Bayesian Loss-based Approaches for Heterogeneous Data

Laurentiu Catalin Hinoveanu

School of Mathematics, Statistics and Actuarial Science

University of Kent

This dissertation is submitted for the degree of

Doctor of Philosophy

November 2019

I would like to dedicate this thesis to my wonderful and loving parents, Elena and Sorin Hinoveanu, my maternal aunt and uncle, Daniela and Dumitru (Mitică) Duţă, my cousins, Anidora and Laurenţiu Duţă, my relatives, Florentina, Amalia, Ionela and Sergiu Hornoiu, as well as the memories of my maternal grandparents, Alecsandra and Ion Duţă, my paternal grandparents, Ecaterina and Nicolae Hinoveanu, my paternal aunt and uncle, Liliana Ursu-Hinoveanu and Stelian Ursu, and my godmother, Constanţa Solomon. You will forever be in my heart.

Acknowledgements

I would like to acknowledge the great support of my supervisors, Dr Fabrizio Leisen and Dr Cristiano Villa, who were more than extraordinary and helpful mentors, they were, above all, great friends. I would also like to appreciate the efforts of Claire Carter, who offered great help in many times of need and Derek Baldwin whose IT advice was always on point. Furthermore, I am thankful to my colleagues and friends, Alan, Aniketh, José and Michele for the thoughtful discussions during the lunch breaks and for their unwavering optimism and good mood. Last but not least, I am extremely grateful to my parents for their never-ending positive attitude when dealing with the personal tragedies which affected us during these years.

Table of contents

List of figures

List of tables

1. Introduction

The objective Bayesian analysis has essentially started once the principle of insufficient reason was advocated by Laplace during the early 19^{th} century. This principle states that if there is not enough knowledge to subjectively build a prior distribution, one should consider all possible cases as equally likely, thus encouraging an uniform prior. Since the second half of the 20^{th} century there has been an increase in the number of methodologies which consider priors suitable for scenarios of no information or of insufficient prior information. Amongst such priors, we would like to mention the Jeffreys's prior (Jeffreys, 1961), the probability matching prior (Welch and Peers, 1963), the reference prior (Bernardo, 1979), the fractional prior (O'Hagan, 1995, 1997) and the loss-based prior (Villa and Walker, 2015b). According to Kass and Wasserman (1996) and Ghosh (2011), the Jeffreys's prior is part of the priors invariant under the action of a group, whilst the reference prior is included in the more general class of divergence priors. The probability matching prior is a prior designed to conform to some frequentist properties, whereas the fractional prior deals with a certain fraction of the data called training sample and is used to tackle the problems created by improper priors for the Bayes Factors. The loss-based prior is based on decision-theoretic arguments. More details about these priors are provided in the following chapter.

The need for objective procedures may stem from the lack of sufficient prior information or from the impracticability of using it. For example, if a model has a large number of parameters, then prior elicitation is not feasible. As such, there have been developed

automated procedures to obtain prior distributions. The collection of all these procedures falls under the name of *Objective Bayes*.

Before embarking in more detail on the objective prior method we have considered together with our contribution to the objective Bayesian literature, we would like to do a short introduction of the Bayesian principles in the context of model selection. Let $M = \{f(x|\theta), \pi(\theta)\}$ represent a Bayesian model comprised of the data generating distribution $f(x|\theta)$ and the prior π on the model parameter (possibly a vector of parameters) θ. The prior can be specified both subjectively and objectively. As it is not the focus of this work, we would like to avoid reviewing subjective priors. More information about them can be found in the works of Ramsey (1926), de Finetti (1937), Lindley (1972), Goldstein (2006), amongst others. In regards to the objective prior, we have used the loss-based prior of Villa and Walker (2015b), which has also been applied in a model selection context by Villa and Walker (2015a). The main idea of the above prior is the link between the misspecification of a model in terms of the limiting behaviour of the posterior distribution as shown by Berk (1966) and the idea of *self-information loss*. In the Bayesian context, we are interested in the posterior distribution which has the form

$$\pi(\theta|x) \propto f(x|\theta) \cdot \pi(\theta). \tag{1.1}$$

From equation (1.1), we utilise the loss-based prior or the principle behind it to compute the posterior through the Bayesian updating process and address various issues in the area of *change point analysis*, as well as *Gaussian Graphical Models* (GGMs) and *proper binary trees*. These represent our contribution to the objective Bayesian literature.

