Contributions to Bayesian Computing for Complex Models

Clara Grazian

THÈSE DE DOCTORAT

de l'Université de recherche Paris Sciences et Lettres
PSL Research University

Préparée dans le cadre d'une cotutelle entre l'Université Paris-Dauphine et Sapienza Università di Roma

Contributions to Bayesian Computing for Complex Models

École Doctorale de Dauphine — ED 543

Spécialité **Sciences**

COMPOSITION DU JURY :

M. Christian ROBERT
Université Paris-Dauphine
Directeur de thèse

M. Brunero LISEO
Sapienza Università di Roma
Co-Directeur de thèse

M.me Kerrie MENGERSEN
Queensland University of Technology
Rapporteure

M. Pierre PUDLO
Université Aix-Marseille
Rapporteur

M.me Maria Maddalena BARBIERI
Università degli Studi di Roma Tre
Présidente du jury

M. Fabrizio LEISEN
University of Kent
Membre du jury

Soutenue le 15.04.2016
par Clara Grazian

Dirigée par **Christian Robert**

Brunero Liseo

DAUPHINE UNIVERSITÉ PARIS | PSL RESEARCH UNIVERSITY PARIS | SAPIENZA UNIVERSITÀ DI ROMA

Résumé

Récemment, la grande complexité des applications modernes, par exemple dans la génétique, l'informatique, la finance, les sciences du climat, etc. a conduit à la proposition des nouveaux modèles qui peuvent décrire la réalité. Dans ces cas, méthodes MCMC classiques ne parviennent pas à rapprocher la distribution *a posteriori*, parce qu'ils sont trop lents pour étudier le space complet du paramètre. Nouveaux algorithmes ont été proposés pour gérer ces situations, où la fonction de vraisemblance est indisponible. Nous allons étudier nombreuses caractéristiques des modèles complexes: comment éliminer les paramètres de nuisance de l'analyse et faire inférence sur les quantités d'intérêt, dans un cadre bayésienne et non bayésienne et comment construire une distribution *a priori* de référence.

Mots Clés

ABC, modèles de mélange, loi a priori de Jeffreys, vraisemblance intégrée, modèles copula.

Abstract

Recently, the great complexity of modern applications, for instance in genetics, computer science, finance, climatic science etc., has led to the proposal of new models which may realistically describe the reality. In these cases, classical MCMC methods fail to approximate the posterior distribution, because they are too slow to investigate the full parameter space. New algorithms have been proposed to handle these situations, where the likelihood function is unavailable. We will investigate many features of complex models: how to eliminate the nuisance parameters from the analysis and make inference on key quantities of interest, both in a Bayesian and not Bayesian setting, and how to build a reference prior.

Keywords

ABC, mixture models, Jeffreys prior, integrated likelihood, copula Models.

SAPIENZA
UNIVERSITÀ DI ROMA

DAUPHINE
UNIVERSITÉ PARIS

UNIVERSITÉ FRANCO ITALIENNE | UNIVERSITÀ ITALO FRANCESE

Dottorato di Ricerca in Statistica Metodologica

Tesi di Dottorato XXVIII Ciclo – anno 2016

Dipartimento di Statistica, Probabilità e Statistiche Applicate

École doctorale de Dauphine

Programme Doctoral en Mathématique – Année 2016

CEntre de REcherche en MAthématiques de la DÉcision

Contributions to Bayesian computing for complex models

Contributions computationnelles à la statistique bayésienne pour modèles complexes

Clara Grazian

Jury :

Maria Maddalena Barbieri	Univerisità degli Studi di Roma Tre	*Examinateur*
Fabrizio Leisen	University of Kent	*Examinateur*
Brunero Liseo	Sapienza Università di Roma	*Directeur de Thèse*
Kerrie Mengersen	Queensland University of Technology	*Rapporteur*
Pierre Pudlo	Aix-Marseille Université	*Rapporteur*
Christian Robert	Université Paris Dauphine	*Directeur de Thèse*

Contents

Summary

Bayesian inference is based on procedures which describe the posterior distribution for the parameter θ after having observed the data set $\mathbf{y}$, $\pi(\theta|\mathbf{y})$, which is available in closed form just in few cases. Therefore, computational methods have been proposed to approximate it.

Recently, the great complexity of modern applications, for instance in genetics, computer science, finance, climatic science etc., has led to the proposal of new models which may realistically describe the reality. For example, the data density may have the form

$$f(\mathbf{y};\,\theta) = \int_{\mathcal{Z}} f(\mathbf{y},\,\mathbf{z};\,\theta)\,d\mathbf{z} = \int_{\mathcal{Z}} f(\mathbf{y} \mid \mathbf{z},\,\theta)\;f(\mathbf{z};\,\theta)\,d\mathbf{z}$$

where $\mathbf{z}$ plays the role of a latent nonobservable structure, then the likelihood function may be unavailable, because of a too large dimension of $\mathbf{z}$. This is the case of stochastic volatility models or genetical models (where the role of the latent variable is played by the complete genealogical tree). In these cases, classical MCMC methods fail to approximate the posterior distribution, because they are too slow to investigate the full parameter space.

New algorithms have been proposed to handle these situations. In particular, approximate Bayesian computation (ABC) allows to manage models where the likelihood function may be considered intractable. The main idea of this class of algorithms is that if one simulates a proposed value for the parameter from a known distribution (for instance, the prior distribution) and then simulates a new data set from the model by fixing the parameter equal to the simulated value and the simulated data set is similar in some sense to the observed one, then the proposed value is likely to have generated the observed data and is included in the sample which will approximate the posterior distribution. Therefore, it is only necessary to be able to simulate from the model to provide an approximation of the posterior

distribution and no manipulation of the likelihood function is required.

The ABC methodology has been proposed in a Bayesian setting as a way to approximate the posterior distribution. Nevertheless it can also be used in other situations. In Chapter 1 we will propose a way to approximate the (integrated) likelihood function of a parameter of interest when the model considers many (potentially infinite) nuisance parameters, to perform inference in a classical setting (where the prior distribution is intended as a weight function for the integration).

Chapter 1 deals with a key point in complex models: nuisance parameters, which usually lack of physical meaning, are introduced to define flexible and realistic models, however the interest of the analysis stays in few parameters. For instance, in multivariate analysis, the concept of dependence is crucial, but it is quite difficult to work with models without introducing a normality assumption. Copula models are a way to separate the information for the marginal univariate distributions from the information on the dependence structure, which is captured by a copula function. They are flexible tools, nevertheless it is by now clear that a misspecification of the shape of the copula function leads to not reliable results, but nonparametric approaches are not yet fully developed in the literature. Chapter 2 proposes a way to make inference on indexes of dependence (as the Spearman's ρ, the Kendall's τ or tail dependence coefficients) without making strong assumptions on the shape of the copula function and via the ABC methodology.

While inference for complex models is more developed, problems of model choice have not yet general solutions. In Chapter 3 we will analyze a recent proposal to redefine the Bayes factor, show its weaknesses and present an alternative approach applicable in situations where the likelihood function is unavailable.

Finally, another type of problems with complicated models is the definition of a prior distribution because of the lack of physical meaning of many parameters. Mixture models are an example of complicated models which allow to describe kurtotic, multimodal and asymmetric data by considering a composition of known distributions:

$$\sum_{i=1}^{k} p_i \, f_i(x|\theta_i) \,, \quad \sum_{i=1}^{k} p_i = 1 \,.$$

Mixture models have an ill-defined nature (non-identifiability, multimodality, unbounded likelihood, etc.) and this leads to some difficulties in defining a prior dis-

tribution, in particular if information on all the parameters is unavailable. Many works have shown that improper priors are likely to produce improper posterior. In Chapter 4, we will analyze the Jeffreys approach to define a noninformative prior in this setting and propose an alternative which consists in a redefinition of the model.

Keywords: ABC, mixture models, Jeffreys prior, integrated likelihood, copula models

Résumé

Le paradigme bayésien a été proposé dans le 18ème siècle avec les travaux de Thomas Bayes (1702-1761), qui a d'abord prouvé le théorème de Bayes dans un cas particulier, et Pierre-Simon Laplace (1749-1827), qui a prouvé le théorème de Bayes plus généralement et introduit les distributions a priori conjuguées et noninformatives et l'idée de la croyance subjective dans la définition de la probabilité d'un événement. Dans les premières décennies du 20eme siècle, les statisticiens Ronald A. Fisher, Jerzy Neyman et Egon Pearson, parmi d'autres, ont proposé un paradigme nouveau, l'approche dite fréquentiste, basée sur l'idée que les procédures statistiques doivent être jugés par leur comportement dans les répétitions hypothétiques de l'expérience. Alors que des grands statisticiens ont travaillé sur le développement de l'inférence bayésienne au cours des années au milieu du 20eme siècle (comme Harold Jeffreys, Leonard J. Savage, Dennis Lindley, parmi beaucoup d'autres), les procédures bayésiennes sont restées un domaine de la recherche théorique, trop difficiles à appliquer.

La raison pour laquelle l'analyse bayésienne n'a pas été utilisée pour des applications réelles pendant une longue période reste dans la définition de la distribution *a posteriori* d'un paramètre $\theta \in \Omega$ sachant un ensemble de données $y \in \mathcal{Y}$ par le théorème de Bayes:

$$\pi(\theta|y) = \frac{\pi(\theta)L(\theta;y)}{\int_{\mathcal{Y}} \pi(\theta)L(\theta;y)dy}$$

où $\pi(\theta)$ est la distribution résumant l'information préalable sur le paramètre θ et $L(\theta;y)$ est la fonction de vraisemblance, qui fournit les informations disponibles avec l'expérience. En théorie, après la distribution *a posteriori* a été définie, l'inférence sur le paramètre θ est basé sur les descriptions de cette distribution. Dans la pratique, la composition des informations disponibles avec les données et les informations disponibles *a priori* ne fournit pas une distribution connue, sauf dans quelques

cas (comme dans le cas des lois conjuguées) et, par conséquent, cette distribution ne peut être gérée analytiquement.

Seulement dans les années 1980 il y a eu une croissance spectaculaire dans la recherche et dans les applications des méthodes bayésiennes, grâce au développement de la technologie informatique, qui a permis la mise en œuvre des méthodes Monte Carlo (et MCMC) pour approcher la distribution *a posteriori* dans des cas plus généraux et de gérer des situations plus complexes (et réalistes).

Récemment, les méthodes proposées dans ces années sont devenues obsolètes dans nombreux cas. La grande complexité des applications modernes, comme dans la génétique, l'informatique, la finance, la science climatique, etc., a conduit à la proposition des nouveaux modèles qui peuvent décrire la réalité de manière plus fidèle. Par exemple, le modèle peut avoir la forme

$$f(\mathbf{y};\,\theta) = \int_{\mathcal{Z}} f(\mathbf{y},\,\mathbf{z};\,\theta)\,d\mathbf{z} = \int_{\mathcal{Z}} f(\mathbf{y} \mid \mathbf{z},\,\theta)\;f(\mathbf{z};\,\theta)\,d\mathbf{z}$$

où $\mathbf{z}$ joue le rôle d'une structure non observable latente; alors la fonction de vraisemblance peut être indisponible, à cause d'une trop grande dimension de $\mathbf{z}$. Tel est le cas des modèles à volatilité stochastique, où l'intégration est par rapport de tout le temps d'observation, ou des modèles génétiques, où le rôle de la variable latente est joué par l'arbre généalogique complet. Dans ces cas, les méthodes MCMC classiques ne parviennent pas à approcher la distribution *a posteriori*, parce qu'ils sont trop lents pour enquêter l'espace totale des paramètres. Dans d'autres cas, la fonction de vraisemblance est indisponible, car il est impossible de travailler analytiquement avec elle, comme dans le cas des champs de Gibbs, où la fonction de vraisemblance est indisponible en raison d'une constante de normalisation en fonction du paramètre.

Des nouveaux algorithmes ont été proposés pour gérer ces situations. En particulier, une nouvelle classe d'algorithmes est la soi-disante classe des algorithmes likelihood-free ou calcul bayésien approché (approximate Bayesian computation, ABC), qui permettent de gérer les modèles où la fonction de vraisemblance peut être considérée comme insoluble. ABC a été proposé dans les dernières années 1990 dans un cadre appliqué; l'idée principale est que si une valeur proposée pour le paramètre est générée à partir d'une distribution connue (par exemple, la distribution *a priori*), puis un nouvel ensemble de données est simulé à partir du modèle, en fixant le paramètre correspondant à la valeur proposée, et les données simulées sont similaires dans un certain sens à celles observées, c'est probable que la valeur pro-

posée ait généré les données observées et est incluse dans l'échantillon qui approchera la distribution *a posteriori*. Il y a nombreux points clés de cette méthodologie, par exemple, la manière de définir la notion de similitude entre les ensembles de données, quand-même c'est seulement nécessaire d'être capable de simuler des données à partir du modèle pour fournir une approximation de la distribution *a posteriori* et aucune manipulation de la fonction de vraisemblance est demandée.

Sachant un modèle statistique avec une densité générique $p(x|\boldsymbol{\theta})$, $\boldsymbol{\theta} \in \Theta \subset \mathbb{R}^d$, on est souvent intéressé par une fonction de faible dimension $\boldsymbol{\psi}$ du vecteur de paramètre $\boldsymbol{\theta}$, telle que $\boldsymbol{\psi} = \boldsymbol{\psi}(\boldsymbol{\theta}) \in \mathbb{R}^k$, avec $k < d$. Les théories statistiques paramétriques ou semi-paramétriques modernes, au moins les approches basés sur la vraisemblence et les théories bayésiennes, visent à construire une fonction de vraisemblance qui dépend seulement de $\boldsymbol{\psi}$. Il y a beaucoup de littérature sur le problème de l'élimination des paramètres de nuisance. Les lecteurs intéressés peuvent se référer à Berger et al. (1999) et Liseo (2005) pour une perspective bayésienne, à Pace and Salvan (1997), Severini (2000) ou Lancaster (2000) pour un point de vue plus classique.

Dans un cadre bayésien le problème de l'élimination des paramètres de nuisance est, à moins en principe, trivial. Soit $\boldsymbol{\lambda} = \boldsymbol{\lambda}(\boldsymbol{\theta})$ la transformation complémentaire, telle que $\boldsymbol{\theta} = (\boldsymbol{\psi}, \boldsymbol{\lambda})$ et soit

$$\pi(\boldsymbol{\theta}) = \pi(\boldsymbol{\psi}, \boldsymbol{\lambda}) = \pi(\boldsymbol{\psi})\pi(\boldsymbol{\lambda}|\boldsymbol{\psi}) \tag{1}$$

la distribution *a priori*. Puis, en supposant que nous observons un ensemble de données $\mathbf{x} = (x_1, \ldots, x_n)$ du modèle avec fonction de vraisemblance $L(\boldsymbol{\psi}, \boldsymbol{\lambda}; \mathbf{x}) \propto p(\mathbf{x}; \boldsymbol{\psi}, \boldsymbol{\lambda})$, la distribution *a posteriori* marginale de $\boldsymbol{\psi}$ est

$$\begin{aligned} \pi(\boldsymbol{\psi}|\mathbf{x}) &= \frac{\int_\Lambda \pi(\boldsymbol{\psi}, \boldsymbol{\lambda}) L(\boldsymbol{\psi}, \boldsymbol{\lambda}; \mathbf{x}) d\boldsymbol{\lambda}}{\int_\Psi \int_\Lambda \pi(\boldsymbol{\psi}, \boldsymbol{\lambda}) L(\boldsymbol{\psi}, \boldsymbol{\lambda}; \mathbf{x}) d\boldsymbol{\lambda} d\boldsymbol{\psi}} \\ &\propto \pi(\boldsymbol{\psi}) \int_\Lambda \pi(\boldsymbol{\lambda}|\boldsymbol{\psi}) L(\boldsymbol{\psi}, \boldsymbol{\lambda}; \mathbf{x}) d\boldsymbol{\lambda}. \end{aligned} \tag{2}$$

L'intégrale dans le côté droit de (2) est, par définition, la vraisemblance intégrée du paramètre d'intérêt $\boldsymbol{\psi}$, où pour "intégration" on entend par rapport à la distribution $\pi(\boldsymbol{\lambda}|\boldsymbol{\psi})$; cette vraisemblence intégrée va être notée par $\tilde{L}(\boldsymbol{\psi}; \mathbf{x})$.

L'utilisation des vraisemblances intégrées est devenu populaire aussi parmi les statisticiens non bayésiens; il y a plusieurs exemples dans lesquels son utilisation est

nettement supérieure, ou au moins équivalente, même du point de vue frequentiste, pour l'estimation de l'incertitude. On se voit, par exemple, Severini (2007), Severini (2010) et Severini (2011).

Toutefois, le calcul explicite de l'intégrale ci-dessus pourrait ne pas être facile, en particulier lorsque la dimension $(d-k)$ est grande. Notez que la dimension d peut également comprendre une possible structure latente qui, à partir d'un point de vue strictement probabiliste, n'est pas différente d'un vecteur de paramètre.

Dans le Chapitre 1, nous sommes intéressés à explorer l'utilisation du ABC comme méthode d'approximation de la fonction de vraisemblance intégrée, dans les situations où une expression analytique de $\tilde{L}(\boldsymbol{\psi};\mathbf{x})$ n'est pas disponibles, ou il est trop coûteuse évaluer la fonction de vraisemblance "globale" $L(\boldsymbol{\psi},\boldsymbol{\lambda};\mathbf{x})$, comme, par exemple, dans des nombreux appllications génétiques ou dans autres situations où les méthodes MCMC classiques ne sont pas satisfaisantes et totalement fiables.

Une autre classe de problèmes où la vraisemblence intégrée est d'intérêt est celle des problèmes semi-paramétriques, où le paramètre d'intérêt est un scalaire - ou un vecteur - et le paramètre de nuisance est représenté par la partie non paramétrique du modèle; dans ces cas, l'intégration sur l'espace Λ est de dimension infinie, et c'est très souvent infaisable d'être résolue analytiquement.

Dans le Chapitre 1, nous allons discuter, à travers plusieurs exemples de complexité croissante, comment la vraisemblance intégrée produite par des algorithmes ABC se comporte en comparaison avec les méthodes existantes. Nous explorerons également son utilisation dans des exemples particuliers où d'autres méthodes manquent de produire une fonction de vraisemblance utile et facile à utiliser pour la paramètre d'intérêt.

L'objectif principal du Chapitre 1 est d'obtenir une approximation de la vraisemblence intégrée $\tilde{L}(\boldsymbol{\psi};\mathbf{x})$, $\boldsymbol{\psi}=\boldsymbol{\psi}(\boldsymbol{\theta})\in\mathbb{R}^k$.

C'est facile voir que

$$\tilde{L}(\boldsymbol{\psi};\mathbf{x}) \propto \frac{\pi(\boldsymbol{\psi}|\mathbf{x})}{\pi(\boldsymbol{\psi})}, \tag{3}$$

c'est à dire que la fonction de vraisemblance intégrée peut être interprétée comme l'évidence empirique qui transforme notre connaissance *a priori* sur le paramètre d'intérêt en connaissance *a posteriori*: dans cette perspective, nous pouvons interpréter $\tilde{L}(\boldsymbol{\psi};\mathbf{x})$ comme la définition bayésienne de la fonction de vraisemblance intégrée.

On suppose que c'est difficile, peut-être impossible, d'obtenir $\tilde{L}(\boldsymbol{\psi}; \mathbf{x})$ sous une forme connue. Par exemple, la paramètre de nuisance $\boldsymbol{\lambda}$ pourrait être de dimension infinie (cas semi-paramétrique) ou il peut représenter une structure latente non-observable associée au modèle statistique comme dans le cas des modèles de Markov chachés ou des modèles semi-Markoviens.

Dans ces situations, si $\tilde{L}(\boldsymbol{\psi}; \mathbf{x})$ n'est pas disponibles, ni $\pi(\boldsymbol{\psi}|\mathbf{x})$ sera disponible. Cependant, il est possible d'obtenir un loi *a posteriori* approchée $\tilde{\pi}(\boldsymbol{\psi}|\mathbf{x})$, en utilisant un algorithme ABC; dans la Section 1.3, nous allons discuter de certaines questions liées à la précision de cette approximation; après, nous allons décrire la mise en œuvre pratique de la méthode.

Comme dans toutes les approches ABC pour l'estimation de la distribution *a posteriori*, on doit

- sélectionner une statistique sommaire $\eta_1(\mathbf{x}), \ldots, \eta_h(\mathbf{x})$;
- sélectionner une distance $\rho(\cdot, \cdot)$ pour mesurer la distance entre les "vraies" données et le données simulées, ou leurs statistiques sommaires;
- sélectionner un seuil de tolérance ε
- choisir un algorithme (MC)MC qui propose des valeurs pour le vecteur de paramètre $\boldsymbol{\theta}$.

Une fois que la loi *a posteriori* est approchée par un échantillon ABC de taille M $(\boldsymbol{\theta}_1^*, \boldsymbol{\theta}_2^*, \ldots \boldsymbol{\theta}_M^*)$, on peut produire une approximation non paramétrique de la densité en fonction de la distribution marginale *a posteriori* de $\boldsymbol{\psi}$, $\tilde{\pi}^{ABC}(\boldsymbol{\psi}|\mathbf{x})$.

Une opération similaire peut être réalisée avec la loi marginale *a priori* $\pi(\boldsymbol{\psi})$, en effectuant une autre - pas chère - simulation de $\pi(\boldsymbol{\psi})$ pour obtenir une autre approximation, $\tilde{\pi}(\boldsymbol{\psi})$. Notez que on est obligé d'utiliser lois *a priori* propres pour tous les paramètres.

Ensuite, on peut définir la vraisemblence intégrée ABC

$$\tilde{L}^{ABC}(\boldsymbol{\psi}; \mathbf{x}) \propto \frac{\tilde{\pi}^{ABC}(\boldsymbol{\psi}|\mathbf{x})}{\tilde{\pi}(\boldsymbol{\psi})}. \tag{4}$$

Le Chapitre 1 est organisé de la manière suivante. Dans la Section 1.2, nous décrivons notre proposition en détail. La Section 1.3 discute les problèmes théoriques liés à la précision de l'approximation ABC. La Section 1.4 compare la vraisemblence

intégrée approchée par ABC avec les autres approches disponibles dans une série d'exemples. La Section 1.5 se termine avec une discussion des avantages et des inconvénients de la méthode.

Après, dans le Chapitre 2, nous allons examiner un exemple particulier de modèles complexes: les modèles de copule. Les modèles de copule sont aujourd'hui largement utilisés dans l'analyse multivariée des données. Les principaux domaines d'application comprennent l'économétrie (Huynh et al. 2015), la géophysique (Scholzel and Friederichs 2008), la mécanique quantique (Resconi and Licata 2015), la science du climat, (Schefzik et al. 2013), la génétique (He et al. 2012), la science actuarielle et la finance (Cherubini et al. (2004)), parmi les autres.

Une copule est un outil probabiliste flexible qui permet au chercheur de modéliser la distribution conjointe d'un vecteur aléatoire en deux étapes distinctes: les distributions marginales et une fonction copule qui capture la structure de dépendance entre les composantes du vecteur.

Du point de vue statistique, alors qu'il est généralement simple produire des estimations fiables des paramètres pour les distributions marginales, le problème de l'estimation de la structure de dépendance est crucial et souvent complexe, en particulier dans des situations de grande dimension.

D'autre part, la dépendance est l'une des caractéristiques les plus fondamentales dans la statistique (appliquée), l'économie et la probabilité. Une énorme liste d'applications importantes peuvent être trouvée dans la récente monographie de Joe (2014).

Dans une approche fréquentiste aux modèles de copule, il n'y a pas de méthodes généralement satisfaisantes pour l'estimation conjointe des paramètres marginaux et de la copule. La mèthode la plus populaire est la soi-disante "Inference from the margins" (IFM), où les paramètres des distributions marginales sont estimés d'abord, puis des pseudo-données sont obtenues en branchant les estimations des paramètres marginales. Ensuite, l'inférence sur les paramètres de la copule est effectuée en utilisant les pseudo-données: cette approche évidemment ne tient pas compte de l'incertitude sur l'estimation des paramètres marginaux. Les solutions alternatives bayésiennes ne sont pas encore pleinement développées, bien que Min and Czado (2010), Craiu and Sabeti (2012), Smith (2013), Wu et al. (2014) sont des exceptions remarquables.

Dans le Chapitre 2, nous allons considérer le problème général de l'estimation

des quantités spécifiques d'intérêt d'une copule générique (comme, par exemple, les coefficients de dépendance de queue ou le ρ de Spearman) en adoptant une approche bayésienne approchée similaire auquel de Mengersen et al. (2013). En particulier, nous allons discuter de l'utilisation de l'algorithme BC_{EL}, sur la base de la vraisemblence empirique qui approche la vraisemblence marginale de la quantité d'intérêt.

Notre approche est approchée en deux aspects:

1. l'explicitation de la distribution *a priori* est requise uniquement pour la quantité d'intérêt. Sa distribution est combinée avec la vraisemblance empirique afin de produire une approximation de la distribution *a posteriori* "vraie".

2. nous n'utilisons pas la "vraie" fonction de risque, mais plutôt une approximation basée sur la théorie de la vraisemblence empirique (Owen 2010). Nous espérons que cela permettra de réduire le biais potentiel des hypothéses incorrectes sur la distribution.

On peut noter, cependant, que le mot "vrai" dans la liste ci-dessus devrait être mieux defini comme "vrai-sous-l'-assuntion-du-modèle". Dans les situations où un vrai modèle est trop difficile à préciser, ou trop complexe à traiter, la vraisemblence empirique peut être un outil extrêmement précieux.

Notre approche peut être adaptée à la modélisation paramétrique et nonparamétrique des distributions marginales. La méthode décrite dans le Chapitre 2 est dans l'esprit de Hoff (2007), mais elle est basée sur un autre type d'approchement; les résultats, bien que dans une perspective différente, peuvent aussi être interprétés à la lumière de Schennach (2005), où une interprétation bayésienne nonparamétrique de la vraisemblence empirique est fournie.

Un modèle de copule est une façon de représenter la distribution conjointe d'un vecteur aléatoire $\mathbf{X} = (X_1, \ldots, X_m)$. Étant donnée une fonction cumulative m-variate $\mathbf{F}$, il est possible de montrer (Sklar 1959) qu'il existe toujours une fonction m-variate $C : [0,1]^m \to [0,1]$, telle que $\mathbf{F}(x_1, \ldots, x_m) = C(F_1(x_1), \ldots, F_m(x_m))$, où F_j est la fonction cumulative marginale de X_j.

En autres termes, la fonction copule C est une fonction de répartition avec des marginales uniformes sur $[0,1]$: elle lie ensemble les fonctions de répartition univariées $F_1, F_2, \ldots, F_m$ afin de produire la fonction de répartition m-variate $\mathbf{F}$. La

fonction copule C ne dépend pas des distributions marginales de $\mathbf{F}$, mais plutôt de la dépendance potentielle entre les composantes du vecteur aléatoire $\mathbf{X}$.

Pour chaque paire de composantes de $\mathbf{X}$, on dit X_i et X_j, on suppose qu'elles ont fonctions de répartition continues F_i et F_j. Il est bien connu que les deux variables transformées $U_i = F_i(X_i)$ et $U_j = F_j(X_j)$ ont distributions marginales uniformes. Un modèle de copule semiparamétrique consiste d'un modèle paramétrique pour la distribution conjointe de (U_i, U_j) et aucune hypothèse sur les distributions marginales. Une copule nonparamétrique est introduite lorsque la distribution conjointe de (U_i, U_j) dépend d'un paramètre de dimension infinie. Dans le Chapitre 2, nous allons permettre aux distributions marginales de F_j de suivre soit un modèle paramétrique soit un modèle nonparamétrique. Pour la fonction de copule nous ne ferons pas des hypothèses paramétriques. Au contraire, nous allons limiter nos objectifs à l'estimation d'une fonction d'intérêt de la copule C. Une discussion sur les approches classiques de l'estimation semiparamétrique des modèles de copule peut être trouvée dans Genest et al. (1995).

La vraisemblance empirique a été introduite par Owen: Owen (2010) est un étudie complet et récent; il est un moyen de production d'une vraisemblence nonparamétrique pour une quantité d'intérêt dans un modèle statistique non spécifié. En particulier, il est utile quand une vraie fonction de risque n'est pas facilement disponible, soit parce qu'il est trop coûteuse à évaluer soit lorsque le modèle est pas complètement spécifié.

On suppose que l'ensemble des données se compose de n observations indépendantes $(x_1, \ldots, x_m)$ d'un vecteur aléatoire $\mathbf{X}$ avec fonction de répartition $\mathbf{F}$ et densité correspondante $\mathbf{f}$.

Plutôt que définir la fonction de vraisemblance habituelle en termes de $\mathbf{f}$, la vraisemblence empirique est définie par rapport à une quantité d'intérêt calculée sur les données, on dit φ, exprimée en fonction de $\mathbf{F}$, $\varphi(\mathbf{F})$, puis une sorte de vraisemblence profil pour φ est calculée en manière nonparamétrique. Plus précisément, on considère un ensemble de conditions sur les moments généralisés de forme

$$E_{\mathbf{F}}\left(h(X, \varphi)\right) = 0, \tag{5}$$

où $h(\cdot)$ est une fonction connue, et φ est la quantité d'intérêt. La vraisemblence empirique résultante est définie comme

$$L_{EL}(\varphi; \mathbf{x}) = max_{\mathbf{p}} \prod_{i=1}^{n} p_i,$$

où le maximum est recherché sur l'ensemble des vecteurs $\mathbf{p}$ tel que $0 \leq p_i \leq 1$, $\sum_{i=1}^{n} p_i = 1$, et

$$\sum_{i=1}^{n} h(\mathbf{x}_i, \varphi) p_i = 0.$$

Les deux premières conditions sont évidentes et indépendante de φ, mais la troisième dépend de l'information des données vers la quantité d'intérêt et peut être définie comme une sorte de condition de non biais.

Dans le Chapitre 2, nous proposons d'adapter l'algorithme BC_{EL} de Mengersen et al. (2013) à une situation où le modèle statistique est partiellement spécifié et l'objectif principal est l'estimation d'une paramètre d'intérêt de dimension finie. En pratique, cela représente le prototype semiparamétrique, où on est principalement intéressé par certaines caractéristiques de la population, bien que le modèle statistique peut contenir des paramètres de nuisance qui sont souvent mis en place afin de produire des modèles plus flexibles qui pourraient mieux décrire les données. Afin de rendre l'inférence robuste pour la quantité d'intérêt, un modèle raisonnable devrait tenir compte de l'incertitude sur les paramètres de nuisance. Même si certains de ces paramètres supplémentaires ne sont pas particulièrement importants en termes d'estimation - ils manquent souvent d'une signification physique précise - leurs estimations peuvent considérablement affecter les inférences sur le paramètre d'intérêt. Dans ces circonstances, il pourrait être plus raisonnable et robuste spécifier partiellement le modèle et adopter une approche semiparamétrique.

La mèthode proposée dans le Chapitre 2 permet de ne pas définir des statistiques sommaires dans l'analyse et ainsi d'éviter la perte d'information typique de la méthode ABC. Cette perte est particulièrement importante dans un autre typologie de problèmes statistiques: les problèmes du choix du modèle, qui sont traités dans le Chapitre 3.

Quand des statistiques sommaires son introduites, l'approximation de la distribution *a posteriori* $\pi(\theta|\mathbf{y})$ est la distribution conjointe

$$\pi_\varepsilon(\theta, \mathbf{z}|\mathbf{y}) = \frac{\pi(\theta) f(\mathbf{z}|\theta) \mathbb{I}_{A_{\varepsilon,\mathbf{y}}}(\mathbf{z})}{\int_{A_{\varepsilon,\mathbf{y}} \times \Theta} \pi(\theta) f(\mathbf{z}|\theta) d\mathbf{z} d\theta} \tag{6}$$

où $\mathbb{I}_{A_{\varepsilon,\mathbf{y}}}$ est la fonction indicatrice de l'espace $\{\mathbf{z} \in \mathcal{Y}\ tel\ que\ \rho(\eta(\mathbf{z}), \eta(\mathbf{y})) < \varepsilon\}$ ($\rho(\cdot)$ est une distance appropriée entre les statistiques sommaires $\eta(\cdot)$ calculées sur les données observées et les données simulées et ε est le niveau de tolérance qui fournisse le degré désiré de similitude). Si les statistiques sommaires $\eta(\cdot)$ sont suffisantes, la distribution *a posteriori* approchée par ABC est une approximation de la vraie distribution *a posteriori* $\pi(\theta|\mathbf{y})$ quand ε va à 0.

Depuis que ABC est utilisé dans des situations complexes, il est peu probable qu'une statistique suffisante de petite dimension d existe. Fearnhead and Prangle (2012) prouve comment le choix de $\eta(\cdot)$ et sa dimension affecte l'erreur Monte Carlo. Dans un certain sens, le choix des statistiques sommaires à utiliser est spécifique au problème, cependant il existe quelques travaux pour le rendre automatique (voir par exemple Nunes and Balding (2010)).

La perte d'informations qui derive de l'utilisation des statistiques sommaires non suffisantes est, en général, considérée acceptable dans les problèmes d'inférence, parce qu'ABC permet de gérer des modèles complexes qui sont autrement intractable, en particulier quand on peut trouver une statistique de synthèse informative pour le paramètre θ.

Néanmoins, la perte d'information est, en quelque sorte, arbitraire lorsque le but de l'expérimentateur est le choix du modèle au lieu d'estimer le paramètre d'intérêt, comme c'est montré par Robert et al. (2011) et dans le Chapitre 3 qui présente également notre proposition au problème du choix de modèle avec ABC.

Si on considère deux modèles, l'approximation du facteur de Bayes à partir de l'algorithm ABC est

$$\hat{B}_{12}(\mathbf{y}) = \frac{\pi(\mathcal{M}=2)\sum_{i=1}^{N}\mathbb{I}_{m^{(i)}=1}\mathbb{I}_{\rho(\eta(\mathbf{z}),\eta(\mathbf{y}))<\varepsilon}}{\pi(\mathcal{M}=1)\sum_{i=1}^{N}\mathbb{I}_{m^{(i)}=2}\mathbb{I}_{\rho(\eta(\mathbf{z}),\eta(\mathbf{y}))<\varepsilon}} \tag{7}$$

qui va approcher

$$B_{12}^{\eta}(\mathbf{y}) = \frac{\int_{\Theta_1}\pi(\theta_1)f_1^{\eta}(\eta(\mathbf{y}|\theta_1))d\theta_1}{\int_{\Theta_2}\pi(\theta_2)f_2^{\eta}(\eta(\mathbf{y}|\theta_2))d\theta_2}. \tag{8}$$

Dans ce contexte, le facteur de Bayes approché par ABC est incompatible avec le vrai facteur de Bayes, à l'exception de très rares cas: même dans le cas d'existence d'une statistique suffisante, $f_m(\mathbf{y}|\theta_m) = g_m(\mathbf{y})f^{\eta}(\eta(\mathbf{y})|\theta_m)$, on a que

$$B_{12}(\mathbf{y}) = \frac{g_1(\mathbf{y})}{g_2(\mathbf{y})} B^{\eta}_{12}. \tag{9}$$

Il est donc impossible d'évaluer la différence entre les deux facteurs sans connaître l'entité du rapport $g_1(\mathbf{y})/g_2(\mathbf{y})$.

L'écart entre le facteur de Bayes approché par l'algorithme ABC et le facteur de Bayes cible est motivé par le fait que, une fois que le modèle a été choisi dans la première étape de l'algorithme, une acceptation ou un rejet est fait en comparant les données observées et les données simulées sur la base d'une statistique de synthèse définie conditionnellement à ce modèle particulier. Donc, même s'il est possible de trouver des statistiques sommaires suffisantes pour tous les modèles de l'analyse, ces statistiques ne seront pas suffisantes pour le problème du choix du modèle.

Notre proposition, décrite dans le Chapitre 3, est de changer la perspective et d'envisager une statistique de synthèse qui est informative pour le problème du choix du modèle et pas nécessairement pour les modèles considérés. Après avoir défini notre proposition, Section 3.3.2 étudie le comportement de la mèthode proposée dans certains exemples réels et simulés: tests d'hypothèses simples (Section 3.3.2), test d'hypothèses composites (Section 3.3.2), test sur modèles de régression (Section 3.3.2) et test pour modèles dynamiques (Section 3.3.2).

La dernière partie de cet travail traite de la définition d'un distribution non-informative *a priori* pour un type particulier de modèle complexe, les modèles de mélange.

Définir une distribution *a priori* est particulièrement délicat pour les modèles complexes, car il est difficile de définir l'influence des choix *a priori* sur les résultats de l'inférence pour des paramètres qui ne sont pas directement liées aux quantités physiques et ont structure de dépendence inconnue avec les autres paramètres.

L'inférence bayésienne dans le cas des mélanges de distributions a été étudié assez largement dans la littérature. Voir, par exemple, MacLachlan and Peel (2000) et Frühwirth-Schnatter (2006) pour des livres et Lee et al. (2009) pour un des nombreuses études.

Du point de vue bayésien, une des nombreuses difficultés avec ce type de distribution,

$$\sum_{i=1}^{k} p_i \, f(x|\theta_i) \, , \quad \sum_{i=1}^{k} p_i = 1 \, ,$$

est que sa nature mal définie (non-identifiabilité, multimodalité, vraisemblence non-bornée, etc.) conduit à une modélisation *a priori* restrictive puisque la plupart des distributions *a priori* impropres ne sont pas acceptables. Ceci est motivé en particulier par le fait qu'une ou plusieurs composantes $f(\cdot|\theta_i)$ peuvent contenir aucun sous-ensemble de l'échantillon (Titterington et al. 1985). Bien que la probabilité d'un tel événement est décroissante rapidement vers zéro lorsque la taille de l'échantillon augmente, elle empêche néanmoins le recours à distributions *a priori* impropres indépendants, à moins que tels événements sont interdits (Diebolt and Robert 1994).

De même, le caractère interchangeable des composantes induit souvent la multimodalité dans la distribution *a posteriori* et à difficultés de convergence comme exemplifié par le phénomène du "label-switching" qui est maintenant tout à fait bien documenté (Celeux et al. 2000, Stephens 2000, Jasra et al. 2005, Frühwirth-Schnatter 2006, Geweke 2007, Puolamäki and Kaski 2009). Ce trait est caractérisée par un manque de symétrie dans le résultat d'un algorithme Monte Carlo Markov Chaîne (MCMC), en ce que la densité *a posteriori* est échangeable dans les composantes de la mélange, mais l'échantillon MCMC ne présente pas cette symétrie. En outre, la plupart des échantillonneurs MCMC ne se concentrent pas autour d'un seul mode de la densité *a posteriori*, mais explorent plusieurs modes, ce qui rend la construction des estimateurs de Bayes des composantes beaucoup plus difficile.

Lors de la spécification d'une loi *a priori* sur les paramètres du modèle de mélange, il est donc tout à fait délicat produire une version noninformative gérable et raisonnable et certains ont parlé contre l'utilisation des lois *a priori* noninformatives dans ce contexte (par exemple, MacLachlan and Peel (2000) ont soutenu qu'il est impossible obtenir une distribution *a posteriori* propre à partir de lois *a priori* entièrement noninformatives), sur la base que les modèles de mélange sont des objets qui nécessitaient d'information *a priori* pour donner un sens à la notion de composante. Par exemple, la distance entre les deux composantes doit être bornée dessous pour éviter de répéter la même composante, encore et encore.

En plus, les composantes doivent être toutes informées par les données, fait qui est illustré dans Diebolt and Robert (1994) où les Authors ont imposé un régime d'achévement (un modèle commun sur les paramètres et les variables latentes) de telle sorte que au moins deux observations ont été répartis à toutes les composantes,

assurant ainsi que la loi *a posteriori* soit bien définie. Wasserman (2000) a prouvé dix ans aprè que cette troncature conduit à des estimateurs cohérents et en outre que seuls ce type de loi *a priori* pourrait produire une loi *a posteriori* cohérente. Alors que la contrainte sur les allocations n'est pas entièrement compatibles avec la représentation i.i.d. d'un modèle de mélange, elle exprime naturellement une exigence de modélisation que toutes les composantes aient un sens en termes de données, que toutes les composantes véritablement contribuent à la génération d'une partie des données. Cela se traduit par une forme d'information *a priori* faible sur combien on veut croire au modèle et combien chaque composante est significative.

Dans le Chapitre 4, nous étudierons d'abord si la loi *a priori* de Jeffreys peut être considérée comme une loi *a priori* de référence dans le contexte des mélanges, même si pas définitive.

Dans la Section 4.2 nous allons fournir la caractérisation formelle de la distribution *a posteriori* pour les paramètres d'un modèle de mélange, en particulier avec des composantes gaussiennes, quand la loi de Jeffreys est utilisée pour eux. Dans la Section 4.3 nous allons analyser la distribution *a priori* et *a posteriori* quand la loi de Jeffreys est utilisée: seulement lorsque les poids des composantes (qui sont définis dans un espace compact) sont les seuls paramètres inconnus il se révèle que la loi de Jeffreys (et donc la loi *a posteriori* relative) est propre; d'autre part, lorsque les autres paramètres ne sont pas connus, la Jeffreys sera demontrée être impropre; dans une seule situation elle fournit une distribution *a posteriori* propre. Dans la Section 4.4 nous allons proposer un moyen de réaliser une analyse noninformative des modèles de mélange et introduire des lois *a priori* impropres pour au moins certains paramètres.

Le Chapitre 4 est une grande amélioration de la littérature actuelle des mélanges, car il permet l'utilisation de lois *a priori* impropres sans passer par une reparametrization du modèle qui peut entraîner des difficultés computationelles.

Nous nous sommes concentrés sur les modèles de mélange, qui démontrent bien la difficulté de choisir une distribution *a priori* pour les modèles complexes. Ce problème est souvent spécifique au modèle utilisé, mais la recherche à venir sera concentrée sur le tentative de généraliser la méthodologie présentée.

Dans ce travail, nous avons étudié les aspects plus méthodologiques, mais les méthodes proposées ont toujours été testé sur des données simulées et des applications réelles (sauf dans le Chapitre 4). Il y a deux applications principales:

écologiques et financières.

Dans le Chapitre 1, nous avons utilisé des données d'une étude sur la faune d'une région entre le Queensland et la Grande Barrière d'Australie. Nous avons analysé comme variable de réponse un indice composite, en échelle logarithmique, qui combine des informations entre les espèces et est considéré comme dépendant de la latitude de manière linéaire et de la longitude de manière inconnue; pour une description plus détaillée, voir Bowman and Azzalini (1997).

Dans les Chapitres 2 et 3 nous nous sommes concentrés sur les applications financières, en particulier pour les données de performance (en échelle logarithmique) de plusieurs institutes italiens. Ces types de données ont deux caractéristiques importantes: ils présentent des queues de distribution lourdes et les structures de dépendance sont non linéaire. Pour cettes raisons, dans le Chapitre 2 des modèles GARCH marginaux pour des distributions t et des modèles de copules nonparamétriques ont été proposées, tandis que dans le Chapitre 3, on introduit des modèles marginaux en fonction des distributions quantiles. Ces distributions sont définies sur la base de leur fonction quantile définie comme une transformation non linéaire d'un quantile d'une loi gaussienne standard; pour cette raison, les distributions quantile sont un cas classique de distributions pour lesquelles la fonction de vraisemblance n'est pas disponibles et sont typiques dans la littérature sur ABC.

Dans ce travail, nous allons étudier nombreux problèmes de l'approche bayésienne pour les modèles complexes: la définition d'une fonction de vraisemblance pour les paramètres d'intérêt quand il y a nombreux paramètres de nuisance, l'inférence pour quelques paramètres d'intérêt lorsque le modèle est pas entièrement spécifié, des techniques de sélection de modèles pour des situations compliquées et la définition d'une distribution *a priori* dans le cas particulier des modèles de mélange. Dans les Chapitres et les Conclusions seront présentées les points les plus délicats qui méritent de plus amples recherches.

Pour conclure, les ordinateurs modernes permettent un développement des procédures bayésiennes impossibles dans le passé. Dans ce travail, nous allons essayér d'utiliser des nouvelles et anciennes méthodes pour les problèmes théoriques et appliqués modernes, grâce à des nouveaux outils de calcul.

Dans la première partie du travail, nous allons montrer que la classe des algorithmes appelée "calcul bayésien approché" peut être utilisée pour résoudre certains problèmes complexes soit dans un cadre bayésien soit dans un cadre classique.

Dans un cadre non-bayésienne, nous allons d'abord montrer qu'il peut être utilisé comme un outil pour approcher la fonction de vraisemblance pour un paramètre d'intérêt en présence de paramètres de nuisance. Le problème de l'élimination des paramètres de nuisance est crucial dans toutes les approches: en particulier, dans beaucoup des applications modernes nombreux paramètres sont introduits pour construire des modèles flexibles et réalistes, néanmoins leur manque d'un sens physique est un problème en termes d'inférence et en termes de construction d'une distribution *a priori* raisonnable.

Dans les problèmes semiparamétriques, où l'intérêt de l'analyse est en quelques paramètres et où il est préférable de limiter les hypothèses sur la forme complète du modèle, nous allons montrer que la mèthode ABC peut également être utilisée; par exemple, elle permet de gérer les modèles de copule et d'étudier la structure de dépendance des variables aléatoires multivariées sans faire des hypothèses fortes sur les distributions univariées ou sur la fonction copule. Lignes futures de recherche seraient axées sur la généralisation de l'approche présentée dans le Chapitre 2, par exemple en introduisant des covariates dans l'analyse et en tenant compte d'autres types de modèles.

Définir les moyens pour le choix du modèle dans le cas des modèles complexes est particulièrement difficile: les approches standards ont tendance à échouer de trouver le bon modèle. Dans le Chapitre 3, nous allons proposer un moyen de rapprocher le facteur de Bayes lorsque la fonction de vraisemblance est indisponible, en utilisant le calcul bayésien approché dans le cas spécifique des distributions quantile. Les travaux futurs seront concentrés sur la comparaison des méthodes avec celles proposées dans la littérature et l'extension de la méthode à modèles plus généraux.

Dans la dernière partie du travail nous allons essayer de construire une analyse noninformative pour les modèles de mélange. Comme dit précédemment, décrire les informations disponibles *a priori* avec loi de probabilité est difficile, en particulier pour les modèles complexes, principalement parce que pas tous les paramètres ont une signification physique. La définition d'une loi *a priori* pour les paramètres d'un modèle de mélange a une longue histoire dans la littérature. Tout d'abord, nous allons analysé la méthode de Jeffreys de définition d'une loi *a priori* noninformative et allons montrer que cette loi ne peux pas être utilisée. Ensuite, nous allons proposer de changer le point de vue en définissant une hiérarchie qui crée une structure de corrélation entre les composants de la mélange et permet d'utiliser lois *a priori* noninformatives impropres au plus haut niveau de la hierarchie.

Mot clés: ABC, modéles de mélange, distribution *a priori* de Jeffreys, vraisemblence integrée, modèles copula

Introduction

The Bayesian paradigm was proposed in the 18th century with the work of Thomas Bayes (1702–1761), who first proved the Bayes' theorem in a special case, and Pierre-Simon Laplace (1749–1827), who proved the Bayes' theorem more generally and introduced conjugate and noninformative prior distributions and introduced the idea of subjective belief in defining the probability of an event. In the first decades of the 20th century some statisticians, as Ronald A. Fisher, Jerzy Neyman and Egon Pearson among others, proposed a new and opposite paradigm, the so-called frequentist approach, based on the idea that the statistical procedures have to been judged by their behavior in hypothetical repetitions of the experiment. While great statistiticians worked on the development of Bayesian inference during the middle years of the 20th century (as Harold Jeffreys, Leonard J. Savage, Dennis Lindley, among many others), the Bayesian procedures remains a field of theoretical research, too hard to be applied.

Bayesian inference is completely based on descriptions of the posterior distribution: once the information available before conducting the experiment and the information obtained after it are merged together to compose the posterior distribution, the inferential part ends and the methodologies to address all the inferential problems (point and interval estimation, hypothesis testing, model choice, etc.) are based on description of it. One of the points of strength of Bayesian analysis is its inner coherence: since the parameter is considered as a random variable with its own distribution and since the inferential procedures are based on this distribution, some of the typical paradoxes of frequentist theory are avoided (for example, the possibility to estimate a non-negative parameter with a negative quantity).

Unfortunately, while Bayesian inference presents some clear advantages, it has some critical points as well. In this work we will address some theoretical and methodological issues in Bayesian analysis by evaluating and proposing new com-

putational solutions.

