] M. Aitkin. *Posterior Bayes factors (with discussion).* J. Royal Statist. Society Series B, vol. 53, pages 111–142, 1991. (Cited on pages 5, 34 and 35.)

[Aitkin 2010] M. Aitkin. Statistical inference: A Bayesian/likelihood approach. CRC Press, Chapman & Hall, New York, 2010. (Cited on pages 35 and 70.)

[Akaike 1973] H. Akaike. *Information theory and an extension of the maximum likelihood principle.* In Second International Symposium on Information Theory, Petrox, B.N. and Caski, F. (eds). Budapest: Akademiai Kiado, pages 267–281, 1973. (Cited on page 5.)

[Altekar 2004] G. Altekar, S. Dwarkadas, P. J. Huelsenbeck and F. Ronquist. *Parallel Metropolis coupled Markov Chain Monte Carlo for Bayesian phylogenetic inference.* Bioinformatics, vol. 20, no. 3, pages 407–415, 2004. (Cited on page 136.)

[Balasubramanian 1997] V. Balasubramanian. *Statistical Inference, Occam's Razor, and Statistical Mechanics on the Space of Probability Distributions.* Neural Computat., vol. 9, no. 2, pages 349–368, 1997. (Cited on page 34.)

[Barnard 2000] J. Barnard, R. McCulloch and X. L. Meng. *Modeling Covariance Matrices in Terms of Standard Deviations and Correlations with Application to Shrinkage.* Statistica Sinica, vol. 10, no. 4, pages 1281–1312, 2000. (Cited on pages 14, 15, 16 and 17.)

[Bayarri 2007] M.J. Bayarri and G. Garcia-Donato. *Extending conventional priors for testing general hypotheses in linear models.* Biometrika, vol. 94, pages 135–152, 2007. (Cited on page 34.)

[Behboodian 1970] J. Behboodian. *On the modes of a mixture of two normal distributions.* Technometrics, vol. 12, no. 1, pages 131–139, 1970. (Cited on page 127.)

[Benaglia 2009] T. Benaglia, D. Chauveau, D. Hunter and D. Young. *mixtools: An r package for analyzing finite mixture models.* Journal of Statistical Software, vol. 32, no. 6, pages 1–29, 2009. (Cited on pages 120 and 149.)

[Berger 1979] J. Berger and J. Bernardo. *Estimating a product of means Bayesian analysis with reference priors.* Journal of the American Statistical Association, vol. 89, pages 200–207, 1979. (Cited on page 3.)

[Berger 1980] J. O. Berger. Statistical Decision Theory: Foundations, Concepts, and Methods. Springer-Verl., New York, 1980. (Cited on page 11.)

[Berger 1985] J.O. Berger. Statistical decision theory and Bayesian analysis. Springer-Verlag, New York, second édition, 1985. (Cited on pages 34, 35 and 37.)

[Berger 1987] J.O. Berger and T. Sellke. *Testing a point-null hypothesis: the irreconcilability of significance levels and evidence (with discussion).* J. American Statist. Assoc., vol. 82, pages 112–122, 1987. (Cited on pages 33 and 34.)

[Berger 1992] J.O. Berger and W.H. Jefferys. *Ockham's razor and Bayesian analysis.* Amer. Scientist, vol. 80, pages 64–72, 1992. (Cited on page 34.)

[Berger 1994] J.O. Berger. *An overview of robust Bayesian analysis (with discussion).* TEST, vol. 3, pages 5–124, 1994. (Cited on pages 11 and 13.)

[Berger 1996] J.O. Berger and L.R. Pericchi. *The Intrinsic Bayes Factor for Model Selection and Prediction.* J. American Statist. Assoc., vol. 91, pages 109–122, 1996. (Cited on pages 4, 5 and 35.)

[Berger 1997] J.O. Berger, B. Boukai and Y. Wang. *Unified frequentist and Bayesian testing of a precise hypothesis (with discussion).* Statistical Science, vol. 12, pages 133–160, 1997. (Cited on page 34.)

[Berger 1998] J.O. Berger, L.R. Pericchi and J. Varshavsky. *Bayes factors and marginal distributions in invariant situations.* Sankhya A, vol. 60, pages 307–321, 1998. (Cited on pages 35, 39 and 44.)

[Berger 1999] J.O. Berger, B. Boukai and Y. Wang. *Simultaneous Bayesian-Frequentist Sequential Testing of Nested Hypotheses.* Biometrika, vol. 86, pages 79–92, 1999. (Cited on page 34.)

[Berger 2001] J.O. Berger and L.R. Pericchi. *Objective Bayesian methods for model selection: introduction and comparison.* In P. Lahiri, editeur, Model Selection, volume 38 of *Lecture Notes – Monograph Series*, pages 135–207, Beachwood Ohio, 2001. Institute of Mathematical Statistics. (Cited on page 35.)

[Berger 2003a] J.O. Berger. *Could Fisher, Jeffreys and Neyman have agreed on testing?* Statistical Science, vol. 18, no. 1, pages 1–32, 2003. (Cited on pages 33 and 34.)

[Berger 2003b] J.O. Berger, J.K. Ghosh and N. Mukhopadhyay. *Approximations and consistency of Bayes factors as model dimension grows.* Journal of Statistical Planning and Inference, vol. 112, no. 1-2, pages 241–258, 2003. (Cited on page 35.)

[Berkhof 2003] J. Berkhof, I. van Mechelen and A. Gelman. *A Bayesian approach to the selection and testing of mixture models.* Statistica Sinica, vol. 13, pages 423–442, 2003. (Cited on pages 38 and 105.)

[Bernardo 1979] J. Bernardo. *Reference posterior distributions for Bayesian inference.* Journal of the Royal Statistical Society. Series B (Methodological), vol. 41, pages 113–147, 1979. (Cited on page 3.)