2. Preliminaries

This chapter focusses on introducing the literature about objective methods for continuous and discrete parameters. We also present the loss-based prior of Villa and Walker (2015b). The last sections of the chapter comprise some of the literature about change point problems, Gaussian graphical models and proper binary trees.

2.1 Objective Methods for Continuous Parameters

In this section, we are going to discuss some objective Bayesian procedures for continuous parameters.

Jeffreys's prior

Jeffreys's prior (Jeffreys, 1961) exhibits invariance in regards to bijective transformations of its arguments and is related to the Fisher information. For the unidimensional case, the Fisher information is

$$I(\theta) = \mathbb{E}_\theta \left[\left(\frac{\partial \log f(x|\theta)}{\partial \theta} \right)^2 \right]. \tag{2.1}$$

Equation (2.1) represents the information that the model under the data generation process $f(x|\theta)$ contains about the parameter θ. Under some regularity conditions, the Fisher

information can be defined as

$$I(\theta) = -\mathbb{E}_\theta \left[\left(\frac{\partial^2 \log f(x|\theta)}{\partial \theta^2} \right) \right].$$

In the unidimensional setting, the Jeffreys's prior has the form:

$$\pi_J(\theta) \propto I(\theta)^{1/2}.$$

The choice of the prior related to the Fisher information has the property of being invariant under one-to-one (bijective) transformations. For a bijective transformation h of θ, we get through the change-of-variables formula

$$\pi(\theta) = \pi(h(\theta)) \frac{\partial h(\theta)}{\partial \theta}$$

and this translates to the following result

$$I(\theta) = I(h(\theta)) \left(\frac{\partial h(\theta)}{\partial \theta} \right)^2,$$

which is the change-of-variables rule applied to the Fisher information. A characteristic of the Jeffreys's prior is to put relatively more mass on those regions of the parameter space where the Fisher information is highly concentrated.

For the multidimensional parameter $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_k)$, the Jeffreys's prior has the form:

$$\pi_J(\boldsymbol{\theta}) \propto \det(\boldsymbol{I}(\boldsymbol{\theta}))^{1/2}, \tag{2.2}$$

where $\boldsymbol{I}(\boldsymbol{\theta})$ is the $k \times k$ Fisher information matrix, which must be positive-definite, with the $(i,j)^{\text{th}}$ element given by

$$\boldsymbol{I}_{(i,j)}(\boldsymbol{\theta}) = \mathbb{E}_{\boldsymbol{\theta}} \left[\left(\frac{\partial \log f(x|\boldsymbol{\theta})}{\partial \theta_i} \right) \left(\frac{\partial \log f(x|\boldsymbol{\theta})}{\partial \theta_j} \right) \right].$$

Clearly for the univariate case, the determinant from equation (2.2) simplifies to the Fisher information under the single parameter θ.

In the multidimensional case, the Jeffreys's prior may lead to incoherent results, as outlined by Robert (2007). To address the issue, Jeffreys recommends to assume the parameters as a priori independent. As such, the Jeffreys's prior $\pi_J(\boldsymbol{\theta})$ simply becomes $\pi_J(\boldsymbol{\theta}) = \prod_{i=1}^{k} \pi_J(\theta_i)$ and is called Jeffreys's independence prior. For more details and examples, refer to Kass and Wasserman (1996).