Firstly, the reason why Bayesian analysis has not been used for real applications for a long time stays in the definition of posterior distribution of a parameter $\theta \in \Omega$ given an observed data set $y \in \mathcal{Y}$ through the Bayes' theorem:

$$\pi(\theta|y) = \frac{\pi(\theta)L(\theta; y)}{\int_{\mathcal{Y}} \pi(\theta)L(\theta; y)dy} = \frac{\pi(\theta)L(\theta; y)}{m(y)}$$

where $\pi(\theta)$ is the distribution summarizing the prior information on the parameter θ and $L(\theta; y)$ is the likelihood function, which provides the information available with the experiment. In theory, after the posterior distribution has been defined, inference on the parameter θ is based on descriptions of it. In practice, the composition of the information available with the data and the information available *a priori* does not provide a distribution known in closed form, except in few cases (as in the case of conjugate priors) and, therefore, this distribution can not be analytically managed.

Only in the 1980s there was a dramatic growth in research and applications of Bayesian methods, thanks to the development of computer technology, which allowed for the implementation of Monte Carlo (and MCMC) methods to approximate the posterior distribution in more general cases and to manage more complicated (and realistic) situations.

Recently, also the methods proposed in those years have become obsolete in many cases. The great complexity of modern applications, as in genetics, computer science, finance, climatic science etc., has led to the proposal of new models which may realistically describe the reality. For example, the model may have the form

$$f\left(\mathbf{y};\, \theta\right) = \int_{\mathcal{Z}} f\left(\mathbf{y},\, \mathbf{z};\, \theta\right) d\mathbf{z} = \int_{\mathcal{Z}} f\left(\mathbf{y} \mid \mathbf{z},\, \theta\right)\, f\left(\mathbf{z};\, \theta\right) d\mathbf{z}$$

where $\mathbf{z}$ plays the role of a latent nonobservable structure; then the likelihood function may be unavailable, because of a too large dimension of $\mathbf{z}$. This is the case of stochastic volatility models, where the integration is with respect to all the time of observation, or genetical models, where the role of the latent variable is played by the complete genealogical tree. In these cases, classical MCMC methods fail to approximate the posterior distribution, because they are too slow to investigate the full parameter space. In other cases, the likelihood function is unavailable, because it is impossible to analytically work with it, as in the case of Gibbs random fields, where the likelihood function is unavailable because of a normalizing constant depending

on the parameter.

New algorithms have been proposed to handle these situations. In particular, a new class of algorihtms is the so-called likelihood-free algorithms or approximate Bayesian computation (ABC), which allows to manage models where the likelihood function may be considered intractable. ABC has been proposed in the last years of the 1990s in an applied setting; the main idea is that if a proposed value for the parameter is generated from a known distribution (for instance, the prior distribution) and then a new data set is simulated from the model by fixing the parameter equal to the proposed value and the simulated data set is similar in some sense to the observed one, then the proposed value is likely to have generated the observed data and is included in the sample which will approximate the posterior distribution. There are many key points in this methodology, for example how to define the concept of similarity between data sets, nevertheless in this way it is only necessary to be able to simulate from the model to provide an approximation of the posterior distribution and no manipulation of the likelihood function is required.

The ABC methodology has been proposed in a Bayesian setting as a way to approximate the posterior distribution. Nevertheless it can also be used in other situations. In Chapter 1 we will propose a way to approximate the (integrated) likelihood function of a parameter of interest when the model includes many (potentially infinite) nuisance parameters, to perform inference in a classical setting (where the prior distribution is intended as a weight function for the integration).

Chapter 1 deals with a key point in complex models: in many situations, the experimenter introduces nuisance parameters which lack of a physical meaning, but are necessary to define flexible and realistic models. Sometimes unknown functions (with an infinite number of parameters) are introduced in the model, to reduce the a priori assumptions. Nevertheless, the interest remains in few parameters. For example, in multivariate analysis, the concept of dependence is crucial, but it is very difficult to work with complicated models (with no assumptions of normality, for instance). In particular, the goal of copula models is to make inference on a copula function which captures all the dependence structure among the variables. It is by now clear that a misspecification of the shape of the copula function leads to not reliable results, but nonparametric approaches are not yet fully developed. Chapter 2 will propose a way to make inference on indexes of dependence (as the Spearman's ρ, the Kendall's τ or tail dependence coefficients) without making assumptions on the shape of the copula function and via the ABC methodology.

Secondly, hypothesis testing may be seen has a problem of determining posterior odds

$$\frac{\Pr(\theta \in \Omega_0|y)}{\Pr(\theta \in \Omega_1|y)} = \frac{\Pr(\theta \in \Omega_0)}{\Pr(\theta \in \Omega_1)} \frac{\int_{\Omega_0} f(y;\theta)\pi_0(\theta)d\theta}{\int_{\Omega_1} f(y;\theta)\pi_1(\theta)d\theta}$$

where the last term, called Bayes factor, is the ratio between the integrated likelihood with respect to the subset of the parameter space relative to each hypothesis (a generalization to the problem of model choice is straightforward). Therefore, the Bayes factor may be seen as a tool to address the problem of hypothesis testing and can be defined as the ratio between marginal distributions

$$B^{\pi}_{01} = \frac{\int_{\Omega_0} f(y;\theta)\pi_0(\theta)d\theta}{\int_{\Omega_1} f(y;\theta)\pi_1(\theta)d\theta}$$

where Ω_i for $i = 0, 1$ is the subspace of Ω relativi to hypothesis H_i and $\pi_i(\cdot)$ is the prior distribution under hypothesis H_i. Again, the theoretical definition of the Bayes factor is straightforwars, nevertheless is computation may be challenging, in particular because of the choice of the prior distribution.

While inference for complex models is more developed, problems of model choice have not yet general solutions. In Chapter 3 we will analyze a recent proposal to redefine the Bayes factor, show its weaknesses and present an alternative applicable in situations where the likelihood function is unavailable.

Thirdly, one of the most delicate and controversial aspects of Bayesian inference is how to compose the prior information in order to form a distribution. Even if prior information is available, defining a distribution (known in closed form) summarizing it is not obvious. Moreover, in the case that little information or no information is available, Bayesian inference still requires to choose a distribution. Many Authors have worked in order to define a "noninformative" prior distribution, i.e. a default procedure to define the prior distribution when prior information is not available. As examples, one may cite the work of Box and Tiao, who consider the Laplace case of constant priors on the parameter, the work of Jeffreys, who considers a model-based prior based on the expected Fisher information, and the work of Berger and Bernardo, who propose an automatic procedure based on maximizing the information obtained with the posterior distribution, i.e. the Kullblack-Leiber divergence between the posterior and the prior distribution. One of the greatest disadvantages of the proposed procedures is that they often lead to improper priors, i.e. not in-

tegrable quantities. While some have argued against the use of improper priors, on several grounds, one may argue in favor of them by considering the posterior distribution as a limiting measure of the posterior function obtained by using an improper prior. As a practical note aside, improper priors are, in general, the only way to interpret the maximum likelihood estimators as Bayesian estimators. Even if one accept the use of improper priors, it could be difficult to assess their influence in the setting of complicated models, where the parameter has no physical meaning.

Mixture models are an example of complicated model which allows to describe kurtotic, multimodal and asymmetric data by considering a composition of known distributions:

$$\sum_{i=1}^{k} p_i \, f_i(x|\theta_i) \,, \quad \sum_{i=1}^{k} p_i = 1 \,.$$

The literature for mixture models is huge, both for inferential problems and for problems of model choice. Mixture models have an ill-defined nature (non-identifiability, multimodality, unbounded likelihood, etc.) and this leads to some difficulties in defining a prior distribution. In many cases, it is difficult to give a meaning to all the parameters of the mixture, therefore a noninformative analysis should be preferred. Nevertheless many works have shown that improper priors are likely to produce improper posterior. In Chapter 4, we will analyze the Jeffreys approach to define a noninformative prior in this setting and propose an alternative which consists in a redefinition of the model.

While the proposal of Chapter 4 is thought for mixture models, it can be generalized to other type of complicated models, in order to use improper priors for parameters lacking of a physical meaning.

A discussion of the results found throughout this thesis is presented in the conclusive Chapter.

Chapter 1

Approximate Integrated Likelihood via ABC methods

1.1 Introduction

Given[1] a statistical model with generic density $p(x|\boldsymbol{\theta})$, with $\boldsymbol{\theta} \in \Theta \subset \mathbb{R}^d$, one is often interested in a low dimensional function $\boldsymbol{\psi}$ of the parameter vector $\boldsymbol{\theta}$, say $\boldsymbol{\psi} = \boldsymbol{\psi}(\boldsymbol{\theta}) \in \mathbb{R}^k$, with $k < d$. Modern parametric or semi-parametric statistical theories, at least the approaches based on likelihood and Bayesian theories, aim at constructing a likelihood function which depends on $\boldsymbol{\psi}$ only. There is much literature on the problem of eliminating nuisance parameters, and we do not even try to summarize it. Interested readers may refer to Berger et al. (1999) and Liseo (2005) for a Bayesian perspective, and to the comprehensive books by Pace and Salvan (1997) and Severini (2000) or to Lancaster (2000) for a more classical point of view. In a Bayesian framework the problem of eliminating the nuisance parameters is, at least in principle, trivial. Let $\boldsymbol{\lambda} = \boldsymbol{\lambda}(\boldsymbol{\theta})$ the complementary parameter transformation, such that $\boldsymbol{\theta} = (\boldsymbol{\psi}, \boldsymbol{\lambda})$ and let

$$\pi(\boldsymbol{\theta}) = \pi(\boldsymbol{\psi}, \boldsymbol{\lambda}) = \pi(\boldsymbol{\psi})\pi(\boldsymbol{\lambda}|\boldsymbol{\psi}) \tag{1.1}$$

the prior distribution. Then, after assuming we observe a data set $\mathbf{x} = (x_1, \ldots, x_n)$ from our working model, and computed the likelihood function $L(\boldsymbol{\psi}, \boldsymbol{\lambda}; \mathbf{x}) \propto p(\mathbf{x}; \boldsymbol{\psi}, \boldsymbol{\lambda})$, the marginal posterior distribution of $\boldsymbol{\psi}$ is

[1] joint wort with Prof. Brunero Liseo, MEMOTEF, Sapienza Università di Roma

$$\begin{aligned}\pi(\boldsymbol{\psi}|\mathbf{x}) &= \frac{\int_{\Lambda} \pi(\boldsymbol{\psi},\boldsymbol{\lambda}) L(\boldsymbol{\psi},\boldsymbol{\lambda};\mathbf{x}) d\boldsymbol{\lambda}}{\int_{\Psi}\int_{\Lambda} \pi(\boldsymbol{\psi},\boldsymbol{\lambda}) L(\boldsymbol{\psi},\boldsymbol{\lambda};\mathbf{x}) d\boldsymbol{\lambda} d\boldsymbol{\psi}} \\ &\propto \pi(\boldsymbol{\psi}) \int_{\Lambda} \pi(\boldsymbol{\lambda}|\boldsymbol{\psi}) L(\boldsymbol{\psi},\boldsymbol{\lambda};\mathbf{x}) d\boldsymbol{\lambda}.\end{aligned} \tag{1.2}$$

The integral in the right-hand side of (1.2) is, by definition, the integrated likelihood for the parameter of interest $\boldsymbol{\psi}$, where "integration" is meant with respect to the conditional prior distribution $\pi(\boldsymbol{\lambda}|\boldsymbol{\psi})$; it will be denoted by $\tilde{L}(\boldsymbol{\psi};\mathbf{x})$. The use of integrated likelihoods has become popular also among non Bayesian statisticians; there are several examples in which its use is clearly superior, or at least equivalent, even from a repeated sampling perspective, in reporting the actual uncertainty associated to the estimates. See for example, Severini (2007), Severini (2010) and Severini (2011).

However the explicit calculation of the above integral might not be so easy, especially when the dimension $d-k$ is large. Notice that the dimension d may also include a possible latent structure which, from a strictly probabilistic perspective, is not different from a parameter vector. In this Chapter we are interested to explore the use of approximate Bayesian computation (ABC, henceforth) methods in producing an approximate integrated likelihood function, in situations where a closed form expression of $\tilde{L}(\boldsymbol{\psi};\mathbf{x})$ is not available, or it is too costly even to evaluate the "global" likelihood function $L(\boldsymbol{\psi},\boldsymbol{\lambda};\mathbf{x})$, like, for example, in many genetic applications or in the hidden (semi)-Markov literature. These are situations where MCMC methods may not be satisfactory and completely reliable.

Another class of problems where an integrated likelihood would be of primary interest is that of semi-parametric problems, where the parameter of interest is a scalar - or a vector - quantity and the nuisance parameter is represented by the nonparametric part of the model; in such cases the integration step over the Λ space would be infinite dimensional, and very often infeasible to be solved in a closed form; we will discuss this issue in Section 1.4.

Approximate Bayesian computation has now become an essential tool for the analysis of complex stochastic models when the likelihood function is unavailable. It can be considered as a (class of) popular algorithms that achieves posterior simulation by avoiding the computation of the likelihood function (see Beaumont (2010) for a recent survey). A crucial condition for the use of ABC algorithms is that it must be relatively easy to generate new pseudo-observations from the working

model, for a fixed value of the parameter vector. In its simplest form, the ABC algorithm is as follows (Algorithm 1 in Marin et al. (2012))

Algorithm 1 Likelihood-free Rejection algorithm

for $i = 1$ **to** N **do**

repeat

Generate θ from the prior distribution $\pi(\cdot)$

Generate z from the likelihood function $f(\cdot \mid \theta)$

until $z = y$ (or some statistics η is such that $\eta(z) \approx \eta(y)$)

set $\theta_i = \theta$

end for

In this Chapter we will argue, through several examples of increasing complexity, how the approximate integrated likelihood produced by ABC algorithms performs when compared with the existing methods. We will also explore its use in particular examples where other methods simply fail to produce a useful and easy-to-use likelihood function for the parameter of interest. The Chapter is organized as follows. In the next section we describe our proposal in detail. Section 1.3 discusses some theoretical issues related to the precision of the ABC approximation. Section 1.4 compares the ABC integrated likelihood with other available approaches in a series of examples. Section 1.5 concludes with a final discussion of pros and cons of the method.

1.2 The proposed method

The main goal of this Chapter is to obtain an approximation of the integrated likelihood $\tilde{L}(\boldsymbol{\psi}; \mathbf{x})$, for $\boldsymbol{\psi} = \boldsymbol{\psi}(\boldsymbol{\theta}) \in \mathbb{R}^k$. From expression (1.2) it is easy to see that

$$\tilde{L}(\boldsymbol{\psi}; \mathbf{x}) \propto \frac{\pi(\boldsymbol{\psi}|\mathbf{x})}{\pi(\boldsymbol{\psi})}, \tag{1.3}$$

that is the integrated likelihood function may be interpreted as the amount of experimental evidence which transforms our prior knowledge into posterior knowledge about the parameter of interest: from this perspective, we can interpret (1.3) as the Bayesian definition of the integrated likelihood function.

Suppose that $\tilde{L}(\boldsymbol{\psi}; \mathbf{x})$ is hard or impossible to obtain in a closed form. For example the nuisance parameter $\boldsymbol{\lambda}$ might be infinite dimensional (see Example 4.4)

or it may represent the non observable latent structure associated to the statistical model as in Hidden Markov or semi-Markov set-ups.

In these situations one can exploit the alternative expression (1.3) of $\tilde{L}(\boldsymbol{\psi}; \mathbf{x})$. Of course, if $\tilde{L}(\boldsymbol{\psi}; \mathbf{x})$ is not available, neither $\pi(\boldsymbol{\psi}|\mathbf{x})$ will be. However it is possible to obtain an approximate posterior distribution $\tilde{\pi}(\boldsymbol{\psi}|\mathbf{x})$, by using some standard ABC algorithm; in Section 1.3 we will discuss some issues related to the precision of this approximation; for now, we describe the practical implementation of the method. As in any ABC approach for the estimation of the posterior distribution, one has to

- select a number of summary statistics $\eta_1(\mathbf{x}), \ldots, \eta_h(\mathbf{x})$;
- select a distance $\rho(\cdot, \cdot)$ to measure the distance between "true" and proposed data, or their summary statistics;
- select a tolerance threshold ε
- choose a (MC)MC algorithm which proposes values for the parameter vector $\boldsymbol{\theta}$.

Once the posterior is approximated by a size M ABC posterior sample $(\boldsymbol{\theta}_1^*, \boldsymbol{\theta}_2^*, \ldots \boldsymbol{\theta}_M^*)$, one can produce a non parametric kernel based density approximation of the marginal posterior distribution of $\boldsymbol{\psi}$, say $\tilde{\pi}^{ABC}(\boldsymbol{\psi}|\mathbf{x})$. A similar operation can be done with the marginal prior $\pi(\boldsymbol{\psi})$, by performing another - cheap - simulation from $\pi(\boldsymbol{\psi})$ to get another density approximation, say $\tilde{\pi}(\boldsymbol{\psi})$. Notice that one is bound to use proper priors for all the involved parameters.

Then one can define the ABC integrated likelihood

$$\tilde{L}^{ABC}(\boldsymbol{\psi}; \mathbf{x}) \propto \frac{\tilde{\pi}^{ABC}(\boldsymbol{\psi}|\mathbf{x})}{\tilde{\pi}(\boldsymbol{\psi})}. \tag{1.4}$$

1.3 The quality of approximation

The gist of this note is to propose an approximate method for producing a likelihood function for a quantity of interest when the usual road of integrating with respect to the nuisance parameters cannot be followed. There are two sources of error in (1.4). The first type of approximate error is introduced by the ABC approximation in the numerator so the level of accuracy of (1.4) is of the same order of any ABC-type approximation. We believe that the main difficulty with ABC methods is the

choice of summary statistics. However, while generic ABC methods have the goal of producing a "global" approximation to the posterior distribution, our particular use of the ABC approximation may suggest some alternative strategies for the choice of summary statistics. Classical statistical theory on the elimination of nuisance parameters can be in fact of some guidance in the selection of summary statistics which are partially or conditionally sufficient for the parameter of interest. Basu (1977) represents an excellent reading on these topics. In particular, his Definition 5 of "Specific Sufficiency" can be used in semi-parametric set-ups, like Example 4.4 below, where the selected summary statistics are oriented towards the preservation of information about the parameter of interest. In our notation a statistic T is specific sufficient for $\boldsymbol{\psi}$ if, for each fixed value of the nuisance parameter $\boldsymbol{\lambda}$, T is sufficient for the restricted statistical model in which $\boldsymbol{\lambda}$ is held fixed and known.

Another source of error in ABC is given by the tolerance threshold ε. As stressed in Marin et al. (2012), the choice of the tolerance level is mostly a matter of computational power: smaller ε's are associated with higher computational costs and more precision. It is enough to reproduce the argument in Section 1.2 of Sisson and Fan (2011) to see that for $\varepsilon \to 0$, the error in (1.4), which is due to the tolerance, vanishes.

Then, there is a balance between the fact that ε has to be small and the fact that the simulation has to be practicable. It could be useful to choose ε in a recursive way, by realizing a first simulation with a high tolerance level and then by choosing it in the left tail of the thresholds related to the accepted values. However, it is always recommended to compare different levels.

The second main source of error is due to the kernel approximation step. A second order expansion for a Gaussian kernel estimator provides that

$$\mathbb{E}\left[\tilde{\pi}^{ABC}\left(\boldsymbol{\psi}|\mathbf{x}\right)\right] = \pi\left(\boldsymbol{\psi}|\mathbf{x}\right) + \frac{1}{2}\frac{\partial^2}{\partial\psi^2}\pi\left(\boldsymbol{\psi}|\mathbf{x}\right)h_x^2 k_2 + \mathcal{O}\left(h_x^4\right) \tag{1.5}$$

where h_x is the bandwidth and $k_2 = 1$ in the case of Gaussian kernel.

A similar approximation holds for the prior distribution. Then, using general results on a first order approximation for the ratio of functions of random variables (Kendall et al. (1987), pag. 351), one has

$$\mathbb{E}\left[\frac{\tilde{\pi}^{ABC}\left(\boldsymbol{\psi}|\mathbf{x}\right)}{\tilde{\pi}\left(\boldsymbol{\psi}\right)}\right]=\frac{\pi\left(\boldsymbol{\psi}|\mathbf{x}\right)+\frac{1}{2}\frac{\partial^2}{\partial\psi^2}\pi\left(\boldsymbol{\psi}|\mathbf{x}\right)h_x^2+\mathcal{O}\left(h_x^4\right)}{\pi\left(\boldsymbol{\psi}\right)+\frac{1}{2}\frac{\partial^2}{\partial\psi^2}\pi\left(\boldsymbol{\psi}\right)h_\pi^2+\mathcal{O}\left(h_\pi^4\right)} \tag{1.6}$$

where h_x is the bandwidth chosen for the approximation of the posterior distribution and h_π is the one chosen for the approximation of the prior. The prior distribution is often known in closed form or may be easily approximated with a higher accuracy than the posterior distribution.

The previous formula ensures that our estimator will be consistent provided that a sample size dependent bandwidth h_n, converging to 0, is adopted.

It is a matter of calculation to show that the variance of the estimator is

$$\begin{aligned}&\mathbb{V}\left[\frac{\tilde{\pi}^{ABC}\left(\boldsymbol{\psi}|\mathbf{x}\right)}{\tilde{\pi}\left(\boldsymbol{\psi}\right)}\right]\\ =&\left[\frac{\pi\left(\boldsymbol{\psi}|\mathbf{x}\right)+C_{\mathbf{x}}}{\pi\left(\boldsymbol{\psi}\right)+C}\right]^2\\ \times&\left[\frac{\frac{\pi(\boldsymbol{\psi}|\mathbf{x})}{2nh_x\sqrt{\pi}}+\mathcal{O}\left(n^{-1}\right)}{\left[\pi\left(\boldsymbol{\psi}|\mathbf{x}\right)+C_{\mathbf{x}}\right]^2}+\frac{\frac{\pi(\boldsymbol{\psi})}{2nh_\pi\sqrt{\pi}}+\mathcal{O}\left(n^{-1}\right)}{\left[\pi\left(\boldsymbol{\psi}\right)+C\right]^2}\right]\end{aligned} \tag{1.7}$$

where

$$C_{\mathbf{x}}=\frac{h_x^2}{2}\frac{\partial^2}{\partial\psi^2}\pi\left(\boldsymbol{\psi}|\mathbf{x}\right)+\mathcal{O}\left(h_x^4\right)$$

and

$$C=\frac{h_\pi^2}{2}\frac{\partial^2}{\partial\psi^2}\pi\left(\boldsymbol{\psi}\right)+\mathcal{O}\left(h_\pi^4\right)$$

.

Again, using a bandwidth h_n, such that $h_n\to 0$, as $n\to\infty$, one can see that the first factor of the variance is asymptotically equal to the square of the true unknown value, while the second factor vanishes like n^{-1}.

In conclusion, the ABC approximation of the integrated likelihood function mainly depends on the ABC approximation and the kernel density estimate of the posterior distribution, whereas the prior distribution may be considered known, in general. Blum (2010) shows that the asymptotic variance of the kernel density estimator of the posterior distribution inversely depends on the number of simulations n and on the kernel bandwidth, while the bias is proportional to the bandwidth. The mean squared error is minimized by

$$h_n = \mathcal{O}\left(n^{-\frac{1}{d+5}}\right) \tag{1.8}$$

where d is the dimension of the summary statistics. Then the minimal MSE is

$$MSE^* = \mathcal{O}\left(n^{-\frac{4}{d+5}}\right) \tag{1.9}$$

which shows that the accuracy in the approximation decreases as the dimension of the summary statistics increases. This result may be used to define the number of simulations (and the burn-in) needed to reach the desired level of accuracy.

1.4 Examples

In this Section we illustrate our proposal throughout several examples of increasing complexity. The first one is a toy example and it is included only to show - in a very simple situation - which are the crucial steps of the algorithm.

Example 4.1. [Poisson means]. Suppose we observe a sample of size n from $X \sim \text{Poi}(\theta_1)$ and, independently of it, another sample of size n from $Y \sim \text{Poi}(\theta_2)$. The parameter of interest is $\psi = \theta_1/\theta_2$. This is considered a benchmark example in partial likelihood literature since the conditional likelihood (see Kalbfleisch and Sprott (1970)), the profile likelihood and the integrated likelihood obtained using the conditional reference prior (Berger et al. (1999)) are all proportional to

$$\tilde{L}(\psi; \mathbf{x}, \mathbf{y}) \propto \frac{\psi^{n\bar{x}}}{(1+\psi)^{n(\bar{x}+\bar{y})}},$$

with the obvious meaning of the symbols above. Without loss of generality, set $\lambda = \theta_2$ as the nuisance parameter.

In this situation, the ABC approximation of the integrated likelihood is, in some sense, not comparable with the "correct" integrated likelihood because the latter is obtained through the use of an improper conditional reference prior on λ given ψ, and, as already stressed, it is not possible to use improper priors in the ABC approach. A solution may be using a prior which mimics the reference prior: we have taken $\theta_1, \theta_2 \overset{\text{iid}}{\sim} \text{Ga}(0.1, 0.1)$. Notice that, in the economy of the method, only the prior on λ, not on ψ is important. The ABC algorithm has been implemented to obtain approximations for the posterior distributions of θ_1 and θ_2. The distance ρ has been taken as the Euclidean distance, different tolerance levels have been

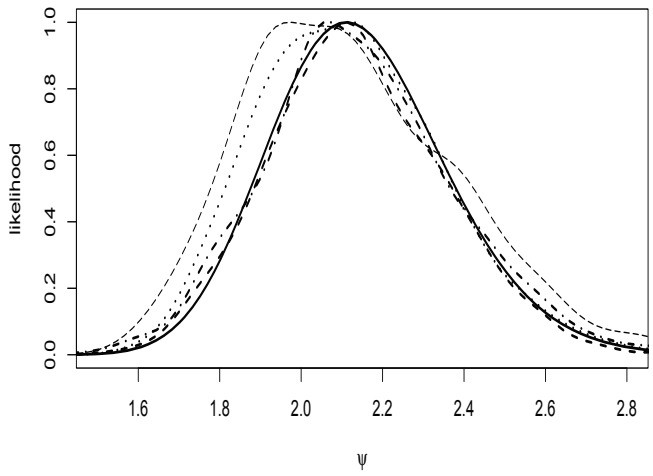

Figure 1.1:

The integrated likelihood of ψ (solid line) and its approximations for different tolerance levels: $\varepsilon = 0.001$ (dashed), $\varepsilon = 0.01$ (dotted), $\varepsilon = 0.1$ (dotdashed) and $\varepsilon = 0.5$ (longdashed).

compared - $\varepsilon = (0.001, 0.01, 0.1, 0.5)$ - and the sample means of the two samples have been taken as summary (sufficient) statistics. Samples of $1,000$ simulations have been obtained to approximate the posterior distributions. The approximation to the posterior distribution of ψ is then simply obtained as the ratio between the accepted values for θ_1 and θ_2 via ABC. Given a sample from the prior distribution of ψ, the approximation of its integrated likelihood is obtained through the ratio between the kernel density estimates of both the prior and the posterior distribution.

Figure 1.1 shows the approximations with different choices of the tolerance level: the approximations are close together and they are all close to the integrated likelihood; the choice for the tolerance level does not seem to have a strong influence; it is mostly a matter of computational power: the acceptance rate is generally very low (often under 1%), nevertheless it grows with the tolerance level. As the threshold goes to zero, the approximation is closer to the integrated likelihood, although the computational time increases.

Simulations have been repeated for different scenarios, by changing the sample size and the number of simulations, however the results do not seem to change in a significant way. In particular, as expected, the algorithm does not depend on the (induced) prior on ψ.

Example 4.2. [Neyman and Scott's class of problems]. This is a famous class

of problems, where the number of parameters increases with the sample size (Neyman and Scott 1948, Lancaster 2000). Here we consider a specific example, already discussed in Davison (2003) and Liseo (2005), namely matched pairs of Bernoulli observations: every subject is assigned to treatment or control group and the randomization occurs separately within each pair, i.e. each data point in one data set is related to one and only one data point in the other data set. Let Y_{ij}'s be Bernoulli random variables, where $i = 1, \ldots, k$ represents the stratum and $j = 0, 1$ indicates the observation within the pair. The probability of success p_{ij} follows a logit model:

$$\operatorname{logit} p_{ij} = \lambda_i + \psi_j \tag{1.10}$$

For identifiability reasons, ψ_0 is set equal to 0, while $\psi_1 = \psi$ is considered constant across the k strata; ψ is the parameter of interest. To formalize the problem, assume $(R_{i0},\, R_{i1})$ are k independent matched pairs such that, for each i:

$$R_{i0} \sim \text{Be}\left(\frac{e^{\lambda_i}}{1+e^{\lambda_i}}\right), \quad R_{i1} \sim \text{Be}\left(\frac{e^{\lambda_i+\psi}}{1+e^{\lambda_i+\psi}}\right). \tag{1.11}$$

The complete likelihood for $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_k)$ and ψ is

$$L\left(\psi,\, \boldsymbol{\lambda}\right) = \frac{e^{\sum_{i=1}^{k} \lambda_i S_i + \psi T}}{\prod_{i=1}^{k} \left(1+e^{\lambda_i}\right)\left(1+e^{\lambda_i+\psi}\right)} \tag{1.12}$$

where $S_i = R_{i0} + R_{i1}$ for $i = 1, \ldots, k$ and $T = \sum_{i=1}^{k} R_{i1}$ is the number of successes among the cases. It is easy to show that the conditional maximum likelihood estimate of λ_i is infinite when $S_i = 0$ or $S_i = 2$. The classical solution to this problem is to eliminate the pairs where $S_i = 0$ or $S_i = 2$ from the analysis. Nevertheless this is certainly a loss of information, because the fact that a pair gives the same result under both treatments may suggest a "not-so-big" difference between groups.

It is easy to show that the conditional maximum likelihood estimator is

$$\left[\hat{\lambda}_{i,\psi} \mid (S_i = 1)\right] = -\frac{\psi}{2};$$

also, let b be the number of pairs with $S_i = 1$. The profile likelihood of ψ is

$$\tilde{L}\left(\psi \mid S_i = 1\right) = \frac{e^{\psi T}}{\left(1+e^{\frac{\psi}{2}}\right)^{2b}} \tag{1.13}$$

This likelihood function is not useful, since the maximum likelihood estimate for ψ is inconsistent (see Davison (2003), Example 12.13): as b increases, $\hat{\psi} \to 2\psi$. The modified version of the profile likelihood, proposed by Barndorff-Nielsen (1983) uses a multiplying factor:

$$M(\psi) = \left| J_{\lambda\lambda}\left(\psi,\, \hat{\lambda}_\psi\right)\right|^{-\frac{1}{2}} \left|\frac{\partial\hat{\lambda}}{\partial\hat{\lambda}_\psi}\right| = \frac{e^{\frac{b\psi}{4}}}{\left(1+e^{\frac{\psi}{2}}\right)^b} \tag{1.14}$$

where $J_{\lambda\lambda}\left(\psi,\, \hat{\lambda}_\psi\right)$ is the lower right corner of the observed Fisher information matrix.

The conditional distribution of T given $S_1 = S_2 = \cdots = S_b = 1$ is Binomial and depends on ψ only. That is $T \mid [S_1 = S_2 = \cdots = S_b = 1,\, \psi] \sim \text{Bin}\left(b,\, \frac{e^\psi}{1+e^\psi}\right)$; we can use it to get a conditional likelihood function:

$$L_C(\psi) \propto \binom{b}{T} \frac{e^{\psi T}}{(1+e^\psi)^b} \tag{1.15}$$

which leads to a consistent maximum conditional likelihood estimator.

A Bayesian approach has the advantage that it does not need to discard the pairs with $S_i = 0$ or 2. The likelihood contribution for the i-th pair is simply

$$L(\psi,\, \lambda_i) = \frac{e^{\lambda_i S_i + \psi R_{i1}}}{(1+e^{\lambda_i})\,(1+e^{\lambda_i+\psi})}. \tag{1.16}$$

With a change of parametrization $\omega_i = e^{\lambda_i}/(1+e^{\lambda_i})$ and using a (proper) Jeffreys' prior for $\omega_i|\psi$ (namely a $Beta\left(\frac{1}{2},\, \frac{1}{2}\right)$), the integrated likelihood is

$$\tilde{L}_i(\psi) = e^{\psi R_{i1}} \int_0^1 \frac{\omega_i^{S_i-\frac{1}{2}}\,(1-\omega_i)^{\frac{3}{2}-S_i}}{1-\omega_i\,(1-e^\psi)} d\omega_i \tag{1.17}$$

where the integral is one of the possible representation of the Hypergeometric or Gauss series, as shown in Abramowitz and Stegun (1964) (formula 15.3.1, pag. 558). Therefore, the integrated likelihood is proportional to

$$\tilde{L}_i(\psi) \propto {}_2F_1\left(1,\, S_i+\frac{1}{2},\, 3,\, 1-e^\psi\right) e^{\psi R_{i1}}. \tag{1.18}$$

Define $\tilde{L}_{jl}(\psi)$ as the integrated likelihood function associated with the i-th pair

for which $(R_{i0},\ R_{i1}) = (j,\ l)$ and n_{jl} the number of pairs for which $(R_{i0},\ R_{i1}) = (j,\ l)$, then the integrated likelihood function for ψ is

$$\tilde{L}(\psi) \propto \prod_{j,l=0,1} \tilde{L}_{jl}(\psi)^{n_{jl}}. \tag{1.19}$$

It is worthwhile to notice that this likelihood is not, in some sense, comparable with profile and conditional likelihoods, because it also considers the pairs discarded by non-Bayesian methods.

The ABC approach has been used with simulated data, with a sample size n equal to 30. Simulations were performed by setting $\psi = 1$, a value which is quite frequent in applications, when similar treatments are compared. The values of $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_n)$ have been generated by setting $\xi_i = \lambda_i/(1+\lambda_i)$ and drawing the ξ's from a $\mathrm{U}(0,1)$ distribution. Again, we have used the Euclidean distance between summary statistics and different tolerance levels $\varepsilon = (0.001,\ 0.01,\ 0.1,\ 0.5)$. The summary statistics are the sample means for $\mathbf{R}_0$ and $\mathbf{R}_1$. We have also assumed a normal prior for ψ with zero mean and standard deviation equal to 10. The proposed values for λ_i's have been generated from a $Beta\left(\frac{1}{2},\ \frac{1}{2}\right)$ distribution for the above defined transformations ξ_i's.

With a sample from the posterior distribution of ψ for each tolerance level and a sample from its prior distribution, we have obtained an approximation of the likelihood of ψ via density kernel estimation. The results are shown in Figure 1.2: the approximations are quite good for tolerance levels below 0.1; on the other hand, when the threshold grows to 0.5 the approximate likelihood function is very flat and multi-modal, i.e. too many proposed values, even very different from the true value of ψ, are misleadingly accepted; for example, a value of ψ around 43 has been accepted in one of our simulations.

Once again, a fair comparison between Bayesian and non-Bayesian approaches is not strictly possible, nevertheless the various proposals are shown in Figure 1.3: all the proposed solutions are concentrated relatively close to the true value, although the profile likelihood seems to be biased towards large values of ψ: this behavior is also present, although to a minor extent in the modified profile and the integrated likelihood solutions. The ABC approximation closely mimics the integrated likelihood, obtained via a saddle-point approximation of the Hypergeometric series (see Butler and Wood (2002)).

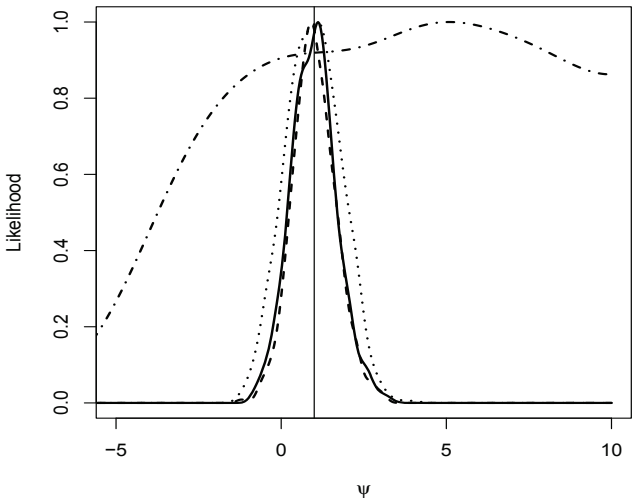

Figure 1.2:

ABC approximations of the integrated likelihood for ψ with different tolerance levels: $\varepsilon = 0.001$ (solid line), $\varepsilon = 0.01$ (dashed), $\varepsilon = 0.1$ (dotted), $\varepsilon = 0.5$ (dotdashed).

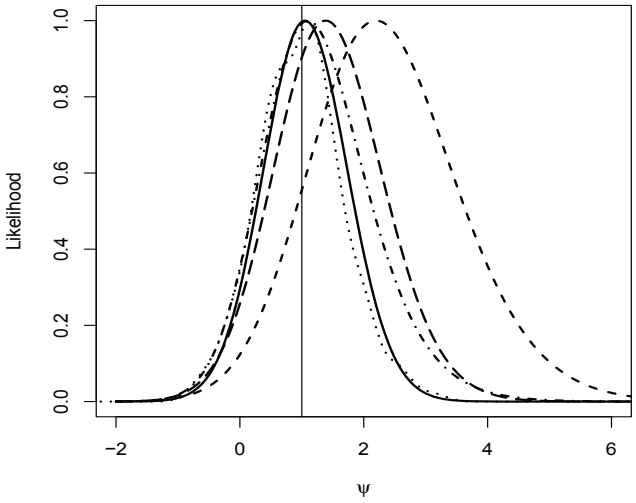

Figure 1.3:

Likelihood functions for ψ based on different solutions ($n = 30$): the profile likelihood (dashed line), the conditional likelihood (dotdashed line), the modified profile likelihood (with the Barndorff-Nielsen correction, longdashed line), the integrated likelihood (solid line, drawn by using the Laplace approximation of ${}_2F_1$ by Butler and Wood (2002) and the ABC approximation ($n_{sim} = 1000$, prior $\psi \sim N(1, 10)$ and tolerance level $\varepsilon = 0.001$, dotted line).

Similar conclusions are valid for different choices of the prior distribution, different sample sizes, and different numbers of simulations. Just like in Example 4.1, the acceptance rates are typically very low (always under 1% for tolerance levels under 0.01 and about 5% for a tolerance level of 0.1). Acceptance rates dramatically increase to about 60%, for $\varepsilon = 0.5$; however in these cases, approximations get much worse.

Example 4.3 [Likelihood function for the quantiles of a g-and-k distribution]. Quantile distributions are, in general, defined by the inverse of their cumulative distribution function. They are characterized by a great flexibility of shapes obtained by varying parameters values. They may easily model kurtotic or skewed data with the great advantage that they typically have a small number of parameters, unlike mixture models which are usually adopted to describe this kind of data. An advantage of quantile distributions is that it is extremely easy to simulate from them by means of a simple inversion. However, there are no free lunches, and the above advantages are paid with the fact that their probability density functions (and therefore, the implied likelihood functions) are often not available in a closed form expression.

One of the most interesting examples of quantile class of distributions is the so-called g-and-k distribution, described in Haynes et al. (1997), whose quantile function Q is given by

$$Q\left(u;\, A,\, B,\, g,\, k\right) = A + B\left[1 + c\frac{1-\exp\left\{-g\, z\left(u\right)\right\}}{1+\exp\left\{-g\, z\left(u\right)\right\}}\right]\cdot \left\{1+z\left(u\right)^2\right\}^k z\left(u\right) \tag{1.20}$$

where $z\left(u\right)$ is the u-th quantile of the standard normal distribution; parameters A, B, g and k represent location, scale, skewness and kurtosis respectively; c is an additional parameter which measures the overall asymmetry and it is generally fixed at 0.8, following Rayner and MacGillivray (2002a). The class of normal distributions is a proper subset of this class; it is obtained by setting $g = k = 0$. Suppose we are interested in one or more quantiles using this model. There is no easy solution to the problem of constructing a partial likelihood for these quantiles. The fact that the likelihood function is not available makes any classical approach practically impossible to implement. Rayner and MacGillivray (2002b) propose a numerical

maximum likelihood approach; however they also explain that very large sample sizes are necessary to obtain reliable estimates of the parameters. On the other hand, even though the quantile distributions have no explicit likelihood, simulation from these models is easy, and an approximate Bayesian computation approach, also for producing an integrated likelihood of the parameters of interest, seems reasonable.

For this specific problem, two types of ABC algorithms have been compared: the former is the usual ABC algorithm based on simulations from the prior distributions (with 10^3 iterations); the latter is an ABC-MCMC algorithm (10^6 iterations, with a burn-in of 10^5 simulations). Two versions of ABC-MCMC have been used, the former described in Marin et al. (2012) (see Algorithm 2) and the latter described in Allingham et al. (2009) (see Algorithm 3). The main difference between these two versions of ABC-MCMC algorithm is that, in the first case, there is no rejection step; at each iteration a value is accepted (either the new proposed value or the value accepted in the previous iteration); in the second case, instead, it is possible to discard the current value and to propose a new one, so the chain always "moves".

Data have been simulated from a g-and-k distribution with parameters $A = 3$, $B = 1$, $g = 2$ and $k = 0.5$. As previously said, c is considered known and set equal to 0.8. The sample size, has been set equal to $n = 1000$. The empirical cumulative distribution function and the histogram of the simulated data are shown in Figure 1.4.

The transition kernel of the ABC-MCMC algorithm needs to be chosen having in mind two conflicting objectives: on one hand, full exploration of the parameter space, and, on the other hand, a reasonably high acceptance rate, which increases for proposals mostly concentrated where the posterior mass is present. As described in Allingham et al. (2009) uniform priors with bounds (0, 10) have been chosen for each parameter and a random walk-normal kernel with variance 0.1 has been used together with a large number of iterations (10^6) so that the parameter space is likely to be fully investigated. The vector of summary statistics consists of the sample mean, the standard deviation, and the sample skewness and kurtosis indexes. The Euclidean distance has been used to compare summary statistics.

The tolerance level ε has been chosen in a recursive way: first, a very large value has been selected, and a histogram of all the distances has been drawn. A reasonable value has been taken from the 5% left tail of this histogram. Then, the chosen threshold has been compared with smaller values. In particular, a threshold

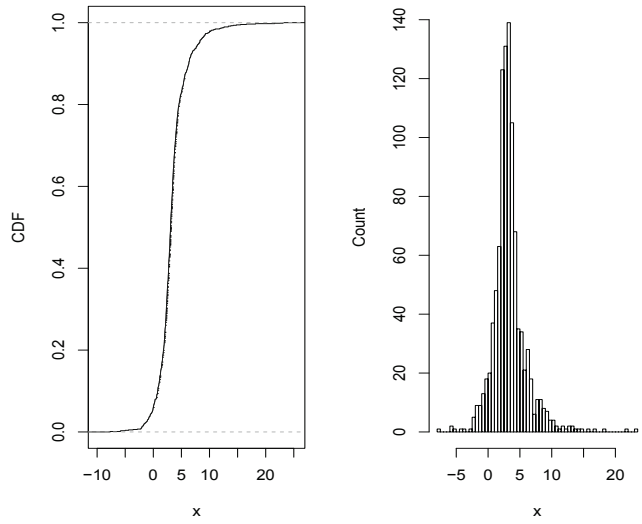

Figure 1.4:

Empirical cumulative distribution function (left) and histogram (right) of the simulated data from a g-and-k distribution (with $A = 3$, $B = 1$, $g = 2$, $k = 0.5$).

equal to 3 corresponds to 3.9% left tail. This has been compared with tolerance levels equal to 2 and 0.5.

Algorithm 2 Likelihood-free MCMC sampling

Initialization

A1) Generate θ' from the prior distribution $\pi(\cdot)$

A2) Generate a data set $\mathbf{z}' \sim f(\cdot \mid \theta')$, where f is the model of the data

A3) If $\rho\{\eta(\mathbf{y}), \eta(\mathbf{z}')\} \leq \varepsilon$, set $\left(\theta^{(0)}, \mathbf{z}^{(0)}\right) = (\theta', \mathbf{z}')$, otherwise return to A1)

MCMC-step

for $t = 1, \ldots, T$ {

1) Generate θ^{prop} from the Markov kernel $q\left(\cdot \mid \theta^{(t-1)}\right)$

2) Generate $\mathbf{z}'$ from the model $f(\cdot \mid \theta^{prop})$

3) Calculate $h\left(\theta^{(t-1)}, \theta^{prop}\right) = \min(1, \frac{\pi(\theta^{prop})\, q\left(\theta^{(t-1)} \mid \theta^{prop}\right)}{\pi\left(\theta^{(t-1)}\right) q\left(\theta^{prop} \mid \theta^{(t-1)}\right)})$

4) **if** $\rho\{\eta(\mathbf{y}), \eta(\mathbf{z}')\} \leq \varepsilon$, set $\left(\theta^{(t)}, \mathbf{z}^{(t)}\right) = (\theta^{prop}, \mathbf{z}')$ with probability h,

else $\left(\theta^{(t)}, \mathbf{z}^{(t)}\right) = \left(\theta^{(t-1)}, \mathbf{z}^{(t-1)}\right)$

}

The analysis of the approximate posterior distributions shows that three out of four parameters (A, B and k) are well identified, while the posterior distribution of g is rather flat. In general, as the tolerance level decreases, results improve and

Algorithm 3 Likelihood-free MCMC sampling

Initialization

A1) Generate θ' from the prior distribution $\pi(\cdot)$

A2) Generate a data set $\mathbf{z}' \sim f(\cdot \mid \theta')$, where f is the model of the data

A3) If $\rho\{\eta(\mathbf{y}), \eta(\mathbf{z}')\} \leq \varepsilon$, set $\left(\theta^{(0)}, \mathbf{z}^{(0)}\right) = (\theta', \mathbf{z}')$, otherwise return to A1)

MCMC-step

for $t = 1, \ldots, T$ {

1) Generate θ^{prop} from the Markov kernel $q\left(\cdot \mid \theta^{(t-1)}\right)$

2) Generate $\mathbf{z}'$ from the model $f(\cdot \mid \theta^{prop})$

3) Calculate $h\left(\theta^{(t-1)}, \theta^{prop}\right) = \min(1, \frac{\pi(\theta^{prop})\, q\left(\theta^{(t-1)} \mid \theta^{prop}\right)}{\pi\left(\theta^{(t-1)}\right) q\left(\theta^{prop} \mid \theta^{(t-1)}\right)})$

4) **if** $\rho\{\eta(\mathbf{y}), \eta(\mathbf{z}')\} \leq \varepsilon$, set $\left(\theta^{(t)}, \mathbf{z}^{(t)}\right) = (\theta^{prop}, \mathbf{z}')$ with probability h,

else return to 1)

}

posterior distributions tend to be more concentrated. Nevertheless, even using the lowest tolerance level the posterior distribution of g does not seem to concentrate around any value. This suggests that the algorithm needs an even smaller value of the threshold. A simulation with tolerance level equal to 0.25 has been then performed using Algorithm 2: the approximation of the posterior distribution of g is still not centered around its true value, even if there is a mode around it; nevertheless the problem with this so low tolerance level and this type of algorithm is that the acceptance rate of new proposed values is very low and the chain does not move too much. This tolerance level is also so low to make the application of the other algorithms prohibitive in terms of computational time.

Our main goal of the analysis was to find an approximation of the integrated likelihood function for a given quantile: in particular, we have considered the percentiles of order 0.05, 0.10, 0.25 and 0.50. Notice that, in the g-and-k distribution model, the median is always equal to A.

The results are shown in Figure 1.5, 1.6 and 1.7. The performance is in general very good: the approximations are always concentrated around the true values.

The ABC algorithm with simulations from the prior distribution has some apparent problems of multi-modality, which are however absent using Algorithm 2. However, in this case, the obtained approximations are not very smooth, and they show more irregularities as the tolerance level decreases: as we have already re-

marked, a too low threshold leads to very low acceptance rates and this means that the chains do not move too much.

In this example, Algorithm 3 has the best overall performance: the approximations are smooth and all concentrated around the true quantile values. As the tolerance level decreases, the likelihood approximations are more concentrated; obviously the computational time gets larger.