[Bernardo 1980] J.M. Bernardo. *A Bayesian analysis of classical hypothesis testing.* In J.M. Bernardo, M. H. DeGroot, D. V. Lindley and A.F.M. Smith, editeurs, Bayesian Statistics. Oxford University Press, 1980. (Cited on page 34.)

[Bernardo 1994] J.M. Bernardo and A.F.M. Smith. Bayesian theory. John Wiley, New York, 1994. (Cited on page 3.)

[Bernardo 2009] J. M. Bernardo and A. F. Smith. Bayesian Theory. (Vol. 405), John Wiley & Sons, New York, 2009. (Cited on page 11.)

[Box 2011] G. E. Box and G. C. Tiao. Bayesian Inference in Statistical Analysis. (Vol. 40), John Wiley & Sons, New York, 2011. (Cited on pages 3 and 11.)

[Carlin 1995] B.P. Carlin and S. Chib. *Bayesian model choice through Markov chain Monte Carlo.* J. Royal Statist. Society Series B, vol. 57, no. 3, pages 473–484, 1995. (Cited on page 38.)

[Carlin 1996] B. P. Carlin and T. A. Louis. Bayes and empirical Bayes methods for data analysis. Chapman & Hall, London, 1996. (Cited on page 3.)

[Casella 1987] G. Casella and R. Berger. *Reconciling Bayesian and frequentist evidence in the one-sided testing problem.* J. American Statist. Assoc., vol. 82, pages 106–111, 1987. (Cited on page 33.)

[Casella 2002] G. Casella, K.L. Mengersen, C.P. Robert and D.M. Titterington. *Perfect Slice Samplers for Mixtures of Distributions.* J. Royal Statist. Society Series B, vol. 64(4), pages 777–790, 2002. (Cited on page 104.)

[Celeux 2000] G. Celeux, M.A. Hurn and C.P. Robert. *Computational and inferential difficulties with mixture posterior distributions.* J. American Statist. Assoc., vol. 95(3), pages 957–979, 2000. (Cited on pages 38, 40, 104, 111 and 116.)

[Chen 2000] M.H. Chen, Q.M. Shao and J.G. Ibrahim. Monte Carlo methods in Bayesian computation. Springer-Verlag, New York, 2000. (Cited on pages 35 and 38.)

[Chib 1995] S. Chib. *Marginal likelihood from the Gibbs output.* J. American Statist. Assoc., vol. 90, pages 1313–1321, 1995. (Cited on pages 35 and 105.)

[Chopin 2010] N. Chopin and C.P. Robert. *Properties of nested sampling.* Biometrika, vol. 97, pages 741–755, 2010. (Cited on page 35.)

[Choudhury 2007] A. Choudhury, S. Ray and P. Sarkar. *Approximating the cumulative distribution function of the normal distribution.* Journal of Statistical Research, vol. 41, pages 59–67, 2007. (Cited on page 53.)

[Christensen 2011] R. Christensen, W.O. Johnson, A.J. Branscum and T.E. Hanson. Bayesian ideas and data analysis: an introduction for scientists and statisticians. CRCPress, New York, 2011. (Cited on pages 22, 28 and 34.)

[Consonni 2013] Guido Consonni, Jonathan J. Forster and Luca La Rocca. *The Whetstone and the Alum Block: Balanced Objective Bayesian Comparison of Nested Models for Discrete Data.* Statistical Science, vol. 28, no. 3, pages 398–423, 2013. (Cited on page 34.)

[Cowles 2002] M. K. Cowles. *Bayesian Estimation of the Proportion of Treatment Effect Captured by Surrogate Marker.* Statistics in Medicine, vol. 21, no. 6, pages 811–834, 2002. (Cited on page 19.)

[Csilléry 2010] K. Csilléry, M. G. Blum, O. E. Gaggiotti and O. Francois. *Approximate Bayesian Computation (ABC) in practice.* Trends in ecology & evolution, vol. 25, no. 7, pages 410–418, 2010. (Cited on page 5.)

[Csiszár 2000] I. Csiszár and P. Shields. *The consistency of the BIC Markov order estimator.* Ann. Statist., vol. 28, pages 1601–1619, 2000. (Cited on page 34.)

[De Santis 1997] F. De Santis and F. Spezzaferri. *Alternative Bayes Factors for model selection.* Canadian J. Statist., vol. 25, pages 503–515, 1997. (Cited on page 34.)

[DeGroot 1970] M.H. DeGroot. Optimal statistical decisions. McGraw-Hill, New York, 1970. (Cited on pages 35 and 37.)

[DeGroot 1973] M.H. DeGroot. *Doing what comes naturally: Interpreting a tail area as a posterior probability or as a likelihood ratio.* J. American Statist. Assoc., vol. 68, pages 966–969, 1973. (Cited on pages 35, 37, 39, 44 and 71.)

[DeGroot 1982] M.H. DeGroot. *Discussion of Shafer's 'Lindley's paradox'.* J. American Statist. Assoc., vol. 378, pages 337–339, 1982. (Cited on page 34.)

[Dickey 1978] J.M. Dickey and E. Gunel. *Bayes factors from mixed probabilities.* Journal of the Royal Statistical Society. Series B (Methodological), vol. 40, no. 1, pages 43–46, 1978. (Cited on page 35.)

[Diebolt 1990] J. Diebolt and C.P. Robert. *Estimation des paramètres d'un mélange par échantillonnage bayésien.* Notes aux Comptes–Rendus de l'Académie des Sciences I, vol. 311, pages 653–658, 1990. (Cited on pages 104, 110 and 111.)

[Diebolt 1993] J. Diebolt and C.P. Robert. *Discussion of "Bayesian computations via the Gibbs sampler" by A.F.M. Smith and G.O. Roberts.* Journal of Royal Statistical Society, Series B, vol. 55, pages 71–72, 1993. (Cited on page 104.)