Reference prior

Reference priors have been introduced by Bernardo (1979) and the general idea is to maximise in expectation the difference in information between the prior and the posterior, where the difference is measured via the Kullback–Leibler (KL) divergence (Kullback and Leibler, 1951) defined as:

$$D_{KL}(p||q) = \begin{cases} \displaystyle\sum_{x_i \in \mathscr{X}} p(x_i) \cdot \log \left[\frac{p(x_i)}{q(x_i)} \right], & \text{if } \mathscr{X} \text{ is discrete} \\ \\ \displaystyle\int_{\mathscr{X}} p(x) \cdot \log \left[\frac{p(x)}{q(x)} \right] \mathrm{d}x, & \text{if } \mathscr{X} \text{ is continous,} \end{cases}$$

where $p, q : \mathscr{X} \to (0, +\infty)$ are two probability distributions. The *expected information* (Shannon (1948) and Lindley (1956)), is the expectation with respect to the marginal density $f(\boldsymbol{x})$ of the KL divergence between the posterior distribution $\pi(\theta|\boldsymbol{x})$ and the prior distribution $\pi(\theta)$, that is

$$I(\pi|M) = \int_{\mathscr{X}} D_{KL}(\pi(\theta|\boldsymbol{x})||\pi(\theta)) f(\boldsymbol{x})\, \mathrm{d}\boldsymbol{x},$$

where $M = \{f(\boldsymbol{x}|\theta), \boldsymbol{x} \in \mathscr{X}, \theta \in \Theta\}$ is the statistical model from which the data is generated, under a certain configuration for parameter θ, and $\boldsymbol{x}$ is the observable data vector. To provide a formal definition of the reference prior we need to specify the meaning behind two characteristics which it needs to obey, namely maximising missing information and prior permissibility. Consider $\Theta \subset \mathbb{R}$. Let us denote with M^k the number of independent k realizations $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_k$ obtained under the previously defined model M with a continuous parameter space. Denote by $\mathscr{P}$ the set of priors of θ such that $\int_\Theta f(\boldsymbol{x}|\theta)\pi(\theta)\, \mathrm{d}\boldsymbol{x} < \infty$. If $\forall \Theta_0 \subset \Theta$ and $\forall \tilde{\pi} \in \mathscr{P}$, with Θ_0 compact, we have

$$\lim_{k\to\infty} \left\{ I(\pi_0|M^k) - I(\tilde{\pi}_0|M^k) \right\} \geq 0,$$

where π_0 and $\tilde{\pi}_0$ are the renormalized restrictions of the priors $\pi(\theta)$ and $\tilde{\pi}(\theta)$ to Θ_0, then the prior π has the maximising missing information property for model M given $\mathscr{P}$. Intuitively, as we repeatedly run model M a very large k number of times, the expected information related to any prior should exist and provide a measure of the missing information about parameter of interest θ associated to the specific prior, because the sequence of observations $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_k$ should give more information about the parameter of interest as k increases. In other words, when $k \to \infty$, $I(\pi|M^k)$ represents the missing information about θ related to the prior $\pi(\theta)$. Furthermore, we would like that the missing information corresponding to prior $\pi(\theta)$ to be larger compared to the missing information associated with any alternative prior $\tilde{\pi}(\theta)$. A prior $\pi(\theta)$ is permissible for model M if it satisfies the following two conditions.