The acceptance rates of these algorithms are in general very low:

- the basic ABC algorithm has an acceptance rate of 0.138% when the threshold is equal to 3, and it goes down to 0.041% and 0.007% with tolerance levels of 2 and 0.5 respectively;
- the ABC-MCMC Algorithm 2 needs, respectively, 187, 1487, about $500K$ and more than 3 millions of simulations for the initialization step for the different tolerance levels 3, 2, 0.5 and 0.25.The acceptance rates of the proposed values are also very low: 18.41%, 9.90%, 0.47% and 0.046% respectively; it is clear that the acceptance rates relative to the smaller thresholds cannot lead to smooth approximations;
- the ABC-MCMC Algorithm 3 needs 1104, 4383 and about $400K$ simulations for the initialization step for tolerance levels 3, 2 and 0.5 respectively; in this case every accepted value is a "new" value, and this solves the problems in Algorithm 2.

In conclusion, ABC-MCMC seems to perform better, although the versions we have implemented present some cons: the algorithm in Marin et al. (2012) is faster but it must be calibrated in terms of the tolerance level, which has to be low in order to achieve good approximations, and the MCMC acceptance rate, which has to be sufficiently high in order to allow the chains to move.

Example 4.4 [Semiparametric regression]. Consider the following model

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \gamma(\mathbf{z}) + \varepsilon, \tag{1.21}$$

where $\mathbf{Y} = (Y_1, Y_2, ..., Y_n)'$ is a vector of n real-valued variables, and $\mathbf{x} = (x_1, x_2, \ldots, , x_n)'$ and $\mathbf{z} = (z_1, z_2, \ldots, z_n)'$ are observed constants respectively taking values in $\mathbb{R}^p$ and $\mathcal{Z}$, ε is the usual random component that we assume having multivariate normal

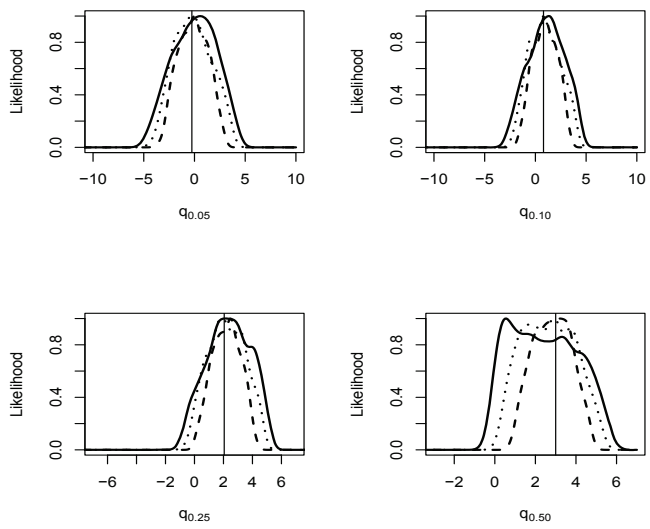

Figure 1.5:

Likelihood approximations of the quantiles of a g-and-k distribution parameters for simulated data, obtained with an ABC algorithm which simulates proposal values from the prior distributions ($\mathrm{U}(0, 10)$ for each parameter): tolerance levels equal to 3 (solid line), 2 (dashed line) and 0.5 (dotted line).

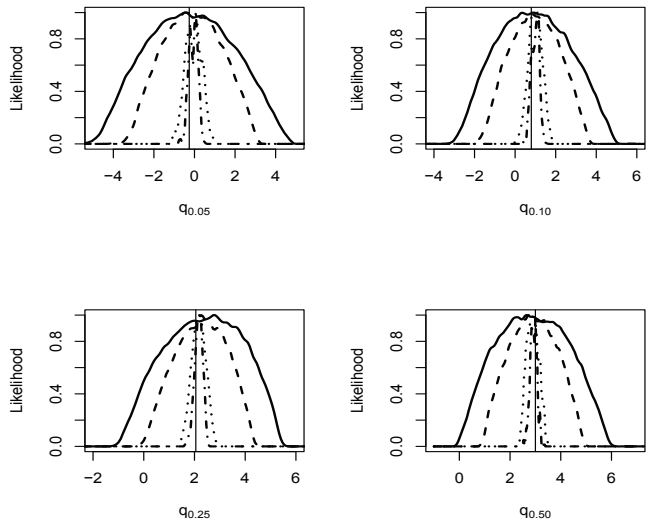

Figure 1.6:

Likelihood approximations of the quantiles of a g-and-k distribution parameters for simulated data, obtained with an ABC-MCMC Algorithm 2 with Gaussian kernel: tolerance levels equal to 3 (solid line), 2 (dashed line) and 0.5 (dotted line) and 0.25 (dotdashed line).

distribution with mean $\mathbf{0}$ and covariance matrix $\mathbf{\Omega}_\phi$ which depends on some param-

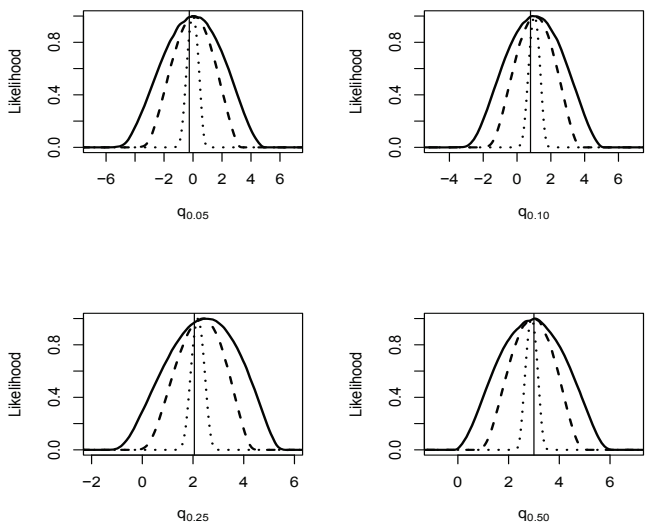

Figure 1.7:

Likelihood approximations of the quantiles of a g-and-k distribution parameters for simulated data, obtained with an ABC-MCMC Algorithm 3 with Gaussian kernel: tolerance levels equal to 3 (solid line), 2 (dashed line) and 0.5 (dotted line).

eters ϕ, $\boldsymbol{\beta}$ is a vector of unknown parameters taking values in $\mathbb{R}^p$ and $\gamma : \mathcal{Z} \to \mathbb{R}$ is an unknown function.

If the analysis is focused on $\boldsymbol{\beta}$ or $\boldsymbol{\Omega}_\phi$, γ may be considered a nuisance parameter and a method to remove it from the analysis is needed. In particular, if a weight function for γ based on a zero-mean Gaussian stochastic process with covariance function $K_\lambda(\cdot,\cdot)$ with parameter λ is used, the vector $(\gamma(z_1), ..., \gamma(z_n))$ has a multivariate normal distribution with mean $\mathbf{0}$ and covariance $\boldsymbol{\Sigma}_\lambda$ and the integrated likelihood function of $\boldsymbol{\beta}$ is

$$|\boldsymbol{\Omega}_\phi + \boldsymbol{\Sigma}_\lambda|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(\mathbf{Y} - \mathbf{X}\beta)'(\boldsymbol{\Omega}_\phi + \boldsymbol{\Sigma}_\lambda)^{-1}(\mathbf{Y} - \mathbf{X}\beta)\right\} \tag{1.22}$$

where $\boldsymbol{\Sigma}_\lambda$ is the $n \times n$ matrix with $K_\lambda(z_i, z_j)$ in the (i,j) element. This form may be obtained because of the assumption on the normal distribution of the errors and the use of a Gaussian process weight function for γ; more general cases are not so straightforward to handle outside the normal set-up.

In He and Severini (2013) the Authors show that, for a given choice of $K_\lambda(\cdot,\cdot)$, when the dispersion parameter, say $\boldsymbol{\eta} = (\phi, \lambda)$, is known, $\boldsymbol{\beta}$ can be estimated by the generalized least-squares estimator: $\hat{\boldsymbol{\beta}} = \mathbf{X}^T\left(\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{V}^{-1}\mathbf{Y}$ where $V =$

$\boldsymbol{\Omega}_\phi + \boldsymbol{\Sigma}_\lambda$; if the dispersion parameter is unknown, β can be estimated as a function of an estimator of $\boldsymbol{\eta}$, $\hat{\boldsymbol{\beta}}(\hat{\boldsymbol{\eta}})$.

The method has been used with data from a survey of the fauna on the sea bed lying between the Queensland coast and the Great Barrier Reef; the response variable analysed is a score, on a log weight scale, which combines information across the captured species; this score value is considered dependent on the latitude $\mathbf{x}$ in a linear way and on the longitude $\mathbf{z}$ in an unknown way; see Bowman and Azzalini (1997) for more details. The model is

$$Y_j = \beta_0 + x_j\beta_1 + \gamma(z_j) + \varepsilon_j, \quad j = 1, ..., n \tag{1.23}$$

where $\varepsilon_1, ..., \varepsilon_j$ are independent normal errors with mean 0 and constant variance σ^2_ε. Using the integrated likelihood approach, a Gaussian covariance function

$$K(z, \tilde{z}) = \tau^2 \exp\left(-\frac{1}{2}\frac{|z-\tilde{z}|^2}{\alpha}\right) \tag{1.24}$$

and a restricted maximum likelihood estimate REML (Harville 1977) for the nuisance parameters, the estimates of β_1 is 1.020, with a standard error of 0.356 (see He and Severini (2013)).

We have used our ABC approximation in order to find an integrated likelihood for $\boldsymbol{\beta}$. It is then necessary to define proper prior distributions for all the parameters of the model, i.e. $\boldsymbol{\beta}$, σ^2_ε and the parameter of the covariance function of the Gaussian process, α and τ^2.

For $\boldsymbol{\beta}$ a g-prior has been chosen such that $\boldsymbol{\beta} \sim \mathrm{N}_2\left(\mathbf{0}, g\sigma^2_\varepsilon\left(\mathbf{X^T X}\right)^{-1}\right)$, where $g \sim \mathrm{U}(0, 2n)$ and $\sigma^2_\varepsilon \sim IG(a, b)$ with a, and b suitably small (as an approximation of the Jeffreys prior). A Gaussian process with squared exponential covariance function has been used as prior process for the function $\gamma(\cdot)$. The hyper-parameters of the Gaussian process have the following prior distributions: $\tau^2 \sim IG(a, b)$, with $a = b = 0.01$ and $\alpha \sim IG(2, \nu)$ with $\nu = \rho_0/(-2\log(0.05))$ and $\rho_0 = \max_{i,j=1...n}|z_i - z_j|$; see Schmidt and Gelfand (2003) and Banerjee et al. (2004) for more details.

The choice of the summary statistics is not straightforward, because it is necessary to find statistics that take into account both the parametric and the nonparametric parts of the model, nevertheless sufficiency is not guaranteed. A function of $\mathbf{z}$ has been considered and the maximum likelihood estimates of the coefficients of the new model have been used as summary statistics. In particular, two choices

of function has been considered: $g(z_j) = z_j$ and $h(z_j) = z_j^2$ for $j = 1, ..., n$. An analysis of the maximum likelihood estimates has shown that the estimate of the constant β_0 is particularly unstable, therefore only the estimates for the predictor variables' coefficients contribute to the approximation as summary statistics.

In the MCMC step, normal transitional kernels have been used for all the parameters of the model, centered at the values accepted on the previous step and with small variance.

The results are shown in Figure 1.8: the ABC approximation with 10^6 simulations are concentrated around the estimates obtained by maximizing the integrated likelihood of the model. In this case, the ABC approach may be seen as a way to properly account for the uncertainty on the nuisance parameters that is not considered when REML estimates are used. Figure 1.8 compares different choices of summary statistics and prior distributions for the variance σ_ε^2: on the left a $\mathrm{U}(0, 10)$ is used and on the right a proper approximation of the Jeffreys prior is used ($\mathrm{Ga}(a, b)$ with a, b small). All the approximations are smooth and concentrated around the maximum likelihood estimate. Moreover, Figure 1.8 shows that using the summary statistics based on a quadratic approximation of $\gamma(\cdot)$ leads to better results, because they are all smooth. On the other hand the approximations obtained by considering a linear model with respect to $\mathbf{z}$ present slight multimodality problems.

The number of simulations for the initialization step depends on the choice of the tolerance level: the approximation of the likelihood of β_1 by using a Uniform prior for σ_ε^2 needs 368, 2053 and 10945 simulations to accept the first value for tolerance levels of 1, 0.5 and 0.25 respectively; the approximation with Gamma prior with small parameters for σ_ε^2 needs 40, 34 and 81 simulations to accept the first value. These results refer to the summary statistics obtained with the quadratic approximation of $\gamma(\cdot)$, the other choice of summary statistics considered has shown similar values.

The acceptance rates of ABC-MCMC algorithm are in general low, in particular with the lowest tolerance levels; they are around 25% for the highest thresholds considered.

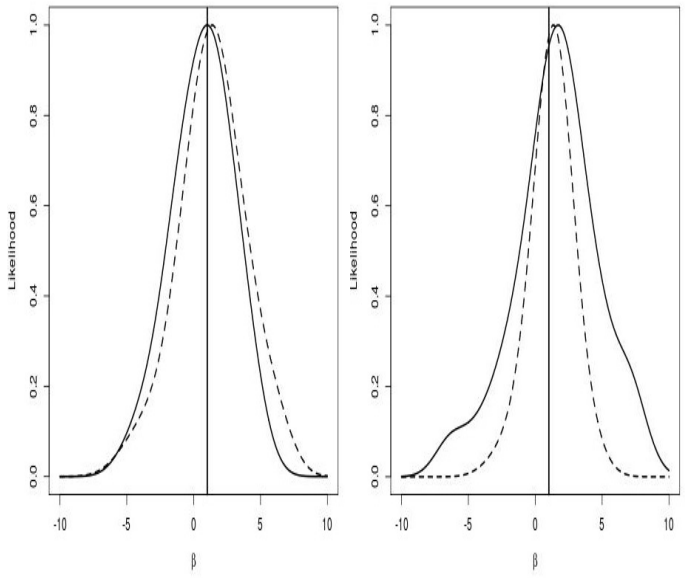

Figure 1.8:

ABC approximation of the integrated likelihood of β_1 in the semiparametric model. The approximations are obtained by using a Uniform prior (left) and an approximation of the Jeffreys prior (right) for σ_ε^2. Two different choices of summary statistics are compared: the maximum likelihood estimates of the model, with linear (solid lines) and quadratic (dashed lines) approximations of $\gamma(\cdot)$.

1.5 Discussion

We have explored the use of the ABC methodology, relatively new computational tools for Bayesian inference in complex models, in a rather classical inferential problem, namely the elimination of nuisance parameters. We stress the fact that there are many situations where it is practically impossible to obtain a likelihood function for the parameter of interest in a closed form: in those cases the proposed method can be a competitive alternative to numerical methods.

As a technical aside one should note that, in many situations the prior $\pi(\boldsymbol{\psi})$ might be available in a closed form, so the kernel approximation of the prior is not necessary, and the accuracy of our method is even better. However, we have preferred to present the method in its generality.

Another issue related to this last point is the approximation of the marginal posterior density of some components of the parameter. Also in this case, the problem is made simpler by the fact that no approximation is needed for the prior and standard asymptotic arguments for kernel estimators hold for the approximation

obtained from the posterior sample. The main drawback of the present approach is that it requires the use of proper prior densities. This can be a problem, especially when the nuisance parameter is high-dimensional and the elicitation process would be difficult. A practical solution in these case is to adopt proper priors which approximate the appropriate improper noninformative prior for that model.

Chapter 2

Approximate Bayesian Computation for Copula Estimation

2.1 Introduction

Copula[1] models are nowadays widely used in multivariate data analysis. Major areas of application include econometrics Huynh et al. (2015), geophysics Scholzel and Friederichs (2008), quantum mechanics Resconi and Licata (2015) climate prediction Schefzik et al. (2013) genetics He et al. (2012), actuarial science and finance (Cherubini et al. (2004), among the others). A copula is a flexible probabilistic tool that allows the researcher to model the joint distribution of a random vector in two separate steps: the marginal distributions and a copula function which captures the dependence structure among the vector components.

From a statistical perspective, whereas it is generally simple to produce reliable estimates of the parameters of the marginal distributions of the data, the problem of estimating the dependence structure, however it is modelled, is crucial and often complex, especially in high dimensional situations. On the other hand, dependence is one of the most fundamental features in (applied) statistics, economics and probability. A huge list of important applications can be found in the recent monograph by Joe (2014).

[1] joint wort with Prof. Brunero Liseo, MEMOTEF, Sapienza Università di Roma

In a frequentist approach to copula models, there are no broadly satisfactory methods for the joint estimation of marginal and copula parameters. The most popular method is the so called Inference From the Margin (IFM) method, where the parameters of the marginal distributions are estimated first, and then pseudo data are obtained by plugging in the estimates of the marginal parameters. Then inference on the copula parameters is performed using the pseudo-data: this approach obviously does not account for the uncertainty on the estimation of the marginal parameters. Bayesian alternative are not yet fully developed, although Min and Czado (2010), Craiu and Sabeti (2012), Smith (2013) and Wu et al. (2014) are remarkable exceptions.

In this work we consider the general problem of estimating some specific quantities of interest of a generic copula (such as, for example, tail dependence index or Spearman's ρ) by adopting an approximate Bayesian approach along the lines of Mengersen et al. (2013). In particular, we discuss the use of the BC_{el} algorithm, based on the empirical likelihood approximation of the marginal likelihood of the quantity of interest. Our approach is approximate in two aspects:

i. elicitation of the prior distribution is required only on the quantity of interest. Its prior distribution is combined with the empirical likelihood in order to produce an approximation to the "true" posterior distribution.

ii. we do not use the "true" likelihood function, but rather an approximation based on empirical likelihood theory Owen (2010). Hopefully, this will reduce the potential bias for incorrect distributional assumptions.

Note, however, that the word "true" in the above list should be better spelled as "true-under-the-assumed-model". In situations where a true model is too hard to specify, or too complex to deal with, the empirical likelihood can be an extremely valuable tool.

Our approach can be adapted both to parametric and nonparametric modelling of the marginal distributions. The method described in this paper is in the spirit of Hoff (2007), but it is based on a different kind of approximation; the results, although from a different perspective, can be also interpreted in the light of Schennach (2005), where a Bayesian nonparametric interpretation of a tilted version of the empirical likelihood is provided.

2.2 Preliminaries: Copulae and Empirical Likelihood

A copula model is a way of representing the joint distribution of a random vector $\mathbf{X} = (X_1, \ldots, X_m)$. Given an m-variate cumulative distribution function (CDF) $\mathbf{F}$, it is possible to show Sklar (1959) that there always exists an m-variate function $C : [0,1]^m \to [0,1]$, such that $\mathbf{F}(x_1, \ldots, x_m) = C(F_1(x_1), \ldots, F_m(x_m))$, where F_j is the marginal CDF of X_j. In other terms, the copula function C is a CDF with uniform margins on $[0,1]$: it binds together the univariate CDF's $F_1, F_2, \ldots, F_m$ in order to produce the m-variate CDF $\mathbf{F}$. The copula function C does not depend on the marginal distributions of $\mathbf{F}$, but rather it accounts for potential dependence among the components of the random vector $\mathbf{X}$.

For each pair of components of $\mathbf{X}$, say X_i and X_j, let us assume that they have continuous CDF's F_i and F_j. It is well known that both the transformed variables $U_i = F_i(X_i)$ and $U_j = F_j(X_j)$ have uniform marginal distributions. A semiparametric copula model consists of a parametric model for the joint distribution of (U_i, U_j) and no assumptions on the marginal CDF's. A nonparametric copula is assumed when the joint distribution of (U_i, U_j) depends on an infinite dimensional parameter. In this paper we will allow the marginal distributions F_j's to follow either a parametric or a non parametric model. For the copula function we will not make any parametric assumption. Rather, we will limit our goal to the estimation of a particular function of interest of the copula C. A discussion on the classical approaches to semiparametric estimation of copula models can be found in Genest et al. (1995).

Empirical likelihood has been introduced by Owen: Owen (2010) is a complete and recent survey; it is a way of producing a nonparametric likelihood for a quantity of interest in an otherwise unspecified statistical model. It is particularly useful when a true likelihood is not readily available either because it is too expensive to evaluate or when the model is not completely specified. Assume that our dataset is composed of n independent replicates $(x_1, \ldots, x_m)$ of some random vector $\mathbf{X}$ with distribution $\mathbf{F}$ and corresponding density f. Rather than defining the usual likelihood function in terms of f, the empirical likelihood is constructed with respect to a given quantity of interest, say φ, expressed as a functional of $\mathbf{F}$, i.e. $\varphi(\mathbf{F})$, and then a sort of profile likelihood of φ is computed in a nonparametric way. More precisely, consider a given

set of generalized moment conditions of the form

$$E_{\mathbf{F}}\left(h(X,\varphi)\right) = 0, \tag{2.1}$$

where $h(\cdot)$ is a known function, and φ is the quantity of interest. The resulting empirical likelihood is defined as

$$L_{EL}(\varphi;\mathbf{x}) = max_{\mathbf{p}} \prod_{i=1}^{n} p_i,$$

where the maximum is searched over the set of vectors $\mathbf{p}$ such that $0 \leq p_i \leq 1$, $\sum_{i=1}^{n} p_i = 1$, and

$$\sum_{i=1}^{n} h(\mathbf{x}_i,\varphi)p_i = 0.$$

Whereas the first two conditions are obvious and independent of φ, the third one induces a profiling of the information towards the quantity of interest, through a sort of unbiasedness condition. This representation of the empirical likelihood has no probabilistic interpretation so far; however, a modification exists in the literature (Schennach 2005) which naturally arises from a nonparametric Bayesian procedure which places a noninformative prior on the space of distributions. The resulting Bayesian exponentially tilted empirical likelihood is defined as

$$L_{BEL}(\varphi;\mathbf{x}) = \max_{(p_1,\ldots,p_n)} \sum_{i=1}^{n} \left(-p_i \log p_i\right),$$

under the constraints $0 \leq p_i \leq 1$, $\sum_{i=1}^{n} p_i = 1$, and $\sum_{i=1}^{n} h(\mathbf{x}_i,\varphi)p_i = 0$.

2.3 ABC and EL

Approximate Bayesian computation has now become an essential tool for the analysis of complex stochastic models, in the case where the likelihood function is unavailable in closed form or it is too expensive to be repeatedly evaluated Marin et al. (2012). It can be considered as a class of popular algorithms that achieves posterior simulation by avoiding the computation of the likelihood function. A crucial condition for the use of ABC algorithms is that it must be relatively easy to generate new pseudo-observations from the working model, for a fixed value of the parameter vector. In its simplest form, the ABC algorithm "proposes" a (pseudo)-randomly drawn parameter value θ^* from the prior distribution and a new data set is generated, conditionally on θ^*; then the value is accepted only if the new data are "similar

enough" to the actual observed data. It can be proved that the set of accepted values represents a sample from an approximation of the posterior distribution of θ Sisson and Fan (2011). However, it is often highly inefficient to propose values from the prior distribution, since it is generally much more diffuse than the posterior distribution. Many more sophisticated computational strategies are available in order to avoid generating values from the prior distribution, see Marin et al. (2012) for example; here we will not discuss these issues and we rather concentrate on a different ABC approach, which can avoid the most expensive step in computational time, that is the proposal of new data sets. This method has been proposed by Mengersen et al. (2013) and it represents a re-sampling scheme where the proposed values are re-sampled with weights proportional to their empirical likelihood. In practice, the algorithm belongs to the family of "sampling importance re-sampling" - SIR, Rubin (1988) - methods for models in which the "true likelihood" evaluation is out of reach and the "true" weights are approximated by their empirical likelihood.

Algorithm 4 BC_{EL} algorithm Mengersen et al. (2013)

for $i = 1$ **to** M **do**
repeat
Generate θ_i from the prior distribution $\pi(\theta)$
Set the weight for θ_i as $\omega_i = L_{EL}(\theta_i; \text{data})$.
end for
for $i = 1$ **to** M **do**
Draw, with replacement, a value θ_i from the previous set of M values using weights ω_i, $i = 1, \ldots, M$.
end for

2.4 The proposed approach

In this Chapter we propose to adapt the BC_{EL} algorithm of Mengersen et al. (2013) to a situation where the statistical model is only partially specified and the main goal is the estimation of a finite dimensional quantity of interest. In practice this represents the prototypical semiparametric set-up, where one is mainly interested

in some meaningful characteristic of the population, although the statistical model may contain nuisance parameters which are often introduced in order to produce more flexible models that might better fit the data at hand. In order to make robust inference on the quantity of interest, a reasonable model should account for the uncertainty on the nuisance parameters, in some way. Even if some of these additional parameters are not particularly important in terms of estimation - they often lack of a precise physical meaning - their estimates can dramatically affect inferences on the parameter of interest. In these circumstances it might be more reasonable and *robust* to partially specify the model and adopt a semiparametric approach.

Oh and Patton (2013) consider, in a frequentist perspective, a Simulated Method of Moments estimation for copula models. Their paper is very close in spirit to what we are proposing, although their main goal is the analysis of partially specified models rather than models with an intractable likelihood.

2.4.1 The algorithm in full detail

We assume that a data set is available in the form of a size $n \times m$ matrix $\mathbf{X}$, where n is the sample size and m is the number of variables, that is

$$\mathbf{X} = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1m} \\ x_{21} & x_{22} & \dots & x_{2m} \\ \dots & \dots & x_{ij} & \dots \\ x_{n1} & x_{n2} & \dots & x_{nm} \end{pmatrix}.$$

In the following, $X_{[\cdot,j]}$ will denote the j-th column (variable) and $X_{[i,\cdot]}$ the i-th row of $\mathbf{X}$, respectively. For each $j = 1, \dots, m$, we consider the available data information in $X_{[\cdot,j]}$ to produce an estimate of the marginal CDF of $X_{[\cdot,j]}$. Let $\boldsymbol{\lambda}_j = (\lambda_j^{(1)}, \lambda_j^{(2)}, \dots \lambda_j^{(S)})'$, $j = 1, 2, \dots m$ be the posterior sample obtained from *some* Bayesian inference method for the distribution of $X_{[\cdot,j]}$. Notice that the vector $\boldsymbol{\lambda}_j$ can be either a sample from the posterior distribution of the parameters of the model we have adopted for $X_{[\cdot,j]}$ or a posterior sample of CDF's in a nonparametric set-up. Then we use a copula representation for estimating the multivariate dependence structure of the random vector $\mathbf{X}$,

$$\mathbf{F}(x_1, \dots, x_m) = C_\theta\big(F_1(x_1), F_2(x_2), \dots, F_m(x_m)\big),$$

where θ is the parameter related to the copula function. Since we are assuming that one has already *estimated* the marginal $F_j(x_j)$'s, $j = 1, \ldots, m$, one now needs to consider the copula $C_\theta(\cdot)$ only. This step can be managed either using some parametric model for the copula (such as Clayton, Gaussian, Skew-t, Gumbel, etc.) or using a nonparametric approach.

Parametric copulae in Bayesian inference have been already investigated in several papers. Here we shoud mention Hoff (2007), Silva and Lopes (2008), Min and Czado (2010), Smith et al. (2012) and Craiu and Sabeti (2012). In this paper, we take a nonparametric route and we concentrate on some specific function of $C_\theta(\cdot)$, say $\varphi = T(C_\theta)$. This is particularly useful and meaningful in those situations where there is no theoretical or empirical evidence that a given copula should be preferred and we are mainly interested in some specific synthetic measure of the multivariate dependence, like for example, the upper tail dependence index between two components of $\mathbf{X}$, that is

$$\chi = \lim_{u \to 1} P(U_j > u | U_h > u) \approx \lim_{u \to 1} \left[2 - \frac{\log P(U_j < u, U_h < u)}{\log P(U_h < u)} \right]$$

where $U_i = F_i(x_i)$, $i = j, h$. Another popular quantity, which we will consider in the final section is the Spearman's measure of association ρ between two components of $\mathbf{X}$, say X_h and X_j, which is defined as the correlation coefficient among the transformed values $U_i = F_i(x_i)$, $i = j, h$ or, in a copula language, as

$$\begin{aligned} \rho &= 12 \int_0^1 \int_0^1 \big(C(u_j, u_h) - u_h u_j\big) du_j du_h \\ &= 12 \int_0^1 \int_0^1 C(u_j, u_h) du_j du_h - 3. \end{aligned} \tag{2.2}$$

We now describe the algorithm in a pseudo-language:

The final output of the above algorithm is then a posterior sample drawn from an approximation of the posterior distribution of the quantity of interest φ. There are several critical issues both in the practical implementation of the method and in its theoretical properties. First, the empirical likelihood is based on moment conditions of the form (2.1). In practical applications these conditions might hold only asymptotically. This is the case, for example, of the Spearman's ρ, which we discuss in the next session. Its sample counterpart ρ_n is only an asymptotically unbiased estimator of ρ so the moment condition is strictly valid only for large samples. Also, prior information is only provided for the marginal distributions and for φ: this, of course, has advantages and, on the other hand, poses theoretical issues.

Algorithm 5 ABCOP algorithm

[1:] For $s = 1, \ldots, S$, use the s-th row of the posterior simulation $\lambda_1^{(s)}, \lambda_2^{(s)}, \ldots, \lambda_m^{(s)}$ to create a matrix of uniformly distributed *pseudo*-data

$$\mathbf{m}u^{(s)} = \begin{pmatrix} u_{11}^{(s)} & u_{12}^{(s)} & \ldots & u_{1m}^{(s)} \\ u_{21}^{(s)} & u_{22}^{(s)} & \ldots & u_{2m}^{(s)} \\ \ldots & \ldots & u_{ij}^{(s)} & \ldots \\ u_{n1}^{(s)} & u_{n2}^{(s)} & \ldots & u_{nm}^{(s)} \end{pmatrix}$$

with $u_{ij}^{(s)} = F_j\big(x_{ij}; \lambda_j^{(s)}\big)$.
[2:] Given a prior distribution $\pi(\varphi)$ for the quantity of interest φ,
for $b = 1, \ldots, B$,

1. draw $\varphi^{(b)} \sim \pi(\varphi)$;
2. compute $EL\left(\varphi^{(b)}; \mathbf{m}u^{(s)}\right) = \omega_{bs}; s = 1, \ldots, S$.
3. take the average weight $\boldsymbol{\omega}_b = S^{-1} \sum_{s=1}^{S} \omega_{bs}$

end for
[3:] re-sample - with replacement - from $\{\big(\varphi^{(b)}, \boldsymbol{\omega}_b\big), b = 1, \ldots, B\}$.

The main advantage is the ease of elicitation: one need not to elicit unnecessary aspects of the prior distribution. This is mainly in the spirit of the partially specified models, quite popular in the econometric literature. Another obvious advantage of the proposed approach is the implied robustness of the method, with respect to different prior opinions about non-essential aspects of the dependence structure. The most important disadvantage of the method is its inefficiency when compared to a parametric copula, under the assumption that the parametric copula is the true model. The practical implementation of the algorithm is quite simple in **R**; it use some functions contained in the suite `gmm`: see for example Chauss (2010).

From a computational perspective the above algorithm is quite demanding, since one needs to run a BC_{EL} algorithm *for each* row of the posterior sample from the marginals. Even though the estimation of the marginal densities of the $X_{[\cdot,j]}$'s might not require a huge values of iterations S, still it might be very expensive to

run S different BC_{EL} algorithms. To avoid this computational burden, we propose to modify the above algorithm by simply performing a single run of the BC_{EL} algorithm, where, for each iteration $b = 1, \ldots, B$, a randomly selected (among the S rows) row $\boldsymbol{\lambda}^s$ is used to transform the actual data into pseudo-data lying in $[0,1]^m$. With this modification the above algorithm gets transformed into Algorithm 6.

Algorithm 6 Modified ABCOP algorithm

[1:] For $j = 1, \ldots, m$, produce a posterior sample for the parameters of the marginal distributions of the $X_{[\cdot,j]}$'s, say $\boldsymbol{\lambda}_j = \lambda_j^{(1)}, \lambda_j^{(2)}, \ldots, \lambda_j^{(S)},\ j = 1, \ldots, m$. Store them into a $S \times k$ matrix $\boldsymbol{\lambda} = (\boldsymbol{\lambda}_1, \ldots, \boldsymbol{\lambda}_j, \ldots, \boldsymbol{\lambda}_m)$ where k is the sum of the dimensions of the parameter spaces of the marginal distributions.
[2:] Given a prior distribution $\pi(\varphi)$ for the quantity of interest φ,
for $b = 1, \ldots, B$,

1. draw a random uniform integer $t(b)$ in $\{1, 2, \ldots, S\}$.
2. use the $t(b)$-th row of $\boldsymbol{\lambda}$ to create a matrix of uniformly distributed *pseudo*-data
$$\mathbf{m}u^{(t(b))} = \begin{pmatrix} u_{11}^{(t(b))} & u_{12}^{(t(b))} & \ldots & u_{1m}^{(t(b))} \\ u_{21}^{(t(b))} & u_{22}^{(t(b))} & \ldots & u_{2m}^{(t(b))} \\ \ldots & \ldots & u_{ij}^{(t(b))} & \ldots \\ u_{n1}^{(t(b))} & u_{n2}^{(t(b))} & \ldots & u_{nm}^{(t(b))} \end{pmatrix}$$
with $u_{ij}^{(t(b))} = F_j\big(x_{ij}; \lambda_j^{(t(b))}\big)$.
3. draw $\varphi^{(b)} \sim \pi(\varphi)$;
4. compute
$$EL\left(\varphi^{(b)}; \mathbf{m}u^{(t(b))}\right) = \omega_b;$$

end for
[3.] store the values $\big(\varphi^{(b)}, \omega_b\big), \quad b = 1, \ldots, B$.
[4.] re-sample - with replacement - from $\{\big(\varphi^{(b)}, \omega_b\big), b = 1, \ldots, B\}$.

2.5 Asymptotics

A possible criticism about the proposed method is that the inferential step has been split into two parts: first, the marginal distributions of the multivariate random variable are estimated; then, pseudo-data are created in order to provide a semi-parametric estimate of the dependence quantity of interest. The "two-step" issue is at the core of the often unsatisfactory behaviour of estimation procedures based on the *Inference from the Margins* methods proposed in Joe (2014) (section 10.1): the main drawback of that approach is that it fails to properly account for the uncertainty on the estimates of the parameters of the marginal distributions. However this problem is much less serious in our setting; in fact, we produce, for each single coordinate of the multivariate distribution, a sample from the joint posterior distribution of the parameters which appear in that marginal distribution. So the actual degree of information on those parameters is completely transferred to the *second step* of the procedure, which creates, for each run of the posterior simulation, a different set of pseudo-data and then *takes averages* on them. Provided that the estimation procedure for the marginals is consistent, we are consistently creating "pseudo-data".

These arguments can be made more precise in the following way. The infinite dimensional parameter space for a copula model is $(C, F_1, \cdots, F_d)$. The parameter of interest is a function $\varphi = T(C)$. Then the copula is defined as $C = (\varphi, C^*)$, with $\varphi \in R^k$ for some $k \geq 1$ and $C^* \in H$ where (H, d_H) is an infinite-dimensional metric space.

We can consider a Bayesian nonparametric approach for the estimation of $(F_1, \cdots, F_d)$ which is asymptotically based on the marginal empirical CDF's; for example, one can use a Dirichlet process mixture. Theorems 5 and 10 in Fermanian et al. (2004) guarantee that an inferential procedure based on the empirical copula C_n is such that $(C_n - C)$ is weakly convergent to a Gaussian process in $\ell^\infty[0,1]$.

Then, the use of the exponentially tilted empirical likelihood for the pseudo-data is justified by the results in Schennach (2005), where, in addition, a computationally convenient reformulation of the problem in semiparametrric Bayesian terms is provided.

However, for finite sample sizes, there may still be a problem: the objects created by the estimated cumulative distributions function (the ones which produce

the pseudo-data) may have a dependence structure which can be potentially very different from the *true* cumulative distribution function transforms. In other terms, if we are using a wrong model on the marginals, the entire posterior sample we use may be misleading and the subsequent step might be biased. This problem is of course common to any parametric statistical procedure and we strongly suggest, in absence of specific information on the marginals, to adopt a nonparametric approach for their estimation.

2.6 A simple illustration: Spearman's ρ

We first illustrate the method in a simple situation, wehn $m = 2$, and assuming that the two marginal distributions of the data are known: without loss of generality we can then assume that they are both uniform in $[0, 1]$; in this case there are no practical differences between Algorithm 5 and Algorithm 6.

The Spearman's ρ measure of dependence has been defined in (2.2). Starting from a sample of size n from a bivariate distribution, say (x_i, y_i), $i = 1, \ldots, n$, the sampling counterpart of ρ, say ρ_n, is the correlation among ranks and it can be written as

$$\rho_n = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{12}{n^2 - 1} R_i Q_i - 3 \frac{n+1}{n-1} \right), \tag{2.3}$$

where

$$R_i = \text{rank}(x_i) = \sum_{k=1}^{n} I(x_k \leq x_i), \quad Q_i = \text{rank}(y_i) = \sum_{k=1}^{n} I(y_k \leq y_i), \quad i = 1, \ldots, n.$$

We consider the ranks (R_i, Q_i) of the original values and compute ρ_n. Also we are able to evaluate the empirical likelihood of ρ for a given value of ρ_n as

$$\max_{p_i} EL(\rho; \rho_n) = \prod_{i=1}^{n} n p_i(\rho)$$

under the constraints $\sum_{i=1}^{n} p_i = 1$, $0 \leq p_i \leq 1$, $i = 1, \ldots, n$ and

$$\sum_{i=1}^{n} p_i \left(\frac{12 R_i Q_i}{n^2 - 1} - 3 \frac{n+1}{n-1} - \rho \right) = 0.$$

From general results on empirical likelihood Owen (2010), one has

$$EL(\rho; \rho_n) = \prod_{i=1}^{n} \left(1 + \eta g(R_i, Q_i; \rho) \right)^{-1}$$

where η is the Lagrange multiplier which can be explicitly obtained from

$$\sum_{i=1}^{n} \frac{g(R_i, Q_i; \rho)}{1 + \eta g(R_i, S_i; \rho)} = 0,$$

where

$$g(R_i, S_i; \rho) = \frac{12 R_i Q_i}{n^2 - 1} - 3\frac{n+1}{n-1} - \rho.$$

We can then use Algorithm 5, with $S = 1$, to produce a posterior sample for the quantity of interest ρ.

The frequentist properties of the estimator (2.3) have been considered in Borkowf (2002), where the Authors show taht the asymptotic variance of ρ_n is

$$\sigma^2(\rho_n) = 144(-9\theta_1^2 + \theta_2 + 2 * \theta_3 + 2 * \theta_4 + 2 * \theta_5), \tag{2.4}$$

where the θ_i's are term linked with the moments of the marginal and joint distributions of (X_1, Y_1) and (X_2, Y_2) i.i.d random variables with distribution $F(x, y)$. In particular

$$\begin{aligned}
\theta_1 &= \mathbb{E}[F_1(X_1)F_2(Y_1)] \\
\theta_2 &= \mathbb{E}[(1 - F_1(X_1))^2(1 - F_2(Y_1))^2] \\
\theta_3 &= \mathbb{E}[(1 - F(X_1, Y_2))(1 - F(X_2))(1 - F(Y_1))] \\
\theta_4 &= \mathbb{E}[(1 - F_1(\max\{X_1, X_2\}))(1 - F_2(Y_1))(1 - F_2(Y_2))] \\
\theta_5 &= \mathbb{E}[(1 - F_1(X_1))(1 - F_1(X_2))(1 - F_2(\max\{Y_1, Y_2\}))].
\end{aligned}$$

Natural estimates of these quantities are given by Genest and Favre (2007), where

$$\theta_{1n} = \frac{1}{n}\sum_{i=1}^{n}\frac{R_i}{n+1}\frac{S_i}{n+1}$$
$$\theta_{2n} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{R_i}{n+1}\right)^2\left(\frac{S_i}{n+1}\right)^2$$
$$\theta_{3n} = \frac{1}{n^3}\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{n}\frac{R_i}{n+1}\frac{S_i}{n+1}\mathbb{I}\{R_k \leq R_i, S_k \leq S_j\} + \frac{1}{4} - \theta_{1n}$$
$$\theta_{4n} = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\frac{S_i}{n+1}\frac{S_j}{n+1}\max\left\{\frac{R_i}{n+1}, \frac{R_j}{n+1}\right\}$$
$$\theta_{5n} = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\frac{R_i}{n+1}\frac{R_j}{n+1}\max\left\{\frac{S_i}{n+1}, \frac{S_j}{n+1}\right\}$$

However, it is important to notice that, in the case of perfect rank agreement, by plugging the above θ_{jn}, $j = 1, \cdots, 5$ into expression (2.4), one obtains a negative number. This phenomenon also appeared in our simulations when data were generated from values of ρ close to 1.

2.6.1 Simulated non uniform data

Here we show an example of bivariate data with non uniform marginal distribution. Data were generated from a Clayton copula with $\theta = 1.076$, and the two marginal distributions were transformed into an exponential distribution with mean 1/3 (for X_1) and a Gaussian distribution with mean 3 and variance 1 (for X_2). Figure 2.1 shows the scatterplot of raw and transformed data. In this particular case the observed value for ρ_n was 0.568.

Figure 2.2 shows the histogram of the BC_{EL} posterior sample for ρ obtained from Algorithm 5. One can notice that the posterior mass is practically entirely on the right of zero, and the posterior mean is 0.56, very close to the observed ρ_n.

2.6.2 An alternative estimator

From a purely pragmatic perspective, it might be tempting to follow an unconventional and "hybrid" route, which we now describe. For each $s = 1, \ldots, S$,

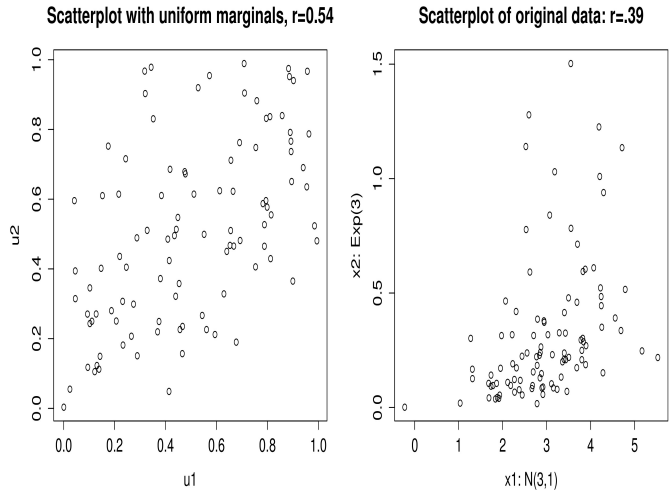

Figure 2.1: Scatterplot of the simulated data and pseudo-data: $X_1 \sim \text{Exp}(3)$; $X_2 \sim N(3, 1)$

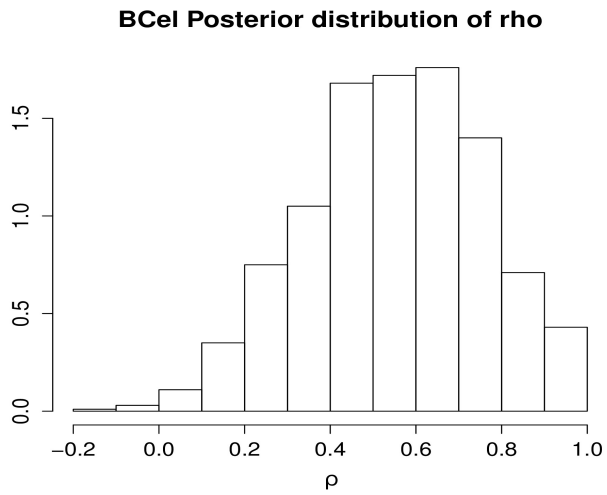

Figure 2.2: Histogram of the posterior sample of values of ρ, using Algorithm 5.

1. Provide an estimate of φ, using $\hat{\varphi}^{(s)}$ from the *plugged-in* model

$$p\big(\mathbf{x};\ \text{marginals}, \varphi, \lambda_1^{(s)}, \ldots, \lambda_m^{(s)}\big) \tag{2.5}$$

   In particular, one could use a sort of maximum likelihood estimate of φ, assuming that the sampling distribution is given by (2.5).

2. Use the distribution of the $\hat{\varphi}^{(s)}$'s as a surrogate of the posterior distribution of φ.

This approach is a further approximation in many ways. First, the distribution of $\hat{\varphi}^{(s)}$'s in step 2 of the above procedure could not properly be treated as a posterior distribution, since we have not introduced any prior distribution on φ. Second, the distribution in step 2 is not a distribution on φ: rather, it can be interpreted as the posterior distribution of the following quantity

$$\hat{\varphi}(\Lambda) = \text{argmax}_{\varphi} p\left(\mathbf{x}|\ \text{marginals}, \varphi, \Lambda\right). \tag{2.6}$$

Notice that

$$\begin{aligned} \mathrm{E}^{\Lambda}\left(\hat{\varphi}(\Lambda)\right) &\neq \text{argmax}_{\varphi}\, \mathrm{E}^{\Lambda}\left(p(\mathbf{x}|\ \text{marginals}, \varphi, \Lambda\right) \\ &= \text{argmax}_{\varphi} IL(\varphi; \mathbf{x}) = \hat{\varphi}^{(IL)}, \end{aligned}$$

where the above expectation is taken with respect to the posterior distribution of the marginal parameters Λ, based on the "marginal" samples and suitable prior information, and IL represents the "correct" integrated likelihood,

$$IL(\varphi; \mathbf{x}) = \int_{\Lambda} p(\mathbf{x}; \boldsymbol{\lambda}, \varphi)\pi(\boldsymbol{\lambda}|\varphi)d\boldsymbol{\lambda}.$$

Also, $\mathrm{Var}\left(\hat{\varphi}(\Lambda)\right)$ under-reports the variability of the estimator, since

$$MSE = \mathrm{Var}\left(\hat{\varphi}(\Lambda)\right) + \left(\mathrm{E}^{\Lambda}\left(\hat{\varphi}(\Lambda)\right) - \hat{\varphi}^{(IL)}\right)^2.$$

However, in practical applications this method works better than the IFM approach, described in Section 2.1. Figure 2.3 shows the behavior of this method with the data used in Figure 2. One can notice a slight bias towards larger values of ρ and an incorrect report of uncertainty.

2.6.3 A small scale simulation

As an illustration we have simulated 500 samples of size $n = 1000$ from some bivariate copulas, in particular from a Clayton copula (with $\rho = 0.50$), a Frank copula

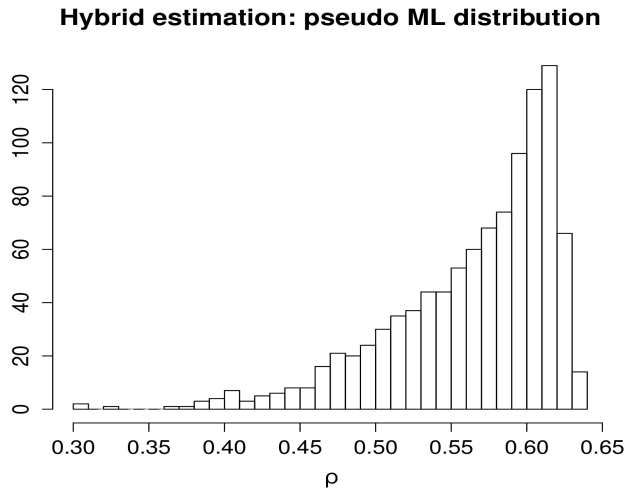

Figure 2.3: Hybrid method: "posterior" distribution of $\hat{\varphi}(\lambda)$

(with true $\rho = 0.50$), a Gumbel copula (with true $\rho = 0.688$) and a Gaussian copula with normal marginals (with true $\rho = 0.80$).

For comparative purposes we have also implemented the nonparametric frequentist procedure described in Genest and Favre (2007), where a confidence interval for the Spearman's ρ is constructed based on the asymptotic sampling distribution of ρ_n. Figure 2.4 and 2.5 compare the frequentist behavior of confidence procedure and our proposal, for different choices of copula density and level of dependence.

The Figures show the sampling (over the 500 generated samples) distribution of i) the lower limit of the equal-tail confidence interval with nominal coverage set at 0.95, ii) the point estimate ρ_n, and iii) the upper limit of the equal-tail confidence interval with nominal coverage set at 0.95 and the sampling distribution of some specific quantiles (namely the 2.5th, the median and the 97.5th percentiles) of the approximated posterior distribution. The prior distribution for ρ has been taken uniform in $(-1, 1)$. Computations were done in **R**, using libraries `copula` and `gmm`.