[Diebolt 1994] J. Diebolt and C.P. Robert. *Estimation of finite mixture distributions by Bayesian sampling.* Journal of Royal Statistical Society, Series B, vol. 56, pages 363–375, 1994. (Cited on pages 7, 40, 71 and 104.)

[Earl 2005] D. J. Earl and M. W. Deem. *Parallel tempering: Theory, applications, and new perspectives.* Physical Chemistry Chemical Physics, vol. 7, no. 23, pages 3910 –3916, 2005. (Cited on pages 135 and 136.)

[Escobar 1995] M.D. Escobar and M. West. *Bayesian prediction and density estimation.* J. American Statist. Assoc., vol. 90, pages 577–588, 1995. (Cited on page 104.)

[Firth 1993] D. Firth. *Bias Reduction of Maximum Likelihood Estimates.* Biometrika, vol. 80, no. 1, pages 27–38, 1993. (Cited on page 14.)

[Fraley 2002] C. Fraley and A. E. Raftery. *Model-based Clustering, Discriminant Analysis and Density Estimation.* Journal of the American Statistical Association, vol. 97, pages 611–631, 2002. (Cited on page 149.)

[Frühwirth-Schnatter 2001] S. Frühwirth-Schnatter. *Markov chain Monte Carlo estimation of classical and dynamic switching and mixture models.* J. American Statist. Assoc., vol. 96, pages 194–209, 2001. (Cited on pages 104 and 111.)

[Frühwirth-Schnatter 2004] S. Frühwirth-Schnatter. *Estimating marginal likelihoods for mixture and Markov switching models using bridge sampling techniques.* The Econometrics Journal, vol. 7, no. 1, pages 143–167, 2004. (Cited on pages 104 and 111.)

[Frühwirth-Schnatter 2006] S. Frühwirth-Schnatter. Finite mixture and markov switching models. Springer-Verlag, New York, New York, 2006. (Cited on pages 6, 7, 36, 38, 103, 105, 132 and 152.)

[Gallier 2011] J. Gallier. Geometric methods and applications: for computer science and engineering (vol. 38). Springer Science & Business Media, 2011. (Cited on page 129.)

[Gelfand 1992] A. E. Gelfand, D. K. Dey and H. Chang. *Model determination using predictive distributions with implementation via sampling-based methods.* In Bayesian Statistics. 4 (J. Bernardo, et al., eds.). Oxford University Press, pages 147–167, 1992. (Cited on page 4.)

[Gelman 1990] Andrew Gelman and Gary King. *Estimating the electoral consequences of legislative redistricting.* J. American Statist. Assoc., vol. 85, no. 410, pages 274–282, 1990. (Cited on pages 104 and 110.)

[Gelman 1992] A. Gelman and D. Rubin. *Inference from iterative simulation using multiple sequences (with discussion).* Statist. Science, pages 457–472, 1992. (Cited on pages 80, 111, 113, 150 and 151.)

[Gelman 1998] A. Gelman and L. X. Meng. *Simulating normalizing constants: From importance sampling to bridge sampling to path sampling.* Statistical science, pages 163–185, 1998. (Cited on pages iii, v and 175.)

[Gelman 2002] A. Gelman. *Prior distribution.* Encyclopedia of Environmentrics, vol. 3, pages 1634–1637, 2002. (Cited on page 2.)

[Gelman 2003] A. Gelman, J.B. Carlin, H.S. Stern and D.B. Rubin. Bayesian data analysis. Chapman and Hall, New York, New York, second édition, 2003. (Cited on page 87.)

[Gelman 2008] A. Gelman. *Objections to Bayesian statistics.* Bayesian Analysis, vol. 3(3), pages 445–450, 2008. (Cited on page 33.)

[Gelman 2013a] A. Gelman, J.B. Carlin, H.S. Stern, D.B. Dunson, A. Vehtari and D.B. Rubin. Bayesian Data Analysis. Chapman and Hall, New York, New York, third édition, 2013. (Cited on pages 3, 12, 34, 35 and 38.)

[Gelman 2013b] A. Gelman, C.P. Robert and J. Rousseau. *Inherent difficulties of non-Bayesian likelihood-based inference, as revealed by an examination of a recent book by Aitkin (with a reply from the author).* Statistics & Risk Modeling, vol. 30, pages 1001–1016, 2013. (Cited on page 35.)

[Geweke 2007] J. Geweke. *Interpretation and Inference in Mixture Models: Simple MCMC Works.* Comput. Statist. Data Analysis, vol. 51, no. 7, pages 3529–3550, 2007. (Cited on pages 111 and 116.)

[Geyer 1991] C. J. Geyer. *Markov Chain Monte Carlo maximum likelihood.* Computing Science and Statistics, no. 23, pages 156–163, 1991. (Cited on pages 123 and 136.)

[Geyer 2011] C. J. Geyer. *Importance sampling, simulated tempering, and umbrella sampling.* In the Handbook of Markov Chain Monte Carlo, S. P. Brooks, et al (eds), Chapman & Hall/CRC, pages 295–311, 2011. (Cited on page 136.)

[Ghosal 2000] Subhashis Ghosal, Jayanta K. Ghosh and Aad W. van der Vaart. *Convergence rates of posterior distributions.* Ann. Statist., vol. 28, no. 2, pages 500–531, 2000. (Cited on page 65.)

[Gigerenzer 1991] G. Gigerenzer. *The Superego, the Ego and the Id in statistical reasoning.* In G. Keren and C. Lewis, editeurs, Methodological and Quantitative Issues in the Analysis of Psychological Data. Erlbaum, Hillsdale, New Jersey, 1991. (Cited on page 33.)