Firstly, $\forall \boldsymbol{x} \in \mathscr{X}$, the posterior $\pi(\theta|\boldsymbol{x})$ is proper. Secondly, for an increasing sequence of compact subsets of Θ, $\{\Theta_i\}_{i=1}^{\infty}$, converging to Θ, the posteriors corresponding to the parameters located in those subsets are *expected logarithmically convergent* to the posterior $\pi(\theta|\boldsymbol{x})$. For a sequence of posterior distributions $\{\pi_i(\theta|\boldsymbol{x})\}_{i=1}^{\infty}$ to be expected logarithmically convergent to the posterior $\pi(\theta|\boldsymbol{x})$, it means that $\lim_{i\to\infty} \int_{\mathscr{X}} D_{KL}(\pi_i(\theta|\boldsymbol{x})||\pi(\theta|\boldsymbol{x}))f_i(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} = 0$, where $f_i(\boldsymbol{x}) = \int_{\Theta_i} f(\boldsymbol{x}|\theta)\pi_i(\theta)\,\mathrm{d}\theta$ is the prior predictive distribution of $\boldsymbol{x}$ marginalised over the compact parameter subset Θ_i. So a permissible prior $\pi(\theta)$ in this context simply means that its posterior needs to obey the expected logarithmically convergent condition. This condition is necessary to guarantee that in the case of improper priors we obtain proper posteriors which arise as suitable limits of posteriors derived under proper priors. Specifically, the prior obeying the aforementioned condition simply leads to a posterior which marginalised over $\boldsymbol{x}$ is approximately equivalent to a proper posterior obtained when we restrict the parameter space to a large compact subset of it. Therefore, the formal definition of a reference prior $\pi(\theta)$ for a parametric model M given a $\mathscr{P}$ class of prior functions is a prior which is both permissible and maximizes the missing information.

Berger et al. (2009) provide a specific way to build the reference prior numerically under some mild conditions. First, we need to define the idea of standard parametric model and standard class of priors. A standard class of priors, $\mathscr{P}_s$, is the collection of strictly positive and continuous priors $\pi(\theta)$ on Θ such that $f(\boldsymbol{x}) = \int_{\Theta} f(\boldsymbol{x}|\theta)\pi(\theta) < \infty$. A parametric model M where the parameters are continuous is considered standard if $\forall \pi \in \mathscr{P}_s$ and $\forall \Theta_0$, we have $I(\pi_0|M^k) < \infty$, where π_0 is the standard prior $\pi(\theta)$ restricted to Θ_0 and $I(\pi_0|M^k)$ is the expected information for k independent realizations $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_k$ under M. Due to the k independent realizations and a specific property of the reference prior, we may consider computationally convenient to work with sufficient statistics for the construction of the reference prior. According to Bernardo (2005), the sufficient statistics are actually *asymptotically sufficient statistics*, that is a function of the data $\boldsymbol{t}_k = \boldsymbol{t}_k(\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_k)$, such

that $\forall \theta \in \Theta \subset \mathbb{R}$ and $\forall \boldsymbol{x}_1, \ldots, \boldsymbol{x}_k \in \mathscr{X}$, we have $\lim_{k \to +\infty} \dfrac{f(\theta|\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_k)}{f(\theta|\boldsymbol{t}_k)} = 1$. Therefore, when we define the reference prior in terms of sufficient statistics, there would be no loss of generality compared with using the entire sample. Furthermore, as Bernardo (2005) and Berger et al. (2009) outline, the reference prior under the entire data is the same with the reference prior under the relevant sufficient statistics, due to the invariance of the expected information to the transformations involved in creating those sufficient statistics.

Let $\boldsymbol{t}_k = \boldsymbol{t}_k(\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_k) \in \mathscr{T}_k$ be a sufficient statistic for the k replications under model M. Under certain mild conditions discussed by Berger et al. (2009), the reference prior for a standard model M given a standard class of priors $\mathscr{P}_s$ is

$$\pi(\theta) = \lim_{k \to \infty} g_k(\theta)/g_k(\theta_0),$$

with

$$g_k(\theta) = \exp\left\{ \int_{\mathscr{T}_k} f(\boldsymbol{t}_k|\theta) \log \left[\frac{f(\boldsymbol{t}_k|\theta)\pi^*(\theta)}{\int_\Theta f(\boldsymbol{t}_k|\theta)\pi^*(\theta)\,\mathrm{d}\theta} \right] \mathrm{d}\boldsymbol{t}_k \right\},$$

where $\pi^*(\theta)$ is an arbitrary fixed prior and θ_0 is an interior point of Θ.