One can see that our procedure produces more precise estimates in terms of intervals. The empirical estimate and the posterior median behave very similarly in all the cases. The average length of the confidence interval and of the Bayesian intervals are shown in Table 2.1: the average length of the frequentist intervals is larger than the length of the corresponding Bayesian intervals when the frequentist procedure is valid, i.e. it has the expected coverage, in particular for simulations from the Clayton and the Frank copulas with $\rho = 0.5$. On the other hand, when

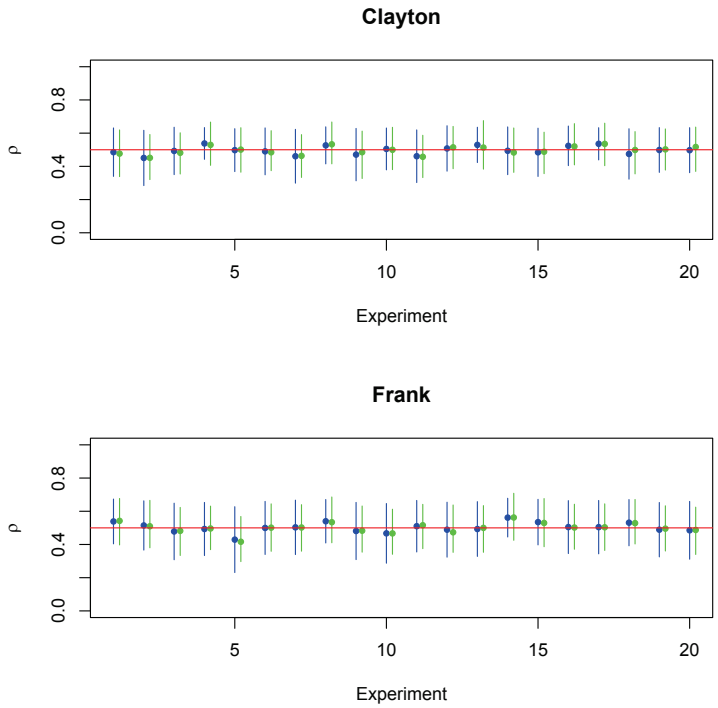

Figure 2.4: 20 out of 500 simulations from a Clayton copula and Frank copula: sample size is $1,000$; the true value of ρ is equal to 0.5 in both cases (red line), comparison between frequentist approach (blue line) and Bayesian approach (green line).

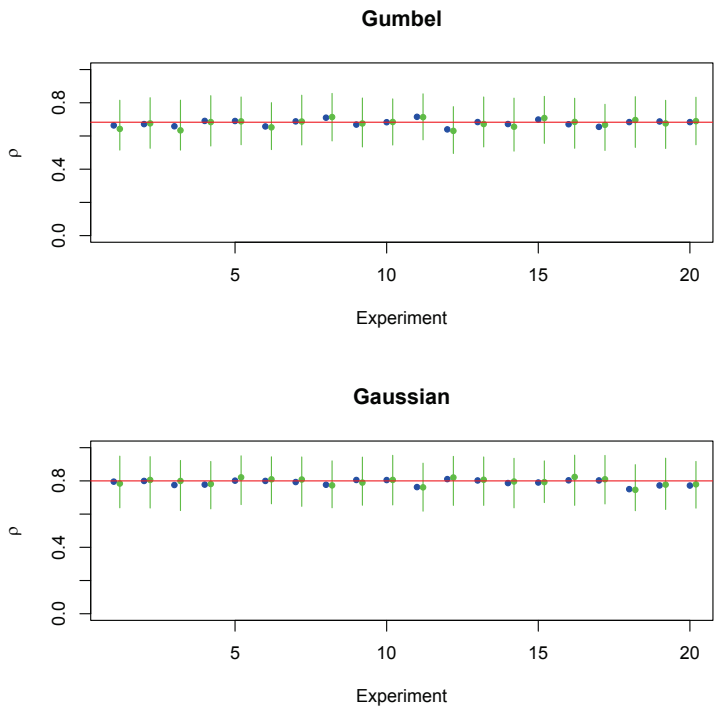

Figure 2.5: 20 out of 500 simulations from a Gumbel copula with $\rho = 0.683$ and a Gaussian copula with $\rho = 0.8$: sample size is $1,000$, comparison between frequentist approach (blue line) and Bayesian approach (green line). The frequestist estimates of the variance are, in these cases, negative, then the frequentist intervals can not be computed.

Table 2.1: Simulations from different copulas: average length and empirical coverage based of the intervals obtained both via frequentist and Bayesian methods, based on 500 repetitions of the experiment

		Ave. Length	**Coverage**
Clayton ($\rho = 0.5$)	*Freq.*	0.2664	0.998
	Bayes.	0.2597	1.000
Frank ($\rho = 0.5$)	*Freq.*	0.3172	1.000
	Bayes.	0.2735	1.000
Gumbel ($\rho = 0.68$)	*Freq.*	-	-
	Bayes.	0.2966	1.000
Gaussian ($\rho = 0.8$)	*Freq.*	-	-
	Bayes.	0.2931	1.000

the true Sperman's ρ increases, the estimate of the variance tends to be negative (98.4% of the experiments for the Gumbel copula with $\rho = 0.68$ and 100% of the experiments for the Gaussian copula with $\rho = 0.80$), then the estimate are not reliable, while the ABC procedure to approximate the Bayesian intervals does not show a different behavior in these situations, the intervals are just slightly larger.

The proportion of frequentist intervals with larger length than the correponding Bayesian intervals is 0.564 for the Clayton copula (with $\rho = 0.5$) and 0.892 for the Frank copula (with $\rho = 0.5$), while the coverage in the other two cases cannot be evaluated because of the negative estimate of the variance.

2.7 Multivariate Analysis

The extension of the proposed procedure to the multivariate case is straightforward, and no further theoretical issues arise. On the other hand, a broadly satisfactory solution in the frequentist approach has not yet been fully developed.

Formula (2.2) provides one of the possible way to express the Spearman's ρ and it suggests its interpretation as a measure of expected distance between the actual copula and the independence copula $\Pi(u, v) = uv$. From this perspective, its extension to the general d-dimensional setting is straightforward: the multivariate ρ becomes

$$\begin{aligned}\rho_1 &= \frac{\int_{[0,1]^d} C(\mathbf{u})d\mathbf{u} - \int_{[0,1]^d} \Pi(\mathbf{u})d\mathbf{u}}{\int_{[0,1]^d} M(\mathbf{u})d\mathbf{u} - \int_{[0,1]^d} \Pi(\mathbf{u})d\mathbf{u}} \\ &= \frac{d+1}{2^d-(d+1)}\left\{2^d \int_{[0,1]^d} C(\mathbf{u})d\mathbf{u} - 1\right\}, \end{aligned} \tag{2.7}$$

where $M(\mathbf{m}u) = \min(u_1, u_2, \ldots, u_d)$ is the upper Fréchet- Hoeffding bound.

Other definitions of the Spearman's ρ exist in the literature (see Schmid and Schmidt (2007)), for instance:

$$\rho_2 = \frac{d+1}{2^d-(d+1)}\left\{2^d \int_{[0,1]^d} \Pi(\mathbf{u})dC(\mathbf{u}) - 1\right\} \tag{2.8}$$

where $\Pi(\mathbf{u}) = \prod_{j=1}^d u_j$ is the independent copula. Finally, a third generalization of ρ is possible as the arithmetic mean of the bivariate ρ's. In particular, one has

$$\rho_3 = 12\binom{d}{2}^{-1} \sum_{k<l} \int_{[0,1]^2} C_{kl}(u,v)dudv - 3, \tag{2.9}$$

where $C_{kl}(u,v)$ is the bivariate marginal copula of C corresponding to the k-th and l-th marginals. This expression appears as ν in Joe (1990); its rationale is different from the one of (2.7) and (2.8), and we will no longer consider it.

If $d = 2$, then $\rho_1 = \rho_2$, but this relation need not to hold in the general $d > 2$ case.

The natural multivariate extension of the empirical copula can be expressed as

$$\widehat{C}_n(\mathbf{u}) = \frac{1}{n}\sum_{j=1}^n \prod_{i=1}^d \mathbb{I}_{\{\hat{U}_{ijn} \le u_i\}}, \qquad \mathbf{u} = (u_1, u_2, \cdots, u_n) \in [0,1]^d$$

where $\hat{U}_{ijn} = \hat{F}_i(X_{ij})$ for $i = 1, \ldots, d$ and $j = 1, \ldots, n$ and $\hat{F}_i(\cdot)$ is the empirical marginal distribution function for the i-th component. Consequently, non-parametric estimators of the multivariate ρ_i for $i = 1, 2, 3$ are

$$\hat{\rho}_{1n} = h(d)\left\{2^d \int_{[0,1]^d} \hat{C}_n(\mathbf{u})d\mathbf{u} - 1\right\} = h(d)\left\{\frac{2^d}{n}\sum_{j=1}^n \prod_{i=1}^d (1-\hat{U}_{ijn}) - 1\right\}$$

$$\hat{\rho_2}_n = h(d)\left\{2^d \int_{[0,1]^d} \Pi(\mathbf{u})d\hat{C}_n(\mathbf{u}) - 1\right\} = h(d)\left\{\frac{2^d}{n}\sum_{j=1}^n \prod_{i=1}^d \hat{U}_{ij,n} - 1\right\}$$

where $h(d) = (d+1)/(2^d - d - 1)$.

Asymptotic properties of these estimators are explored and assessed in Schmid and Schmidt (2007). In particular it is known that

$$\sqrt{n}(\rho_{i,n} - \rho_i) \dot{\sim} \mathcal{N}(0, \sigma_i^2), \qquad i = 1, 2.$$

The expressions for σ_i^2, $i = 1, 2$ are given in Schmid and Schmidt (2007). The variances of the above estimators can be analytically computed only in very few cases. In general they will depend on unknown quantities which must be estimated, for example via bootstrap methods Schmid and Schmidt (2006).

Bootstrap estimators of ρ_1 and ρ_2 have been proved to be consistent: on the other hand the bootstrap estimators of the variances tend to dramatically underestimate the variability of $\hat{\rho}_{in}$, $i = 1, 2$. We have performed several simulation experiments and our results always indicate that the coverage of the resulting confidence intervals for both ρ_1 and ρ_2 may be quite far from the nominal value, and that the severity of the problem will in typically depend on the specific copula seat we sampled from.

On our approximate Bayesian side, once an estimator of the multivariate version of ρ is available, it is possible to apply the procedure presented in Section 2.4, with no particular modifications.

Figure 2.6, 2.7, 2.8 and 2.9 show the results of a simulation study from a Clayton ($\rho = 0.5$), Frank ($\rho = 0.5$), Gumbel ($\rho = 0.7$) and Gaussian ($\rho = 0.8$) copula respectively.

The frequentist intervals obtained via a boostrap estimate of the variance of the $\hat{\rho}_i$, $i = 1, 2$ are always too narrow to produce reliable estimates; in the case of Clayton copula the estimated coverage is about 5.8% and it tends to further decrease as the degree of dependence increases. The case of other copula is even worse, since the coverage is invariably close to 0%.

The problem of estimating the standard error of the estimates in the frequentist approach does not seem to be dependent on the dimensionality of the problem, as one may see from Table 2.2, which shows the average length of the estimated confidence intervals for $\hat{\rho}_1$ and $\hat{\rho}_2$ and the average length of the corresponding (approximated) Bayesian credible intervals: while the average length is always very low, even for high dimensions, the Bayesian intervals show a length decreasing with the dimension d.

We conclude this section with the natural comment that what we have presented here of the Spearman's ρ can be applied, in principle, to any other summary of the

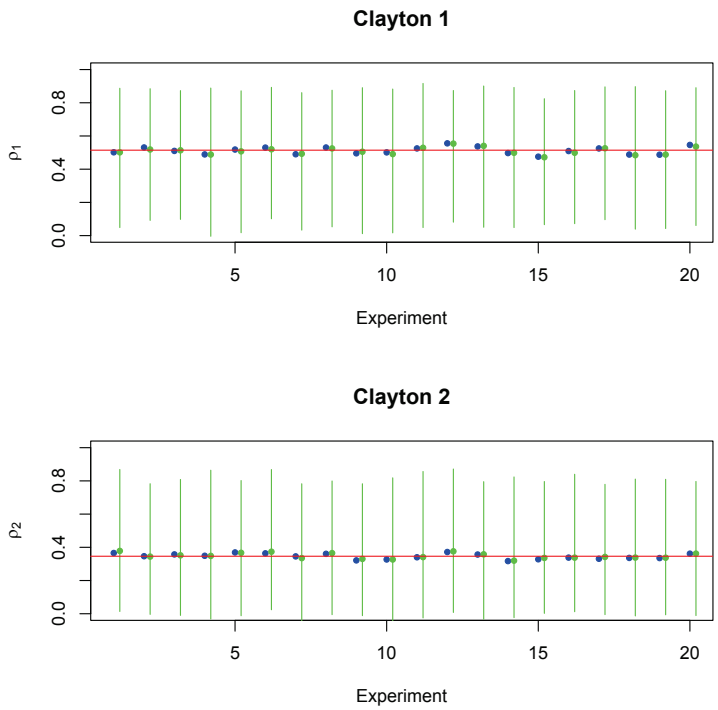

Figure 2.6: 20 out of 500 simulations from a Clayton copula: sample size is 1000; the true value of ρ is equal to 0.5 (red line). The results for frequentist (blue) and Bayesian (green) procedures. The solid lines represent the point estimates of ρ_1 (left) and ρ_2 (right), the dotted lines represent the corresponding intervals of level 0.95. The red line represents the true value

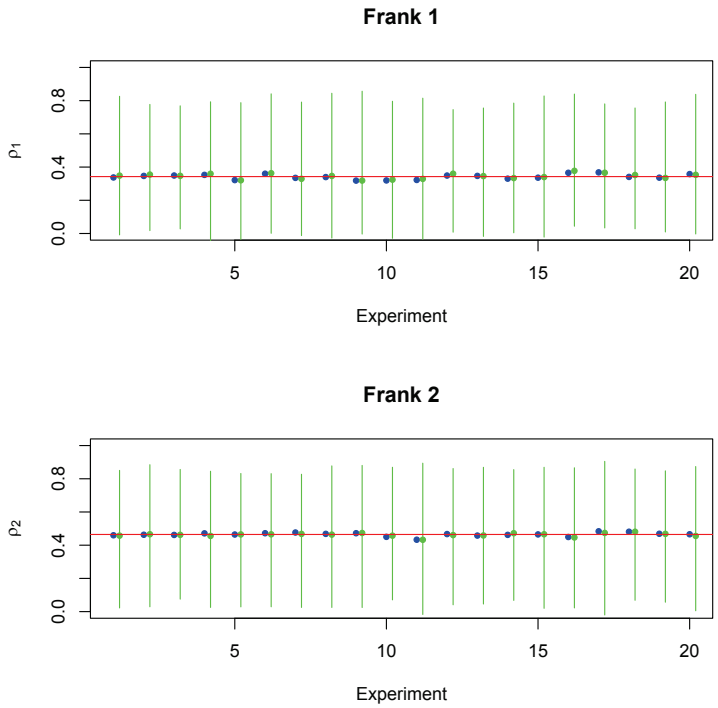

Figure 2.7: As in Figure 2.6, simulations from a Frank copula with true values of ρ_1 (left) and ρ_2 (right) in red.

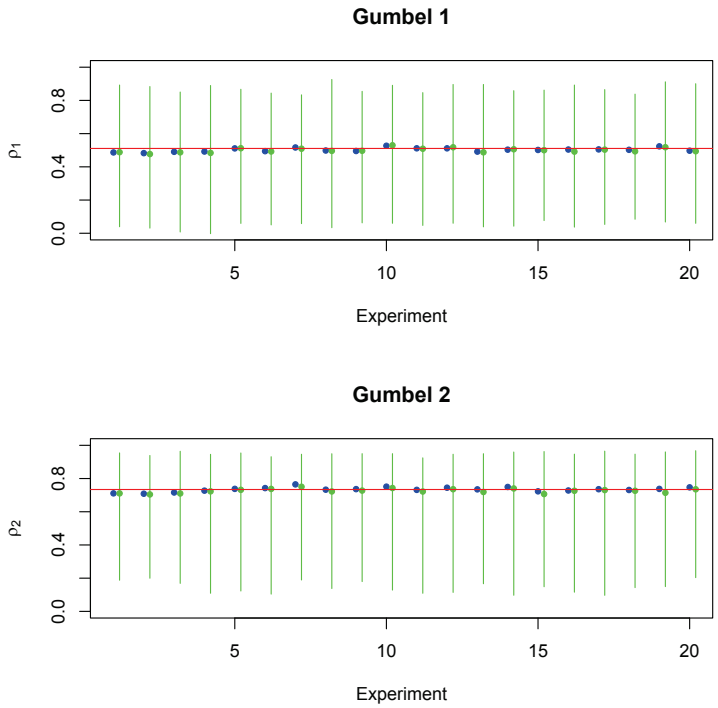

Figure 2.8: As in Figure 2.6, simulations from a Gumbel copula with true values of ρ_1 (left) and ρ_2 (right) in red.

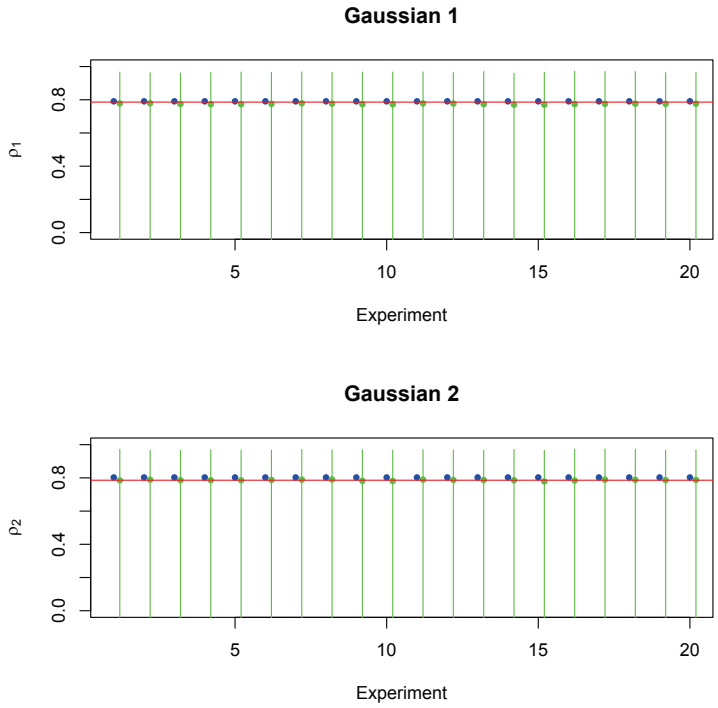

Figure 2.9: As in Figure 2.6, simulations from a Gaussian copula with true values of ρ_1 (left) and ρ_2 (right) in red.

Table 2.2: Average lengths of the confidence intervals (based on a bootstrap estimator of the variance of the estimates) and of the corresponding Bayesian credible intervals obtained in 50 repetitions of each experiment of dimension d by simulating data from a Clayton copula with $\theta = 1.076$.

	$\hat{\rho}_1{}^{freq}$	$\hat{\rho}_2{}^{freq}$	$\hat{\rho}_1{}^{Bayes}$	$\hat{\rho}_2{}^{Bayes}$
$d = 2$	0.0032	0.0032	1.1933	1.1801
$d = 3$	0.0026	0.00260	1.0844	1.0853
$d = 4$	0.0026	0.0026	0.9495	0.9594
$d = 5$	0.0027	0.0027	0.8728	0.8914
$d = 6$	0.0027	0.0027	0.8211	0.8224
$d = 7$	0.0030	0.0030	0.8022	0.7882
$d = 8$	0.0031	0.0031	0.7828	0.7541
$d = 9$	0.0032	0.0032	0.7680	0.7492
$d = 10$	0.0035	0.0035	0.7558	0.7439
$d = 25$	0.0047	0.0047	0.7462	0.7480
$d = 50$	0.0073	0.0073	0.7299	0.7634

multivariate copula, such as the Kendall's τ, or, as illustraed in the next section, some measure of tail dependence.

2.8 Tail Dependence

Measures of dependence as the Spearman's ρ or the Kendall's τ are not always suited to explain the dependence structure. In particular, dependencies between extreme events such as extreme negative stock returns or large portfolio losses are better explained by alternative dependence measures.

For example, tail dependence coefficients (see, for example, Sibuya (1960)) have been proposed to better capture dependence among extreme values. The upper and lower tail dependence coefficients have a definition based on the survival function: given a two-dimensional vector $X = (X_1, X_2)$ with marginal distribution functions F_1 and F_2 respectively, the upper tail dependence coefficient is defined as

$$\lambda_U = \lim_{v \to 1} \Pr\{X_1 > F_1^{-1}(v) | X_2 > F_2^{-1}(v)\} \tag{2.10}$$

and the lower tail dependence coefficients is defined as

$$\lambda_L = \lim_{v \to 0} \Pr\{X_1 < F_1^{-1}(v) | X_2 < F_2^{-1}(v)\} \tag{2.11}$$

Tail dependence coefficients can also be represented in terms of the underlying copula, in the following way

$$\lambda_U = \lim_{v \to 1} \frac{1 - 2v - C(v, v)}{1 - v}$$
$$\lambda_L = \lim_{v \to 0} \frac{C(v, v)}{v}$$

For definitions and properties one may refers to Schmid and Stadtmüller (2006).

A review of the parametric and non-parametric estimators for the tail dependence coefficients is given in Frahm et al. (2005). Among the various proposal we consider the one proposed by Joe et al. (1992):

$$\hat{\lambda}_U = 2 - \frac{1 - \hat{C}_n\left(\frac{n-k}{n}, \frac{n-k}{n}\right)}{1 - \frac{n-k}{n}}. \tag{2.12}$$

where $\hat{C}_n$ is the empirical copula, and $0 < k \leq n$ is a parameter tuned by the experimenter. A typical choice, suggested in Joe et al. (1992) is $k = \sqrt{n}$. A similar

estimator is proposed for λ_L. Schmid and Stadtmüller (2006) proves strong consistency and asymptotic normality for these estimators. The same Authors have also derived the asymptotic variance of $\hat{\lambda}_L$ and $\hat{\lambda}_U$. However these expression are of limited use since they depend on unknown quantities. They propose to use the variance of the tail dependence coefficient of a copula for which the same quantities are easy to compute. Nevertheless this method does not provide any quantification of the potential error.

A peculiar problem of the analysis of the extreme dependence is that it is based on few data, and this result in non reliable standard errors. In contrast, within the BC_{EL} approach, we are able to provide an approximation of the entire posterior distribution of the index, which can be summarized in various ways. Figures 2.10, 2.11 and 2.12 show the approximated intervals for the frequentist (obtained via a bootstrap estimation of the variance) and the Bayesian procedure; Bayesian intervals are always wider than the corresponding frequentist ones. Nevertheless, the coverage of the frequentist intervals is, on average, around 0.10, far from the nominal 0.95, which is reached by the Bayesian estimates, the length of frequentist intervals is too small to be observed in the Figures.

Multivariate extensions of tail dependence coefficients are not fully developed. An interesting proposal is discussed in Di Bernardino and Rullière (2015). They consider a random vector $X := (X_1, X_2, \cdots, X_d)$ and denote by I the set $\{1, 2, \ldots, d\}$. For a given subset of indices $I_h \subset I$, and its corresponding complement set $\bar{I}_h$ a multivariate version of the tail dependence coefficients can be expressed as

$$\lambda_U^{I_h}(v, \mathbf{v}) = \lim_{v \to 1} \Pr\{X_i > F_i^{-1}(v), i \in I_h | \mathbf{X_j} > F_j^{-1}(\mathbf{v}), j \in \bar{I}_h\} \tag{2.13}$$

$$\lambda_L^{I_h}(v, \mathbf{v}) = \lim_{v \to 1} \Pr\{X_i \leq F_i^{-1}(v), i \in I_h | \mathbf{X_j} \leq F_j^{-1}(\mathbf{v}), j \in \bar{I}_h\} \tag{2.14}$$

or, in copula notation

$$\lambda_U^{I_h}(\mathbf{v}) = \lim_{v \to 1} \Pr\left\{ \frac{\mathbf{X} \in \prod_{i=1}^{d}(\mathbf{v}, 1)}{\mathbf{X} \in \prod_{i \in \bar{I}_h}(\mathbf{v}, 1)} \right\} \tag{2.15}$$

$$\lambda_L^{I_h}(\mathbf{v}) = \lim_{v \to 1} \Pr\left\{ \frac{\mathbf{X} \in \prod_{i=1}^{d}(0, \mathbf{v})}{\mathbf{X} \in \prod_{i \in \bar{I}_h}(0, \mathbf{v})} \right\} \tag{2.16}$$

In this situation, theory is not fully developed to evaluate the performance of the relative estimators; Di Bernardino and Rullière (2015) proposes to estimate the

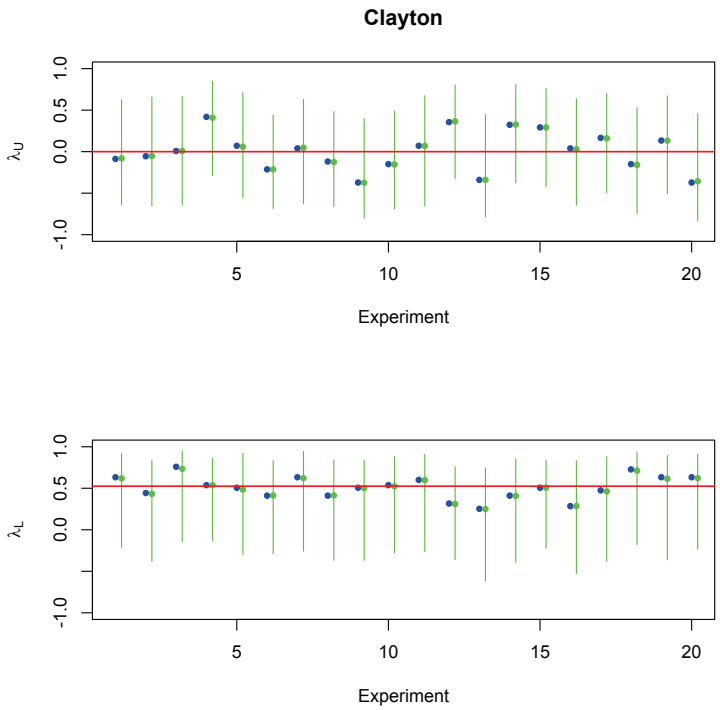

Figure 2.10: 20 out of 500 simulations from a Clayton copula with $\theta = 1.076$: sample size is 1000; comparison between the frequentist intervals (blue) and the approximated Bayesian credible intervals (green) of level 0.95 for the upper tail coefficient λ_U (above) and the lower tail dependence coefficient λ_L (below); for the Clayton copula, $\lambda_U^{true} = 0$ and $\lambda_L^{true} = 2^{-\frac{1}{\theta}}$.

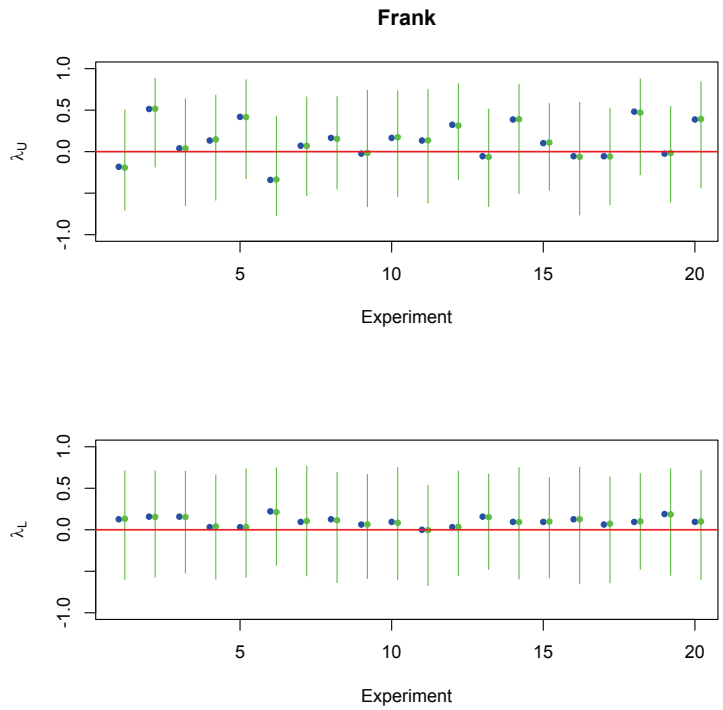

Figure 2.11: 20 out of 500 simulations from a Frank copula with $\theta = 3.45$: sample size is 1000; comparison between the frequentist intervals (blue) and the approximated Bayesian credible intervals (green) of level 0.95 for the upper tail coefficient λ_U (above) and the lower tail dependence coefficient λ_L (below); for the Frank copula, $\lambda_U^{true} = 0$ and $\lambda_L^{true} = 0$.

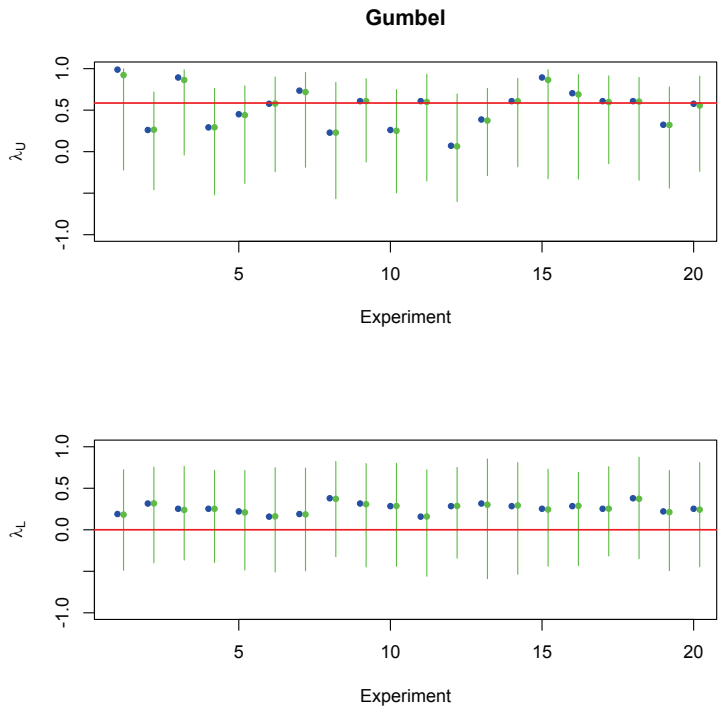

Figure 2.12: 20 out of 500 simulations from a Gumbel copula with $\theta = 2.00$: sample size is 1000; comparison between the frequentist intervals (blue) and the approximated Bayesian credible intervals (green) of level 0.95 for the upper tail coefficient λ_U (above) and the lower tail dependence coefficient λ_L (below); for the Gumbel copula, $\lambda_U^{true} = 2 - 2^{\frac{1}{theta}}$ and $\lambda_L^{true} = 0$.

multivariate tail dependence coefficients through estimation of the generator of the generator, however if we assume to have no information about the shape of the copula function, it is difficul to assess the estimation error in this way.

On the other hand, our approach may be easily extended to the multivariate setting.

2.9 Conditional measures of dependence

In some cases the analysis may be focused on measures of dependence as functions of conditioning variables. In the case of two response variables X_1 and X_2, both depending on the same covariate Z, the observations (x_{1i}, x_{2i}, z_i) follow a distribution $F_{X_1,X_2|Z}(\cdot|z)$. Gijbels et al. (2011) and Acar et al. (2011) have proposed classical procedure based on local smoothing techniques, to handle this kind of problems. In particular, Acar et al. (2011) proposes a nonparametric approach based on local likelihood to estimate the relationship between the copula parameter and the covariate; in this case, the choice of the parametric copula function is crucial, then the method may not be considered completely nonparametric and, moreover, even if the Authors propose a method to select the copula function based on cross-validation, the results still deeply depend on the choice of the copula function with a high error in case of wrong choice.

On the other hand, Gijbels et al. (2011) proposes nonparametric estimators of the conditional copula. After having defined the conditional copula as:

$$C_z(u_1, u_2) = H_z(F^{-1}_{X_1|Z}(u_1|z), F^{-1}_{X_2|Z}(u_2|z))$$

where $F^{-1}_{X_1|Z}(\cdot|z)$ and $F^{-1}_{X_2|Z}(\cdot|z)$ represent the conditional marginal quantile function of X_1 and X_2 respectively, and estimator of $H_z(\cdot|x)$ is provided by

$$H_{zh}(x_1, x_2) = \sum_{i=1}^{n} w_{ni}(z, h_n)\mathbb{I}(X_{1i} \leq x_1, X_{2i} \leq x_2) \tag{2.17}$$

where $\{wni(z, h_n)\}$ is a sequence of weights which smooth over the covariate space with a degree of smoothing depending on the bandwidth h_n going to zero as the sample size increases, i.e. the weigths go to zero as the observed point z_i is far from the point z and the speed depends on the bandwidth h_n which is $O(n^{-1/5})$. Example

of kernel-based weigths may be found in Nadaraya (1964) or Watson (1964) or in Fan and Gijbels (1996):

$$w_{ni}(z) = \frac{\frac{1}{nh_n}K(\frac{Z_i - z}{h_n})\left(S_{2n} - \frac{Z_i - z}{h_n}S_{1n}\right))}{S_{0n}S_{2n} - S_{1n}^2}$$

where $S_{jn} = \frac{1}{nh_n}\sum_{i=1}^{n}\left(\frac{Z_i - z}{h_n}\right)^j K\left(\frac{Z_i - z}{h_n}\right)$ for $j = \{0, 1, 2\}$. The function $K(\cdot)$ is a kernel density integrating to one, for instance

$$K(y) = \frac{35}{32}(1 - y^2)\mathbb{I}_{[-1,1]}(y)$$

Unfortunately, estimator 2.17 and his modified version available in Gijbels et al. (2011) are biased (even if, the second version is able to reduce the bias, while fixing the variance at the same level of estimator 2.17).

Algorithm 5 produces an approximation of the posterior distribution of any summary of the multivariate dependence, once a multivariate estimator is available, as in the case of the Spearman's ρ. In some cases the analysis may be focused on measures of dependence as functions of some available conditioning variables. In the case of two response variables X_1 and X_2, both depending on the same covariate Z, the observations (x_{1i}, x_{2i}, z_i) follow a distribution $F_{X_1,X_2|Z}(\cdot|z)$. Gijbels et al. (2011) proposes the following estimator for the Spearman's ρ.

$$\hat{\rho}_n(x) = 12\sum_{i=1}^{n} w_{ni}(x, h_n)(1 - \hat{U}_{i1})(1 - \hat{U}_{i2}) \tag{2.18}$$

where $\hat{U}_{i,j} = \sum_{i'=1}^{n} w_{i'}(x, h_n)\mathbb{I}(U_{i'j} \leq u_{ij})$ for $j = 1, 2$, $U_{ij} = F_j(x_{ij})$ and $w_{ij}(x, h_n)$ are appropriately chosen weights depending on x_{ij} and a bandwidth h_n, for example kernel-based weights as the Nadaraya-Watson. Unfortunately, estimator (2.18) is based on an estimator of the conditional copula, given in Gijbels et al. (2011), which is biased. A first simulation study implemented for $10,000$ simulations of the function $\rho(x)$ (see Figure 2.13) shows that, while the estimator 2.18 is not able to capture the true function (it underestimates the dependence among values), the Bayesian estimate obtained via Algorithm 5 can recover it, even if the variance increases as the value of the covariate increases. Further research will be focused on trying to understand why this happens and on producing more stable estimates.

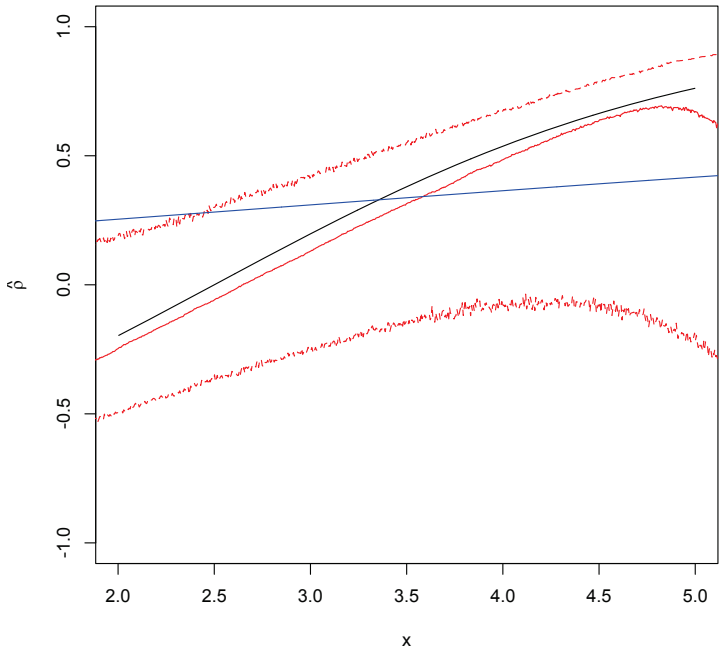

Figure 2.13: Simulations from the conditional Clayton copula based on $10,000$ ABC simulations of $\rho(x)$ and 100 data points: true function $\rho(x)$ in black, Bayesian estimates in red (median, 0.05 and 0.95 credible bands), frequentist estimate in blue.

2.10 Example: Spearman's ρ for Student-t log-returns

We now analyze a real data-set containing the log-returns FTSE-MIB of two Italian banks, Monte dei Paschi di Siena (BMPS) and Banco Popolare (BP), by assuming that the log-returns for each bank may be described by a GARCH(1,1) model with Student-t innovations for the log-returns $\{y_t\}$ from 01/07/2013 to 30/06/2014 (only weekdays) available on the web page *https://it.finance.yahoo.com*.

The GARCH(1,1) model for Student-t innovation may be rewritten via data augmentation, following Geweke (1993):

$$
\begin{aligned}
y_t &= \varepsilon_t \sqrt{\frac{\nu-2}{\nu}\,\omega_t h_t} \quad t = 1, \cdots, T \\
\varepsilon_t &\sim \mathcal{N}(0,1) \\
\omega_t &\sim IG\left(\frac{\nu}{2}, \frac{\nu}{2}\right) \\
h_t &= \alpha_0 + \alpha_1 y_{t-1}^2 + \beta h_{t-1} \quad t = 1, \cdots, T
\end{aligned}
$$

where $\alpha_0 > 0$, $\alpha_1, \beta >= 0$ and $\nu > 2$, $\mathcal{N}(0,1)$ denotes the standard normal distribution and $IG(a,b)$ denotes the inverted gamma distribution with shape parameter a and scale parameter b. Figure 5 shows the scatterplot of the log-returns and the transformed version of them, using, as a point estimate, the posterior mean of each parameter.

For each bank, the posterior distribution of the model parameters $(\alpha_0, \alpha_1, \beta, \nu)$ may be approximated by using the **R** package `bayesGARCH` Ardia and Hooherheide (2010). Once a sample from the approximated distribution is simulated for each parameter and for each bank, Algorithm 6 is applied as follows as in Algorithm 7.

The output of Algorithm 6 relative to the log-returns of Monte dei Paschi di Siena and Banco Popolare are shown in Figure 2.15: the estimated posterior mean of ρ is 0.614.

Moving to a multidimensional setting and consider a number k of investments, where $k > 2$ is straightforward in a Bayesian setting, while a frequentist approach to the problem is not fully developed yet.

Figure 2.16 shows the results of the Bayesian procedure based on Algorithm 6 and

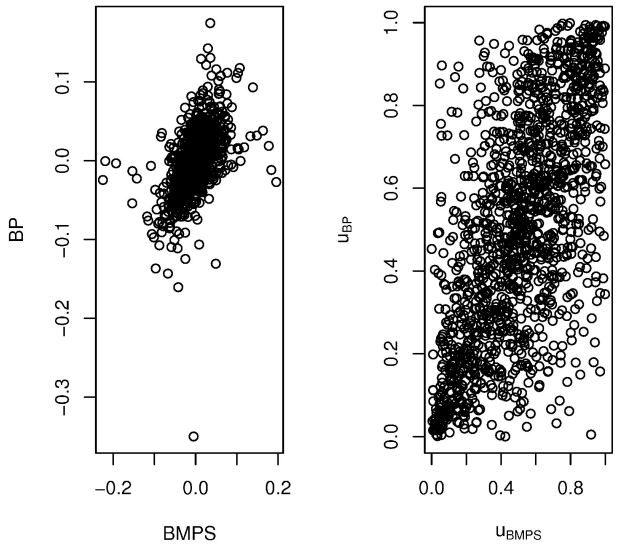

Figure 2.14: Scatterplot of the log-returns of the investments of Monte dei Paschi di Siena (BMPS) and Banco Popolare (BP) on the left and of the transformed data on the right.

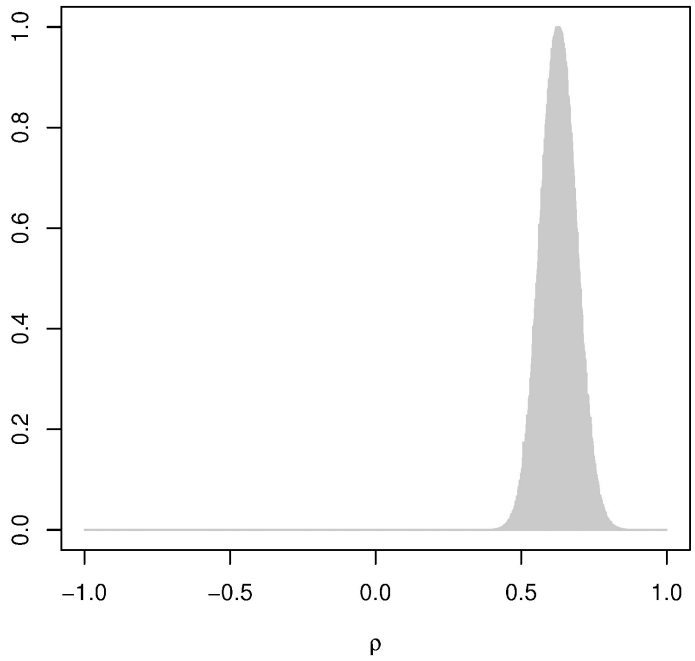

Figure 2.15: Approximation of the posterior distribution of the Spearman's ρ for the log-returns of the investments of two Italian institutes based on $10,000$ simulations.

Algorithm 7 ABCOP algorithm: Application

for $m = 1, \cdots, M$

1: Simulate a value $\rho^{(m)} \sim Unif(-1, 1)$.

2: Sample two integer values $b_j^{(m)}$ $(j = 1, 2)$ in $\{1, \cdots, S\}$, where S is the number of posterior simulations.

3: Consider the $b_j^{(m)}$-th row of the MCMC output for the parameters of the j-th marginal (i.e. $\alpha_{0j}, \alpha_{1j}, \beta_j, \nu_j$), for $j = 1, 2$).

4: Compute pseudo-data $u_{ij}^{(m)}$ for $i = 1, \cdots, T$ and $j = 1, 2$ as

$$u_{ij}^{(m)} = \mathbf{F}_{\nu_j}\left(y_i; \frac{\nu_j^{(m)} - 2}{\nu_j^{(m)}} h_{ij}^{(m)}\right)$$

where $\mathbf{F}_\nu(x, d)$ is the CDF of a Student-t distribution with ν degrees of freedom and scale parameter d.

5: Compute the estimated sample Spearman's $\rho_n^{(m)}$ as in (2.3) and the weight relative to the simulated $\rho^{(m)}$ as $\omega^{(m)} = EL(\rho_n^{(m)}; \mathbf{u}_1^{(m)}, \mathbf{u}_2^{(m)})$ as in Owen (2010).

three different estimators chosen in the empirical likelihood step (see Section 2.7 for details). While in a frequentist approach, the three estimators have different properties and may lead to different estimates, the Bayesian procedure based on Algorithm 6 provide similar approximations of the posterior distribution of the Spearman's ρ, no matter what estimator has been used to define the empirical likelihood.

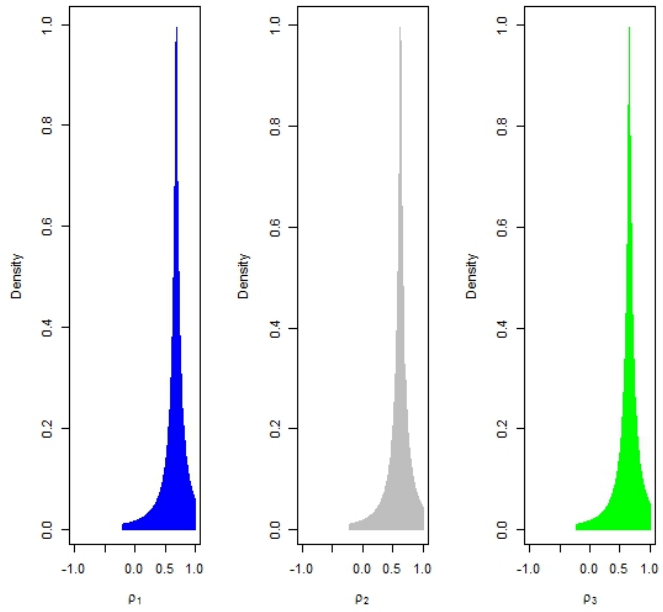

Figure 2.16: Approximation of the posterior distribution of the Spearman's ρ for the log-returns of the investments of five Italian institutes (Monte dei Paschi di Siena, Banco Popolare, Unicredit, Intesa-Sanpaolo and Mediobanca) during the same period as Figure 2.15 based on $10,000$ simulations.

Chapter 3

New approaches in Bayesian Model Choice

3.1 On proper scoring rules for Bayesian model selection

The Bayesian approach to model choice presents some difficulties, in particular when using improper noninformative priors (for which the Bayes factor in not computable) and in terms of computational approximation. Dawid and Musio (2015) is an attempt to replace the traditional log-score by a proper local scoring rule, with the purpose of finding a general method to avoid these difficulties.

Let $\mathcal{M}$ be a finite class of statistical models for the same observable $\mathbf{Y} \in \mathcal{Y} \subseteq \mathcal{R}^n$; each $M \in \mathcal{M}$ is a parametric family, with parameter $\theta_M \in \Theta_M$. Let P_{θ_M} be the distribution of $\mathbf{Y}$ under model M, with density $p_M(\mathbf{y}|\theta_M)$. Within each model M, the parameter θ_M is given a prior distribution Π_M, with density function $\pi(\theta_M)$ with respect to Lebesgue measure $d\theta_M$ over Θ_M. The predictive density function of $\mathbf{Y}$ under model M is

$$p_M(\mathbf{y}) = \int_{\Theta_M} p_M(\mathbf{y}|\theta_M)\pi_M(\theta_M)d\theta_M$$

and it is equivalent to the marginal likelihood for model M, $p(\mathbf{y}|M)$. Using the Bayes' theorem, the posterior probability $\pi(M|\mathbf{y})$ for model M is given by

$$\pi(M|\mathbf{y}) \propto \pi(M) \times p(\mathbf{y}|M)$$

where $\pi(M)$ is the prior probability that the true model is $M \in \mathcal{M}$, and the missing

constant is to ensure that $\sum_{M\in\mathcal{M}}\pi(M|\mathbf{y}) = 1$. Then, the posterior odds ratio in favor of one model M_1 against another model M_2 is equal to

$$\frac{\pi(M_1|\mathbf{y})}{\pi(M_2|\mathbf{y})} = \frac{\pi(M_1)}{\pi(M_2)}\frac{p(\mathbf{y}|M_1)}{p(\mathbf{y}|M_2)} = \frac{\pi(M_1)}{\pi(M_2)}\mathrm{B}_{12}(\mathbf{y}).$$

The Bayes factor is the coefficient by which it is needed to multiply the prior odds ratio in order to obtain the posterior odds ratio, given $\mathbf{Y} = \mathbf{y}$. It depends on the prior densities π_{M_1}, π_{M_2} and this may lead to problems when using, for example, improper priors.

The *log score* is defined as

$$S_L(\mathbf{y}, Q) = -\log q(\mathbf{y}) \tag{3.1}$$

for any distribution Q with density function $q(\cdot)$ over $\mathcal{Y}$, and $\mathbf{y} \in \mathcal{Y}$. Then, the log-Bayes factor for model M_1 against model M_2

$$\log p(\mathbf{y}|M_1) - \log p(\mathbf{y}|M_2)$$

can be seen as a comparison of the *log scores* of the two marginal densities, given $\mathbf{Y} = \mathbf{y}$ (Good 1952).

Let

$$S(P, Q) = \int S(\mathbf{y}, Q)p(\mathbf{y})d\mathbf{y}$$

be defined as the expected score when $\mathbf{Y}$ has distribution P, with density function p; a scoring rule S has the property of being *proper* if $S(P, Q)$ is minimized, for any given P, by the choice $Q = P$.