[Gill 2004] J. Gill and G. Casella. *Dynamic tempered transitions for exploring multimodal posterior distributions*. Political Analysis, vol. 12, no. 4, pages 425–443, 2004. (Cited on pages 135 and 136.)

[Good 1950] I. J. Good. Probability and the Weighting of Evidence. London: Charles Griffin, 1950. (Cited on page 4.)

[Grazian 2015] C. Grazian and C.P. Robert. Jeffreys priors for mixture estimation, volume 126, pages 37–48. Springer Verlag, 2015. (Cited on page 107.)

[Green 1995] P.J. Green. *Reversible jump MCMC computation and Bayesian model determination*. Biometrika, vol. 82, no. 4, pages 711–732, 1995. (Cited on pages 5 and 35.)

[Griffin 2010] J. E. Griffin. *Default priors for density estimation with mixture models*. Bayesian Analysis, vol. 5, no. 1, pages 45–64, 2010. (Cited on page 104.)

[Gruen 2008] B. Gruen and F. Leisch. *FlexMix Version 2: Finite mixtures with concomitant variables and varying and constant parameters*. Journal of Statistical Software, vol. 28, no. 4, pages 1–35, 2008. (Cited on page 149.)

[Gruen 2015] Bettina Gruen. *bayesmix: Bayesian Mixture Models with JAGS*, 2015. R package version 0.7-4. (Cited on page 149.)

[Hamze 2010] F. Hamze, N. Dickson and K. Karimi. *Robust parameter selection for parallel tempering*. International Journal of Modern Physics C, vol. 21, no. 5, pages 603–615, 2010. (Cited on page 136.)

[Hasselblad 1966] V. Hasselblad. *Estimation of parameters for a mixture of normal distributions*. Technometrics, vol. 8, no. 3, pages 431–444, 1966. (Cited on page 6.)

[Hastie 2001] T. Hastie, R. Tibshirani and J. Friedman. The elements of statistical learning. Springer-Verlag, New York, 2001. (Cited on page 113.)

[Jasra 2005] A. Jasra, C.C. Holmes and D.A. Stephens. *Markov Chain Monte Carlo Methods and the Label Switching Problem in Bayesian Mixture Modeling*. Statist. Sci., vol. 20, no. 1, pages 50–67, 2005. (Cited on pages 38, 104 and 111.)

[Jaynes 1980] E. T. Jaynes. *Marginalization and prior probabilities*. In Bayesian Analysis in Econometrics and Statistics, A. Zellner (ed.). North Holland, Amsterdam, 1980. (Cited on page 2.)

[Jaynes 1983] E. T. Jaynes. Papers on probability, statistics and statistical physics. R.D. Rosencrantz (ed.) Reidel, Dordrecht, 1983. (Cited on page 2.)

[Jaynes 2003] E.T. Jaynes. Probability theory. Cambridge University Press, Cambridge, 2003. (Cited on page 11.)

[Jefferys 1992] W. Jefferys and J.O. Berger. *Ockham's razor and Bayesian analysis*. American Scientist, vol. 80, pages 64–72, 1992. (Cited on page 36.)

[Jeffreys 1939] H. Jeffreys. Theory of probability. The Clarendon Press, Oxford, first édition, 1939. (Cited on pages 3, 11, 34, 35, 44 and 107.)

[Johnson 2010] V.E. Johnson and D. Rossell. *On the use of non-local prior densities in Bayesian hypothesis tests*. J. Royal Statist. Society Series B, vol. 72, pages 143–170, 2010. (Cited on pages 34 and 36.)

[Johnson 2013a] V.E. Johnson. *Revised standards for statistical evidence*. Proc Natl Acad Sci USA, 2013. (Cited on pages 36 and 70.)

[Johnson 2013b] V.E. Johnson. *Uniformly most powerful Bayesian tests*. J. Royal Statist. Society Series B, vol. 41, pages 1716–âĂŞ1741, 2013. (Cited on pages 36 and 70.)

[Kadane 1980] J. B. Kadane and J. M. Dickey. *Bayesian decision theory and the simplification of models*. In Evaluation of Econometric Models, Kmenta, J. and Ramsey, J. (eds). New York: Academic Press, pages 245–268, 1980. (Cited on page 4.)

[Kamary 2014] K. Kamary, K.L. Mengersen, C.P. Robert and J. Rousseau. *Testing hypotheses as a mixture estimation model*. arxiv:1214.4436, 2014. (Cited on page 105.)

[Kamary 2015] Kaniav Kamary and Kate Lee. *Ultimixt: Bayesian Analysis of a Non-Informative Parametrization for Gaussian Mixture Distributions*, 2015. R package version 2.0. (Cited on pages 113, 114 and 120.)

[Kamary 2016a] K. Kamary. *R code: Reflecting about Selecting Noninformative Priors*. https://docs.google.com/viewer?a=v&pid=sites&srcid=ZGVmYXVsdGRvbWFpbnxrYW5pYXZrYW1hcnl8Z3g6M2I0NTQyNGRmOGVlYzRjZA, January 2016. (Cited on pages 26 and 30.)

[Kamary 2016b] K. Kamary. *R code: Testing hypotheses as a mixture estimation model*. https://docs.google.com/viewer?a=v&pid=sites&srcid=ZGVmYXVsdGRvbWFpbnxrYW5pYXZrYW1hcnl8Z3g6NWZhZDUwNTFkZTE2NzgwYg, January 2016. (Cited on pages 74, 75, 78, 79, 83, 86, 91, 95 and 98.)

[Kass 1995] R.E. Kass and A.E. Raftery. *Bayes factors*. J. American Statist. Assoc., vol. 90, pages 773–795, 1995. (Cited on pages 4 and 35.)

[Kass 1996] R. Kass and P. Wasserman. *The selection of prior distributions by formal rules*. Journal of the American Statistical Association, vol. 91, no. 431, pages 1343–1370, 1996. (Cited on page 3.)

[Lad 2003] F. Lad. *Appendix: the Jeffreys–Lindley Paradox and its Relevance to Statistical Testing*. In Conference on Science and Democracy, Palazzo Serra di Cassano, Napoli, 2003. (Cited on page 34.)

[Laplace 1820] P. S. Laplace. Théorie Analytique des Probabilités. Courcier, Paris, 1820. (Cited on page 10.)

[Lavine 1999] Michael Lavine and Mark J. Schervish. *Bayes Factors: What They Are and What They Are Not*. American Statist., vol. 53, no. 2, pages 119–122, 1999. (Cited on page 35.)

[Lee 2009] K. Lee, J.-M. Marin, K.L. Mengersen and C.P. Robert. *Bayesian Inference on Mixtures of Distributions*. In N.S. Narasimha Sastry, M. Delampady and B. Rajeev, editeurs, Perspectives in Mathematical Sciences I: Probability and Statistics, pages 165–202. World Scientific, Singapore, 2009. (Cited on pages 38, 40, 105 and 111.)

[Lewandowski 2014] K. Lewandowski, P. Knychala and M. Banaszak. *Parallel-tempering Monte-Carlo simulation with feedback-optimized algorithm applied to a coil-to-globule transition of a lattice homopolymer*. arXiv preprint arXiv:1410.3778, 2014. (Cited on page 136.)

[Li 2009] Y. Li, V. A. Protopopescu, N. Arnold, X. Zhang and A. Gorin. *Hybrid parallel tempering and simulated annealing method*. Applied Mathematics and Computation, vol. 212, no. 1, pages 216–228, 2009. (Cited on page 136.)

[Liang 2008] F. Liang, R. Paulo, G. Molina, M. A. Clyde and J. O. Berger. *Mixtures of G-priors for Bayesian Variable Selection*. J. American Statist. Assoc., vol. 103, no. 481, pages 410–423, 2008. (Cited on page 13.)

[Lindley 1957] D.V. Lindley. *A statistical paradox*. Biometrika, vol. 44, pages 187–192, 1957. (Cited on page 34.)

[Lopes 2011] H. F. Lopes and J. L. Tobias. *Confronting Prior Convictions: On Issues Prior Sensitivity and Likelihood Robustness Bayesian Analysis*. Annu. Rev. of Economics, vol. 3, no. 1, pages 107–131, 2011. (Cited on page 10.)

[MacKay 2002] David J. C. MacKay. Information theory, inference & learning algorithms. Cambridge University Press, Cambridge, UK, 2002. (Cited on page 34.)

[Madigan 1994] D. Madigan and A.E. Raftery. *Model selection and accounting for model uncertainty in graphical models using Occam's Window*. J. American Statist. Assoc., vol. 89, pages 1535–1546, 1994. (Cited on page 34.)

[Marin 2005] J.-M. Marin, K.L. Mengersen and C.P. Robert. *Bayesian Modelling and Inference on Mixtures of Distributions*. In C.R. Rao and D. Dey, editeurs, Handbook of Statistics, volume 25, pages 459–507. Springer-Verlag, New York, 2005. (Cited on pages 71 and 105.)

[Marin 2006] J.M. Marin and C.P. Robert. The bayesian core. Springer-Verlag, New York, 2006. To appear. (Cited on pages 6 and 127.)

[Marin 2007] J.-M. Marin and C.P. Robert. Bayesian core. Springer-Verlag, New York, 2007. (Cited on pages 13, 14, 52, 53, 95 and 124.)

[Marin 2011] J.-M. Marin and C.P. Robert. *Importance sampling methods for Bayesian discrimination between embedded models.* In M.-H. Chen, D.K. Dey, P. Müller, D. Sun and K. Ye, editeurs, Frontiers of Statistical Decision Making and Bayesian Analysis. Springer-Verlag, New York, 2011. (Cited on pages 35 and 38.)

[Marin 2014] J.-M. Marin, N. Pillai, C.P. Robert and J. Rousseau. *Relevant statistics for Bayesian model choice.* J. Royal Statist. Soc. Series B, vol. 76, no. 5, pages 833£–859, 2014. (Cited on pages 36 and 50.)

[Marinari 1992] E. Marinari and G. Parisi. *Simulated tempering: A new Monte Carlo schemes.* Europhysics letters, vol. 19, pages 451–458, 1992. (Cited on page 122.)

[Mayo 2006] D.G. Mayo and D.R. Cox. *Frequentist statistics as a theory of inductive inference.* In Javier Rojo, editeur, Optimality: The Second Erich L. Lehmann Symposium, Lecture Notes-Monograph Series, pages 77–97. Institute of Mathematical Statistics, Beachwood, Ohio, USA, 2006. (Cited on page 33.)

[McVinish 2009] R. McVinish, J. Rousseau and K. Mengersen. *Bayesian Goodness-of-Fit Testing with Mixtures of Triangular Distributions.* Scandinavian Journ. Statist., vol. 36, pages 337–354, 2009. (Cited on page 35.)

[Mengersen 1996] K.L. Mengersen and C.P. Robert. *Testing for mixtures: A Bayesian entropic approach (with discussion).* In J.O. Berger, J.M. Bernardo, A.P. Dawid, D.V. Lindley and A.F.M. Smith, editeurs, Bayesian Statistics 5, pages 255–276. Oxford University Press, Oxford, 1996. (Cited on pages 104, 106 and 107.)

[Mengersen 2011] K.L. Mengersen, C.P. Robert and D.M. Titterington. Mixtures: Estimation and applications. John Wiley, 2011. (Cited on page 105.)

[Miasojedow 2013] B. Miasojedow, E. Moulines and M. Vihola. *An adaptive parallel tempering algorithm.* J. Comput. Graphical Statist., vol. 22, no. 3, pages 649–664, 2013. (Cited on pages 122 and 136.)