A scoring rule $S(\mathbf{y}, Q)$ is called *local (of order m)* if it can be express as a function of $\mathbf{y}$, $q(\mathbf{y})$, and derivatives of $q(\cdot)$, up to the mth order, evaluated at $\mathbf{y}$. For example, the log-score $S_L(\mathbf{y}, Q)$ in (3.1) is local of order 0.

When the sample space $\mathcal{Y}$ is an interval on the real line, Parry et al. (2012) showed that all proper local scoring rules can be expressed as a linear combination of the log-score and a "key local" scoring rule. The "key local" is a proper local scoring rule which is *homogeneous*, which means that its value does not change if q and all its derivatives are multiplied by some constant $c > 0$.

Therefore, the usual log-Bayes factor can be replaced by some key local scoring rule, so that the dependence on the normalizing constant will disappear and the problems with the normalizing constant in a Bayesian analysis will be avoided.

For the observed data $\mathbf{Y} = \mathbf{y}$, where $\mathbf{Y} \in \mathcal{Y} \subseteq \mathcal{R}^n$, the key local scoring order-2 rule of Hyvärinen (2005) is:

$$S_H(\mathbf{y}, Q) = 2\Delta \ell_n q(\mathbf{y}) + \|\nabla \ell_n q(\mathbf{y})\|^2 \tag{3.2}$$

for any $n > 1$, where Δ is the Laplacian operator $\sum_{i=1}^{n} \partial^2/(\partial y_i)^2$, ∇ is the gradient, and $q(\mathbf{y})$ is a continuous function of $\mathbf{y}$ (the factor 2 is added for convenience).

Note that, in order to apply this scoring rule, the function of $\mathbf{y}$ must be continuous, which means that the proposed approach cannot be applied to discrete distributions.

3.1.1 Linear model

The method will be applied to the case of the normal linear model for a data vector $\mathbf{Y} = (Y_1, ..., Y_n)^T$, assuming to be

$$\mathbf{Y} \sim N(X\boldsymbol{\theta}, \sigma^2 I_n) \tag{3.3}$$

where X is a known $(n \times p)$ design matrix of rank p, $\boldsymbol{\theta} \in \mathcal{R}^p$ is an unknown parameter vector, I_n is the identity matrix of size n and σ^2 is assumed to be known.

Let

$$\boldsymbol{\theta} \sim N(\mathbf{m}, V) \tag{3.4}$$

be a normal prior distribution on $\boldsymbol{\theta}$; then, the marginal distribution Q of $\mathbf{Y}$ is

$$\mathbf{Y} \sim N(X\mathbf{m}, XVX^T + \sigma^2 I_n)$$

with precision matrix derived from applying the matrix lemma in equation (10) of Lyndley and Smith (1972):

$$\begin{aligned} \Phi &= (XVX^T + \sigma^2 I_n)^{-1} \\ &= \sigma^{-2} \left\{ I_n - X(X^T X + \sigma^2 V^{-1})^{-1} X^T \right\}. \end{aligned} \tag{3.5}$$

Since we are interested on using improper priors, one of these can be generated by letting $V^{-1} \to 0$. This will lead to $\Phi = \sigma^{-2}\Pi$, where

$$\Pi = I_n - X(X^T X)^{-1} X^T$$

is the projection matrix onto the space of residuals, so that $\mathrm{tr}\Phi = \sigma^{-2}(n-p)$.

When a normal case is considered, it is easy to show that

$$\nabla \ell_n q(\mathbf{y}) = -\Phi(\mathbf{y} - \boldsymbol{\mu})$$

$$\Delta \ell_n q(\mathbf{y}) = -\mathrm{tr}\Phi,$$

then by applying these formulas to the linear model, we have:

$$\begin{aligned} S_H(\mathbf{y}, Q) &= -2\sigma^{-2}(n-p) + \frac{RSS}{\sigma^4} \\ &= \frac{1}{\sigma^4}\{RSS - 2\nu\sigma^2\} \end{aligned} \tag{3.6}$$

where RSS is the residual sum of squares for model (3.3), on $\nu = n - p$ degrees of freedom. It is important to notice that, because of the homogeneity of the Hyvärinen score (3.2), there is no need to specify the normalizing constant for the improper prior density.

Furthermore, if $\mathrm{rank}(X) < p$, the integral defining the marginal density of $\mathbf{Y}$ is not finite at each $\mathbf{y}$, so that no marginal joint density can be defined. Then, the above analysis is not applicable when $n < p$. On the contrary, when X is of rank p, this integral is finite for each $\mathbf{y}$, even if the resulting density is improper.

Finally, using the criterion (3.6) means comparing different normal linear models in terms of their penalized scaled residual sum of squares, which is equivalent to the Akaike's Information Criterion (Akaike 1974), for the known variance case. However, here the known residual variance σ^2 is the same across all models, if it varies the criterion (3.6) and AIC are no longer equivalent.

3.2 A discussion of "Bayesian model selection based on proper scoring rules" by A.P. Dawid and M. Musio

The[1] frustrating issue of Bayesian model selection preventing improper priors (DeGroot 1982) and hence most objective Bayes approaches has been a major impediment to the development of Bayesian statistics in practice (Marin and Robert 2007), as the failure to provide a "reference" answer is an easy entry for critics who point out the strong dependence of posterior probabilities on prior assumptions. This was presumably not forecasted by the originator of the Bayes factor, Harold Jeffreys, who customarily and informaly used improper priors on nuisance parameters in his construction of Bayes factors (Robert et al. 2009). It is therefore a very welcome item of news that a truly Bayesian approach can allow for improper priors.

As also pointed out in the paper, there exist a wide range of "objective Bayes" solutions in the literature (Robert 2001), all provided with validating arguments of sorts, but this range by itself implies that such solutions are doomed in that they cannot agree for a given dataset and a given prior.

Finding a criterion that does not depend on the normalising constant of the predictive possibly is the unravelling key to handle improper priors and we congratulate the Authors for this finding of the Hyvärinen score and related proper scoring rules. Some difficulties deriving from the use of improper prior distributions in model choice may be solved by applying the approach proposed in the paper. There are nonetheless some issues with this solution:

- calibration difficulties: once the score value is computed, the calibration of its strength very loosely relates to a loss function, hence makes decision in favour of a model difficult;
- a clear dependence on parameterisation: changing y into the tranform $\mathfrak{h}(y)$ produces a different score;
- a dependence on the dominating measure: as exhibited in the case of exponential families, changing the dominating measure modifies the score function;

[1] This work is based on a Master research thesis written by Ilaria Masiani under the supervision of the candidate and Christian P. Robert, at Université Paris-Dauphine. A reply to this discussion is not publicly available

- the arbitrariness of the Hyvärinen score, which is indeed independent of the normalising constant, but offers limited arguments in favour of this particular combination of derivatives. Since there exists a immense range of possible score functions, a stronger connection with inferential properties is a clear requirement;

- as noted above, consistency is not a highly compelling argument for the layperson, as it does not help in the calibration and selection of the score. Having the multivariate score being inconsistent, is highlighting this difficulty.

Furthermore, the only application of the method presented in the paper is within the setting of the normal linear model and we worry that the approach may not be easily extended to other types of models. In particular, the representation of the precision matrix of the marginal distribution in equation 3.5, based on the Woodbury matrix inversion lemma, is essential to easily apply the proper scoring rule approach to model choice with an improper prior, given that an improper prior may then be seen as a limiting version of a conjugate prior and its influence disappears in the following computations. However, the approach overcomes the singularity of the precision matrix of the marginal distribution.

We first performed some simulation studies when applying the proposed method to models that differ from the normal linear model. When choosing between two different models with no covariates, we observed that the proposed approach can perform well as for instance when a Gamma model is opposed to a normal model (well in the sense of comparing with a standard Bayes factor). However, when the comparison is operated between a Pareto distribution and a normal distribution, the approach does not often select the right model when data are generated from the Pareto distribution, while the Bayes factor always leads the right model. In addition, we came to the realisation that the method based on the Hyvärinen scoring rule may not be applied to some models, for example when data come from a Laplace distribution, which is not differentiable in 0, or for discrete models.

Our simulation studies have also coverered linear models, whether nested or not nested. The details of the simulation models are given in the captions of Figure 3.1–3.3. The performance of the multivariate Hyvärinen score when comparing normal linear models is excellent, as shown in Figure 3.1, even when using an improper prior, provided the sample size is larger than the number of parameters in the model: following repeated simulations, we observed that the method is always able

to choose the right model. We however noticed that, when the true model does not involve covariates, the ability of the method to discriminate between models is reduced. Although this approach shows a consistent behavior and chooses the right model with higher and higher certainty when the sample size increases, our simulations have also shown that the log proper scoring rule tends to infinity more slowly than the Bayes factor or than the likelihood ratio. It is approximately four times slower, all priors being equal, as shown in Figures 3.2 and 3.3, which represent the comparison between the approach based on the log-Bayes factor and the one based on the difference between the score functions for the case of linear models, both nested (Figure 3.3) and non-nested (Figure 3.2).

As a final remark, we would like to point out the alternative and recent proposal of Kamary et al. (2014) for correctly handling partly improper priors in testing settings through the tool of mixture modelling, each model under comparison corresponding to a component of the mixture distribution. Testing is then handled as an estimation problem in an encompassing model. Therein, the Authors show consistency in a wide range of situations. We currently appreciate the approach through mixture estimation as the most compelling for the many reasons advanced in Kamary et al. (2014), in particular because the posterior distribution of the weight of a model is easily interpretable and scalable towards selecting this very model or an alternative one. Furthermore, it returns posterior probabilities for the models under comparison without the need to resort to specific prior probability weights.

3.3 Model Choice with approximate Bayesian Computation

For inferential problems approximate Bayesian computation has been used in many fields, like genetics (Tavaré et al. 1997, Pritchard et al. 1999), finance (Creel and Kristensen 2015), signal processing (Kervrann et al. 2014), etc. thanks to their property to give inferential results for models for which the likelihood function is intractable.

The basic version of the algorithm is available in Section 1.1 (see Algorithm 1). The approximation of the posterior distribution $\pi(\theta|\mathbf{y})$ is the joint distribution

$$\pi_\varepsilon(\theta, \mathbf{z}|\mathbf{y}) = \frac{\pi(\theta) f(\mathbf{z}|\theta) \mathbb{I}_{A_{\varepsilon,\mathbf{y}}}(\mathbf{z})}{\int_{A_{\varepsilon,\mathbf{y}} \times \Theta} \pi(\theta) f(\mathbf{z}|\theta) d\mathbf{z} d\theta} \tag{3.7}$$

where $\mathbb{I}_{A_{\varepsilon,\mathbf{y}}}$ is the indicator function for the space $\{\mathbf{z} \in \mathcal{Y}\ s.t.\ \rho(\eta(\mathbf{z}), \eta(\mathbf{y})) < \varepsilon\}$ ($\rho(\cdot)$ is an appropriate distance between summary statistics $\eta(\cdot)$ computed on both the observed and the simulated data and ε is the tolerance level which provides the desired degree of similarity). If the summary statistics $\eta(\cdot)$ is sufficient, the posterior distribution approximated via ABC can be proved to be an approximation of the true posterior distribution $\pi(\theta|\mathbf{y})$ as ε goes to 0.

Since ABC is used in complex situations, it is unlikely that a sufficient statistics of small dimension d exists. Fearnhead and Prangle (2012) prove how the choice of $\eta(\cdot)$ and its dimension affect the Monte Carlo error. In some sense, the choice of the summary statistics to use is problem-specific, nevertheless some works exist to make it automatic; see for example Nunes and Balding (2010).

The loss of information due to the use of non-sufficient summary statistics is, in general, considered acceptable in inferential problems because ABC allows to manage complex models which are otherwise intractable, in particular when one can find an informative summary statistics for the parameter θ. Nevertheless, the loss of information is in some way arbitrary when the aim of the experimenter is model choice instead of estimating the parameter value, as shown in Robert et al. (2011) and in Section 3.3.1, which also presents our proposal to the problem of model choice with ABC. Section 3.3.2 studies the behavior of the proposed method in some real and simulated examples: simple hypothesis testing, composite hypothesis testing, regression and dynamic models. The Chapter ends with a discussion.

3.3.1 Some difficulties with ABC for model choice

When moving the interest in problems of model choice, the basic Algorithm 1 may be simply modified as in Algorithm 8 (Robert et al. 2011).

Algorithm 8 ABC-MC

Consider a set of possible models and a data set $\mathbf{y}$ from $f_m(\cdot|\theta_m)$

for $i = 1$ to N **do**
  repeat
    Generate m from the prior $\pi(\mathcal{M} = m)$
    Generate θ_m from $\pi_m(\cdot)$
    Simulate $\mathbf{z}$ from the model $f_m(\cdot \mid \theta_m)$
  until $\rho(\eta(\mathbf{z}), \eta(\mathbf{y})) < \varepsilon$
  Set $m^{(i)} = m$ and $\theta^{(i)} = \theta_m$
end for

The output of Algorithm 8 provides the absolute frequencies of the possible models, which are approximation of their posterior probabilities. Some modifications exists of this basic idea: see, for example, Beaumont et al. (2002) and Cornuet et al. (2008).

The need to choose a summary statistic, that is acceptable in inferential problems even if it is not sufficient, leads to some arbitrariness in the setting of model choice. When considering two models, the approximation of the Bayes factor deriving from Algorithm 8 is

$$\hat{B}_{12}(\mathbf{y}) = \frac{\pi(\mathcal{M}=2)\sum_{i=1}^{N}\mathbb{I}_{m^{(i)}=1}\mathbb{I}_{\rho(\eta(\mathbf{z}),\eta(\mathbf{y}))<\varepsilon}}{\pi(\mathcal{M}=1)\sum_{i=1}^{N}\mathbb{I}_{m^{(i)}=2}\mathbb{I}_{\rho(\eta(\mathbf{z}),\eta(\mathbf{y}))<\varepsilon}} \tag{3.8}$$

which approximates

$$B_{12}^{\eta}(\mathbf{y}) = \frac{\int_{\Theta_1}\pi(\theta_1)f_1^{\eta}(\eta(\mathbf{y}|\theta_1))d\theta_1}{\int_{\Theta_2}\pi(\theta_2)f_2^{\eta}(\eta(\mathbf{y}|\theta_2))d\theta_2}. \tag{3.9}$$

In this setting, the Bayes factor approximated by ABC is inconsistent with the target Bayes factor, except for very few cases: even in the fortunate case of existence of a sufficient statistics for each model m, $f_m(\mathbf{y}|\theta_m) = g_m(\mathbf{y})f^{\eta}(\eta(\mathbf{y})|\theta_m)$ and then

$$B_{12}(\mathbf{y}) = \frac{g_1(\mathbf{y})}{g_2(\mathbf{y})}B_{12}^{\eta}; \tag{3.10}$$

therefore it is impossible to evaluate the difference between these two quantities without knowing the entity of the ratio $g_1(\mathbf{y})/g_2(\mathbf{y})$.

A solution for ABC in model choice

The discrepancy between the Bayes factor approximated by Algorithm 8 and the target Bayes factor is due to the fact that, once the model has been chosen in the first step of the Algorithm, an acceptance or a rejection is done by comparing the observed and the simulated data sets via a summary statistics defined conditionally on that particular model. Along these lines, even if it is possible to find sufficient summary statistics for all the models in the analysis, these statistics will not be "sufficient" for the problem of model choice.

Our proposal is to change the perspective and consider a summary statistic that is informative for the problem of model choice and not necessarily for the considered model.

Suppose there are two possible models in the analysis, m_1 and m_2 (the generalization to k models is straightforward), described by $f_1(\mathbf{y}|\theta_1)$ and $f_2(\mathbf{y}|\theta_2)$ respectively, where θ_1 and θ_2 may be scalars or vectors. Suppose also that it is possible to have an approximation of the possible models, for example via expansion or linearization, say $h_1(\mathbf{y}|\theta_1)$ and $h_2(\mathbf{y}|\theta_2)$. In this case, the Bayes factor to compare the simplified versions of m_1 and m_2

$$B^h_{12}(\mathbf{y}) = \frac{\int_{\Theta_1} \pi(\theta_1) h_1(\mathbf{y}|\theta_1) d\theta_1}{\int_{\Theta_2} \pi(\theta_2) h_2(\mathbf{y}|\theta_2) d\theta_2} \tag{3.11}$$

is informative with respect to the comparison between the actual models. Algorithm 9 follows.

Algorithm 9 ABC-MC with BF

Consider two possible models for which simpler approximations exist. Consider a data set $\mathbf{y}$ from $f_m(\cdot|\theta_m)$. Define the Bayes factor for the approximated models $BF^h_{12}(\mathbf{y})$ as in 3.11.

for $i = 1$ to N **do**
  repeat
    Generate m from the prior $\pi(\mathcal{M} = m)$
    Generate θ_m from $\pi_m(\cdot)$
    Simulate $\mathbf{z}$ from the model $f_m(\cdot \mid \theta_m)$
    Compute the Bayes factor $BF^h_{12}(\mathbf{z})$ by considering the approximated models
  until $\rho(BF^h_{12}(\mathbf{y}), BF^h_{12}(\mathbf{z})) < \varepsilon$
  Set $m^{(i)} = m$ and $\theta^{(i)} = \theta_m$
end for

Algorithm 9 is not thought to be used for inference, even if the output includes a sample of parameter values, because the summary statistics used is certainly informative for the parameters values, nevertheless more informative statistics may be chosen.

3.3.2 Applications

Quantile distributions

Quantile distributions are an useful and flexible tool used to model data moving away from the normal target, like kurtotic or asymmetric data, and which are characterized by a small number of parameters, unlike mixture models usually used to model this type of data. Even if they are very flexible, they are not widespread in applications because they are defined by their quantile function which is a non-linear transformation of the quantiles of a normal distribution; therefore the probability density function (and so the likelihood function) is not available, that means that any approach but ABC is difficult to apply (Allingham et al. 2009), even if some attempts exist: for example, Rayner and MacGillivray (2002b) and Su (2007) propose a numerical likelihood approach while Haynes and Mengersen (2005) propose a Bayesian estimation via MCMC.

Tukey (1977) is usually considered the inventor of the g-and-h family of distributions, generalized by MacGillivray (1992) for a more interpretable version. An

example of quantile distribution is the g-and-k distribution, whose quantile function is defined as follows:

$$Q_{gk}(p; A, B, g, k) = A + B\left(1 + c\frac{1 - e^{-gz(p)}}{1 + e^{-gz(p)}}\right) \cdot (1 + z(p)^2)^k z(p) \tag{3.12}$$

where $z(p)$ is the quantile of level p of a standard normal distribution and $\theta = (A, B, g, k)$ are the unknown parameters (c is assumed fixed and equal to 0.8, see Rayner and MacGillivray (2002b) for a justification). It is easy to see that the normal distribution is a special case of the g-and-k distribution, with $g = 0$ and $k = 0$. In general, quantile distributions are very flexible and may describe many known distributions, both exactly (as in the case of the g-and-k distribution and the normal distribution) and approximately.

When the goal of the analysis is inference, some choices of summary statistics have alreaby been proposed in the literature: for example, the sample moments or the complete set of the order statistics $S(\mathbf{y}) = (y_{(1)}, \cdots, y_{(n)})$ as in Allingham et al. (2009) or robust estimates of location, scale, skewness and kurtosis as in Drovandi and Pettitt (2011). Nevertheless, these summary statistics do not take into account the relationship between possible models, therefore they are not appropriate when the goal of the analysis is model choice, as explained in Section 3.3.1.

The setting of quantile distributions is an ideal case to apply the method presented in Section 3.3.1.

Simple Hypothesis Testing

Consider two simple hypotheses

$$\begin{aligned} H_0 &: A = A_0 \\ H_1 &: A = A_1 \end{aligned} \tag{3.13}$$

where A is the location parameter of a g-and-k distribution and every other parameter is known. Computing the Bayes factor for the system of hypotheses (3.13) is complicated, nevertheless it is straightforward if the hypotheses could be related to normal models with location parameters (e.g. mean values) A_0 and A_1 respectively, that is

$$H_0 : \mu = A_0$$
$$H_1 : \mu = A_1 \tag{3.14}$$

where μ is the unknown mean parameter of a normal distribution with $\sigma = B$ known. In this situation, suppose for simplicity $\mu_0 < \mu_1$; the Bayes factor is then

$$B_{01}^{\mathcal{N}}(\mathbf{y}) = \exp\left\{-\frac{n(\mu_0 - \mu_1)}{\sigma^2}\left(\bar{y} - \frac{\mu_0 + \mu_1}{2}\right)\right\}. \tag{3.15}$$

In applying Algorithm 9 the experimenter may use (3.15) to compare the hypotheses and then keep only those values of proposed parameters A which lead to a Bayes factor close to the one computed for normal hypotheses.

Figure 3.4 shows the comparison between two g-and-k distributions which differ for the location parameters only.

The distance has been selected as the Euclidean distance between the Bayes factor for normal approximated model for the observed and the simulated data sets:

$$\rho(BF_{ij}^{\mathcal{N}}(\mathbf{y}), BF_{ij}^{\mathcal{N}}(\mathbf{z})) = \sqrt{(BF_{ij}^{\mathcal{N}}(\mathbf{y}) - BF_{ij}^{\mathcal{N}}(\mathbf{z}))^2}. \tag{3.16}$$

The tolerance level is chosen after a pilot run of the algorithm with a high value of ε, which allows to construct the approximated distribution of the values of $\rho(\eta(\mathbf{y}), \eta(\mathbf{z}))$; the value of ε corresponding to the left-hand tail of this distribution at a desired level of accuracy is then selected (see Allingham et al. (2009) for details). After fixing the tolerance level, the results will be compared with smaller values. Figure 3.5 shows the output of the pilot run for the models described in Figure 3.4 for a target tolerance level $\varepsilon = 100$. The quantile of level 0.05 of the distribution of ρ in pilot run has been found to be 1.95. This tolerance level has been compared to smaller ones, in particular 0.97 and 0.5.

Table 3.1 shows the mean values of the probabilities of choosing the right and the wrong models in 10^3 repetitions of the experiment. The proposed procedure always chooses the right model both when the null hypothesis is true and when it is wrong, conditionally on the selected tolerance levels.

Table 3.1 shows the results when B is fixed and equal to 1, nevertheless the scale of the distribution (parameter B) has an influence on the performance of the algorithm. Table 3.2 shows the approximated probabilities of correctly choosing when

Table 3.1: Simple Hypotheses

ε	**1.95**	**0.97**	**0.50**
$\Pr(\mathcal{M} = m_0 \mid m_0)$	1.0000	1.0000	1.0000
$\Pr(\mathcal{M} = m_1 \mid m_0)$	0.0000	0.0000	0.0000
$\Pr(\mathcal{M} = m_1 \mid m_1)$	0.9967	0.9977	0.9977
$\Pr(\mathcal{M} = m_0 \mid m_1)$	0.0033	0.0023	0.0023

the true model is the one under H_0 and when it is the one under H_1 for un increasing scale parameter (and 10^3 repetitions of the experiment). The method proposed in Section 3.3.1 seems to have a good behavior under H_0: even if the frequency of choosing the right model is decreasing when the scale parameter increases, the method still produces a higher approximated probability for the true model. When the alternative hypothesis represents the true model, there is a substantial indifference between the two models, with a slight preference for H_0. Since the prior distribution for the hypothesis was build to describe *a priori* indifference between the two hypothesis and the model are very close when $B = 10$, as shown in Figure 3.4 by the dashed lines, this behavior could be exaplained by the Lindley's paradox (see Lindley (1957), Jeffreys (1939)).

Table 3.2: Simple Hypotheses: influence of the variance

H_0 **true**	$\Pr(\mathcal{M} = m_0 \mid m_0)$	$\Pr(\mathcal{M} = m_1 \mid m_0)$
B= 1.00	1.000	0.000
B= 3.25	0.999	0.001
B= 5.50	0.878	0.123
B= 7.75	0.724	0.276
B=10.00	0.641	0.358
H_1 **true**	$\Pr(\mathcal{M} = m_1 \mid m_1)$	$\Pr(\mathcal{M} = m_0 \mid m_1)$
B= 1.00	0.992	0.008
B= 3.25	0.552	0.448
B= 5.50	0.460	0.540
B= 7.75	0.443	0.557
B=10.00	0.444	0.556

Similar results are obtained when the tolerance level is decreased. The choice of the parameters of asymmetry and kurtosis seems to have a smaller effect on the

behavior of the method.

Composite Hypothesis Testing

When the hypotheses are simple, the parameter space is constituted by two points, one for each hypothesis. In real applications, this situation is hard to occur. It is more usual to have a system of hypotheses which divides the parameter space into subspaces made of a finite or an infinite set of points. For a g-and-k distribution

$$\begin{aligned} H_0 &: A \in \Theta_0 \\ H_1 &: A \in \Theta_1 \end{aligned} \tag{3.17}$$

where $\Theta = \{\Theta_0 \cup \Theta_1\} = \mathbb{R}$. In a Bayesian setting, each hypothesis has its own prior probability, say $\pi_0 = \Pr(A \in \Theta_0)$ and $\pi_1 = \Pr(A \in \Theta_1)$. The Bayes factor for composite hypotheses is then the ratio between marginal distributions:

$$B_{01}(\mathbf{y}) = \frac{\int_{A\in\Theta_0} f_0(\mathbf{y}|A,B,g,k)\pi_0(A)dA}{\int_{A\in\Theta_1} f_1(\mathbf{y}|A,B,g,k)\pi_1(A)dA} \tag{3.18}$$

where $f_0(\cdot)$ and $f_1(\cdot)$ are the joint distributions of the data conditional on the hypothesis.

The Bayes factor in (3.18) is difficult to compute, because $f_0(\cdot)$ and $f_1(\cdot)$ are unavailable, nevertheless it is possible to compute the Bayes factor for two normal models. Suppose, therefore, the data follow a normal model, e.g. $y_i \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2)$ with σ known and μ unknown, define the system of hypotheses

$$\begin{aligned} H_0^{\mathcal{N}} &: \mu = 0 \\ H_1^{\mathcal{N}} &: \mu \neq 0 \end{aligned} \tag{3.19}$$

and consider a prior distribution conditionally on H_1 $\pi(\mu) = \mathbb{N}(0, h\sigma^2)$. Then the Bayes factor for the hypotheses in (3.19) can be demonstrate to be

$$BF_{01}^{\mathcal{N}} = (nh+1)^{\frac{1}{2}} \exp\left\{-\frac{1}{2\sigma^2}\frac{n^2\bar{y}}{n+h^{-1}}\right\} \tag{3.20}$$

where $\bar{y}$ is the sample mean. When comparing H_0 and H_1 in (3.17) via ABC, one could use the Bayes factor in (3.20) as summary statitic for model choice.

Figure 3.6 shows the histogram for the distance ρ of the accepted values for the pilot run described in Section 3.3.2. In this setting a tolerance level equal to 0.80 has been selected as the quatile of level 0.05 and the results have been compared with smaller tolerance levels, equal to 0.50 and 0.25.

Table 3.3 shows the average posterior probabilities of the possible models for 10^3 repetitions of the experiment, in the case the true model is the one under the null hypothesis and it is the one under the alternative hypothesis. In both the cases, it is evident that the procedure described in 3.3.1 has a good behavior in choosing the right model. Moreover, the results seem to be stable with respect to the tolerance level and they seem to be more confident when the true model is the one under the alternative hypothesis.

Table 3.3: Composite Hypotheses

ε	**0.80**	**0.50**	**0.25**
$\Pr(\mathcal{M} = m_0 \mid m_0)$	0.7820	0.7817	0.7826
$\Pr(\mathcal{M} = m_1 \mid m_0)$	0.2180	0.2183	0.2174
$\Pr(\mathcal{M} = m_1 \mid m_1)$	0.9999	0.9999	0.9999
$\Pr(\mathcal{M} = m_0 \mid m_1)$	0.0001	0.0001	0.0001

Table 3.4 shows how the results are affected by increasing the variance of the data or the variance of the prior (under the alternative hypothesis): there is a substantial stability of the approximated probabilities of the models under analysis; in particular, when the true model is the one under the alternative hypothesis the method always chooses the true model. A similar behavior has been found when the parameters of asymmetry or kurtosis have been changed.

Regression

In some situations, the models under analysis differ in something more than the values of the parameters, for example they may differ in the dimension of the parameters or in the form of the probability distribution of the model. In this case, the Bayes factor may still be used to compare the available models, with a straightforward modification of Equation (3.20).

Table 3.4: Composite Hypotheses - Incresing variances

$\Pr(\mathcal{M} = m_0 \mid m_0)$	*h=3.25*	*h=5.50*	*h=7.75*	*h=10.00*
B= 1.00	0.701	0.788	0.851	0.861
B= 3.25	0.730	0.802	0.840	0.873
B= 5.50	0.712	0.792	0.840	0.885
B= 7.75	0.722	0.829	0.840	0.897
B=10.00	0.711	0.793	0.852	0.850
$\Pr(\mathcal{M} = m_1 \mid m_1)$	*h=3.25*	*h=5.50*	*h=7.75*	*h=10.00*
B= 1.00	1.000	1.000	1.000	1.000
B= 3.25	1.000	1.000	1.000	1.000
B= 5.50	1.000	1.000	1.000	1.000
B= 7.75	1.000	1.000	1.000	1.000
B=10.00	1.000	1.000	1.000	0.556

For general models the Bayes factor is

$$B_{12}(\mathbf{y}) = \frac{\int_{\Theta_1} f_1(\mathbf{y}|\theta_1)\pi_1(\theta_1)d\theta}{\int_{\Theta_2} f_2(\mathbf{y}|\theta_2)\pi_2(\theta_2)d\theta}. \tag{3.21}$$

By applying a Taylor expansion of the log-likelihood around the maximum likelihood estimator $\hat{\theta}_i$ for the i-model, the log-Bayes factor may be approximated by

$$-2\log(BF_{12}) \cong -2\log l - (p_2 - p_1)\log n + A \tag{3.22}$$

where $l = \frac{f_1(\mathbf{y}|\hat{\theta_1})}{f_2(\mathbf{y}|\hat{\theta_2})}$ is the ratio between maximised likelihood functions, p_1 and p_2 are the dimensions of model m_1 and m_2 respectively and $A = \mathcal{O}(1)$ is a part, depending on the prior and on the observed Fisher information matrix, which does not depend on n. From Equation (3.22) the asymptotic consistency of the Bayes factor and its asymptotic equivalence with the BIC criterion (Schwarz 1978) derive.

Suppose the aim of the analysis is to compare two (or more) models which depend on different sets of covariates (the models can be nested) and which can be described by quantile distributions which take into account a certain grade of asymmetry or kurtosis. In this case, the method described in 3.3.1 may be used to compare the available models, by using the Bayes factor or some other model selection criterion (as the BIC) computed for linear approximations of the models as summary statistic.

Data have been simulated by fixing the scale, kurtosis and asymmetry parameters

Table 3.5: Regression

ε	**0.25**	**0.10**	**0.05**
$\Pr(\mathcal{M} = m_1 \mid m_1)$	0.5917	0.5931	0.851
ε	**2.5**	**1.5**	**0.50**
$\Pr(\mathcal{M} = m_2 \mid m_2)$	0.9938	0.9960	0.9961

and defining the location parameter as depending on some covariates. First, two models have been considered: a model without dependence on covariates and a model where the location parameter linearly depends on a variable X simulated as follows: $X \sim Unif(2, 5)$. Then, the choice is between models

$$\begin{aligned} M_1 &: A = \beta_0 \\ M_2 &: A = \beta_0 + \beta_1 \times x \end{aligned} \tag{3.23}$$

with all the other parameters considered as unknown nuisance parameters.

Again, the tolerance level has been fixed via a first pilot run which defines the distribution of the distance $\rho(\cdot)$. Then the parameters have been simulated in the ABC step by fixing flat priors: a g-prior for $\boldsymbol{\beta} \sim \mathcal{N}(\boldsymbol{\beta_0}, hB(X^TX)^{-1})$, with $h = 0.1$, and flat uniform priors in $(0, 10)$ for B, g and k, following Allingham et al. (2009).

The pilot simulation, in this case, has suggested to use two different thresholds when the true model is M_1 and when it is M_2, then we have used $\varepsilon_1 = (0.25, 0.10, 0.05)$ when M_1 is the true model and $\varepsilon_2 = (2.5, 1.0, 0.50)$ when the true model is M_2.

The results for 10^2 repetitions of the experiment are shown in Table 3.5. Again, the method seems to always find the true model, in particular it is very confident when the true model is the one which considers covariates, nevertheless it still has a good behavior when the model is the one which considers only the intercept.

One problem in applying ABC to model choice is the fact the Bayes factor approximated by ABC by using a statistic summarizing the data does not necessarily converge to the correct limit as n goes to infinity as explained in Section 3.3.1 and in Robert et al. (2011). A check of the validity of the method is, therefore, studying the behavior of the procedure presented in Section 3.3.1 when the sample size increases, as in Table 3.6, which shows that as n goes to infinity the probability of choosing the right model goes to 1 (the results in Table 3.6 are based on simulations

Table 3.6: Model choice in regression as the sample size increases

n	$\Pr(\mathcal{M} = m_1 \mid m_1)$	$\Pr(\mathcal{M} = m_2 \mid m_1)$
10	0.711	0.775
20	0.756	0.779
30	0.731	0.778
40	0.843	0.999
50	0.759	1.000
60	0.756	1.000
70	0.933	0.999
80	0.922	1.000
90	0.930	1.000
100	0.915	1.000
200	0.965	1.000
300	0.918	1.000
400	0.989	1.000
500	0.988	1.000
600	0.988	1.000
700	0.978	1.000
800	0.992	1.000
900	0.980	1.000
1000	0.984	1.000

from both the models $M_1 : A = \beta_0 + \beta_1 x_1$ and $M_2 : A = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$ where the covariates have been generated in the following way: $x_1 \sim Unif(2, 5)$, $x_2 \sim Unif(-1, 1)$ and $x_3 \sim Unif(0, 5)$).

Quantile distribution for log-returns

The method is now applied to a real data set, containing the log-returns of the Italian bank Monte dei Paschi di Siena (MPS) daily collected from 1 July 2013 to 30 June 2015 (only weekdays) available on the web page *https://it.finance.yahoo.com*. Data are shown in Figure 3.7.

We model the data using a g-and-k distribution, assuming that the data may be time-correlated and considering a MA(1)-type or a MA(2)-type autocorrelation structure. Other models can be considered. This means that the $z_i(p_i)$ follows a

Table 3.7: Log-returns: MA(1) vs MA(2)

ε	**0.050**	**0.025**	**0.010**
$\Pr(M = MA(1)\|\mathbf{q}(\mathbf{z}))$	0.531	0.535	0.536

MA(1) or a MA(2) model, i.e.

$$M_1 : z_i = \theta_i + \alpha\theta_{i-1} \quad for\, i = 1, \cdots, n$$
$$M_2 : z_i = \theta_i + \alpha_1\theta_{i-1} + \alpha_2\theta_{i-2} \quad for\, i = 1, \cdots, n$$

where n is the number of observations and $\theta_j \sim \mathcal{N}(0,1)$ for $j = 1, \cdots, n$. Each z_i is then divided by $\sqrt{1+\alpha^2}$ under M_1 and by $\sqrt{1+\alpha_1^2+\alpha_2^2}$ under M_2 to ensure it is marginally distributed as a $\mathcal{N}(0,1)$. Then each z_i is used to derive the respective quantile $q_i(z_i)$ of the corresponding quantile distribution as in (3.12).

The parameters have been simulated from their prior distibutions chosen as in Section 3.3.2 and the comparison between log-likelihoods has been chosen as summary statistic and computed via the arima function of the "e1071" R package.

Table 3.7 shows the resulting posterior probabilities that the model for the data is a MA(1). We can see that there is a slight preference for the MA(1), confirmed for all the three tolerance levels considered.

3.3.3 Conclusion

The Chapter has proposed a new method to solve problems of model choice when the considered models have a complicated structure. This method is based on the approximate Bayesian methodology which implies the choice of statistics (sufficient or not) summarizing the data. The key point in model choice with ABC is that both models have to be compatible with the summary statistics in the sense of Marin et al. (2014).

Our proposal is to consider as summary statistic the Bayes factor or other quantities traditionally used to compare models (as AIC, BIC, log-likelihoods, etc.) for approximated and simplified versions of the available models; for example, if the data may be modeled by a normal distribution in first approximation, one can compute the Bayes factor for normal hypotheses on the observed data and then compare it with the Bayes factor for normal hypotheses computed on the data simulated by

using the model under analysis. Even if the Bayes factor computed in such a way is not correct, it could be informative for the problem of model choice.

We have shown through simulations that the proposed method has a good behavior in choosing the right model, with more and more accuracy as the sample size increases. We have also applied the method on a financial data set and compared moving average models. In this case, it is possible to use an approximated Bayes factor as in Koop and Potter (1999) or to use some other quantities informative for model choice, as the log-likelihood values for moving average models, available in many R packages.

Every result based on simulations in the Chapter is obtained via repetitions of the experiments and shows a substantial stability (the standard errors, not shown throughout the Chapter, are always below 0.01 and the approximated confidence intervals never contain the value of indifference 0.5).

In this work, we have worked on quantile distributions, because it is easy to simulate from them and they are a classical example where methods other than ABC are difficult or even impossible to use. Nevertheless, other situations are still possible and may be investigated.

Moreover, the simulation from the prior distributions makes the method computationally demanding and other types of ABC algorithms, for example the ABC-MCMC described in Marjoram et al. (2014) may be investigated to reduce the computational time.

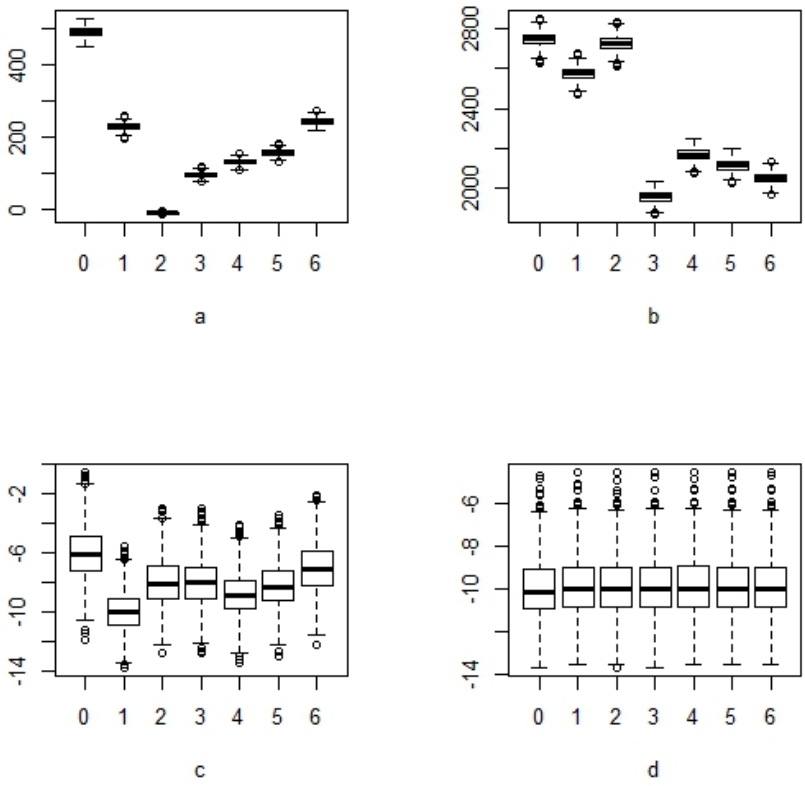

Figure 3.1: *Boxplots over 1,000 simulations of the sample distributions of the scores of seven models under analysis, depending on the true model. Model selection is performed in the case of nested linear models. The data was simulated from one of seven nested linear models with up to six covariates. The design matrix is denoted by* $\mathbf{X}$*. While* M_0 *is the model that uses zero covariate,* M_1 *to* M_6 *use the first, the first two, up to all of the covariates. The values of the covariates were simulated from normal proposals. The data* $\mathbf{y} = (y_1, ..., y_n)$ *have distribution* $\mathbf{y}|\boldsymbol{\theta} \sim N(\mathbf{X}\boldsymbol{\theta}, \sigma^2)$*, with* $n = 100$ *and* $\sigma^2 = 10$*. In (a) the true model used for the generations is* M_2 *(which considers two regressors), in (b) it is* M_3 *which considers the first three regressors, in (c) it is* M_1 *which considers the first regressor while in (d) the correct model is* M_0 *which considers only the constant.*

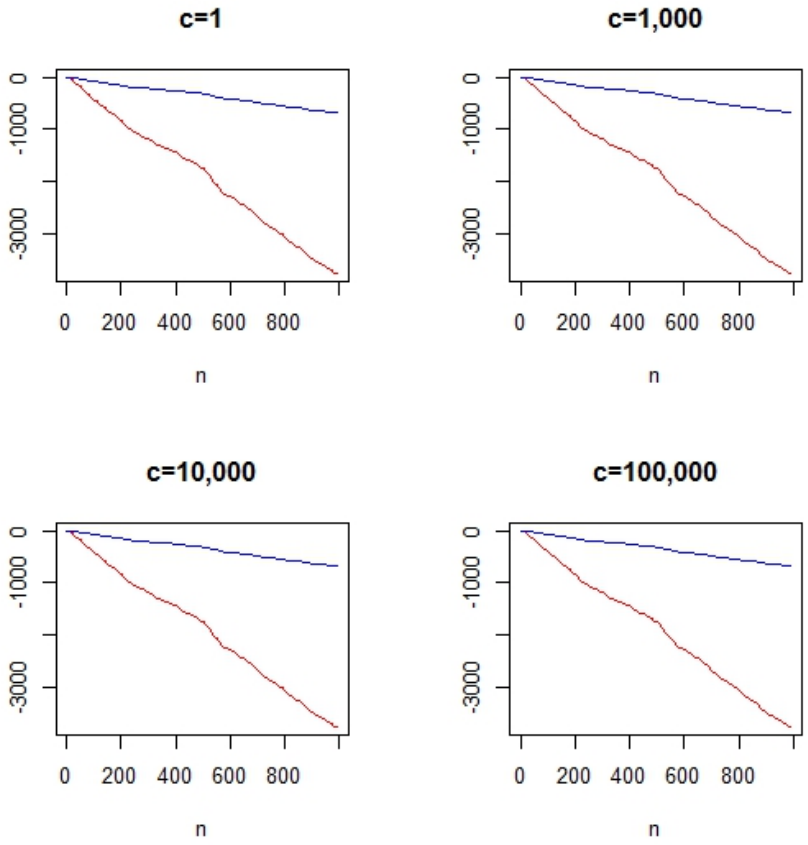

Figure 3.2: *Linear model: log-Bayes factor (red) and difference of the scores (blue) as a function of an increasing sample size* $n = 1, ..., 1000$*, and of the prior variance on* θ*,* $V = c\sigma^2$*, where* $\sigma^2 = 10$ *is known. Given simulated data* $\mathbf{y} = (y_1, ..., y_n)$ *with conditional distribution* $\mathbf{y}|\theta \sim N(\mathbf{X}\theta, \sigma^2)$ *we consider one regressor and two possible models for generating the data:* $M_0 : \theta = 0$ *and* $M_1 : \theta = 1$ *when the true model is* M_1*.*

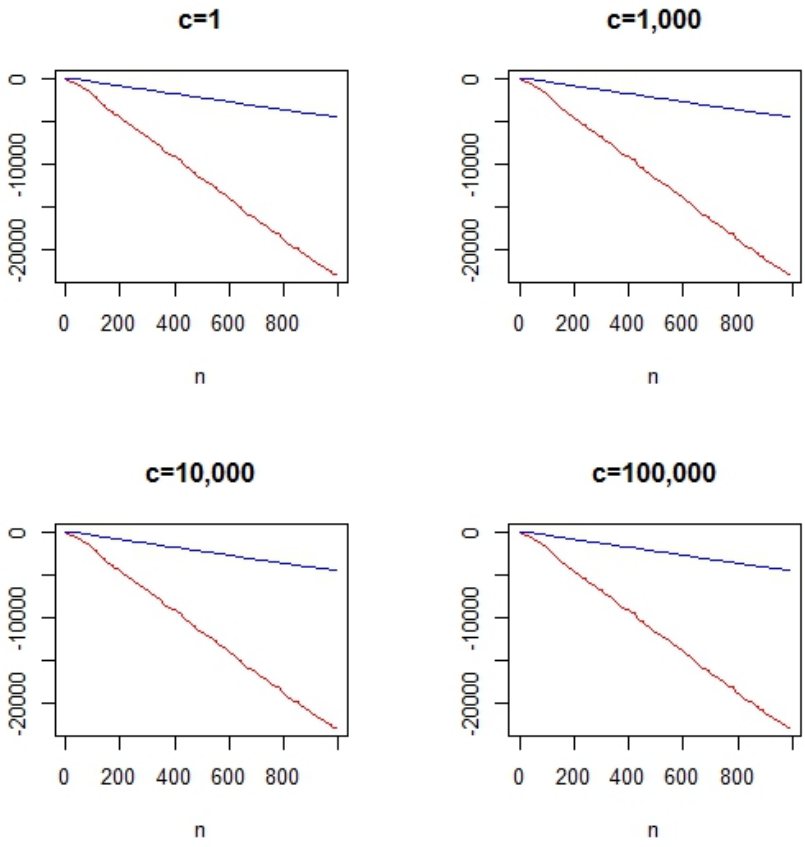

Figure 3.3: *Nested models: log-Bayes factor (red) and difference of the scores (blue) as a function of an increasing sample size* $n = 1, ..., 1000$*, and of the prior variance on* θ*,* $V = c\sigma^2$*, where* $\sigma^2 = 10$*. The setting is the same as Figure 3.1, where we consider six possible regressors and we compare model* M_3 *which considers the first three regressors against model* M_6 *which considers all the regressors (*M_3 *is the true model in our simulations).*

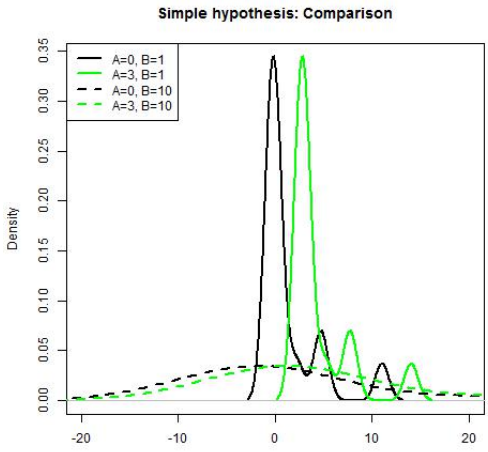

Figure 3.4: Comparison between two g-and-k densities with fixed $B = 1$ (solid lines) and $B = 10$ (dashed lines), $g = 2$ and $k = 0.5$ and $A_1 = 0$ (in black) and $A_2 = 3$ (in green).

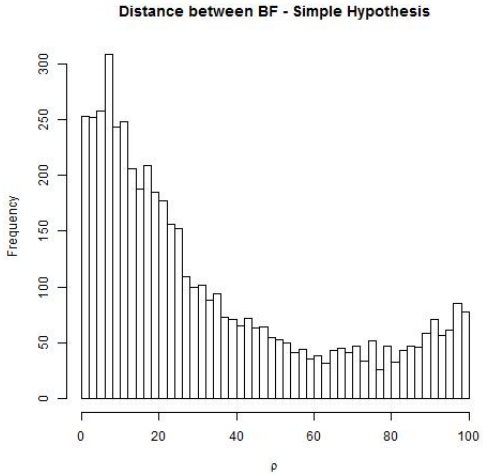

Figure 3.5: Distance ρ as in equation (3.16) for Algorithm 9 with a tolerance level $\varepsilon = 100$ and the system of hypotheses in (3.13).

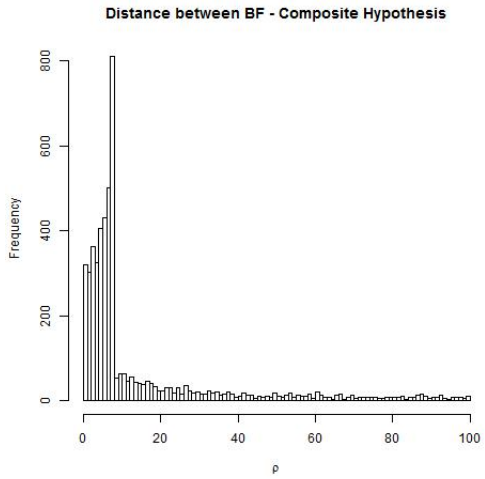

Figure 3.6: Distance ρ as in equation 3.16 for Algorithm 9 with a tolerance level $\varepsilon = 100$ and the system of hypotheses in (3.17).