[Neal 1994] R.M. Neal. *Contribution to the discussion of "Approximate Bayesian Inference with the Weighted Likelihood Bootstrap" by Michael A. Newton and Adrian E. Raftery.* J. Royal Statist. Society Series B, vol. 56 (1), pages 41–42, 1994. (Cited on page 35.)

[Neal 1996] R. M. Neal. *Sampling from multimodal distributions using tempered transitions.* Statistics and Computing, vol. 6, no. 4, pages 353–366, 1996. (Cited on pages 122 and 136.)

[Neal 1999] R.M. Neal. *Erroneous results in "Marginal likelihood from the Gibbs output".* Rapport technique, University of Toronto, 1999. (Cited on pages 35 and 105.)

[Newton 1994] M.A. Newton and A.E. Raftery. *Approximate Bayesian inference by the weighted likelihood bootstrap (with discussion).* J. Royal Statist. Society Series B, vol. 56, pages 1–48, 1994. (Cited on page 35.)

[Neyman 1933] J. Neyman and E.S. Pearson. *The testing of statistical hypotheses in relation to probabilities a priori.* Proc. Cambridge Philos. Soc., vol. 24, pages 492–510, 1933. (Cited on page 33.)

[O'Hagan 1994] A. O'Hagan. Bayesian inference. Numeéro 2B de Kendall's Advanced Theory of Statistics. Chapman and Hall, New York, 1994. (Cited on page 104.)

[O'Hagan 1995] A. O'Hagan. *Fractional Bayes factors for model comparisons.* J. Royal Statist. Society Series B, vol. 57, pages 99–138, 1995. (Cited on pages 5 and 35.)

[O'Neill 2014] P. D. O'Neill and T. Kypraios. *Bayesian model choice via mixture distributions with application to epidemics and population process models.* ArXiv e-prints, 2014. (Cited on pages 5 and 37.)

[Pearson 1894] K. Pearson. *Contributions to the mathematical theory of evolution.* Philosophical Transactions of the Royal Society of London, vol. 185, pages 71–110, 1894. (Cited on page 5.)

[Plummer 2008] M. Plummer. *Penalized loss functions for Bayesian model comparison.* Biostatistics, vol. 9, no. 3, pages 523–539, 2008. (Cited on page 34.)

[R Development Core Team 2006] R Development Core Team. *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria, 2006. (Cited on page 52.)

[Raiffa 1961] H. Raiffa and R. Schlaifer. Applied statistical decision theory. Division of Research, Graduate School of Business Administration, Harvard University, 1961. (Cited on page 2.)

[Rasmussen 2001] C. E. Rasmussen and Z. Ghahramani. *Occam's Razor.* In Advances in Neural Information Processing Systems, volume 13, 2001. (Cited on page 36.)

[Rattan 2013] O. Rattan, A. Camacho, A. Meijer and G. Donker. *Statistical modelling of summary values leads to accurate Approximate Bayesian Computations*. arXiv preprint arXiv:1305.4283, 2013. (Cited on page 5.)

[Richardson 1997] S. Richardson and P.J. Green. *On Bayesian analysis of mixtures with an unknown number of components (with discussion)*. J. Royal Statist. Society Series B, vol. 59, pages 731–792, 1997. (Cited on pages 7, 38, 103, 104, 119 and 132.)

[Rissanen 2012] J. Rissanen. Optimal estimation of parameters. Cambridge University Press, Cambrdige, 2012. (Cited on pages 3 and 11.)

[Robbins 1948] H. Robbins. *Mixture of distributions*. The Annals of Mathematical Statistics, vol. 19, no. 3, pages 360–369, 1948. (Cited on page 6.)

[Robbins 1961] H. Robbins. *Identifiability of mixtures*. The Annals of Mathematical Statistics, vol. 32, no. 1, pages 244–248, 1961. (Cited on page 6.)

[Robert 1993] C.P. Robert. *A Note on Jeffreys-Lindley paradox*. Statistica Sinica, vol. 3, no. 2, pages 601–608, 1993. (Cited on page 34.)

[Robert 1998] C.P. Robert and M. Titterington. *Reparameterisation strategies for hidden Markov models and Bayesian approaches to maximum likelihood estimation*. Statistics and Computing, vol. 8, no. 2, pages 145–158, 1998. (Cited on pages 104 and 107.)

[Robert 2001] C.P. Robert. The Bayesian choice. Springer-Verlag, New York, second édition, 2001. (Cited on pages 2, 3, 4, 5, 34 and 35.)

[Robert 2007] C. P. Robert. The Bayesian Choice. Springer, New York, 2007. (Cited on pages 10 and 11.)

[Robert 2009a] C. Robert and G. Casella. Introducing Monte Carlo Methods with R. Springer Science & Business Media, 2009. (Cited on page 150.)

[Robert 2009b] C.P. Robert, N. Chopin and J. Rousseau. *Theory of Probability revisited (with discussion)*. Statist. Science, vol. 24(2), pages 141–172 and 191–194, 2009. (Cited on page 44.)

[Robert 2011] C.P. Robert, J.-M. Cornuet, J.-M. Marin and N. Pillai. *Lack of confidence in ABC model choice*. Proceedings of the National Academy of Sciences, vol. 108(37), pages 15112–15117, 2011. (Cited on pages 36 and 70.)

[Robert 2014] C.P. Robert. *On the Jeffreys–Lindley paradox*. Philosophy of Science, vol. 5, no. 2, pages 216–232, 2014. (Cited on pages 34 and 35.)

[Roberts 1997] G. O. Roberts, A. Gelman and W. R. Gilks. *Weak convergence and optimal scaling of random walk Metropolis algorithms*. The Annals of Applied probability, vol. 7, no. 1, pages 110–120, 1997. (Cited on pages 110, 116 and 150.)

[Roberts 2001] G. O. Roberts and S. J. Rosenthal. *Optimal scaling for various Metropolis-Hastings algorithms*. Statist. Science, vol. 16, no. 4, pages 351–367, 2001. (Cited on pages 110 and 150.)

[Roberts 2009] G. O. Roberts and S. J. Rosenthal. *Examples of adaptive MCMC*. Journal of Computational and Graphical Statistics, vol. 18, no. 2, pages 349–367, 2009. (Cited on pages 110, 111 and 150.)

[Robertson 1969] C. A. Robertson and J. G. Fryer. *Some descriptive properties of normal mixtures*. Scandinavian Actuarial Journal, vol. 1969, no. 3–4, pages 137–146, 1969. (Cited on page 127.)

[Rodriguez 2014] C.E. Rodriguez and S.G. Walker. *Label Switching in Bayesian Mixture Models: Deterministic Relabeling Strategies*. Journal of Computational and Graphical Statistics, vol. 23, no. 1, pages 25–45, 2014. (Cited on page 38.)

[Roeder 1990] K. Roeder. *Density estimation with confidence sets exemplified by superclusters and voids in galaxies*. Journal of the American Statistical Association, vol. 85, pages 617–624, 1990. (Cited on pages 104 and 119.)

[Rolph 1968] J. E. Rolph. *Bayesian estimation of mixing distributions*. The Annals of Mathematical Statistics, pages 1289–1302, 1968. (Cited on page 6.)

[Rosenthal 2011] S. J. Rosenthal. *Optimal proposal distributions and adaptive MCMC*. In G. Jones S. Brooks A. Gelman and X.L. Meng, editeurs, Handbook of Markov Chain Monte Carlo, pages 93–112. CRC Press, 2011. (Cited on pages 110 and 150.)

[Rossi 2010] P. Rossi and R. McCulloch. *Bayesm: Bayesian inference for marketing/micro-econometrics*. R package version, vol. 2, pages 357–365, 2010. (Cited on pages 104, 120 and 149.)

[Rousseau 2007] J. Rousseau. *Approximating interval hypotheses: p-values and Bayes factors*. In J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, editeurs, Bayesian Statistics 8: Proceedings of the Eigth International Meeting. Oxford University Press, 2007. (Cited on page 35.)

[Rousseau 2011] J. Rousseau and J. Mengersen. *Asymptotic behaviour of the posterior distribution in overfitted mixture models*. Journal of the Royal Statistical Society: Series B (Statistical Methodology), vol. 73, no. 5, pages 689–710, 2011. (Cited on pages 7, 36, 66, 67, 69, 71, 104 and 105.)

[Sayyareh 2011] A. Sayyareh. *A new upper bound for Kullback-Leibler divergence*. Applied Mathematical Sciences, vol. 5, no. 67, pages 3303–3317, 2011. (Cited on page 119.)

[Schwarz 1978] G. Schwarz. *Estimating the dimension of a model*. Ann. Statist., vol. 6, pages 461–464, 1978. (Cited on page 34.)

[Seaman III 2012] J. W. Seaman III, J. W. Seaman Jr and J. D. Stamey. *Hidden Dangers of Specifying Noninformative Priors*. The American Statistician, vol. 66, no. 2, pages 77–84, 2012. (Cited on pages iii, v, 3, 9, 10, 11, 12, 14, 16, 19, 20, 21, 22 and 175.)

[Shafer 1982] G. Shafer. *On Lindley's Paradox (with discussion)*. Journal of the American Statistical Association, vol. 378, pages 325–351, 1982. (Cited on page 34.)

[Skilling 2006] J. Skilling. *Nested sampling for general Bayesian computation*. Bayesian Analysis, vol. 1(4), pages 833–860, 2006. (Cited on page 35.)

[Spanos 2013] A. Spanos. *Who should be afraid of the Jeffreys–Lindley paradox?* Philosophy of Science, vol. 80, pages 73–93, 2013. (Cited on page 34.)

[Spiegelhalter 1998] D. J. Spiegelhalter, N. G. Best and B. P. Carlin. *Bayesian deviance, the effective number of parameters, and the comparison of arbitrarily complex models*. Research Report, pages 98–009, 1998. (Cited on page 5.)

[Spiegelhalter 2002] D. J. Spiegelhalter, N.G. Best, Carlin B.P. and A. Van der Linde. *Bayesian measures of model complexity and fit*. Journal of the Royal Statistical Society, Series B, vol. 64, pages 583–640, 2002. (Cited on page 34.)

[Sprenger 2013] J. Sprenger. *Testing a precise null hypothesis: The case of Lindley's paradox*. Philosophy of Science, vol. 80, no. 5, pages 733–744, 2013. (Cited on page 34.)

[Steele 2006] R. Steele, A.E. Raftery and M. Emond. *Computing Normalizing Constants for Finite Mixture Models via Incremental Mixture Importance Sampling (IMIS)*. Journal of Computational and Graphical Statistics, vol. 15, pages 712–734, 2006. (Cited on page 35.)

[Stephens 2000] M. Stephens. *Dealing with label switching in mixture models*. J. Royal Statist. Society Series B, vol. 62(4), pages 795–809, 2000. (Cited on pages 38, 104 and 111.)

[Stojanovski 2011] E. Stojanovski and D. Nur. *Prior Sensitivity Analysis for a Hierarchical Model*. Proceeding of the Fourth Annual ASEARCH Conference, pages 64–67, 2011. (Cited on page 10.)

[Stone 1972] M. Stone and A. David. *UnBayesian implications of improper Bayes inference in routine statistical problems*. Biometrika, vol. 59, no. 2, pages 369–375, 1972. (Cited on page 3.)

[Stoneking 2014] C. J. Stoneking. *Bayesian inference of Gaussian mixture models with noninformative priors*. arXiv preprint arXiv:1405.4895, 2014. (Cited on page 7.)