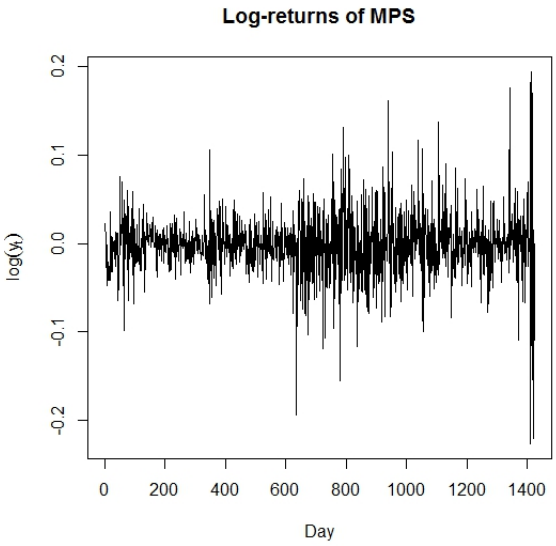

Figure 3.7: Daily log-returns of Monte dei Paschi di Siena between 1 July 2013 and 30 June 2015.

Chapter 4

Jeffreys prior for mixture estimation

4.1 Introduction

Bayesian inference in mixtures of distributions[1] has been studied quite extensively in the literature. See, e.g., MacLachlan and Peel (2000) and Frühwirth-Schnatter (2006) for book-long references and Lee et al. (2009) for one among many surveys. From a Bayesian perspective, one of the several difficulties with this type of distribution,

$$\sum_{i=1}^{k} p_i\, f(x|\theta_i)\,, \quad \sum_{i=1}^{k} p_i = 1\,, \tag{4.1}$$

is that its ill-defined nature (non-identifiability, multimodality, unbounded likelihood, etc.) leads to restrictive prior modelling since most improper priors are not acceptable. This is due in particular to the feature that a sample from (4.1) may contain no subset from one of the k components $f(\cdot|\theta_i)$ (see. e.g., Titterington et al. 1985). Albeit the probability of such an event is decreasing quickly to zero as the sample size grows, it nonetheless prevents the use of independent improper priors, unless such events are prohibited (Diebolt and Robert 1994). Similarly, the exchangeable nature of the components often induces both multimodality in the posterior distribution and convergence difficulties as exemplified by the *label switching* phenomenon that is now quite well-documented (Celeux et al. 2000, Stephens

[1] joint wort with Prof. Christian Robert, CEREMADE, Université Paris-Dauphine, CREST and University of Warwick

2000, Jasra et al. 2005, Frühwirth-Schnatter 2006, Geweke 2007, Puolamäki and Kaski 2009). This feature is characterized by a lack of symmetry in the outcome of a Monte Carlo Markov chain (MCMC) algorithm, in that the posterior density is exchangeable in the components of the mixture but the MCMC sample does not exhibit this symmetry. In addition, most MCMC samplers do not concentrate around a single mode of the posterior density, partly exploring several modes, which makes the construction of Bayes estimators of the components much harder.

When specifying a prior over the parameters of (4.1), it is therefore quite delicate to produce a manageable and sensible non-informative version and some have argued against using non-informative priors in this setting (for example, MacLachlan and Peel (2000) argue that it is impossible to obtain proper posterior distribution from fully noninformative priors), on the basis that mixture models were ill-defined objects that required informative priors to give a meaning to the notion of a component of (4.1). For instance, the distance between two components needs to be bounded from below to avoid repeating the same component over and over again. Alternatively, the components all need to be informed by the data, as exemplified in Diebolt and Robert (1994) who imposed a completion scheme (i.e., a joint model on both parameters and latent variables) such that all components were allocated at least two observations, thereby ensuring that the (truncated) posterior was well-defined. Wasserman (2000) proved ten years later that this truncation led to consistent estimators and moreover that only this type of priors could produce consistency. While the constraint on the allocations is not fully compatible with the i.i.d. representation of a mixture model, it naturally expresses a modelling requirement that all components have a meaning in terms of the data, namely that all components genuinely contributed to generating a part of the data. This translates as a form of weak prior information on how much one trusts the model and how meaningful each component is on its own (by opposition with the possibility of adding meaningless artificial extra-components with almost zero weights or almost identical parameters).

While we do not seek Jeffreys priors as the ultimate prior modelling for non-informative settings, being altogether convinced of the lack of unique reference priors (Robert 2001, Robert et al. 2009), we think it is nonetheless worthwile to study the performances of those priors in the setting of mixtures in order to determine if indeed they can provide a form of reference priors and if they are at least well-defined in such settings. We will show that only in very specific situations the Jeffreys prior

provides reasonable inference.

In Section 4.2 we provide a formal characterisation of properness of the posterior distribution for the parameters of a mixture model, in particular with Gaussian components, when a Jeffreys prior is used for them. In Section 4.3 we will analyze the properness of the Jeffreys prior and of the related posterior distribution: only when the weights of the components (which are defined in a compact space) are the only unknown parameters it turns out that the Jeffreys prior (and so the relative posterior) is proper; on the other hand, when the other parameters are unknown, the Jeffreys prior will be proved to be improper and in only one situation it provides a proper posterior distribution. In Section 4.4 we propose a way to realize a non-informative analysis of mixture models and introduce improper priors for at least some parameters. Section 4.6 concludes the Chapter.

4.2 Jeffreys priors for mixture models

We recall that the Jeffreys prior was introduced by Jeffreys (1939) as a default prior based on the Fisher information matrix

$$\pi^{\mathrm{J}}(\theta) \propto |I(\theta)|^{1/2},$$

whenever the later is well-defined; $I(\cdot)$ stand for the expected Fisher information matrix and the symbol $|\cdot|$ denotes the determinant. Although the prior is endowed with some frequentist properties like matching and asymptotic minimal information (Robert 2001, Chapter 3), it does not constitute the ultimate answer to the selection of prior distributions in non-informative settings and there exist many alternative such as reference priors (Berger et al. 2009), maximum entropy priors (Rissanen 2012), matching priors (Ghosh et al. 1995), and other proposals (Kass and Wasserman 1996). In most settings Jeffreys priors are improper, which may explain for their conspicuous absence in the domain of mixture estimation, since the latter prohibits the use of most improper priors by allowing any subset of components to go "empty" with positive probability. That is, the likelihood of a mixture model can always be decomposed as a sum over all possible partitions of the data into k groups at most, where k is the number of components of the mixture. This means that there are terms in this sum where no observation from the sample brings any amount of information about the parameters of a specific component.

Approximations of the Jeffreys prior in the setting of mixtures can be found, e.g., in Figueiredo and Jain (2002), where the Authors revert to independent Jeffreys priors on the components of the mixture. This induces the same negative side-effect as with other independent priors, namely an impossibility to handle improper priors.

Rubio and Steel (2014) provide a closed-form expression for the Jeffreys prior for a location-scale mixture with two components. The family of distributions they consider is

$$\frac{2\epsilon}{\sigma_1} f\left(\frac{x-\mu}{\sigma_1}\right) \mathbb{I}_{x<\mu} + \frac{2(1-\epsilon)}{\sigma_2} f\left(\frac{x-\mu}{\sigma_2}\right) \mathbb{I}_{x>\mu}$$

(which thus hardly qualifies as a mixture, due to the orthogonality in the supports of both components that allows to identify which component each observation is issued from). The factor 2 in the fraction is due to the assumption of symmetry around zero for the density f. For this specific model, if we impose that the weight ϵ is a function of the variance parameters, $\epsilon = {}^{\sigma_1}/_{\sigma_1+\sigma_2}$, the Jeffreys prior is given by $\pi(\mu, \sigma_1, \sigma_2) \propto {}^{1}/_{\sigma_1\sigma_2\{\sigma_1+\sigma_2\}}$. However, in this setting, Rubio and Steel (2014) demonstrate that the posterior associated with the (regular) Jeffreys prior is improper, hence not relevant for conducting inference. (One may wonder at the pertinence of a Fisher information in this model, given that the likelihood is not differentiable in μ.) Rubio and Steel (2014) also consider alternatives to the genuine Jeffreys prior, either by reducing the range or even the number of parameters, or by building a product of conditional priors. They further consider so-called non-objective priors that are only relevant to the specific case of the above mixture.

Another obvious explanation for the absence of Jeffreys priors is computational, namely the closed-form derivation of the Fisher information matrix is almost inevitably impossible. The reason is that integrals of the form

$$-\int_{\mathcal{X}} \frac{\partial^2 \log\left[\sum_{h=1}^{k} p_h\, f(x|\theta_h)\right]}{\partial\theta_i \partial\theta_j} \left[\sum_{h=1}^{k} p_h\, f(x|\theta_h)\right]^{-1} dx$$

(in the special case of component densities with a single parameter) cannot be computed analytically. We derive an approximation of the elements of the Fisher information matrix based on Riemann sums. The resulting computational expense is of order $\mathrm{O}(d^2)$ if d is the total number of (independent) parameters. Since the elements of the information matrix usually are ratios between the component densities and the mixture density, there may be difficulties with non-probabilistic methods of integration. Here, we use Riemann sums (with 550 points) when the component

standard deviations are sufficiently large, as they produce stable results, and Monte Carlo integration (with sample sizes of 1500) when they are small. In the latter case, the variability of MCMC results seems to decrease as σ_i approaches 0.

4.3 Properness for prior and posterior distributions

Unsurprisingly, most Jeffreys priors associated with mixture models are improper, the exception being when only the weights of the mixture are unknown, as already demonstrated in Bernardo and Giròn (1988).

We will characterize properness and improperness of Jeffreys priors and derived posteriors, when some or all of the parameters of distributions from location-scale families are unknown. These results are established both analytically and via simulations, with sufficiently large Monte Carlo experiments checking the behavior of the approximated posterior distribution.

4.3.1 Characterization of Jeffreys priors

Weights of mixture unknown

A representation of the Jeffreys prior and the derived posterior distribution for the weights of a three-component mixture model is given in Figure 4.1: the prior distribution is much more concentrated around extreme values in the support, i.e., it is a prior distribution conservative in the number of important components.

Lemma 4.3.1. *When the weights p_i are the only unknown parameters in (4.1), the corresponding Jeffreys prior is proper.*

Figure 4.2 shows the boxplots for the means of the approximated posterior distribution for the weights of a three-component Gaussian mixture model.

Proof. The generic element of the Fisher information matrix is (for $i, j = \{1, \ldots, k-1\}$)

$$\int_{\mathcal{X}} \frac{(f_i(x) - f_k(x))(f_j(x) - f_k(x))}{\sum_{l=1}^{k} p_l f_l(x)} dx \tag{4.2}$$

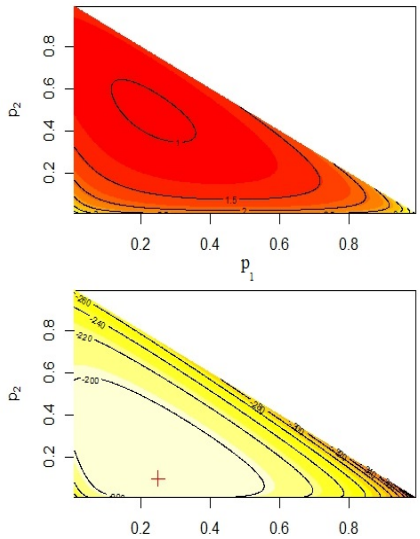

Figure 4.1: Approximations (on a grid of values) of the Jeffreys prior (on the log-scale) when only the weights of a Gaussian mixture model with three-components are unknown (on the top) and of the derived posterior distribution (with known means equal to -1, 0 and 2 respectively and known standard devitations equal to 1, 5 and 0.5 respectively). The red cross represents the true values.

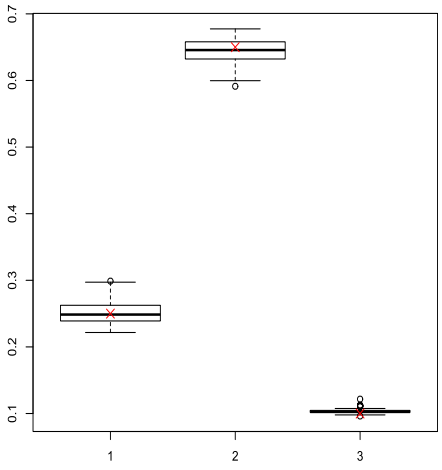

Figure 4.2: Boxplots of the estimated means of the three-component mixture model $0.25\mathcal{N}(-10,1) + 0.65\mathcal{N}(0,5) + 0.10\mathcal{N}(15,0.5)$ for 50 simulated samples of size 100, obtained via MCMC with 10^5 simulations. The red crosses represent the true values of the weights.

when we consider the parametrization in $(p_1, \ldots, p_{k-1})$, with

$$p_k = 1 - p_1 - \cdots - p_{k-1}\,.$$

We remind that, since the Fisher information matrix is a positive semi-definite, its determinant is bounded by the product of the terms in the diagonal, thanks to the Hadamard's inequality. Therefore, we may consider the diagonal term,

$$\begin{aligned}
\int_{\mathcal{X}} \frac{(f_i(x) - f_k(x))^2}{\sum_{l=1}^{k} p_l f_l(x)} dx &= \int_{f_i(x) \geq f_k(x)} \frac{(f_i(x) - f_k(x))^2}{\sum_{l=1}^{k} p_l f_l(x)} dx \\
&\quad + \int_{f_i(x) \leq f_k(x)} \frac{|(f_i(x) - f_k(x))^2|}{\sum_{l=1}^{k} p_l f_l(x)} dx \\
&= \int_{f_i(x) \geq f_k(x)} \frac{f_i(x) - f_k(x)}{\sum_{l=1}^{k} p_l f_l(x)} \{f_i(x) - f_k(x)\} dx \\
&\quad + \int_{f_i(x) \leq f_k(x)} \left| \frac{f_i(x) - f_k(x)}{\sum_{l=1}^{k} p_l f_l(x)} \right| |f_i(x) - f_k(x)| dx \\
&= \frac{1}{p_i} \int_{f_i \geq f_k} \frac{p_i \{f_i(x) - f_k(x)\}}{p_i \{f_i(x) - f_k(x)\} + \sum_{l \neq i,k} p_l \{f_l(x) - f_k(x)\} + f_k(x)} \\
&\qquad \{f_i(x) - f_k(x)\} dx \\
&\quad + \frac{1}{p_i} \int_{f_i \leq f_k} \left| \frac{p_i \{f_i(x) - f_k(x)\}}{p_i \{f_i(x) - f_k(x)\} + \sum_{l \neq i,k} p_l \{f_l(x) - f_k(x)\} + f_k(x)} \right| \\
&\qquad |f_i(x) - f_k(x)| dx \\
&\leq \frac{1}{p_i} \int_{f_i(x) \geq f_k(x)} \{f_i(x) - f_k(x)\} dx + \frac{1}{p_i} \int_{f_i(x) \leq f_k(x)} |f_i(x) - f_k(x)| dx \\
&= \frac{2}{p_i} \int_{f_i(x) \geq f_k(x)} \{f_i(x) - f_k(x)\} dx
\end{aligned}$$

where the last row is motivated by the fact that both integrals are equal.

Therefore, the Jeffreys prior will be bounded by the square root of the product of the terms in the diagonal of the Fisher information matrix

$$\pi^J(\mathbf{p}) \propto \prod_{i=1}^{k} p_i^{-\frac{1}{2}}$$

which is a generalization to k components of the prior provided in Bernardo and Giròn (1988) for $k = 2$ (however, Bernardo and Giròn (1988) find the reference

prior for the limiting case when all the components have pairwise disjoint supports, while for the opposite limiting case where all the components converge to the same distribution, the Jeffreys prior is the uniform distribution on the k-dimensional simplex). □

This reasoning leads Bernardo and Giròn (1988) to conclude that the usual $\mathcal{D}(\lambda_1, \ldots, \lambda_k)$ Dirichlet prior with $\lambda_i \in [1/2, 1]$ for $\forall\, i = 1, \cdots, k$ seems to be a reasonable approximation. They also prove that the Jeffreys prior for the weights p_i is convex, with an argument based on the sign of the second derivative.

As a remark, the configuration shown in proof of Lemma 4.3.1 is compatible with the Dirichlet configuration of the prior proposed by Rousseau and Mengersen (2011).

The shape of the Jeffreys prior for the weights of a mixture model depends on the type of the components. Figure 4.3, 4.4 and 4.5 show the form of the Jeffreys prior for a two-component mixture model for different choices of components. It is always concentrated around the extreme values of the support, however the amount of concentration around 0 or 1 depends on the information brought by each component. In particular, Figure 4.3 shows that the prior is much more symmetric as there is symmetry between the variances of the distribution components, while Figure 4.4 shows that the prior is much more concentrated around 1 for the weight relative to the normal component if the second component is a Student t distribution.

Finally Figure 4.5 shows the behavior of the Jeffreys prior when the first component is Gaussian and the second is a Student t and the number of degrees of freedom is increasing. As expected, as the Student t is approaching a normal distribution, the Jeffreys prior becomes more and more symmetric.

Location and scale parameters of a mixture model unknown

If the components of the mixture model (4.1) are distributions from a location-scale family and the location or scale parameters of the mixture components are unknown, this turns the mixture itself into a location-scale model. As a result, model (4.1) may be reparametrized by following Mengersen and Robert (1996), in the case of Gaussian components

$$p\mathcal{N}(\mu, \tau^2) + (1-p)\mathcal{N}(\mu + \tau\delta, \tau^2\sigma^2) \tag{4.3}$$

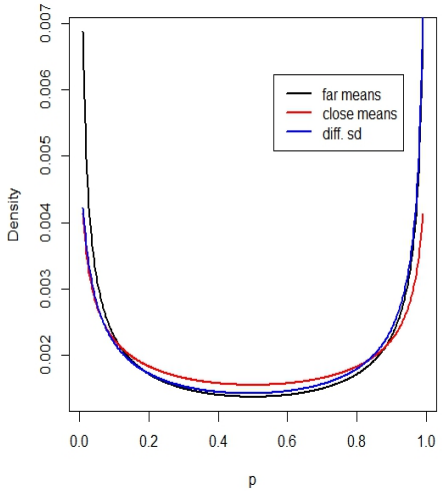

Figure 4.3: Approximations of the marginal prior distributions for the first weight of a two-component Gaussian mixture model, $p\mathcal{N}(-10,1)+(1-p)\mathcal{N}(10,1)$ (black), $p\mathcal{N}(-1,1)+(1-p)\mathcal{N}(1,1)$ (red) and $p\mathcal{N}(-10,1)+(1-p)\mathcal{N}(10,10)$ (blue).

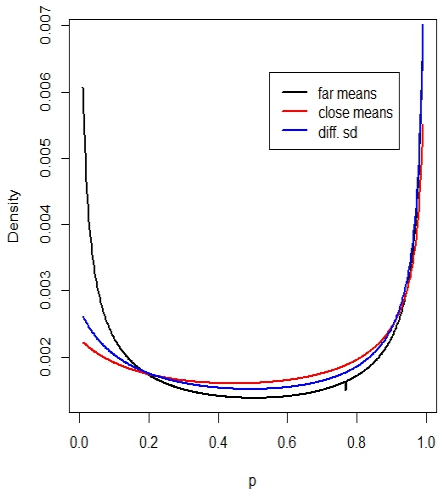

Figure 4.4: Approximations of the marginal prior distributions for the first weight of a two-component mixture model where the first component is Gaussian and the second is Student t, $p\mathcal{N}(-10,1)+(1-p)\,\mathrm{t}(df=1,10,1)$ (black), $p\mathcal{N}(-1,1)+(1-p)\,\mathrm{t}(df=1,1,1)$ (red) and $p\mathcal{N}(-10,1)+(1-p)\,\mathrm{t}(df=1,10,10)$ (blue).

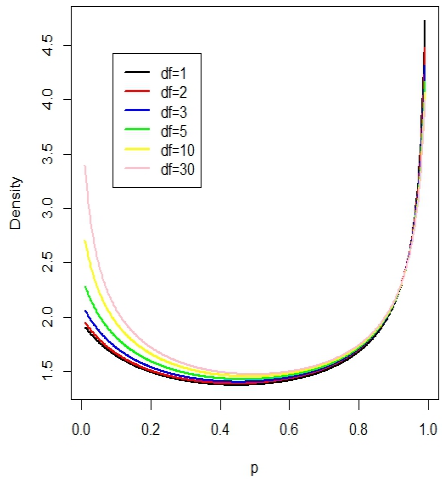

Figure 4.5: Approximations of the marginal prior distributions for the first weight of a two-component mixture model where the first component is Gaussian and the second is Student t with an increasing number of degrees of freedom.

namely using a reference location μ and a reference scale τ (which may be, for instance, the location and scale of a specific component). Equation (4.3) may be generalized to the case of k components as

$$p\mathcal{N}(\mu,\tau^2)+\sum_{i=1}^{k-2}(1-p)(1-q_1)\cdots(1-q_{i-1})q_i\mathcal{N}(\mu+\tau\theta_1+\cdots+\tau\cdots\sigma_{i-1}\theta_i,\tau^2\sigma_1^2\cdots\sigma_i^2)$$
$$+(1-p)(1-q_1)\cdots(1-q_{k-2})\mathcal{N}(\mu+\tau\theta_1+\cdots+\tau\cdots\sigma_{k-2}\theta_{k-1},\tau^2\sigma_1^2\cdots\sigma_{k-1}^2) \tag{4.4}$$

In this way, the mixture model is more clearly a location-scale model, which implies that the Jeffreys prior is flat in the location and powered as $\tau^{-d/2}$ if d is the total number of parameters of the components, respectively (Robert 2001, Chapter 3), as we will see in the following.

Lemma 4.3.2. *If the parameters of the components of a mixture model are either location or scale parameters, the corresponding Jeffreys prior is improper.*

In the proof of Lemma 4.3.2, we will consider a Gaussian mixture model and then extend the results to the general situation of components from a location-scale family.

Unknown location parameters

Proof. We first consider the case where the means are the only unknown parameters of a Gaussian mixture model

$$g_X(x) = \sum_{l=1}^{k} p_l \mathcal{N}(x|\mu_l, \sigma_l^2)$$

The generic elements of the expected Fisher information matrix are, in the case of diagonal and off-diagonal terms respectively:

$$\mathbb{E}\left[-\frac{\partial^2 \log g_X(X)}{\partial \mu_i^2}\right] = \frac{p_i^2}{\sigma_i^4} \int_{-\infty}^{\infty} \frac{[(x-\mu_i)\mathcal{N}(x|\mu_i, \sigma_i^2)]^2}{\sum_{l=1}^{k} p_l \mathcal{N}(x|\mu_l, \sigma_l^2)} dx$$

$$\mathbb{E}\left[-\frac{\partial^2 \log g_X(X)}{\partial \mu_i \partial \mu_j}\right] = \frac{p_i p_j}{\sigma_i^2 \sigma_j^2} \int_{-\infty}^{\infty} \frac{(x-\mu_i)\mathcal{N}(x|\mu_i, \sigma_i^2)(x-\mu_j)\mathcal{N}(x|\mu_j, \sigma_j^2)}{\sum_{l=1}^{k} p_l \mathcal{N}(x|\mu_l, \sigma_l^2)} dx$$

Now, consider the change of variable $t = x - \mu_i$ in the above integrals, where μ_i is thus the mean of the i-th Gaussian component ($i \in \{1, \cdots, k\}$). The above integrals are then equal to

$$\mathbb{E}\left[-\frac{\partial^2 \log g_X(X)}{\partial \mu_j^2}\right] = \frac{p_j^2}{\sigma_j^4} \int_{-\infty}^{\infty} \frac{[(t-\mu_j+\mu_i)\mathcal{N}(t|\mu_j-\mu_i, \sigma_i^2)]^2}{\sum_{l=1}^{k} p_l \mathcal{N}(t|\mu_l-\mu_i, \sigma_l^2)} dx$$

$$\mathbb{E}\left[-\frac{\partial^2 \log g_X(X)}{\partial \mu_j \partial \mu_m}\right] = \frac{p_j p_m}{\sigma_j^2 \sigma_m^2} \int_{-\infty}^{\infty} \frac{(t-\mu_j+\mu_i)\mathcal{N}(x|\mu_j, \sigma_j^2)(t-\mu_m+\mu_i)\mathcal{N}(t|\mu_m-\mu_i, \sigma_m^2)}{\sum_{l=1}^{k} p_l \mathcal{N}(t|\mu_l-\mu_i, \sigma_l^2)} dx$$

Therefore, the terms in the Fisher information only depend on the differences $\delta_j = \mu_i - \mu_j$ for $j \in \{1, \cdots, k\}$. This implies that the Jeffreys prior is improper since a reparametrization in (μ_i, δ) shows the prior does not depend on μ_i.

This feature will reappear whenever the location parameters are unknown.

When considering the general case of components from a location-scale family, this feature of improperness of the Jeffreys prior distribution is still valid, because, once reference location-scale parameters are chosen, the mixture model may be rewritten as

$$p_1 f_1(x|\mu,\tau) + \sum_{i=2}^{k} p_i f_i(\frac{a_i + x}{b_i}|\mu,\tau,a_i,b_i). \tag{4.5}$$

Then the second derivatives of the logarithm of model (4.5) behave as the ones we have derived for the Gaussian case, i.e. they will depend on the differences between each location parameter and the reference one, but not on the reference location itself. Then the Jeffreys prior will be constant with respect to the global location parameter.

□

When considering the reparametrization (4.3), the Jeffreys prior for δ for a fix μ has the form:

$$\pi^J(\delta|\mu) \propto \left[\int_{\mathfrak{X}} \frac{\left[(1-p)x\exp\{-\frac{x^2}{2}\}\right]^2}{p\sigma\exp\{-\frac{\sigma^2(x+\frac{\delta}{\sigma\tau})^2}{2}\} + (1-p)\exp\{-\frac{x^2}{2}\}}dx\right]^{\frac{1}{2}}$$

and the following result may be demonstrated.

Lemma 4.3.3. *The Jeffreys prior of δ conditional on μ when only the location parameters are unknown is improper.*

Proof. The improperness of the conditional Jeffreys prior on δ depends (up to a constant) on the double integral

$$\int_{\Delta}\int_{\mathfrak{X}} c\frac{\left[(1-p)x\exp\{-\frac{x^2}{2}\}\right]^2}{p\sigma\exp\{-\frac{\sigma^2(x+\frac{\delta}{\sigma\tau})^2}{2}\} + (1-p)\exp\{-\frac{x^2}{2}\}}dxd\delta.$$

The order of the integrals is allowed to be changed, then

$$\int_{\mathfrak{X}} x^2 \int_{\Delta} \frac{\left[(1-p)\exp\{-\frac{x^2}{2}\}\right]^2}{p\sigma\exp\{-\frac{\sigma^2(x+\frac{\delta}{\sigma\tau})^2}{2}\} + (1-p)\exp\{-\frac{x^2}{2}\}}d\delta dx.$$

Define $f(x) = (1-p)e^{-\frac{x^2}{2}} = \frac{1}{d}$. Then

$$\int_x x^2 \int_\Delta \frac{1}{d^2 p\sigma \exp\{-\frac{\sigma^2(x+\frac{\delta}{\sigma\tau})^2}{2}\} + d} d\delta dx.$$

Since the behavior of $\left[d^2 p\sigma \exp\{-\frac{\sigma^2(x+\frac{\delta}{\sigma\tau})^2}{2}\} + d\right]$ depends on $\exp\{-\delta^2\}$ as δ goes to ∞, we have that

$$\int_{-\infty}^{+\infty} \frac{1}{\exp\{-\delta^2\} + d} d\delta > \int_A^{+\infty} \frac{1}{\exp\{-\delta^2\} + d} d\delta$$

because the integrand function is positive. Therefore

$$\int_A^{+\infty} \frac{1}{\exp\{-\delta^2\} + d} d\delta > \int_A^{+\infty} \frac{1}{\varepsilon + d} d\delta = +\infty$$

i.e. the conditional Jeffreys prior on δ is improper.

□

Figure 4.6 compares the behavior of the prior and the resulting posterior distribution for the difference between the means of a two-component Gaussian mixture model: the prior distribution is symmetric and it has different behaviors depending on the value of the other parameters, but it always stabilizes for large enough values; the posterior distribution appears to always concentrate around the true value.

Unknown scale parameters

Consider now the second case of the scale parameters being the only unknown parameters.

Proof. First, consider a Gaussian mixture model and suppose the mixture model is composed by only two components; the Jeffreys prior for the scale parameters is defined as

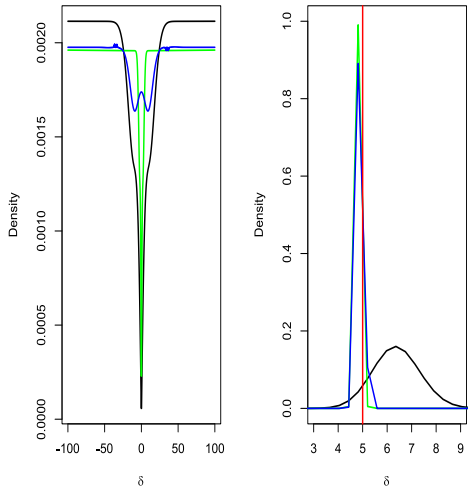

Figure 4.6: Approximations (on a grid of values) of the Jeffreys prior (on the natural scale) of the difference between the means of a Gaussian mixture model with only the means unknown (left) and of the derived posterior distribution (on the right, the red line represents the true value), with known weights equal to $(0.5, 0.5)$ (black lines), $(0.25, 0.75)$ (green and blue lines) and known standard deviations equal to $(5, 5)$ (black lines), $(1, 1)$ (green lines) and $(7, 1)$ (blue lines).

$$\pi^J(\sigma_1,\sigma_2) \propto \left\{ \frac{p_1^2}{\sigma_1^2} \int_{-\infty}^{\infty} \frac{\left[\left(\frac{(x-\mu_1)^2}{\sigma_1^2}-1\right)\mathcal{N}(x|\mu_1,\sigma_1^2)\right]^2}{\sum_{l=1}^{2} p_l \mathcal{N}(x|\mu_l,\sigma_l^2)} dx \right.$$
$$\cdot \frac{p_2^2}{\sigma_2^2} \int_{-\infty}^{\infty} \frac{\left[\left(\frac{(x-\mu_2)^2}{\sigma_2^2}-1\right)\mathcal{N}(x|\mu_2,\sigma_2^2)\right]^2}{\sum_{l=1}^{2} p_l \mathcal{N}(x|\mu_l,\sigma_l^2)} dx$$
$$\left. - \left[\frac{p_1 p_2}{\sigma_1 \sigma_2} \int_{-\infty}^{\infty} \frac{\left(\frac{(x-\mu_1)^2}{\sigma_1^2}-1\right)\left(\frac{(x-\mu_2)^2}{\sigma_2^2}-1\right)\mathcal{N}(x|\mu_1,\sigma_1^2)\mathcal{N}(x|\mu_2,\sigma_2^2)}{\sum_{l=1}^{2} p_l \mathcal{N}(x|\mu_l,\sigma_l^2)} dx \right]^2 \right\}^{\frac{1}{2}}$$

Since the Fisher information matrix is positive definite, it is bounded by the product on the diagonal, then we can write:

$$\pi^J(\sigma_1,\sigma_2) \leq c\frac{p_1p_2}{\sigma_1\sigma_2}\left\{\int_{-\infty}^{\infty}\frac{\left(\frac{(x-\mu_1)^2}{\sigma_1^2}-1\right)^2\frac{1}{\sigma_1^2}\exp\left\{-\frac{(x-\mu_1)^2}{\sigma_1^2}\right\}}{\frac{p_1}{\sigma_1}\exp\left\{-\frac{(x-\mu_1)^2}{2\sigma_1^2}\right\}+\frac{p_2}{\sigma_2}\exp\left\{-\frac{(x-\mu_2)^2}{2\sigma_2^2}\right\}}dx\right.$$
$$\left.\cdot\int_{-\infty}^{\infty}\frac{\left(\frac{(x-\mu_2)^2}{\sigma_2^2}-1\right)^2\frac{1}{\sigma_2^2}\exp\left\{-\frac{(x-\mu_2)^2}{\sigma_2^2}\right\}}{\frac{p_1}{\sigma_1}\exp\left\{-\frac{(x-\mu_1)^2}{2\sigma_1^2}\right\}+\frac{p_2}{\sigma_2}\exp\left\{-\frac{(x-\mu_2)^2}{2\sigma_2^2}\right\}}dx\right\}^{\frac{1}{2}}.$$

In particular, if we reparametrize the model by introducing $\sigma_1 = \tau$ and $\sigma_2 = \tau\sigma$ and study the behavior of the following integral

$$\int_0^{\infty}\int_0^{\infty}c\frac{p_1p_2}{\tau\sigma}\left\{\int_{-\infty}^{\infty}\frac{\left(z^2-1\right)^2\exp\left\{-z^2\right\}}{p_1\exp\left\{-\frac{z^2}{2}\right\}+\frac{p_2}{\sigma}\exp\left\{-\frac{(z\tau+\mu_1-\mu_2)^2}{2\tau^2\sigma^2}\right\}}dz\right.$$
$$\left.\cdot\left\{\int_{-\infty}^{\infty}\frac{\left(u^2-1\right)^2\exp\left\{-u^2\right\}}{p_1\sigma\exp\left\{-\frac{(u\tau\sigma+\mu_2-\mu_1)^2}{2\tau^2}\right\}+p_2\exp\left\{-\frac{u^2}{2}\right\}}du\right\}^{\frac{1}{2}}d\tau d\sigma \quad (4.6)$$

where the internal integrals with respect to z and u converge with respect to σ and τ, then the behavior of the external integrals only depends on $\frac{1}{\tau\sigma}$. Therefore they do not converge.

This proof can be easily extended to the case of k components: the behavior of the prior depends on the inverse of the product of the scale parameters, which implies that the prior is improper.

Moreover this proof may be easily extended to the general case of mixtures of location-scale distributions (4.5), because the second derivatives of the logarithm of the model will depend on factors b_i^{-2} for $i \in 1,\cdots,k$. When the square root is considered, it is evident that the integral will not converge. □

All the parameters unknown

When considering all the parameters unknown, the form of the Jeffreys prior may be partly defined by considering the mixture model as a location-scale model, for which a general solution exists; see Robert (2001).

Lemma 4.3.4. *When all the parameters of a Gaussian mixture model are unknown, the Jeffreys prior is constant in μ and powered as $\tau^{-d/2}$, where d is the total number of components parameters.*

Proof. We have already proved the Jeffreys prior is constant on the global mean (first proof of Lemma 4.3.2).

Consider a two-component mixture model and the reparametrization (4.3). With some computations, it is straightforward to derive the Fisher information matrix for this model, partly shown in Table 4.1, where each term is multiplied for a term which does not depend on τ.

Table 4.1: Factors depending on τ of the Fisher information matrix for the reparametrized model (4.3)

	σ	δ	$\mathbf{p}$	μ	τ
σ	1	1	1	τ^{-1}	τ^{-1}
δ	1	1	1	τ^{-1}	τ^{-1}
$\mathbf{p}$	1	1	1	τ^{-1}	τ^{-1}
μ	τ^{-1}	τ^{-1}	τ^{-1}	τ^{-2}	τ^{-2}
τ	τ^{-1}	τ^{-1}	τ^{-1}	τ^{-2}	τ^{-2}

Therefore, the Fisher information matrix considered as a function of τ is a block matrix. From well-known results in linear algebra, if we consider a block matrix

$$M = \begin{bmatrix} A & B \\ C & D \end{bmatrix}$$

then its determinant is given by $\det(M) = \det(A - BD^{-1}C)\det(D)$. In the case of a two-component mixture model, $\det(D) \propto \tau^{-4}$, while $\det(A - BD^{-1}C) \propto 1$ (always seen as functions of τ only). Then the Jeffreys prior for a two-component location-scale mixture model is proportional to τ^{-2} and, then, not convergent.

This result may be easily generalized to the case of k components.

□

4.3.2 Posterior distributions of Jeffreys priors

We now derive analytical and computational characterizations of the posterior distributions associated with Jeffreys priors for mixture models. Simulated examples are used to support the analytical results.

For this purpose, we have repeated simulations from the models

$$0.50\mathcal{N}(\mu_1, 1) + 0.50\mathcal{N}(\mu_2, 0.5) \tag{4.7}$$

and

$$0.25\mathcal{N}(\mu_1, 1) + 0.65\mathcal{N}(\mu_0, 0.5) + 0.10\mathcal{N}(\mu_2, 5) \tag{4.8}$$

where μ_1 and μ_2 are chosen to be either close ($\mu_1 = -1$, $\mu_2 = 2$) or well separated ($\mu_1 = -10$, $\mu_2 = 15$) and $\mu_0 = 0$.

The Tables shown in the following will analyze the behavior of simulated Markov chains with the goal to approximate the posterior distribution. Even if the output of an MCMC method is not conclusive to assess the properness of the target distribution, it may give a hint on improperness: if the target is improper, an MCMC chain cannot be positive recurrent but instead either null-recurrent or transient (Robert and Casella 2004), then it should show convergence problems, as trends or difficulties to move from a particular region. Therefore, simulation studies will be used to support analytical results on properness or improperness of the posterior distribution. In the following, we will say that the results are stable if they show a convergent behavior, i.e. they move around the true values which have generated the data. In particular, an approximation is stable if the proportion of experiments for which the chains show no trend and acceptance rates around the expected values (20%-40%, which means that there are not regions where the chain have difficulties to move from) is 0.

The following results are based on Gaussian mixture models, anyway, the Jeffreys prior has a behavior common to all the location-scale families, as shown in Section 4.3.1, as well as the likelihood function; therefore the results may be generalized to any location-scale family.

Location parameters unknown

A first numerical study where the Jeffreys prior and its posterior are computed on a grid of parameter values confirms that, provided the means only are unknown, the prior is constant on the difference between the means and takes higher and higher values as the difference between them increases. However, the posterior distribution is correctly concentrated around the true values for a sufficiently high sample size and it exhibits the classical bimodal nature of such posteriors (Celeux et al. 2000).

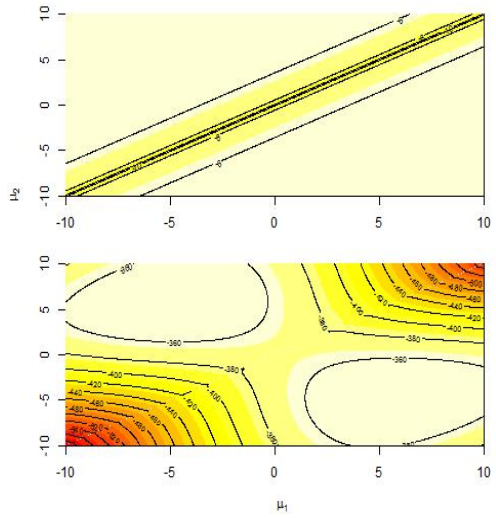

Figure 4.7: Approximations (on a grid of values) of the Jeffreys prior (on the log-scale) when only the means of a Gaussian mixture model with two components are unknown (on the top) and of the derived posterior distribution (with known weights both equal to 0.5 and known standard deviations both equal to 5).

In Figure 4.7, the posterior distribution appears to be perfectly symmetric because the other parameters (weights and standard deviations) have been fixed as identical.

Tables 4.2 and 4.3 show that, when considering a two-component Gaussian mixture model, the results are stabilizing for a sample size equal to 10 if the components are close and they are always stable if the means are far enough; on the other hand, huge sample sizes (around 100 observations) are needed to have always converging chains for a three-component mixture model (even if, when the components are well-separated a sample size equal to 10 seems to be enough to have stable results).

Lemma 4.3.5. *When $k = 2$, the posterior distribution derived from the Jeffreys prior when only the means are unknown is proper.*

Proof. The conditional Jeffreys prior for the means of a Gaussian mixture model is

$$\begin{aligned}\pi^J(\mu|p,\sigma) \propto \frac{p_1 p_2}{\sigma_1^2\sigma_2^2}\Bigg\{&\int_{-\infty}^{+\infty}\frac{[t\mathcal{N}(0,\sigma_1)]^2}{p_1\mathcal{N}(0,\sigma_1)+p_2\mathcal{N}(\delta,\sigma_2)}dt\\ &\times\int_{-\infty}^{+\infty}\frac{[u\mathcal{N}(0,\sigma_2)]^2}{p_1\mathcal{N}(-\delta,\sigma_1)+p_2\mathcal{N}(0,\sigma_2)}du\\ &-\left(\int_{-\infty}^{+\infty}\frac{t\mathcal{N}(0,\sigma_1)(t-\delta)\mathcal{N}(\delta,\sigma_2)}{p_1\mathcal{N}(0,\sigma_1)+p_2\mathcal{N}(\delta,\sigma_2)}dt\right)^2\Bigg\}^{\frac{1}{2}}\end{aligned}$$

where $\delta = \mu_2 - \mu_1$.

The posterior distribution is then defined as

$$\prod_{j=1}^{n} \left[p_1 \mathcal{N}(\mu_1, \sigma_1) + p_2 \mathcal{N}(\mu_2, \sigma_2)\right] \pi^J(\mu_1, \mu_2 | p, \sigma)$$

The likelihood may be rewritten (without loss of generality, by considering $\sigma_1 = \sigma_2 = 1$, since they are known) as

$$\begin{aligned} L(\theta) &= \prod_{j=1}^{n} \left[p_1 \mathcal{N}(\mu_1, 1) + p_2 \mathcal{N}(\mu_2, 1)\right] \\ &= \frac{1}{(2\pi)^{\frac{n}{2}}} \left[p_1^n e^{-\frac{1}{2}\sum_{i=1}^{n}(x_i-\mu_1)^2} + \sum_{j=1}^{n} p_1^{n-1} p_2 e^{-\frac{1}{2}\sum_{i\neq j}(x_i-\mu_1)^2 - \frac{1}{2}(x_j-\mu_2)^2} \right. \\ &+ \sum_{j=1}^{n}\sum_{k\neq j} p_1^{n-2} p_2^2 e^{-\frac{1}{2}\sum_{i\neq j,k}(x_i-\mu_1)^2 - \frac{1}{2}\left[(x_j-\mu_2)^2 + (x_k-\mu_2)^2\right]} \\ &\left. + \cdots + p_2^n e^{-\frac{1}{2}\sum_{j=1}^{n}(x_j-\mu_2)^2} \right] \end{aligned} \tag{4.9}$$

Then, for $|\mu_1| \to \infty$, $L(\theta)$ tends to the term $p_2^n e^{-\frac{1}{2}\sum_{j=1}^{n}(x_j-\mu_2)^2}$ that is constant for μ_1. Therefore we can study the behavior of the posterior distribution for this part of the likelihood to assess its properness.

This explains why we want the following integral to converge:

$$\int_{\mathbb{R}\times\mathbb{R}} p_2^n e^{-\frac{1}{2}\sum_{j=1}^{n}(x_j-\mu_2)^2} \pi^J(\mu_1, \mu_2) d\mu_1 d\mu_2$$

which is equal to (by the change of variable $\mu_2 - \mu_1 = \delta$)

$$\int_{\mathbb{R}\times\mathbb{R}} p_2^n e^{-\frac{1}{2}\sum_{j=1}^{n}(x_j-\mu_1-\delta)^2} \pi^J(\mu_1, \delta) d\mu_1 d\delta$$

We have seen that the prior distribution only depends on the difference between the means δ:

$$\begin{aligned} &\int_{\mathbb{R}} p_2^n \int_{\mathbb{R}} e^{-\frac{1}{2}\sum_{j=1}^{n}(x_j-\mu_1-\delta)^2} d\mu_1 \pi^J(\delta) d\delta \\ &\propto \int_{\mathbb{R}} \int_{\mathbb{R}} e^{-\frac{1}{2}\sum_{j=1}^{n}(x_j-\delta)^2 + \mu_1\sum_{j=1}^{n}(x_j-\delta) - \frac{1}{2}n\mu_1^2} d\mu_1 \pi^J(\delta) d\delta \\ &= \int_{\mathbb{R}} \left[\int_{\mathbb{R}} e^{\mu_1\sum_{j=1}^{n}(x_j-\delta) - \frac{1}{2}n\mu_1^2} d\mu_1\right] e^{-\frac{1}{2}\sum_{j=1}^{n}(x_j-\delta)^2} \pi^J(\delta) d\delta \\ &= \int_{\mathbb{R}} e^{-\frac{1}{2}\sum_{j=1}^{n}(x_j-\delta)^2 + \sum_{j=1}^{n}\frac{(x_j-\delta)}{2n}} \pi^J(\delta) d\delta \\ &\approx \int_{\mathbb{R}} e^{-\frac{n-1}{2}\delta^2} \pi^J(\delta) d\delta. \end{aligned} \tag{4.10}$$

The prior on δ depends on the determinant of the corresponding Fisher information matrix that is positive definite, then it is bounded by the product of the Fisher information matrix diagonal entries:

$$\pi(\delta) \leq \frac{p_1 p_2}{\sigma_1 \sigma_2} \left\{ \int_{-\infty}^{+\infty} \frac{\left[t\mathcal{N}(0,\sigma_1^2)\right]^2}{p_1\mathcal{N}(0,\sigma_1^2) + p_2\mathcal{N}(\delta,\sigma_2^2)} dt \times \int_{-\infty}^{+\infty} \frac{\left[u\mathcal{N}(0,\sigma_2^2)\right]^2}{p_1\mathcal{N}(-\delta,\sigma_1^2) + p_2\mathcal{N}(0,\sigma_2^2)} du \right\}^{\frac{1}{2}} \tag{4.11}$$

where we have used the proof of lemma 4.3.2 and a change of variable $(t-\delta) = u$ in the second integral. As $\delta \to \pm\infty$, this quantity is constant with respect to δ. Therefore the integral (4.10) is convergent for $n \geq 2$.

□

Unfortunately this result cannot be extended to the general case of k components.

Lemma 4.3.6. *When $k > 2$, the posterior distribution derived from the Jeffreys prior when only the means are unknown is improper.*

Proof. In the case of $k \neq 2$ components, the Jeffreys prior for the location parameters is still constant with respect to a reference mean (for example, μ_1). Therefore it depends on the difference parameters $(\delta_2 = \mu_2 - \mu_1, \delta_3 = \mu_3 - \mu_1, \cdots, \delta_k = \mu_k - \mu_1)$.

The Jeffreys prior will be bounded by the product on the diagonal, which is an extension of (4.11):

$$\pi^J(\delta_2, \cdots, \delta_k) \leq c \left\{ \int_{-\infty}^{\infty} \frac{[t\mathcal{N}(0,\sigma_1^2)]^2}{p_1\mathcal{N}(0,\sigma_1^2) + \cdots + p_k\mathcal{N}(\delta_k,\sigma_k^2)} dt \right.$$
$$\left. \cdots \int_{-\infty}^{\infty} \frac{[u\mathcal{N}(0,\sigma_k^2)]^2}{p_1\mathcal{N}(-\delta_k,\sigma_1^2) + \cdots + p_k\mathcal{N}(0,\sigma_k^2)} du \right\}^{\frac{1}{2}}.$$

If we consider the case as in Lemma 4.3.5, where only the part of the likelihood depending on e.g. μ_2 may be considered, the convergence of the following integral has to be studied:

$$\int_{\mathbb{R}} \cdots \int_{\mathbb{R}} e^{-\frac{n-1}{2}\delta_2^2} \pi^J(\delta_2, \cdots, \delta_k) d\delta_2 \cdots d\delta_k$$

In this case, however, the integral with respect to δ_2 may converge, nevertheless the integrals with respect to δ_j with $j \neq 2$ will diverge, since the prior tends to be constant for each δ_j as $|\delta_j| \to \infty$.

This results confirms the idea that each part of the likelihood gives information about at most the difference between the location of the respective components and the reference locations, but not on the locations of the other components.

□

Scale parameters unknown

Lemma 4.3.7. *The posterior distribution derived from the Jeffreys prior when only the standard deviations are unknown is improper.*

Proof. Consider equation (4.9) generalized to the case of σ_1 and σ_2 unknown: then when we integrate the posterior distribution with respect to σ_1 and σ_2, the complete integral may be split into several integrals then summed up. In particular, if we consider the first part of the likelihood (which only depends on the first component of the mixture) and use the change of variable used in (4.6), we have:

$$\int_0^\infty \int_0^\infty c\frac{p_1^n}{\tau^n}\frac{p_1 p_2}{\tau\sigma}\exp\left\{-\frac{1}{2\tau^2}\sum_{i=1}^{n}(x_i-\mu_1)^2\right\} \times \left\{\int_{-\infty}^{\infty}\frac{(z^2-1)^2\exp\{-z^2\}}{p_1\exp\left\{-\frac{z^2}{2}\right\}+\frac{p_2}{\sigma}\exp\left\{-\frac{(z\tau+\mu_1-\mu_2)^2}{2\tau^2\sigma^2}\right\}}dz \times \int_{-\infty}^{\infty}\frac{(u^2-1)^2\exp\{-u^2\}}{p_1\sigma\exp\left\{-\frac{(u\tau\sigma+\mu_2-\mu_1)^2}{2\tau^2}\right\}+p_2\exp\left\{-\frac{u^2}{2}\right\}}du\right\}^{\frac{1}{2}} d\tau d\sigma.$$

The integral with respect to τ in the previous equation converges, nevertheless the likelihood does not provide information for σ, then the integral with respect to σ diverges and the posterior will be improper.

This results may be easily extented to the case of k components: there is a part of the likelihood which only depends on the global scale parameter and is not informative for any other component; the form of the integral will remain the same, with integrations with respect to $\sigma_1, \sigma_2, \cdots, \sigma_{k-1}$ which do not converge.

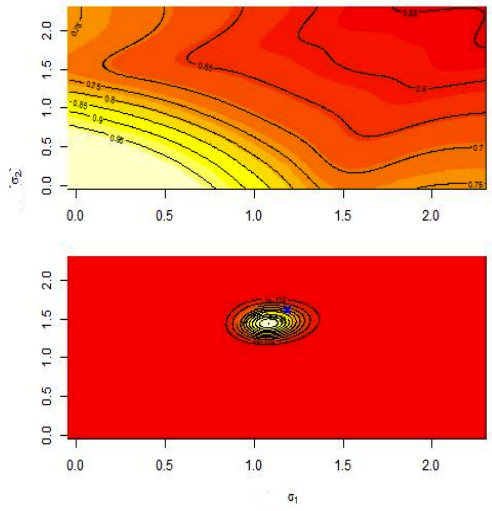

Figure 4.8: Approximations (on a grid of values) of the Jeffreys prior (on the log-scale) when only the standard deviations of a Gaussian mixture model with two components are unknown (on the top) and of the derived posterior distribution (with known weights both equal to 0.5 and known means equal to $(-5, 5)$). The blue cross represents the maximum likelihood estimates.

□

When only the standard deviations are unknown, the Jeffreys prior is concentrated around 0. The posterior distribution shown in Figures 4.8 turns out to be concentrated around the true values of the parameters only for a sufficient high sample size (in the figures, n is always equal to 100).

Figures 4.9 and 4.10 show the prior and the posterior distributions of the scale parameters of a two-component mixture model for some situations with different weights and different means.

Repeated simulations show that, for a Gaussian mixture model with two components, a sample size equal to 10 is necessary to have convergent results, while for a three-component Gaussian mixture model with a sample size equal to 50 it is still possible to have chains stuck to values of standard deviations close to 0.

Table 4.4 and 4.5 show results for repeated simulations in the cases of two-component and three-component Gaussian mixture models with unknown standard deviations, respectively, where the means that generate the data may be close or far from one another. In Table 4.4 it seems that the chains tend to be convergent for

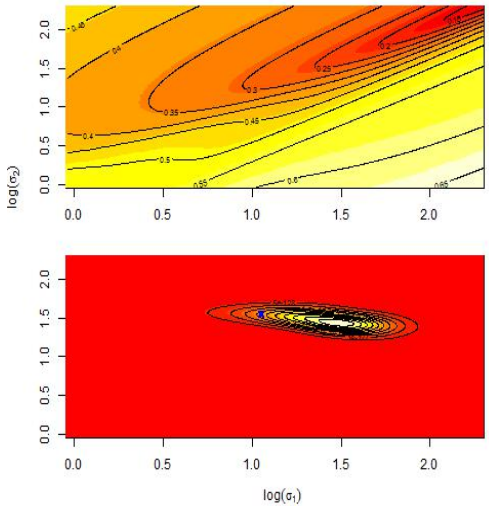

Figure 4.9: Same as Figure 4.8 but with known weights equal to $(0.25, 0.75)$ and known means equal to $(-1, 1)$.

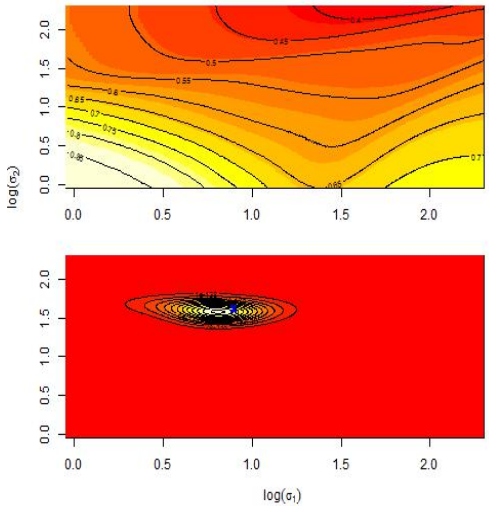

Figure 4.10: Same as Figure 4.8 but with known weights equal to $(0.25, 0.75)$ and known means equal to $(-2, 7)$.

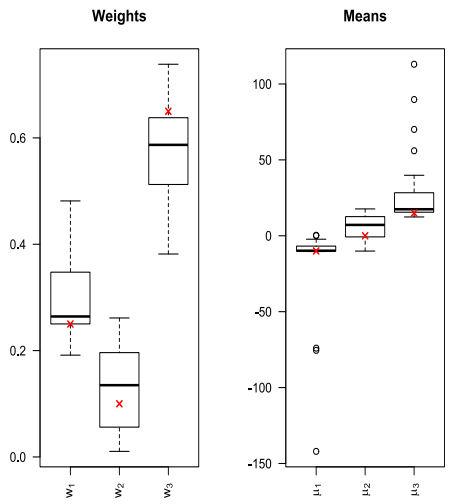

Figure 4.11: Boxplots of posterior means of the weigths and the means of the three-component mixture model $0.25\mathcal{N}(-1,1) + 0.65\mathcal{N}(0,0.5) + 0.10\mathcal{N}(2,5)$ for 50 replications of the experiment, obtained via MCMC with 10^5 simulations. The red cross represents the true value.

sample sizes smaller than 10, but in Table 4.5 one may see that even with a high sample size (equal to 50) it may happens, for $k = 3$, that the chains are stuck to very small values of standard deviations and this fact confirms what we have proved in Lemma 4.3.7.

Location and weight parameters unknown.

Figure 4.11 shows the boxplots of repeated simulations when both the weights and the means are unknown. It is evident that the posterior chains are concentrated around the true values, neverthless some chains (the 14% of the replications) show a drift to very high values (in absolute value) and this behavior suggests improperness of the posterior distribution.

All the parameters unknown

Improperness of the prior does not imply improperness of the posterior, obviously, but requires a careful checking of whether or not the posterior is proper, however the proof of Lemma 4.3.7 gives an hint about the actual properness of the posterior distribution when all the parameters are unknown.

Theorem 4.3.1. *The posterior distribution derived from the Jeffreys prior when all the parameters are unknown is improper.*

Proof. Consider the elements on the diagonal of the Fisher information matrix; again, since the Fisher information matrix is positive definite, the determinant is bounded by the product of the terms in the diagonal.

Consider a reparametrization into $\tau = \sigma_1$ and $\tau\sigma = \sigma_2$. Then it is straightforward to see that the integral of this part of the prior distribution will depend on a term $(\tau)^{-(d+1)}(\sigma)^{-d}$. Again, as in the proof of Lemma 4.3.7, when composing the prior with the part of the likelihood which only depends on the first component, this part does not provide information about the parameters σ and the integral will diverge.

In particular, the integral of the first part of the posterior distribution relative to the part of the likelihood dependent on the first component only and on the product of the diagonal terms of the Fisher information matrix for the prior when considering a two-component mixture model is

$$\begin{aligned}
&\int_0^1 \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} \int_0^{\infty} \int_0^{\infty} c\frac{p_1^n}{\tau^n}\frac{p_1^2 p_2^2}{\tau^3\sigma^2} \exp\left\{-\frac{1}{2\tau^2}\sum_{i=1}^{n}(x_i-\mu_1)^2\right\} \\
&\quad\times \left\{ \int_{-\infty}^{\infty} \frac{\left[\sigma \exp\left\{-\frac{(\tau\sigma y+\delta)^2}{2\tau^2}\right\} - \exp\left\{-\frac{y^2}{2}\right\}\right]^2}{p_1\sigma \exp\left\{-\frac{(\tau\sigma y+\delta)^2}{2\tau^2}\right\} + p_2 \exp\left\{-\frac{y^2}{2}\right\}} dy \right. \\
&\quad\times \int_{-\infty}^{\infty} \frac{z^2\exp(-z^2)}{p_1 \exp\left\{-\frac{z^2}{2}\right\} + \frac{p_2}{\sigma}\exp\left\{-\frac{(z\tau-\delta)^2}{2\tau^2\sigma^2}\right\}} dz \\
&\quad\times \int_{-\infty}^{\infty} \frac{w^2\exp\{-w^2\}}{p_1\sigma \exp\left\{-\frac{(\tau\sigma w+\delta)^2}{2\tau^2\sigma^2}\right\} + p_2 \exp\left\{-\frac{w^2}{2}\right\}} dw \\
&\quad\times \int_{-\infty}^{\infty} \frac{(z^2-1)^2\exp\{-z^2\}}{p_1 \exp\left\{-\frac{z^2}{2}\right\} + \frac{p_2}{\sigma}\exp\left\{-\frac{(z\tau+\mu_1-\mu_2)^2}{2\tau^2\sigma^2}\right\}} dz \\
&\quad\times \left. \int_{-\infty}^{\infty} \frac{(u^2-1)^2\exp\{-u^2\}}{p_1\sigma \exp\left\{-\frac{(u\tau\sigma+\mu_2-\mu_1)^2}{2\tau^2}\right\} + p_2 \exp\left\{-\frac{u^2}{2}\right\}} du \right\}^{\frac{1}{2}} d\tau d\sigma d\mu_1 d\mu_2 dp_1.
\end{aligned}$$

When considering the integrals relative to the Jeffreys prior, they do not represent an issue for convergence with respect to the scale parameters, because exponential terms going to 0 as the scale parameters tend to 0 are present. However, when considering the part out of the previous integrals, a factor σ^-2 whose behavior is

not convergent is present. Then this particular part of the posterior distribution is not integrating.

When considering the case of k components, the integral will always inversily depends on $\sigma_1, \sigma_2, \cdots, \sigma_{k-1}$ and then the posterior will always be improper.

As a note aside, it is worth noting that the usual separation between parameters proposed by Jeffreys himself in multidimensional problems does not change the behavior of the posterior, because even if the Fisher information matrix is decomposed as

$$I(\theta) = \begin{pmatrix} I_1(\theta_1) & 0 \\ 0 & I_2(\theta_2) \end{pmatrix}$$

for any possible combination of the parameters $\theta = (p, \mu_1, \mu_2, \sigma_1, \sigma_2)$ (note that θ_1 and θ_2 are vectors and $I(\theta_1)$ and $I(\theta_2)$ are diagonal or non-diagonal matrices), the product of the elements in the diagonal (considered in the proof) will be the same.

□

For small sample sizes, the chains tend to get stuck when very small values of standard deviations are accepted. Table 4.6 and 4.7 show the results for different sample sizes and different scenarios (in particular, the situations when the means are close or far from each other are considered) for a mixture model with two and three components respectively. The second and the third columns show the reason why the chain goes into trouble: sometimes the chains do not converge and tend towards very high values of means, sometimes the chains get stuck to very small values of standard deviations.

Since the improperness of the posterior distribution is mainly due to the scale parameters, we may use a reparametrization of the problem as in (4.3) and use a proper prior on the parameter σ, for example, by following Robert and Mengersen (1999)

$$p(\sigma) = \frac{1}{2}\mathcal{U}_{[0,1]}(\sigma) + \frac{1}{2}\frac{1}{\mathcal{U}_{[0,1]}(\sigma)}$$

and the Jeffreys prior for all the other parameters $(\mathbf{p}, \mu, \delta, \tau)$ conditionally on σ.

Actually, using a proper prior on σ does not avoid convergence trouble, as demonstrated by Table 4.8, which shows that, even if the chains with respect to the stan-

dard deviations are not stuck around 0 when using a proper prior for σ in the reparametrization proposed by Robert and Mengersen (1999), the chains with respect to the locations parameters demonstrate a divergent behavior.

4.4 A noninformative alternative to Jeffreys prior

The information brought by the Jeffreys prior does not seem to be enough to conduct inference in the case of mixture models. The computation of the determinant creates a dependence between the elements of the Fisher information matrix in the definition of the prior distribution which makes it difficult to find slight modifications of this prior that would lead to a proper posterior distribution. For example, using a proper prior for part of the scale parameters and the Jeffreys prior conditionally on them does not avoid impropriety, as we have demonstrated in Section 4.3.2.

The literature covers attempts to define priors which add a small amount of information that is sufficient to conduct the statistical analysis without overwhelming the information contained in the data. Some of these are related to the computational issues in estimating the parameters of mixture models, as in the approach of Casella et al. (2002), who find a way to use perfect slice sampler by focusing on components in the exponential family and conjugate priors. A characteristic example is given by Richardson and Green (1997), who propose weakly informative priors, which are data-dependent (or empirical Bayes) and are represented by flat normal priors over an interval corresponding to the range of the data. Nevertheless, since mixture models belong to the class of ill-posed problems, the influence of a proper prior over the resulting inference is difficult to assess.

Another solution found in Mengersen and Robert (1996) proceeds through the reparametrization (4.3) and introduces a reference component that allows for improper priors. This approach then envisions the other parameters as departures from the reference and ties them together by considering each parameter θ_i as a perturbation of the parameter of the previous component θ_{i-1}. This perspective is justified by the fact that the $(i-1)$-th component is not informative enough to absorb all the variability in the data. For instance, a three-component mixture model

gets rewritten as

$$p\mathcal{N}(\mu,\tau^2)+(1-p)q\mathcal{N}(\mu+\tau\theta,\tau^2\sigma_1^2)\\ +(1-p)(1-q)\mathcal{N}(\mu+\tau\theta+\tau\sigma\epsilon,\tau^2\sigma_1^2\sigma_2^2)$$

where one can impose the constraint $1 \geq \sigma_1 \geq \sigma_2$ for identifiability reasons. Under this representation, it is possible to use an improper prior on the global location-scale parameter (μ, τ), while proper priors must be applied to the remaining parameters. This reparametrization has been used also for exponential components by Gruet et al. (1999) and Poisson components by Robert and Titterington (1998). Moreover, Roeder and Wasserman (1997) propose a Markov prior which follows the same resoning of dependence between the parameters for Gaussian components, where each parameter is again a perturbation of the parameter of the previous component θ_{i-1}.

This representation suggests to define a global location-scale parameter in a more implicit way, via a hierarchical model that considers more levels in the analysis and choose noninformative priors at the last level in the hierarchy.

More precisely, consider the Gaussian mixture model (4.1)

$$g(x|\boldsymbol{\theta}) = \sum_{i=1}^{K} p_i \mathcal{N}(x|\mu_i, \sigma_i). \tag{4.12}$$

The parameters of each component may be considered as related in some way; for example, the observations have a reasonable range, which makes it highly improbable to face very different means in the above Gaussian mixture model. A similar argument may be used for the standard deviations.

Therefore, at the second level of the hierarchical model, we may write

$$\begin{aligned}\mu_i &\overset{iid}{\sim} \mathcal{N}(\mu_0, \zeta_0)\\ \sigma_i &\overset{iid}{\sim} \frac{1}{2}\mathcal{U}(0,\zeta_0) + \frac{1}{2}\frac{1}{\mathcal{U}(0,\zeta_0)}\\ \mathbf{p} &\sim Dir\left(\frac{1}{2},\cdots,\frac{1}{2}\right)\end{aligned} \tag{4.13}$$

which indicates that the location parameters vary between components, but are likely to be close, and that the scale parameters may be lower or bigger than ζ_0, but

not exactly equal to ζ_0. The weights are given a Dirichlet prior (or in the case of just two components, a Beta prior) independently from the components' parameters.

At the third level of the hierarchical model, the prior may be noninformative:

$$\pi(\mu_0, \zeta_0) \propto \frac{1}{\zeta_0}. \tag{4.14}$$

As in Mengersen and Robert (1996) the parameters in the mixture model are considered tied together; on the other hand, this feature is not obtained via a representation of the mixture model itself, but via a hierarchy in the definition of the model and the parameters.

Theorem 4.4.1. *The posterior distribution derived from the hierarchical representation of the Gaussian mixture model associated with* (4.12), (4.13) *and* (4.14) *is proper.*

Proof. Consider the composition of the three levels of the hierarchical model described in equations (4.12), (4.13) and (4.14):

$$\begin{aligned}\pi(\boldsymbol{\mu}, \boldsymbol{\sigma}, \mu_0, \zeta_0; \mathbf{x}) \propto\ & L(\mu_1, \mu_2, \sigma_1, \sigma_2; \mathbf{x}) p^{-1/2}(1-p)^{-1/2} \\ & \times \frac{1}{\zeta_0} \frac{1}{2\pi\zeta_0^2} \exp\left\{-\frac{(\mu_1-\mu_0)^2(\mu_2-\mu_0)^2}{2\zeta_0^2}\right\} \\ & \times \left[\frac{1}{2}\frac{1}{\zeta_0}\mathbb{I}_{[\sigma_1\in(0,\zeta_0)]}(\sigma_1) + \frac{1}{2}\frac{\zeta_0}{\sigma_1^2}\mathbb{I}_{[\sigma_1\in(\zeta_0,+\infty)]}(\sigma_1)\right] \\ & \times \left[\frac{1}{2}\frac{1}{\zeta_0}\mathbb{I}_{[\sigma_2\in(0,\zeta_0)]}(\sigma_2) + \frac{1}{2}\frac{\zeta_0}{\sigma_2^2}\mathbb{I}_{[\sigma_2\in(\zeta_0,+\infty)]}(\sigma_2)\right]\end{aligned} \tag{4.15}$$

where $L(\cdot; \mathbf{x})$ is given by equation (4.9).

Once again, we can initialize the proof by considering only the first term in the sum composing the likelihood function for the mixture model. Then the product in (4.15) may be split into four terms corresponding to the different terms in the scale parameters' prior. For instance, the first term is

$$\int_0^\infty \int_{-\infty}^\infty \int_{\mathbb{R}} \int_{\mathbb{R}} \int_{\mathbb{R}^+} \int_{\mathbb{R}^+} \int_0^1 \frac{1}{\sigma_1^n} p_1^n \exp\left\{-\frac{\sum_{i=1}^n (x_i-\mu_1)^2}{2\sigma_1^2}\right\} \times \frac{1}{\zeta_0^3} \exp\left\{-\frac{(\mu_1-\mu_0)^2(\mu_2-\mu_0)^2}{2\zeta_0^2}\right\} \times \frac{1}{4}\frac{1}{\zeta_0}\frac{1}{\zeta_0} \mathbb{I}_{[\sigma_1\in(0,\zeta_0)]}(\sigma_1)\mathbb{I}_{[\sigma_2\in(0,\zeta_0)]}(\sigma_2) dp d\sigma_1 d\sigma_2 d\mu_1 d\mu_2 d\mu_0 d\zeta_0$$

and the second one

$$\int_0^\infty \int_{-\infty}^\infty \int_{\mathbb{R}} \int_{\mathbb{R}} \int_{\mathbb{R}^+} \int_{\mathbb{R}^+} \int_0^1 \frac{1}{\sigma_1^n} p_1^n \exp\left\{-\frac{\sum_{i=1}^n (x_i-\mu_1)^2}{2\sigma_1^2}\right\} \times \frac{1}{\zeta_0^3} \exp\left\{-\frac{(\mu_1-\mu_0)^2(\mu_2-\mu_0)^2}{2\zeta_0^2}\right\} \times \frac{1}{4}\frac{1}{\zeta_0}\frac{\zeta_0}{\sigma_2^2} \mathbb{I}_{[\sigma_1\in(0,\zeta_0)]}(\sigma_1)\mathbb{I}_{[\sigma_2\in(\zeta_0,\infty)]}(\sigma_2) dp d\sigma_1 d\sigma_2 d\mu_1 d\mu_2 d\mu_0 d\zeta_0.$$

The integrals with respect to μ_1, μ_2 and μ_0 converge, since the data are carrying information about μ_0 through μ_1. The integral with respect to σ_1 converges as well, because, as $\sigma_1 \to 0$, the exponential function goes to 0 faster than $\frac{1}{\sigma_1^n}$ goes to ∞ (integrals where $\sigma_1 > \zeta_0$ are not considered here because this reasoning may easily extend to those cases). The integrals with respect to σ_2 converge, because they provide a factor proportional to ζ_0 and $1/\zeta_0$ respectively which simplifies with the normalizing constant of the reference distribution (the uniform in the first case and the Pareto in second one). Finally, the term $1/\zeta_0^4$ resulting from the previous operations has its counterpart in the integrals relative to the location priors. Therefore, the integral with respect to ζ_0 converges.

The part of the posterior distribution relative to the weights is not an issue, since the weights belong to the corresponding simplex. □

Table 4.9 shows the results given by simulation from the posterior distribution of the hierarchical mixture model and confirms that the chains always converge.

Figures 4.12–4.19 show the reslts of a simulation study to approximate the posterior distribution of the means of a two or three-component mixture model, compared to the true values (red vertical lines) and for different sample sizes, from $n = 3$ to $n = 1000$.

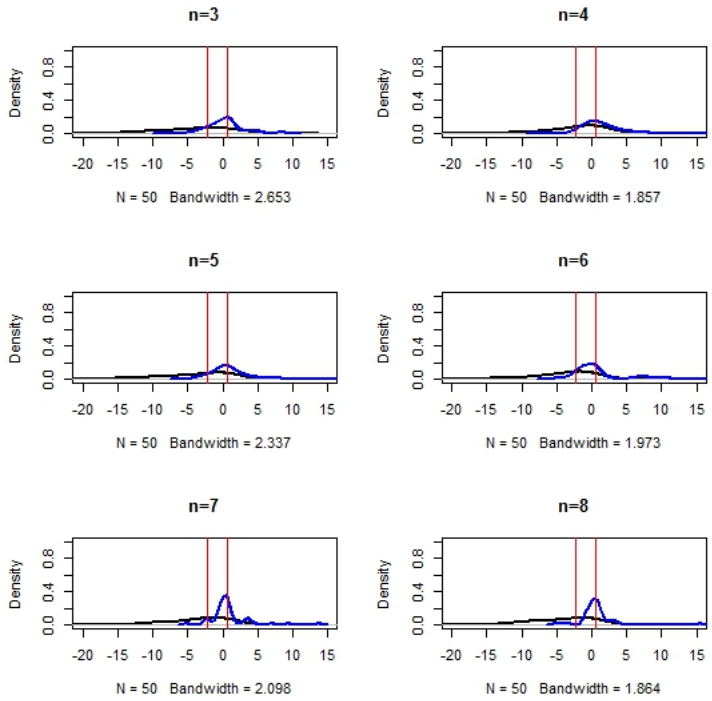

Figure 4.12: Distribution of the posterior means for the hierarchical mixture model with two components, global mean $\mu_0 = 0$ and global variance $\zeta_0 = 5$, based on 50 replications of the experiment with different sample sizes, black and blue lines for the marginal posterior distribution of μ_1 and μ_2 respectively.

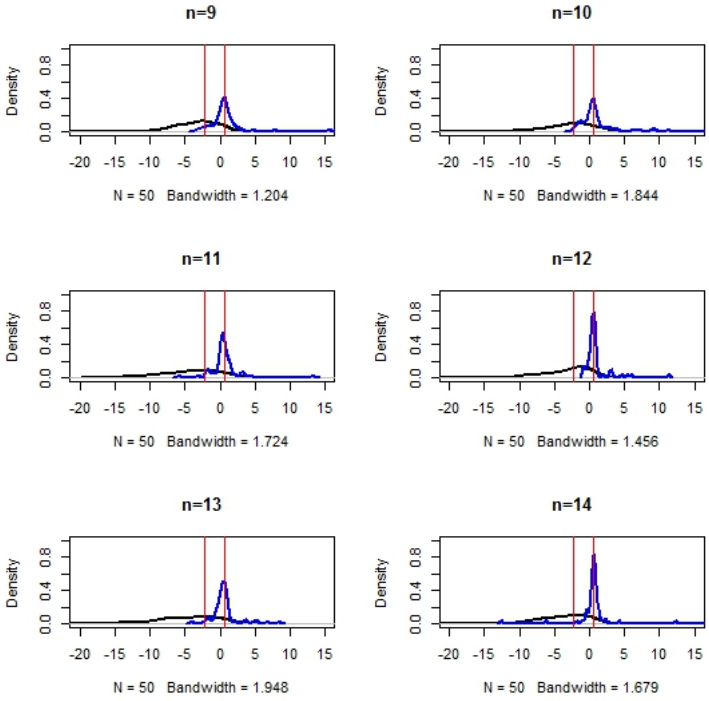

Figure 4.13: Same caption as in Figure 4.12.

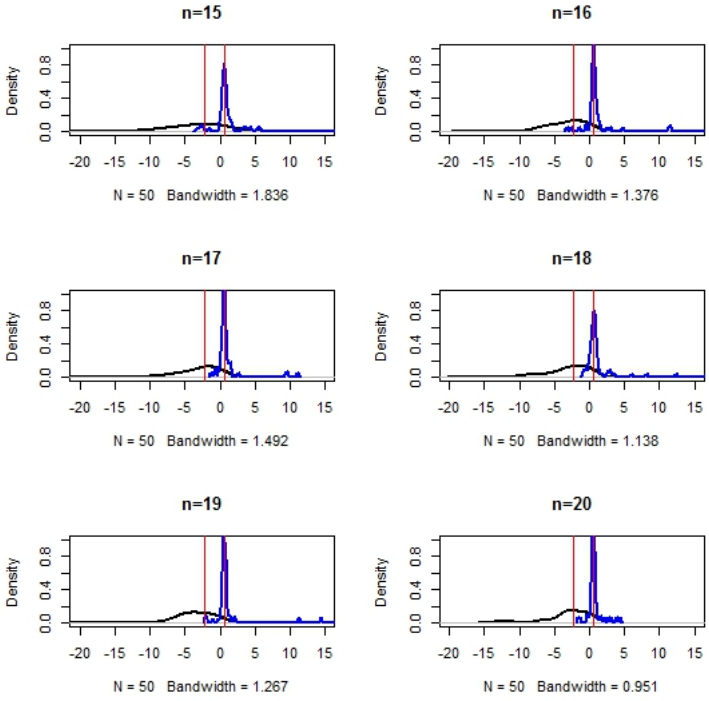

Figure 4.14: Same caption as in Figure 4.12.

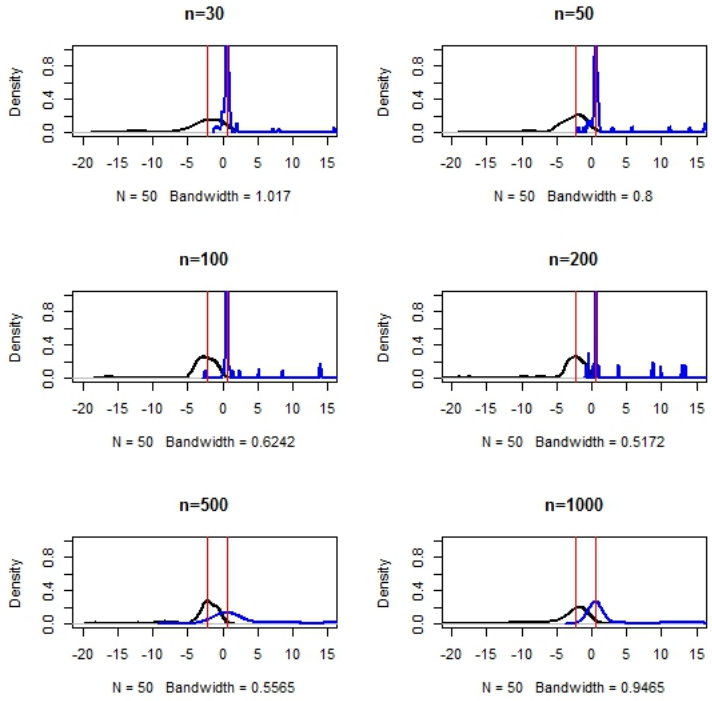

Figure 4.15: Same caption as in Figure 4.12.

4.5 Implementation features

The computing expense due to derive the Jeffreys prior for a set of parameter values is in $\mathcal{O}(d^2)$ if d is the total number of (independent) parameters.

Each element of the Fisher information matrix is an integral of the form

$$-\int_{\mathcal{X}} \frac{\partial^2 \log\left[\sum_{h=1}^{k} p_h\, f(x|\theta_h)\right]}{\partial\theta_i \partial\theta_j} \left[\sum_{h=1}^{k} p_h\, f(x|\theta_h)\right]^{-1} dx$$

which has to be approximated. We have applied both numerical integration and Monte Carlo integration and simulations show that, in general, numerical integration obtained via Gauss-Kronrod quadrature (see Piessens et al. (1983) for details), has more stable results. Neverthless, when one or more proposed values for the standard deviations or the weights is too small, the approximations tend to be very dependent on the bounds used for numerical integration (usually chosen to omit a negligible part of the density) or the numerical approximation may not be even applicable. In this case, Monte Carlo integration seems to have more stable, where the stability of the results depends on the Monte Carlo sample size.

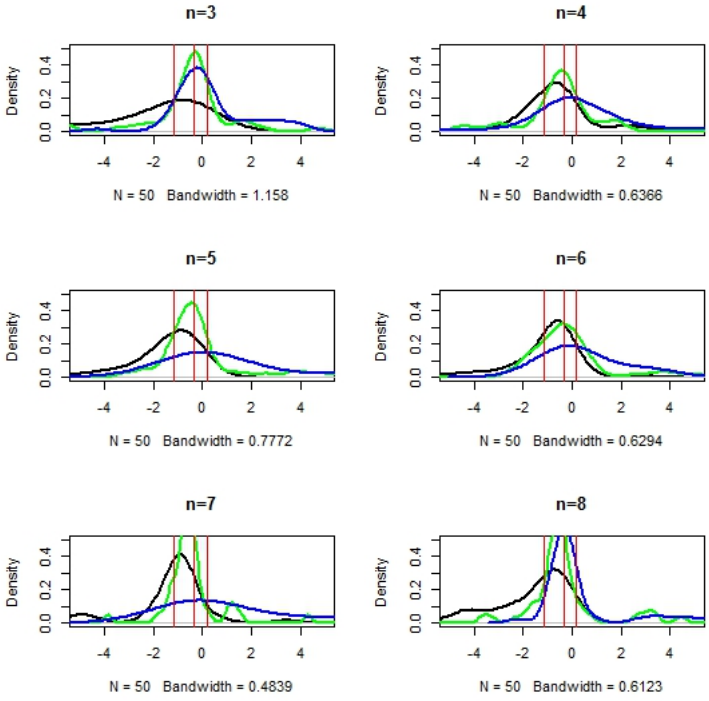

Figure 4.16: Distribution of the posterior means for the hierarchical mixture model with three components, global mean $\mu_0 = 0$ and global variance $\zeta_0 = 5$, based on 50 replications of the experiment with different sample sizes (the red lines stands for the true values, black, green and blue lines for the marginal posterior distributions of μ_1, μ_2 and μ_3 respectively).

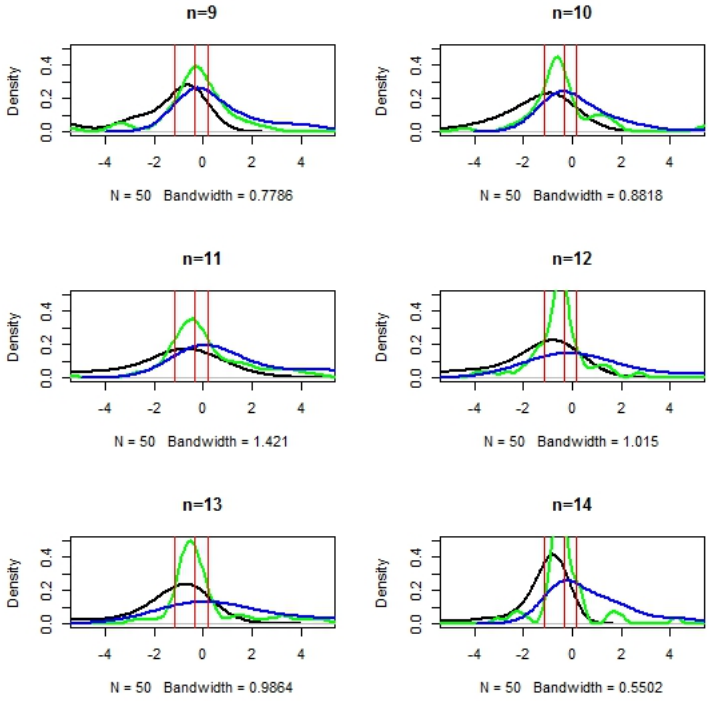

Figure 4.17: Same caption as in Figure 4.16.

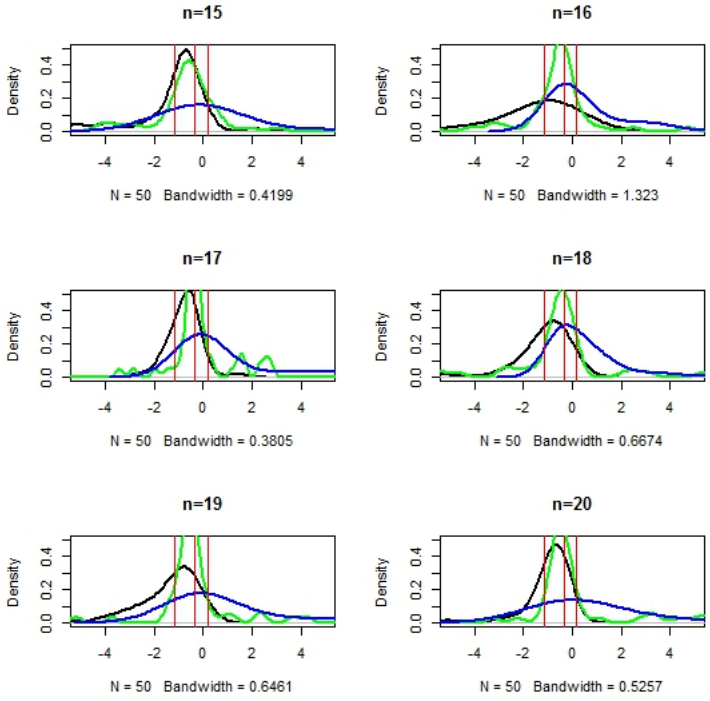

Figure 4.18: Same caption as in Figure 4.16.

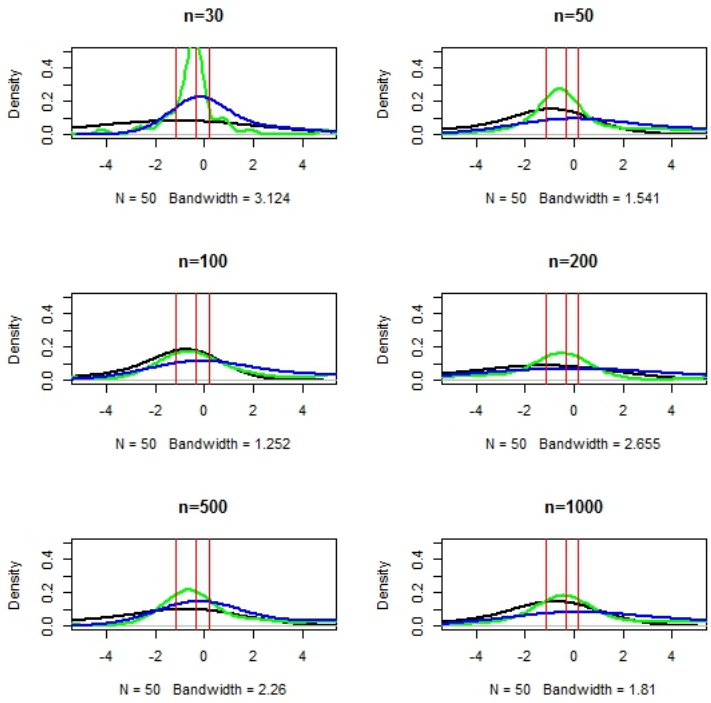

Figure 4.19: Same caption as in Figure 4.16.

Figure 4.20 shows the value of the Jeffreys prior obtained via Monte Carlo integration of the elements of the Fisher information matrix for an increasing number of Monte Carlo simulations both in the case where the Jeffreys prior is concentrated (where the standard deviations are small) and where it assumes low values. The value obtained via Monte Carlo integration is then compared with the value obtained via numerical integration. The sample size relative to the point where the graph stabilizes may be chosen to perform the approximation.

A similar analysis is shown in Figures 4.21 and 4.22 which provide the boxplots of 100 replications of the Monte Carlo approximations for different numbers of simulations (on the x-axis); one can choose to use the number of simulations which lead to a reasonable or acceptable variability of the results.

Since the approximation problem is one-dimensional, another numerical solution could be based on the sums of Riemann; Figure 4.23 shows the comparison between the results of the Gauss-Kronrod quadrature procedure and a procedure based on sums of Riemann for an increasing number of points considered in a region which contain the 99.999% of the data density. Moreover, Figure 4.24 shows the comparison between the approximation to the Jeffreys prior obtained via Monte Carlo integration and via the sums of Riemann: it is clear that the sums of Riemann lead

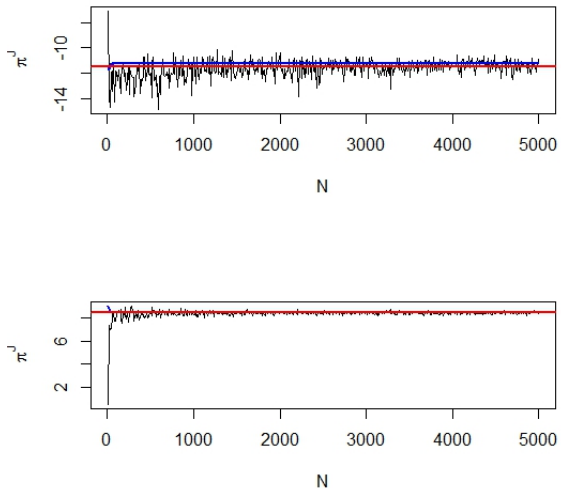

Figure 4.20: Jeffreys prior obtained via Monte Carlo integration (and numerical integration, in *red*) for the model $0.25\mathcal{N}(-10,1)+0.10\mathcal{N}(0,5)+0.65\mathcal{N}(15,7)$ (above) and for the model $\frac{1}{3}\mathcal{N}(-1,0.2)+\frac{1}{3}\mathcal{N}(0,0.2)+\frac{1}{3}\mathcal{N}(1,0.2)$ (below).

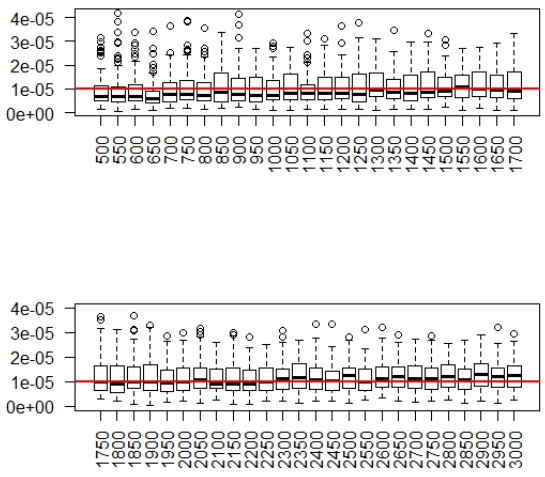

Figure 4.21: Boxplots of 100 replications of the procedure which approximates the Fisher information matrix via Monte Carlo integration to obtain the Jeffreys prior for the model $0.25\mathcal{N}(-10,1)+0.10\mathcal{N}(0,5)+0.65\mathcal{N}(15,7)$ for sample sizes from 500 to 3000. The value obtained via numerical integration is represented by the red line.

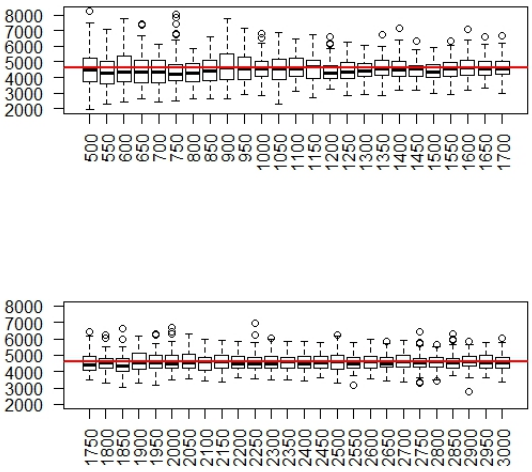

Figure 4.22: Same caption as in Figure 4.21 for the model $\frac{1}{3}\mathcal{N}(-1, 0.2)+\frac{1}{3}\mathcal{N}(0, 0.2)+\frac{1}{3}\mathcal{N}(1, 0.2)$.

to more stable results in comparison with Monte Carlo integration. On the other hand, they can be applied in more situations than the Gauss-Kromrod quadrature, in particular, in cases where the standard deviations are very small (of order 10^{-2}). Nevertheless, when the standard deviations are smaller than this, one has to pay attention on the features of the function to integrate. In fact, the mixture density tends to concentrate around the modes, with regions of density close to 0 between them. The elements of the Fisher informtation matrix are, in general, ratios between the components' densities and the mixture density, then in those regions an indeterminate form of type $\frac{0}{0}$ is obtained; Figure 4.25 represents the behavior of one of these elements when $\sigma_i \to 0$ for $i = 1, 2$.

Thus, we have decided to use the sums of Riemann (with a number of points equal to 550) to approximate the Jeffreys prior when the standard deviations are sufficiently large and Monte Carlo integration (with sample sizes of 1500) when they are too small. In this case, the variability of the results seems to decrease as σ_i approaches 0, as shown in Figure 4.26.

We have chosen to consider Monte Carlo samples of size equal to 1500 because both the value of the approximation and its standard deviations are stabilizing.

An adaptive MCMC algorithm has been used to define the variability of the kernel density functions used to propose the moves. During the burnin, the variability of

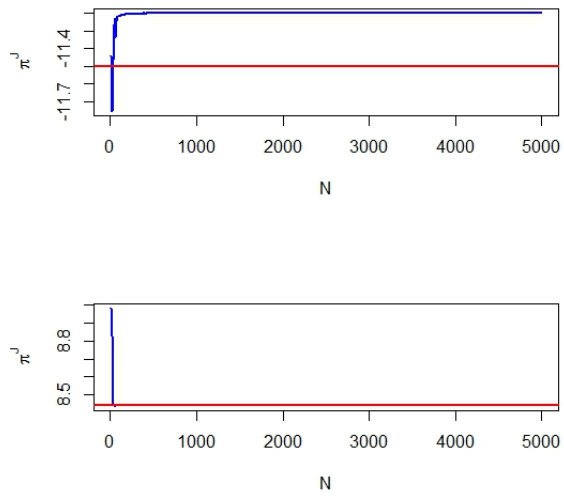

Figure 4.23: Comparison between the Jeffreys prior density obtained via integration in the Fisher information matrix via Gauss-Kronrod quadrature and sums of Riemann for the model $0.25\mathcal{N}(-10,1) + 0.10\mathcal{N}(0,5) + 0.65\mathcal{N}(15,7)$ (above) and $\frac{1}{3}\mathcal{N}(-1,0.2) + \frac{1}{3}\mathcal{N}(0,0.2) + \frac{1}{3}\mathcal{N}(1,0.2)$ (below).

the kernel distributions has been reduced or increased depending on the acceptance rate, in a way such that the acceptance rate stay between 20% and 40%. The transitional kernel used have been truncated normals for the weights, normals for the means and log-normals for the standard deviations (all centered on the values accepted in the previous iteration).

4.6 Conclusion

This thorough analysis of the Jeffreys priors in the setting of Gaussian mixtures shows that mixture distributions can also be considered as an ill-posed problem with regards to the production of noninformative priors. Indeed, we have shown that most configurations for Bayesian inference in this framework do not allow for the standard Jeffreys prior to be taken as a reference. While this is not the first occurrence where Jeffreys priors cannot be used as reference priors, the wide range of applications of mixture distributions weights upon this discovery and calls for a new paradigm in the construction of noninformative Bayesian procedures for mixture inference. Our proposal in Section 4.4 could constitute such a reference, as it simplifies the

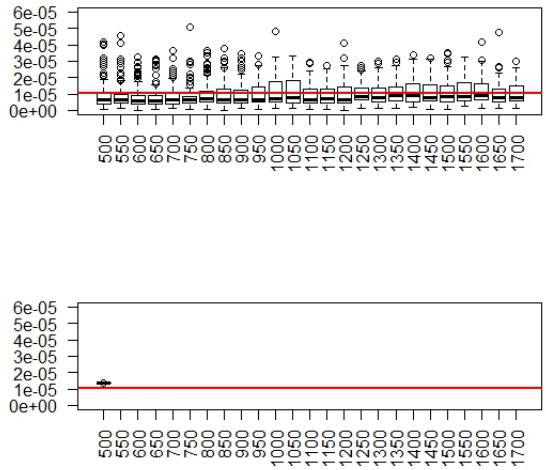

Figure 4.24: Boxplots of 100 replications of the procedure based on Monte Carlo integration (above) and sums of Riemann (below) which approximates the Fisher information matrix of the model $0.25\mathcal{N}(-10, 1) + 0.10\mathcal{N}(0, 5) + 0.65\mathcal{N}(15, 7)$ for sample sizes from 500 to 1700. The value obtained via numerical integration is represented by the red line (in the graph below, all the approximations obtained with more than 550 knots give the same result, exactly equal to the one obtained via Gauss-Kronrod quadrature).

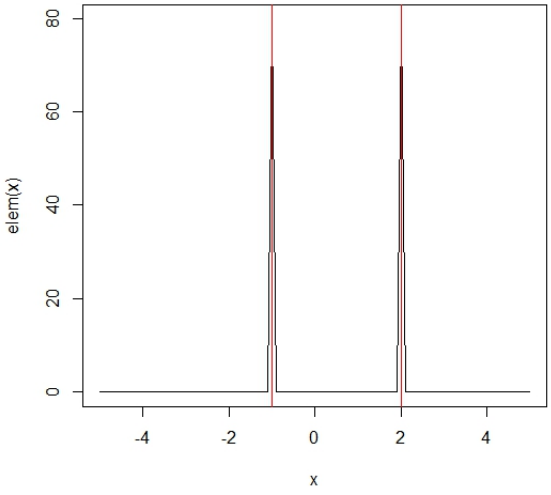

Figure 4.25: The first element on the diagonal of the Fisher information matrix relative to the first weight of the two-component Gaussian mixture model $0.5\mathcal{N}(-1, 0.01) + 0.5\mathcal{N}(2, 0.01)$.

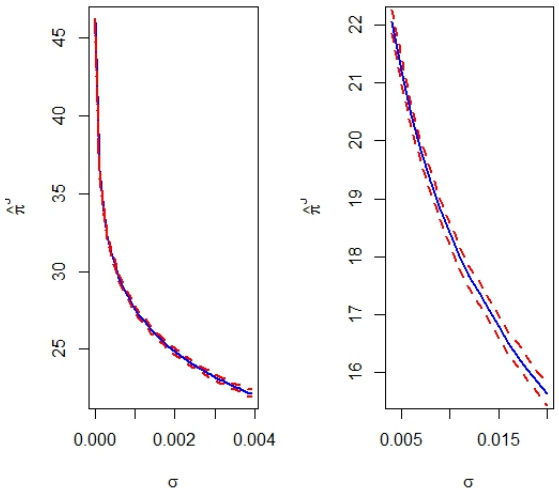

Figure 4.26: Approximation of the Jeffreys prior (in log-scale) for the two-component Gaussian mixture model $0.5\mathcal{N}(-1,\sigma)+0.5\mathcal{N}(2,\sigma)$, where σ is taken equal for both components and decreasing.

representation of Mengersen and Robert (1996).