[Swendsen 1986] R. H. Swendsen and J. S. Wang. *Replica Monte Carlo simulation of spin-glasses*. Physical Review Letters, vol. 57, no. 21, pages 2607–2609, 1986. (Cited on pages 135 and 136.)

[Syversveen 1998] A. R. Syversveen. *Noninformative Bayesian priors. Interpretation and problems with construction and applications*. Preprint Statistics, vol. 3, 1998. (Cited on page 3.)

[Tanner 1987] M. Tanner and W. Wong. *The calculation of posterior distributions by data augmentation*. J. American Statist. Assoc., vol. 82, pages 528–550, 1987. (Cited on page 111.)

[Toni 2010] T. Toni and M. P. Stumpf. *Simulation-based model selection for dynamical systems in systems and population biology*. Bioinformatics, vol. 26, no. 1, pages 104–110, 2010. (Cited on page 5.)

[van Havre 2014] Z. van Havre, K. Mengersen, J. Rousseau and N. White. *Addressing open questions in mixture models*. Rapport technique 1412.08, QUT, Department of Statistics, Technical Report Series, 2014. (Cited on pages 49 and 133.)

[van Havre 2015] Z. van Havre, J. White N. Rousseau and K. Mengersen. *Overfitting Bayesian Mixture Models with an Unknown Number of Components*. arXiv preprint arXiv:1502.05427, 2015. (Cited on page 7.)

[Vehtari 2002] A. Vehtari and J. Lampinen. *Bayesian model assessment and comparison using crossvalidation predictive densities*. Neural Computation, vol. 14, pages 2439–2468, 2002. (Cited on pages 35 and 38.)

[Vehtari 2012] A. Vehtari and J. Ojanen. *A survey of Bayesian predictive methods for model assessment, selection and comparison*. Statistics Surveys, vol. 6, pages 142–228, 2012. (Cited on pages 35 and 38.)

[Wang 2011] Z. Wang and N. Freitas. *Predictive adaptation of hybrid Monte Carlo with Bayesian parametric bandits*. In NIPS Deep Learning and Unsupervised Feature Learning Workshop, 2011. (Cited on page 136.)

[Wasserman 1999] L. Wasserman. *Asymptotic inference for mixture models by using data-dependent priors*. J. Royal Statist. Society Series B, vol. 61, no. 1, pages 159–180, 1999. (Cited on page 104.)

[Welch 1963] B. Welch and H. Peers. *On Formulae for Confidence Points based on Integrals of Weighted Likelihoods*. J. Royal Statist. Society Ser. B, vol. 25, pages 318–329, 1963. (Cited on pages 3 and 11.)

[Yang 1996] R. Yang and J. O. Berger. *A catalog of nonuniformmative priors*. Institute of Statistics and Decision Sciences, Duke University, 1996. (Cited on page 3.)

[Zellner 1986] A. Zellner. *On assessing prior distributions and Bayesian regression analysis with g-prior distribution regression using Bayesian variable selection.* In Bayesian inference and decision techniques: Essays in Honor of Bruno de Finetti, pages 233–243. North-Holland / Elsevier, 1986. (Cited on pages 13 and 56.)

[Ziliak 2008] S.T. Ziliak and D.N. McCloskey. The cult of statistical significance: How the standard error costs us jobs, justice, and lives. Univ of Michigan Pr, 2008. (Cited on page 34.)

Résumé

Dans le cas particulier de l'approche bayésienne, la solution à la comparaison de modèles est le facteur de Bayes. Le facteur de Bayes est très peu utilisé. La première partie de cette thèse traite d'un examen général sur les lois a priori non informatives, et montre la stabilité globale des distributions postérieures. Dans la deuxième partie de la thèse, nous considérons un nouveau paradigme pour les tests bayésiens d'hypothèses en définissant une alternative à la construction traditionnelle de probabilités a posteriori qu'une hypothèse est vraie. Cette méthode se fonde sur l'examen des modèles en compétition en tant que composants d'un modèle de mélange. Dans la dernière partie de la thèse, nous sommes intéressés à la construction d'une analyse bayésienne de référence pour mélanges de gaussiennes par la création d'une nouvelle paramétrisation centrée sur la moyenne et la variance de ces modèles, ce qui nous permet de développer une loi a priori non-informative pour les mélanges avec un nombre arbitraire de composants. Cette partie de la thèse est suivie par une package R nommée Ultimixt qui met en œuvre une description de notre analyse bayésienne générique de mélanges de gaussiennes.

Mots Clés

Distribution de mélange, Loi a priori non-informative, Analyse bayésienne, A priori impropre, Choix du modèle bayésien, Méthodes de MCMC.

Abstract

In the special case of the Bayesian approach, the solution of model comparison is the Bayes factor. The Bayes factor is however problematic in some cases. The first part of this thesis deals with a general review on non-informative priors and demonstrating the overall stability of posterior distributions. In the second part, we consider a novel paradigm for Bayesian testing of hypotheses and Bayesian model comparison. The idea is to define an alternative to the traditional construction of posterior probabilities that a given hypothesis is true and to replace the original testing problem with estimation that focus on the probability weight of a given model within a mixture model. In the last part, we construct a reference Bayesian analysis of mixtures of Gaussian distributions by creating a new parameterization centered on the mean and variance of those models itself. This enables us to develop a genuine non-informative prior for Gaussian mixtures with an arbitrary number of components. This part of the thesis is followed by the description of R package, Ultimixt, which implements a generic reference Bayesian analysis of unidimensional mixtures of Gaussian distributions obtained by a location-scale parameterization of model.

Keywords

Mixture distribution, Non-informative prior, Bayesian analysis, Improper prior, Bayesian model choice, MCMC methods.

Numero national de thèse :