Table 4.2: μ unknown, k=2: results of 50 replications of the experiment for both close and far means with a Monte Carlo approximation of the posterior distribution based on 10^5 simulations and a burn-in of 10^4 simulations. The table shows the average acceptance rate, the proportion of chains diverging towards higher values and the average ratio between the log-likelihood of the last accepted values and the true values in the 50 replications when using the Jeffreys prior (on the left) and a prior constant on the means (on the right).

k=2	**Jeffreys prior** (Close Means)			**Constant prior**		
Sample Size	*Ave. Accept. Rate*	*Chains towards high values*	*Ave. $lik(\theta^{fin})$ / $lik(\theta^{true})$*	*Ave. Accept. Rate*	*Chains towards high values*	*Ave. $lik(\theta^{fin})$ / $lik(\theta^{true})$*
2	0.2505	0.88	1.8182	0.2709	0.72	1.9968
3	0.2656	0.94	1.6804	0.2782	0.58	1.9613
4	0.2986	0.56	1.3097	0.2812	0.18	1.9824
5	0.2879	0.48	1.2918	0.2830	0.14	1.8358
6	0.3066	0.16	1.1251	0.3090	0.00	1.9363
7	0.3052	0.24	1.1205	0.3103	0.02	1.7994
8	0.3181	0.02	1.0149	0.3521	0.00	1.3923
9	0.3101	0.02	1.0244	0.3369	0.00	1.5219
10	0.3460	0.00	0.9914	0.3627	0.00	1.2933
15	0.3418	0.00	1.0097	0.3913	0.00	1.1970
20	0.3881	0.00	0.9948	0.4097	0.00	1.1032
50	0.4556	0.00	1.0005	0.4515	0.00	1.0303
100	0.5090	0.00	1.0008	0.5090	0.00	1.0007
500	0.5603	0.00	1.0006	0.5305	0.00	1.0002
1000	0.4915	0.00	1.0006	0.2327	0.00	1.0042
k=2	(Far means)			**Constant prior**		
2	0.2752	0.00	1.0838	0.2736	0.00	1.0474
3	0.2692	0.00	1.0313	0.2546	0.00	1.0313
4	0.2969	0.00	1.1385	0.3152	0.00	1.0167
5	0.2938	0.00	1.0138	0.2920	0.00	0.9968
6	0.3066	0.00	1.2207	0.3470	0.00	0.9975
7	0.3350	0.00	1.1055	0.3473	0.00	0.9920
8	0.3154	0.00	1.1374	0.3583	0.00	1.0092
9	0.3309	0.00	1.1566	0.3512	0.00	0.9893
10	0.3338	0.00	1.1820	0.3601	0.00	1.0112
15	0.3579	0.00	1.1796	0.3840	0.00	1.0136
20	0.3950	0.00	1.1615	0.4190	0.00	1.0096
50	0.4879	0.00	1.1682	0.4659	0.00	1.0059
100	0.5083	0.00	1.2123	0.4957	0.00	1.0017
500	0.5570	0.00	1.1996	0.4777	0.00	0.9976
1000	0.3463	0.00	1.2161	0.1792	0.00	1.0010

Table 4.3: μ unknown, k=3: as in Table 4.2 for two three-component Gaussian mixture models, with close and far means, only for the Jeffreys prior.

k=3	**Jeffreys prior** (Close Means)		
Sample Size	*Ave. Accept. Rate*	*Chains towards high values*	*Ave. $lik(\theta^{fin})$ / $lik(\theta^{true})$*
2	0.2366	1.00	2.5175
3	0.2608	1.00	2.8447
4	0.2455	0.98	1.3749
5	0.2446	1.00	1.3807
6	0.2330	1.00	1.4062
7	0.2480	0.98	1.2411
8	0.2684	0.94	1.2535
9	0.2784	0.98	1.2744
10	0.2904	0.68	1.1168
15	0.3214	0.74	1.1217
20	0.3819	0.32	1.0616
30	0.3774	0.10	1.0383
50	0.4407	0.04	1.0108
100	0.4935	0.00	1.0018
500	0.5577	0.00	1.0068
1000	0.5511	0.00	1.0006
k=3	(Far means)		
2	0.2641	1.00	2.1786
3	0.2804	1.00	2.1039
4	0.2813	0.82	1.1173
5	0.2840	0.84	1.0412
6	0.2887	0.84	1.1050
7	0.2865	0.82	1.0840
8	0.3248	0.66	1.0982
9	0.3277	0.76	1.1177
10	0.2998	0.00	1.2604
15	0.3038	0.00	1.3149
20	0.2869	0.00	1.3533
30	0.3762	0.00	1.2479
50	0.4283	0.00	1.3791
100	0.5251	0.00	1.2585
500	0.5762	0.00	1.4779
1000	0.4751	0.00	1.2161

Table 4.4: σ unknown, $k = 2$: results of 50 replications of the experiment for both close and far means with a Monte Carlo approximation of the posterior distribution based on 10^5 simulations and a burn-in of 10^4 simulations. The table shows the average acceptance rate, the proportion of chains stuck at values of standard deviations close to 0 and the average ratio between the log-likelihood of the last accepted values and the true values in the 50 replications when using the Jeffreys prior.

k=2	**Jeffreys prior** (Close Means)		
Sample Size	*Ave. Accept. Rate*	*Chains stuck at small values of σ*	*Ave. $lik(\theta^{fin})$ / $lik(\theta^{true})$*
2	0.2414	0.02	1.2245
3	0.1875	0.02	1.1976
4	0.2403	0.00	1.0720
5	0.2233	0.02	1.1269
6	0.2475	0.00	1.0553
7	0.2494	0.02	1.0324
8	0.2465	0.00	1.0093
9	0.2449	0.00	1.0026
10	0.2476	0.00	0.9960
15	0.2541	0.00	0.9959
20	0.2480	0.00	0.9946
30	0.2364	0.00	1.0052
50	0.2510	0.00	0.9981
100	0.3033	0.00	0.9994
500	0.4314	0.00	0.9999
1000	0.4353	0.00	1.0001
k=2	(Far means)		
2	0.2262	0.14	1.09202
3	0.2384	0.10	1.0536
4	0.2542	0.02	1.0281
5	0.2502	0.04	0.9932
6	0.2550	0.00	0.9981
7	0.2554	0.00	0.9569
8	0.2473	0.00	0.9929
9	0.2481	0.00	0.9888
10	0.2402	0.00	0.9969
15	0.2431	0.00	0.9988
20	0.2416	0.00	0.9998
30	0.2453	0.04	1.0016
50	0.2550	0.00	0.9992
100	0.2359	0.00	0.9999
500	0.3000	0.00	1.0001
1000	0.3345	0.00	1.0000

Table 4.5: σ unknown, $k = 3$: as in table 4.4 for two three-components Gaussian mixture models, with close and far means.

k=3	**Jeffreys prior** (Close Means)		
Sample Size	*Ave. Accept. Rate*	*Chains stuck at small values of* σ	*Ave.* $lik(\theta^{fin})$ / $lik(\theta^{true})$
2	0.0441	0.88	0.1206
5	0.0659	0.72	1.0638
6	0.0621	0.70	1.1061
7	0.1013	0.54	1.0655
8	0.0781	0.52	1.0880
9	0.0729	0.60	1.1003
10	0.1506	0.26	1.0516
15	0.1689	0.18	1.0493
20	0.2322	0.10	1.0478
30	0.2366	0.00	1.0125
50	0.4407	0.02	1.0061
100	0.2666	0.00	1.0021
500	0.3871	0.00	1.0003
1000	0.4353	0.00	1.0001
k=3	(Far means)		
2	0.0222	0.78	1.0045
5	0.0610	0.44	1.0427
6	0.0567	0.52	1.0317
7	0.0779	0.46	1.0147
8	0.0862	0.32	1.0244
9	0.1312	0.26	1.0027
10	0.1472	0.18	1.0350
15	0.15884	0.14	1.0170
20	0.2331	0.06	1.0092
30	0.2464	0.04	1.0062
50	0.2498	0.00	1.0017
100	0.2567	0.00	1.0008
500	0.2594	0.00	0.9999
1000	0.3073	0.00	1.2161

Table 4.6: k=2, ($\mathbf{p}$, μ, σ) unknown: results of 50 replications of the experiment for both close and far means with a Monte Carlo approximation of the posterior distribution based on 10^5 simulations and a burn-in of 10^4 simulations. The table shows the average acceptance rate, the proportion of chains diverging towards higher values, the proportion of chains stuck at values of standard deviations close to 0 and the average ratio between the log-likelihood of the last accepted values and the true values in the 50 replications when using the Jeffreys prior.

k=2	**Jeffreys prior** (Close Means)			
Sample Size	*Ave. Accept. Rate*	*Chains stuck at small values of* σ	*Chains towards high values of* μ	*Ave.* $lik(\theta^{fin})$ / $lik(\theta^{true})$
5	0.1119	0.54	0.74	3.5280
6	0.1241	0.56	0.74	3.6402
7	0.0927	0.56	0.70	3.2180
8	0.0693	0.54	0.70	3.1380
9	0.1236	0.42	0.72	3.3281
10	0.1081	0.44	0.84	2.8173
11	0.1172	0.40	0.78	2.1455
12	0.1107	0.40	0.70	1.8998
13	0.1273	0.44	0.74	1.8269
14	0.1253	0.42	0.76	1.2876
15	0.1218	0.36	0.82	1.2949
20	0.1278	0.38	0.66	1.2587
k=2	(Far means)			
5	0.1650	0.18	0.30	3.7712
6	0.2218	0.12	0.20	3.1400
7	0.1836	0.12	0.36	3.1461
8	0.2313	0.08	0.08	3.5102
9	0.1942	0.14	0.12	3.5585
10	0.2290	0.04	0.02	3.0718
11	0.2320	0.04	0.02	2.9825
12	0.2305	0.08	0.02	2.9122
13	0.2264	0.06	0.00	2.9571
14	0.2292	0.08	0.04	1.0612
15	0.2005	0.12	0.04	1.0804
20	0.2343	0.00	0.02	1.0146

Table 4.7: k=3, ($\mathbf{p}$, μ, σ) unknown: as in table 4.6 for two three-component Gaussian mixture models with close and far means.

k=3	**Jeffreys prior** (Close Means)			
Sample Size	*Ave. Accept. Rate*	*Chains stuck at small values of* σ	*Chains towards high values of* μ	*Ave.* $lik(\theta^{fin})$ / $lik(\theta^{true})$
5	0.0302	0.76	0.44	2.9095
6	0.0368	0.76	0.48	3.2507
7	0.0290	0.80	0.30	3.1318
8	0.0578	0.62	0.54	3.0043
9	0.0488	0.74	0.52	2.5798
10	0.0426	0.70	0.44	2.3023
11	0.0572	0.66	0.38	1.7497
12	0.0464	0.66	0.48	1.4032
13	0.0706	0.52	0.44	1.9303
14	0.0556	0.66	0.36	1.3588
15	0.0610	0.74	0.44	1.3588
20	0.0654	0.48	0.46	1.2161
k=3	(Far means)			
5	0.0644	0.60	0.10	5.9707
6	0.0631	0.64	0.18	2.0557
7	0.0726	0.54	0.08	2.9351
8	0.1745	0.22	0.12	2.9193
9	0.1809	0.32	0.04	95.793
10	0.1724	0.28	0.14	2.5938
11	0.1948	0.24	0.14	3.1566
12	0.1718	0.26	0.08	2.8595
13	0.2110	0.16	0.06	1.8595
14	0.1880	0.24	0.10	1.2165
15	0.1895	0.20	0.12	1.2133
20	0.2468	0.08	0.02	1.0146

Table 4.8: k=2, ($\mathbf{p}$, μ, δ, τ,σ) unknown, proper prior on σ: results of 50 replications of the experiment by using a proper prior on σ and the Jeffreys prior for the other parameters conditionally on it for both close and far means with a Monte Carlo approximation of the posterior distribution based on 10^5 simulations and a burn-in of 10^4 simulations. The table shows the average acceptance rate, the proportion of chains diverging towards higher values, the proportion of chains stuck at values of standard deviations close to 0 and the average ratio between the log-likelihood of the last accepted values and the true values in the 50 replications when using the Jeffreys prior.

K=2	**Jeffreys prior** (Close Means)			
Sample Size	*Ave. Accept. Rate*	*Chains stuck at small values of* σ	*Chains towards high values of* μ	*Ave.* $lik(\theta^{fin})$ / $lik(\theta^{true})$
5	0.2094	0.02	0.92	1.4440
6	0.2152	0.00	0.98	1.3486
7	0.2253	0.00	0.92	1.3290
8	0.2021	0.00	0.94	1.2258
9	0.1828	0.00	0.84	1.2666
10	0.2087	0.00	0.88	1.1770
11	0.1854	0.00	0.94	1.2088
12	0.1829	0.00	0.86	1.2153
13	0.1658	0.00	0.92	1.1682
14	0.2017	0.00	0.86	1.2043
15	0.1991	0.00	0.88	1.2002
20	0.1851	0.00	0.76	1.1688
K=2	(Far means)			
5	0.2071	0.00	0.70	1.5741
6	0.2021	0.00	0.68	1.4384
7	0.1947	0.00	0.60	1.3597
8	0.2054	0.00	0.44	1.2869
9	0.2093	0.00	0.46	1.3064
10	0.2271	0.00	0.20	1.1618
11	0.2030	0.00	0.32	1.1996
12	0.2178	0.00	0.24	1.1494
13	0.2812	0.00	0.18	1.1215
14	0.1880	0.00	0.08	1.0717
15	0.2511	0.00	0.06	1.0594
20	0.2359	0.00	0.00	1.0166

Table 4.9: Hierarchical Mixture model: results of 50 replications of the experiment for a two and a three-component Gaussian mixture model with a Monte Carlo approximation of the posterior distribution based on 10^5 simulations and a burn-in of 10^4 simulations. The table shows the average acceptance rate, the proportion of chains diverging towards higher values, the proportion of chains stuck at values of standard deviations close to 0, the mean and the median log-likelihood of the last accepted values and the mean and the median maximum log-likelihood of the accepted values.

k=2							
Sample Size	*Ave. Accept. Rate*	*Chains stuck at small values of σ*	*Chains towards high values of μ*	*Mean $l(\theta^{fin})/l(\theta^{true})$*	*Median $l(\theta^{fin})/l(\theta^{true})$*	*Mean $max(l(\theta))/l(\theta^{true})$*	*Median $max(l(\theta))/l(\theta^{true})$*
3	0.1947	0.00	0.00	1.1034	0.9825	0.0838	0.5778
4	0.2295	0.00	0.00	1.0318	1.0300	0.4678	0.5685
5	0.2230	0.00	0.00	0.9572	0.9924	0.8464	0.7456
6	0.2275	0.00	0.00	0.9870	0.9641	0.6614	0.6696
7	0.2112	0.00	0.00	1.0658	1.0043	0.8406	0.7848
8	0.2833	0.00	0.00	1.0077	1.0284	0.8268	0.8495
9	0.2696	0.00	0.00	1.0741	1.0179	0.8854	0.8613
10	0.2266	0.00	0.00	1.1446	0.9968	0.9589	0.8508
15	0.1982	0.00	0.00	1.0201	0.9959	0.9409	0.9280
20	0.2258	0.00	0.00	1.2023	1.0145	0.9172	0.9400
30	0.2073	0.00	0.00	0.9888	1.0022	1.0424	0.9656
50	0.2724	0.00	0.00	1.0493	1.0043	1.0281	0.9859
100	0.2739	0.00	0.00	1.0932	1.0025	1.0805	0.9932
200	0.3031	0.00	0.00	1.1610	1.0036	1.1519	0.9964
500	0.2753	0.00	0.00	1.1729	1.0023	1.1694	0.9989
1000	0.2317	0.00	0.00	1.1800	1.0021	1.1772	0.9994
k=3							
3	0.2840	0.00	0.00	1.1316	1.0503	0.3432	0.2950
4	0.2217	0.00	0.00	1.0326	0.9452	0.6699	0.6624
5	0.2144	0.00	0.00	1.0610	1.0421	0.6858	0.6838
6	0.2258	0.00	0.00	1.0908	0.9683	0.6472	0.6355
7	0.1843	0.00	0.00	1.0436	0.9915	0.7878	0.8008
8	0.2760	0.00	0.00	1.0276	1.0077	0.7996	0.7958
9	0.2028	0.00	0.00	1.0025	1.0145	0.7830	0.8016
10	0.2116	0.00	0.00	1.0426	1.0015	0.8752	0.8591
15	0.2023	0.00	0.00	1.0247	1.0063	0.8810	0.8871
20	0.2211	0.00	0.00	1.0281	1.0104	0.9290	0.9268
30	0.2242	0.00	0.00	1.1978	1.0123	1.0841	0.9508
50	0.2513	0.00	0.00	1.0543	1.0142	1.0148	0.9775
100	0.2768	0.00	0.00	1.0563	1.0206	1.0324	0.9955
200	0.2910	0.00	0.00	1.0325	1.0118	1.0200	0.9993
500	0.2329	0.00	0.00	1.0943	1.0079	1.0882	1.0002
1000	0.2189	0.00	0.00	1.1068	1.0105	1.1212	1.0110

Conclusions

Modern computers allow for a development of Bayesian procedures impossible in the past. In this work we have tried to use some new and old methodologies for modern theoretical and applied problems, thanks to new computational tools.

In the first part of the work, we have shown that the class of algorithms called "approximate Bayesian computation" may be used to solve some complicated problems both in a Bayesian and in a classical setting. In a non-Bayesian setting (Chapter 1), we have first shown that it can be used as a tool to approximate the likelihood function for a parameter of interest in the presence of nuisance parameters. One of the most interesting open problem of the work is that the proposed approach requires the use of proper priors, unless the marginalisation of the prior can be done analytically. It has also been pointed out that in some cases a simple numerical method based on Gaussian quadrature may be used. However, when the dimension of the parameter of interest increases or when the nuisance parameter is highly dimensional the Gaussian quadrature is unlikely to produce good approximations; so there may be cases where this is the "only" possible approach. A Bayesian method (based on the computation of a posterior distribution) to produce non-Bayesian estimates may also be criticized. In particular, simpler algorithms may be used if only the mode of the likelihood function is of interest. Nevertheless we believe that deriving and approximating the complete form of the likelihood function is actually of interest, for any inference based on the likelihood. This work has been presented at the ISBA World meeting 2014 (Cancun, Mexico, invited talk), at the Bayes in Paris seminar at ENSAE (Paris, France, invited talk), at the Lunch Seminar of Università degli Studi di Torino (Torino, Italy, invited talk) and at BAYSM, BAyesian Young Statistician Meeting, 2013 (Milan, Italy, contributed talk) in various forms. It will be also presented at the 48th Italian Statistical Society (S.I.S.) meeting (Salerno, Italy, 8-10 June 2016) in an invited Specialized session on the interplay between frequentist and Bayesian methods.

The problem of eliminating the nuisance parameters is crucial in all the approaches: in particular, in modern applications many parameters are introduced to construct flexible and realistic models, nevertheless their lack of a physical meaning is a problem both in terms of making inference on them and of constructing reasonable prior distributions. In semiparametric problems, where the interest of the analysis is in few parameters, while it is preferred to limit the assumptions on the complete shape of the model, we have shown that the ABC methodology may also be used to deal with semiparametric problems; for instance, it allows to manage copula models and to study the dependence structure of multivariate random variables without making strong assumptions on the univariate distributions or on the copula function. Future lines of research would be focused on generalizing the approach presented in Chapter 2, for instance, by introducing covariates in the analysis and by considering other type of models. This work represents an important contribution in multivariate analysis, firstly because it allows to work with non-Gaussian distributions and secondly because it provides a way to handle multivariate distributions of arbitrary dimension (while classical proposals have been focused mainly on bivariate distributions). We have analyzed the log-returns from several Italian institutes and created a method to analyze the relationships between institutes, in particular, during crisis periods. The work will be presented at the next ISBA World meeting 2016 (Sardinia, Italy, invited talk) and has been presented at Sixth IMS-ISBA Joint Meeting (Lenzerheide, Switzerland, 4-7 January 2016). A generalization of the method proposed in Chapter 2 is needed: the introduction of covariates in the analysis, even if important for real applications and a realistical modeling, may hide computational and theoretical problems, as exemplified by Craiu and Sabeti (2012). Therefore, further research may be focused on the development of methodologies which consider the introduction of covariates. It has been noted that the choice of the moment equations identifying the parameter of interest (for example, the Spearman's ρ) is quite delicate: it is clearly necessary to choose an estimator which has good asymptotic properties, in order to approximate the true quantity of interest without any bias. However, the generalization to dimension $d \neq 2$ in Section 2.7 shows that this is not always possible. Further research will be focused on studying other types of pseudo-likelihoods to use in the analysis, in order to avoid the moments constraints typical in the empirical likelihood methodology.

It has been also pointed out that, while the practical gain in analyzing each component separately is clear, the object created by the estimated cumulative dis-

tribution function transforms may have a very different correlation structure from the true cumulative distribution function transforms. A further validation of the method is, therefore, needed, maybe based on some consistency conditions on the estimated cumulative distribution functions.

Defining ways for model choice in the case of complex models is particularly challenging: standard approaches tend to fail in choosing the right model. In Chapter 3 we have first analyzed the proposal by Dawid and Musio (2015) and stressed some of its limits, in particular its difficulty to handle non-normal models. We have then proposed a way to approximate the Bayes factor when the likelihood function is unavailable via approximate Bayesian computation. The method has been empirically tested in an example typical in the ABC literature: quantile distributions. The method has been proved to have a good behavior in terms of approximating properties (both in finite sample and asymptotically), however a deeper theoretical treatment is needed to better study its theoretical properties and propose it as a benchmark. The final example of Section 3.3.2 is useful to understand the computational difficulties and to study the behavior of the method in a financial application, however more complicated situations, for example not involving quantile distributions, have to be studied in order to generalize it. Moreover, a deeper comparison with other existing methods is needed. This work has already been presented at CMStatistics 2015 (London, U.K., 12-14 December 2015, contributed talk).

In the last part of the work we have tried to construct a noninformative analysis for mixture models. As said before, describing the information available *a priori* with a distribution is difficult for complex models, mainly because not all the parameters have a physical meaning. The definition of a noninformative prior for the parameters of a mixture models has a long history in the literature. First, we have analyzed a classical approach for defining a noninformative prior, namely the Jeffreys' procedure, thanks to computational methods to approximate the prior which is not available in closed form. Then, we have proposed to change the point of view by defining a hierarchy which creates correlation between the components of the mixture and allows for a noninformative and improper prior at the highest level. This proposal may be seen as a generalization of the classical attempts to use improper priors for reparametrized models (Mengersen and Robert 1996, Wasserman 2000, Robert and Titterington 1998) and is more general. It allows for the use of improper priors at the highest level of the hierarchy, by introducing the implicit information that the components are related: since the distributions composing the

mixture model are used together to represent the data, it seems reasonable that a form of dependence exists among them. This work has been presented at the "Seminaire des Jeunes Chercheurs" (Paris, France, invited talk) and at the 2nd BAYSM 2014 meeting (Vienna, Austria, contributed talk). In this case, a natural extension of the work could be considering an unknown number of components and generilize the hierarchical representation for the parameters of any (complex) model to study if it may be used as a general answer on how to choose a noninformative prior for parameter without a physical meaning.

Statement about the contributions

This thesis is the result of a three-year Double PhD programme between Sapienza Università di Roma, under the supervision of Prof. Brunero Liseo, and Université Paris-Dauphine, under the supervision of Prof. Christian Robert. The interaction between the candidate and her supervisors has been essential in the development of the projects. In general, all the proofs have been developed by the candidate and the programmes have been written in R by her. In the following, a detailed list of the contributions of all the authors is provided.

- **Chapter 1:** This chapter is a joint work by the candidate and Prof. Brunero Liseo. The problem of defining a likelihood function for a parameter of interest in presence of nuisance parameters has been extensively studied by Prof. Liseo and the idea of using the ABC methodology to obtain the integrated likelihood function when the models are complicated may be attributed to him. Section 1.3 is a contribution of the candidate. The first two examples (Poisson means and the Neyman and Scott's class of problems) are known in the literature about the eliminations of nuisance parameters and have been proposed by professor Liseo to test the method. The example with the g-and-k distributions has been proposed by the candidate, while the example on semiparametric regression was proposed by prof. Liseo. In all the cases, the implementation of the algorithm (including the choices of the summary statistics) is due to the candidate.

- **Chapter 2:** This chapter is a joint work by the candidate and Prof. Liseo. Copula estimation is a research interest of prof. Liseo, but the use of the ABC methodology in this setting has been proposed by the candidate. Algorithms 5 and Algorithm 6 have been proposed by the candidate. Prof. Liseo and the candidate have jointly worked in Section 2.5.1, the alternative estimator proposed in Section 2.5.3 has been proposed by Prof. Liseo, Section 2.6, 2.7

and 2.8 are the candidate's contributions.

- **Chapter 3:** Section 3.1 and Section 3.2 are based on a Master research thesis written by Ilaria Masiani under the supervision of the candidate and Prof. Christian Robert. In this case, Figure 3.1, 3.2 and 3.3 have been obtained by Ilaria Masiani and the candidate together. Section 3.3 is a contribution by the candidate.

- **Chapter 4:** This chapter is a joint work by the candidate and Prof. Robert. The idea of using the Jeffreys prior of the weights of a general mixture model has been firstly proposed by Prof. Robert. All the proofs have been developed by the candidate, as well as the simulations. The alternative to the Jeffreys prior proposed in Section 4.4 derives from an idea of the candidate. Section 4.5 is a contribution of the candidate.

Bibliography

Abramowitz, M., and Stegun,I. A. (1964) *Handbook of mathematical functions with formulas, graphs, and mathematical tables*, volume 55 of *National Bureau of Standards Applied Mathematics Series*. U.S. Government Printing Office, Washington, D.C. 16

Acar, E.F., Craiu, R.V., and Yao, F. (2011). Dependence Calibration in Conditional Copulas: A Nonparametric Approach. *Biometrics*, 67:445–453. 60

Allingham, D., King, R. A. R., and Mengersen, K. (2009). Bayesian estimation of quantile distributions. *Statistics and Computing*, 19(2):189–201. 20, 77, 78, 79, 84

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control.*, 19: 716–723. 70

Ardia, D., and Hoogerheide, L. F. (2010). Bayesian Estimation of the GARCH(1,1) Model with Student-t Innovations. *The R Journal*, 2(2): 41–47. 63

Banerjee, S., Gelfand, A. E., and Carlin, B. P. (2004) *Hierarchical Modelling and Analysis for Spatial Data*. Chapman and Hall CRC. 26

Barndorff-Nielsen, O. (1983) On a formula for the distribution of the maximum likelihood estimator. *Biometrika*, 70(2):343–365. 16

Basu, D. (1977) On the elimination of nuisance parameters. *Journal of the American Statistical Association*, 72(358):355–366. 11

Beaumont, M. (2010) Approximate Bayesian computation in evolution and ecology. *Annual Review of Ecology, Evolution, and Systematics*, 41:379–406. 8

Beaumont, M., Zhang, W., and Balding, W. (2002). Approximate Bayesian Computation in Population Genetics. *Genetics*, 162(4):2025–2035. 75

Berger, J., Bernardo, J., and Sun, D. (2009). Natural induction: An objective Bayesian approach. *Rev. Acad. Sci. Madrid*, **A 103** 125–159. (With discussion). 95

Berger, J. O., Liseo, B., and Wolpert, R. L. (1999) Integrating Likelihood Methods for Eliminating Nuisance Parameters. *Statistical Science*, 14(1):1–28. xiii, 7, 13

Bernardo, J., and Giròn, F. (1988). A Bayesian analysis of simple mixture problems. In *Bayesian Statistics 3* (J. Bernardo, M. DeGroot, D. Lindley and A. Smith, eds.). Oxford University Press, Oxford, 67–78. 97, 99, 100

Blum, M. G. B. (2010) Approximate Bayesian Computation: A Nonparametric Perspective. *Journal of the American Statistical Association*, 105(491):1178–1187. 12

C. B. Borkowf. Computing the nonnull asymptotic variance and the asymptotic relative efficiency of Spearma's rank correlation *Computational Statistics & Data Analysis*, 39: 271–286, 2002. 42

Bowman, A. W., and Azzalini, A. (1997) *Applied Smoothing Techniques for Data Analysis*. Oxford Univ. Press. xxiv, 26

Butler, R. W., and Wood, A. T. A. (2002) Laplace approximations for hypergeometric functions with matrix argument. *The Annals of Statistics*, 30(4):1155–1177. 17, 18

Casella, G., Mengersen, K., Robert, C. P., and Titterington, D. (2002). Perfect slice samplers for mixtures of distributions. *J. Royal Statist. Society Series B*, **64(4)** 777–790. 119

Celeux, G., Hurn, M., and Robert, C. P. (2000). Computational and inferential difficulties with mixture posterior distributions. *J. American Statist. Assoc.*, **95(3)** 957–979. xxii, 93, 109

Chaussé, P. (2010). Computing Generalized Method of Moments and Generalized Empirical Likelihood with R. *Journal of Statistical Software*, 11(34): 1–35. 38

Cherubini, U., Luciano, E., and Vecchiato, W. (2004). *Copula Methods in Finance*. John Wiley & Sons, New York, San Francisco, Calif.. xvi, 31

Cornuet, J. M., Santos, F., Beaumont, M. A., Robert, C. P., Marin, J. M., Balding, D. J., Guillemaud, T., and Estoup, A. (2008). Inferring population history with DIY ABC: a user-friendly approach to approximate Bayesian computation. *Bioinformatics*, 24(23):2713–2719. 75

Craiu, V. R., and Sabeti, A. (2012). In mixed company: Bayesian inference for bivariate conditional copula models with discrete and continuous outcomes. *J. Multivariate Anal.*, 110: 106–120. xvi, 32, 37, 144

Creel, M., and Kristensen, D. (2015). ABC of SV: Limited information likelihood inference in stochastic volatility jump-diffusion models. *Journal of Empirical Finance*, 31(C):85–108. 73

Davison, A. C. (2003) *Statistical models*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, Cambridge. 15, 16

Dawid, A. P., and Musio, M. (2015). Bayesian Model Selection Based on Proper Scoring Rules. *Bayesian Anal.*, 10(2): 479–499. 67, 145

DeGroot, M. (1982). Discussion of Shafer's 'Lindley's paradox'. *J. American Statist. Assoc.*, 378: 337–339. 71

Di Bernardino, E., and Rullière, D. (2015). On tail dependence coefficients of transformed multivariate Archimedean copulas *Fuzzy Sets and Systems*, http:dx.doi.org/10.1016/j.fss.2015.08.030, 2015. 56

Diebolt, J., and Robert, C. P. (1994). Estimation of finite mixture distributions by Bayesian sampling. *J. Royal Statist. Society Series B*, **56** 363–375. xxii, 93, 94

Drovandi, C. C., and Pettitt, A. N. (2011). Likelihood-free Bayesian estimation of multivariate quantile distributions. *Computational Statistics and Data Analysis*, 55(9):2541-2556. 78

Fan, J., and Gijbels, I. (1996). *Local Polynomial Modelling and Its Applications*. Chapman & Hall/CRC Press, New York (USA). 61

Fearnhead, P., and Prangle, D. (2012) Constructing summary statistics for approximate Bayesian computation: semi-automatic approximate Bayesian computation. *Journal of the Royal Statistical Society: Series B*, 74(3):419-474. xx, 74

Fermanian, J.D., Radulović, D., and Wegkamp, M. (2004). Weak Convergence of empirical copula processes. *Bernoulli*, 10(5): 847–860. 40

Figueiredo, M., and Jain, A. (2002). Unsupervised learning of finite mixture models. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, **24** 381–396. 96

Frahm, G., Junker, M., and Schmidt, R. (2005). Estimating the tail-dependence coefficient: Properties and pitfalls. *Insurance: Mathematics and Economics*, 37:80–100, 2005. 55

Frühwirth-Schnatter, S. (2006). *Finite Mixture and Markov Switching Models*. Springer-Verlag, New York, New York. xxi, xxii, 93, 94

Genest, C., Ghoudi, K., and Rivest, L. P. (1995). A semiparametric estimation procedure of dependence parameters in multivariate families of distributions. *Biometrika*, 82(3): 543–552. xviii, 33

Genest, C., and Favre, A.C. (2007). Everything you always wanted to know about copula modeling but were afraid to ask. *Journal of Hydrologic Engineering*, pages 347–368. 42, 46

Geweke, J. F. (1993). Bayesian treatment of the independent Student-t linear model. *J. Appl. Econometr.*, S1(8): S19–S40. 63

Geweke, J. (2007). Interpretation and inference in mixture models: Simple MCMC works. *Comput. Statist. Data Analysis*, **51** 3529–3550. xxii, 94

Ghosh, M., Carlin, B. P., and Srivastiva, M. S. (1995). Probability matching priors for linear calibration. *TEST*, **4** 333–357. 95

Gijbels, I., Veraverbeke, N., and Omelka, M. (2011). Conditional copulas, association measures and their applications *Computational Statistics & Data Analysis*, 55(5): 1919–1932. 60, 61

Good, I. J. (1952). Rational Decisions. *Journal of the Royal Statistical Society. Series B* , 14(1):107–114. 68

Gruet, M., Philippe, A., and Robert, C. P. (1999). MCMC control spreadsheets for exponential mixture estimation. *J. Comput. Graph. Statist.*, **8** 298–317. 120

Harville, D. A. (1977) Maximum Likelihood Approaches to Variance Component Estimation and to Related Problems. Journal of the American Statistical Association, 72 (358): 320–338. 26

Haynes, M. A., MacGillivray, H. L., and Mengersen, K. L. (1997) Robustness of ranking and selection rules using generalised g-and-k distributions. *Journal of Statistical Planning and Inference*, 65(1):45–66. 19

Haynes, M., and Mengersen, K. (2005). Bayesian estimation of g-and-k distributions using MCMC. *Computational Statistics*, 20(1): 7-30. 77

He, J., Li, H., Edmondson, A. C., Rader, D. L., and Li, M. (2012). A Gaussian copula approach for the analysis of secondary phenotypes in case–control genetic association studies. *Biostat.*, 13(3): 497–508. xvi, 31

He, H., and Severini, T. A. (2013) Integrated likelihood estimation in semiparametric regression models. *Technical report, Northwestern University Department of Statistics*. 25, 26

Hoff, P. (2007). Extending the rank likelihood for semiparametric copula estimation. *The Annals of Applied Statistics*, 1(1): 265–283. xvii, 32, 37

Huynh, V. N., Kreinovich, V., and Sriboonchitta, S. (2015). *Modeling Dependence in Econometrics*. Springer, New York. xvi, 31

Hyvävarinen, A. (2005). Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning*, 6:695–709. 69

Jasra, A., Holmes, C., and Stephens, D. (2005). Markov Chain Monte Carlo methods and the label switching problem in Bayesian mixture modeling. *Statist. Sci.*, **20** 50–67. xxii, 94

Jeffreys, H. (1939). *Theory of Probability*. 1st ed. The Clarendon Press, Oxford. 80, 95

H. Joe. Multivariate dependence *Journal of Multivariate Analysis*, 35: 12–30, 1990 50

Joe, H. (2014). *Dependence modelling with copulas, Monographs on Statistics and Applied Probability, 134*. Chapman & Hall–CRC Press, London. xvi, 31, 40

Joe, H, Smith, R. L., and Weissman, I. (1992). Bivariate threshold models for extremes.. *Journal of the Royal Statistical Society, Series B*, 54:171–183, 1992. 55

Kalbfleisch, J. D., and Sprott, D. A. (1970) Application of likelihood methods to models involving large numbers of parameters. *Journal of the Royal Statistical Society*,**B**, 32(2):175–208. 13

Kamary, K., Mengersen, K., Robert, C. P., and Rousseau, J. (2014). Testing hypotheses as a mixture estimation model. *arxiv:1214.2044*. 73

Kass, R., and Wasserman, L. (1996). Formal rules of selecting prior distributions: a review and annotated bibliography. *J. American Statist. Assoc.*, **91** 343–1370. 95

Kendall, M., Stuart, A., and Ord, J. K. (1987) *Kendall's advanced theory of statistics. Vol. 1.* The Clarendon Press Oxford University Press, New York, fifth edition. ISBN 0-19-520561-8. Distribution theory. 11

Kervrann, C., Roudot, P., and Waharte, F. (2014). Approximate Bayesian Computation, stochastic algorithms and non-local means for complex noise models. *IEEE International Conference on Image Processing*, Oct 2014, Paris, France. pp.4. 73

Koop, G., and Potter S. M. (1999). Bayes factors and nonlinearity: Evidence from economic time series. *Journal of Econometrics*, 88(2): 251-281. 87

Lancaster, T. (2000) The incidental parameter problem since 1948. *Journal of Econometrics*, 95(2):391–413. ISSN 0304-4076. xiii, 7, 15

Lee, K., Marin, J.-M., Mengersen, K., and Robert, C. P. (2009). Bayesian inference on mixtures of distributions. In *Perspectives in Mathematical Sciences I: Probability and Statistics* (N. N. Sastry, M. Delampady and B. Rajeev, eds.). World Scientific, Singapore, 165–202. xxi, 93

Lindley, D. V. (1957). A Statistical Paradox. *Biometrika*, 44(1-2):187-192. 80

Liseo, B. (2005) The elimination of nuisance parameters. In *Bayesian thinking: modeling and computation*, volume 25 of *Handbook of Statistics*, pages 193–219. Elsevier/North-Holland, Amsterdam. xiii, 7, 15

Lindley, D. V., and Smith, A. F. M. (1972). Bayes estimates for the linear model (with Discussion). *Journal of the Royal Statistical Society, Series B*, 34(1):1–41. 69

MacGillivray, H. L. (1992). Shape properties of the g-and-h and Johnson families. *Communications in Statistics - Theory and Methods*, 21(5): 1233-1250. 77

MacLachlan, G., and Peel, D. (2000). *Finite Mixture Models*. John Wiley, New York. xxi, xxii, 93, 94

Marin, J. M., Pudlo, P., Robert, C. P., and Ryder, R. J. (2012). Approximate Bayesian computational methods. *Statistics and Computing*, 6(22): 1167–1180. 9, 11, 20, 23, 34, 35

Marin, J. M., Pillai, N., Robert, C. P., and Rousseau, J. (2014). Relevant statistics for Bayesian model choice. *Journal of the Royal Statistical Society: Series B*, 76(5): 833-859. 86

Marin, J., and Robert, C. P. (2007). *Bayesian Core*. Springer-Verlag, New York. 71

Marjoram,P., Molitor,J., Plagnol, V., and Tavaré, S. (2003). Markov chain Monte Carlo without likelihoods. *Proceedings of the National Academy of Sciences of the United States of America*, 100(26): 15324-15328. 87

Mengersen, K., Pudlo, P., and Robert, C. P. (2013). Bayesian computation via empirical likelihood. *Proc. of the National Academy of Sciences*, 4(110): 1321–1326. xvii, xix, 32, 35

Mengersen, K., and Robert, C. P. (1996). Testing for mixtures: A Bayesian entropic approach (with discussion). In *Bayesian Statistics 5* (J. Berger, J. Bernardo, A. Dawid, D. Lindley and A. Smith, eds.). Oxford University Press, Oxford, 255–276. 100, 119, 121, 133, 145

Min, A., and Czado, C. (2010). Bayesian Inference for Multivariate Copulas using Pair- copula Constructions. *Journal of Financial Econometrics*, 4(8):511–546. xvi, 32, 37

Nadaraya, E.A. (1964). On estimating regression. *Theor. Prob. Appl.*, 9:141–142. 61

Neyman, J., and Scott, E. L. (1948) Consistent estimates based on partially consistent observations. *Econometrica*, 16(1):1–32. 15

Nunes, M. A., and Balding, D. J. (2010). On optimal selection of summary statistics for approximate Bayesian computation. *Statistical Applications in Genetics and Molecular Biology*, 9(1):1544-1576. xx, 74

Oh, D. H., and Patton, A. J. (2013). Simulated Method of Moments Estimation for Copula-Based Multivariate Models. *J. Amer. Stat. Assoc.*, 502(108):689–700. 36

Owen, A. B. (2010). *Empirical Likelihood*. Chapman & Hall/CRC Press, New York (USA). xvii, xviii, 32, 33, 41, 65

Pace, L., and Salvan, A. (1997) *Principles of statistical inference. From a neo-Fisherian perspective*. World Scientific Publishing Co. Inc., River Edge, NJ. xiii, 7

Parry, M. F., Dawid, A. P., and Lauritzen, S. L. (2012). Proper local scoring rules. *Annals of Statistics*, 40(1):561–592. 68

Piessens, R., deDoncker-Kapenga, E., Uberhuber, C. and Kahaner, D. (1983). *QUADPACK, A subroutine package for automatic integration*. Springer Verlag. 125

Pritchard, J., Seielstad, M., Perez-Lezaun, A., and Feldman, M. (1999). Population growth of human Y chromosomes: a study of Y chromosome microsatellites. *Molecular Biology and Evolution*, 16(12):1791-1798. 73

Puolamäki, K., and Kaski, S. (2009). Bayesian solutions to the label switching problem. In *Advances in Intelligent Data Analysis VIII* (N. Adams, C. Robardet, A. Siebes and J.-F. Boulicaut, eds.), vol. 5772 of *Lecture Notes in Computer Science*. Springer Berlin Heidelberg, 381–392. xxii, 94

Rayner, G. D., and MacGillivray, H. L. (2002a) Weighted quantile-based estimation for a class of transformation distributions. *Computational Statistics and Data Analysis*, 39(4):401–433. 19

Rayner, G., and MacGillivray, H. (2002). Numerical maximum likelihood estimation for the g-and-k and generealized g-and-h distributions. *Statistics and Computing*, 12(1):55-75. 19, 77, 78

Resconi, G., and Licata, I. (2015). Entropy and Copula Theory in Quantum Mechanics. In *Proceedings of the 1st Int. Electron. Conf. Entropy Appl., 3–21 November 2014*, Sciforum Electronic Conference Series. xvi, 31

Richardson, S., and Green, P. (1997). On Bayesian analysis of mixtures with an unknown number of components (with discussion). *J. Royal Statist. Society Series B*, **59** 731–792. 119

Rissanen, J. (2012). *Optimal Estimation of Parameters*. Cambridge University Press. 95

Robert, C. P. (2001). *The Bayesian Choice*. 2nd ed. Springer-Verlag, New York. 71, 94, 95, 102, 107

Robert, C. P., and Casella, G. (2004). *Monte Carlo Statistical Methods*. 2nd ed. Springer-Verlag, New York. 109

Robert, C. P., Chopin, N., and Rousseau, J. (2009). Theory of Probability revisited (with discussion). *Statist. Science*, **24(2)** 141–172 and 191–194. 71, 94

Robert, C. P., Cornuet, J. M., Marin, J. M., and Pillai, N. (2011). Lack of confidence in approximate Bayesian computation model choice. *Proceedings of the National Academy of the United States of America*, 108(37):15112-15117. xx, 74, 84

Robert, C. P., and Mengersen, K. (1999). Reparametrization issues in mixture estimation and their bearings on the Gibbs sampler. *Comput. Statist. Data Analysis*, **29** 325–343. 118, 119

Robert, C. P., and Titterington, M. (1998). Reparameterisation strategies for hidden Markov models and Bayesian approaches to maximum likelihood estimation. *Statistics and Computing*, **8** 145–158. 120, 145

Roeder, K., and Wasserman, L. (1997). Practical Bayesian density estimation using mixtures of normals. *J. American Statist. Assoc.*, **92** 894–902. 120

Rousseau, J., and Mengersen, K. (2011). Asymptotic behaviour of the posterior distribution in overfitted mixture models. *J. Royal Statist. Society Series B*, **73** 689–710. 100

Rubin, D. R. (1988). Using the SIR algorithm to simulate posterior distributions (with discusion). In *Bayesian Statistics 3 J.M. Bernardo, M.H. DeGroot, D.V. Lindley, and A.F.M. Smith, Eds.*, Bayesian Statistics, pages 395–402. Oxford University Press, Oxford (UK). 35

Rubio, F., and Steel, M. (2014). Inference in two-piece location-scale models with Jeffreys priors. *Bayesian Analysis*, **9** 1–22. 96

Schefzik, R., Thorarinsdottir, T. L., and Gneiting, T. (2013). Uncertainty Quantification in Complex Simulation Models Using Ensemble Copula Coupling. *Statist. Sci.*, 28(4): 616–640. xvi, 31

Schennach, S. M. (2005). Bayesian exponentially tilted empirical likelihood. *Biometrika*, 1(92): 31–46. xvii, 32, 34, 40

Schmid, F., and Schmidt, R. (2006). Bootstrapping Spearman's multivariate rho. *Proceedings of COMPSTAT 2006*, 759–766, 2006. 51

Schmid, F., and Schmidt, R. (2007). Multivariate Extensions of Spearman's Rho and Related Statistics. *Statistics and Probability Letters*, 77(4):407–416, 2007. 50, 51

Schmidt, R., and Stadtmüller, U. (2006). Non-parametric Estimation of Tail Dependence. *Scandinavian Journal of Statistics*, 33(2):1467–9469, 2006. 55, 56

Schmidt, A. M., and Gelfand, A. (2003) A Bayesian Coregionalization Model for Multivariate Pollutant Data. *Journal of Geophysics Research*, 108(D24). 26

Scholzel, C., and Friederichs, P. (2008). Multivariate non-normally distributed random variables in climate research: introduction to the copula approach. *Nonlinear Processes in Geophysics*, 15: 761–772. xvi, 31

Schwarz, G. E. (1978). Estimating the dimension of a model. *Annals of Statistics*, 51(6)(2):461-464. 83

Severini, T. A. (2007) Integrated likelihood functions for non-Bayesian inference. *Biometrika*, 94(3):529–542. ISSN 0006-3444. xiv, 8

Severini, T. A. (2010) Likelihood ratio statistics based on an integrated likelihood. *Biometrika*, 97(2):481–496 ISSN 0006-3444. xiv, 8

Severini, T. A. (2011) Frequency properties of inferences based on an integrated likelihood function. *Statistica Sinica*, 21(1):433–447. ISSN 1017-0405. xiv, 8

Severini, T. A. (2000) *Likelihood methods in statistics*. Oxford University Press, Oxford xiii, 7

Sibuya, M. (1960). Bivariate extreme statistics. *Annals of the Institute of Statistical Mathematics*, 11:195–210, 1960. 55

Silva, R. d. S., and Lopes, H. F. (2008). Copula, marginal distributions and model selection: a Bayesian note. *Stat. Comput.*, 18(3): 313–320. 37

Sisson, S. A., and Fan, Y. (2011). Likelihood-free MCMC. In *Handbook of Markov chain Monte Carlo*, Handb. Mod. Stat. Methods, pages 313–335. Chapman & Hall/CRC, London. 11, 35

Sklar, M. (1959). Fonctions de répartition à n dimensions et leurs marges. *Publ. Inst. Statist. Univ. Paris*, 8: 229–231. xvii, 33

Smith, M. S. (2013). Bayesian Approaches to Copula Modelling. In *Hierarchical Models and MCMC: A Tribute to Adrian Smith, P. Damien, P. Dellaportas, N. Polson, and D. Stephens (Eds)*, Bayesian Statistics, pages 395–402. Oxford University Press, Oxford (UK). xvi, 32

Smith, M. S., Gan, Q., and Kohn, R. J. (2012). Modelling dependence using skew-t copulas: Bayesian inference and applications. *Journal of Applied Econometrics*, 3(27): 500–522. 37

Stephens, M. (2000). Dealing with label switching in mixture models. *J. Royal Statist. Society Series B*, **62(4)** 795–809. xxii, 93

Su, S. (2007). Numerical maximum log likelihood estimation for generalized lambda distributions. *Computational Statistics and Data Analysis*, 51(8):3983-3998. 77

Tavaré, S., Balding, D. J., Griffiths, R. C., and Donnelly, P. (1997) Inferring Coalescence Times from DNA Sequence Data. *Genetics*, 145(2):505-518. 73

Titterington, D., Smith, A., and Makov, U. (1985). *Statistical Analysis of Finite Mixture Distributions*. John Wiley, New York. xxii, 93

Tukey, J. W. (1977) Modern techniques in data analysis. *NSFSponsored Regional Research Conference*, Southeastern Massachusetts University, North Dartmouth, Massachusetts. 77

Wasserman, L. (2000). Asymptotic inference for mixture models using data dependent priors. *J. Royal Statist. Society Series B*, **62** 159–180. xxiii, 94, 145

Watson, G.S. (1964). Smooth regression analysis. *Sankhyā*, 26:359–372. 61

Wu, J., Wang, X., and Walker, S. G. (2014). Bayesian Nonparametric Inference for a Multivariate Copula Function. *Methodol. Comput. Appl. Probab.*, 1(16): 747–763. xvi, 32