Across all previous definitions and conditions from this section, we have considered θ as an one-dimensional parameter. The general multivariate case has been treated by Berger and Bernardo (1992a). According to the original authors, this general case of the reference method is notationally quite complex and very hard to implement. As such, for the multidimensional case we restrict our attention to the case of two parameters, that is $\boldsymbol{\theta} = (\omega, \lambda)$, a case also described by Kass and Wasserman (1996). Here, we consider ω as the parameter of interest and λ as a nuisance parameter. Then, the analysis proceeds as following: keeping ω fixed, we first use the standard reference method to derive the reference prior for λ, that is $\pi^{\mathrm{R}}(\lambda|\omega)$. We then use the reference prior for λ to compute $f(\boldsymbol{x}|\omega) = \int f(\boldsymbol{x}|\omega, \lambda)\pi^{\mathrm{R}}(\lambda|\omega)\,\mathrm{d}\lambda$. Next, we employ again the standard reference method to obtain $\pi^{\mathrm{R}}(\omega)$ by using the previously computed $f(\boldsymbol{x}|\omega)$. Now, the reference prior for $\boldsymbol{\theta}$ is

simply $\pi^{\mathrm{R}}(\boldsymbol{\theta}) = \pi^{\mathrm{R}}(\omega)\pi^{\mathrm{R}}(\lambda|\omega)$. As outlined by Kass and Wasserman (1996), when some regularity conditions are obeyed, the reference prior for this particular case $\boldsymbol{\theta} = (\omega, \lambda)$ has a clear form:

$$\pi^{\mathrm{R}}(\omega, \lambda) \propto \pi_J(\lambda_\omega) \exp\left\{ \int \pi_J(\lambda_\omega) \log[S(\omega, \lambda)] \, \mathrm{d}\lambda \right\}, \tag{2.3}$$

where $\pi_J(\lambda_\omega)$ is the standard Jeffreys's prior for λ when ω is fixed and $S = \sqrt{\frac{\det(\boldsymbol{I}(\boldsymbol{\theta}))}{\det(\boldsymbol{I}_{(2,2)}(\boldsymbol{\theta}))}}$ with $\boldsymbol{I}(\boldsymbol{\theta})$ being the multidimensional Fisher information matrix and $\boldsymbol{I}_{(2,2)}(\boldsymbol{\theta})$ representing the $(2,2)^{\mathrm{nd}}$ element of $\boldsymbol{I}(\boldsymbol{\theta})$ that corresponds to the nuisance parameter. Clearly, if we swap the importance of the two parameters around we get a different result for the $\pi^{\mathrm{R}}(\boldsymbol{\theta})$. As such, in the multidimensional case, we have that different orderings of importance for the parameters lead to different reference priors. More details about how to order the parameters for the multidimensional reference analysis is provided by Berger and Bernardo (1992b). For an overview on how to tackle the multidimensional case when more than one nuisance parameter is present we would like to refer to Bernardo (2005). The basic principle outlined by Bernardo (2005) is represented by a simple application of the chain rule of probability and then working backwards in identifying the reference priors for the least significant nuisance parameter to the parameter of interest akin to the way we have identified the reference prior in equation (2.3).

Probability matching prior

This type of prior was developed with the idea of the corresponding Bayesian credible interval *matching* its frequentist analogue in terms of coverage probabilities. It was introduced by Welch and Peers (1963) and further advanced by Datta and Mukerjee (2004) and Ghosh (2011). We are interested in prior $\pi(\theta_1)$ for real one-dimensional parameter θ_1 such that $\Pr(\theta_1 \leq \theta_1^{1-\alpha}(\pi, \boldsymbol{X})) = 1 - \alpha + o(n^{-r/2}), \forall r \in \{1, 2, 3, \ldots\}, \forall \alpha \in (0, 1)$, where $\boldsymbol{X}$ is a vector of n i.i.d. (independent and identically distributed) random variables sampled from distri-

bution $f(\boldsymbol{x}|\boldsymbol{\theta})$, $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_d)^T$, $d = 1$; α is the confidence level and $\theta_1^{1-\alpha}(\pi, \boldsymbol{X})$ is the $1-\alpha$ posterior quantile of θ_1 under the prior $\pi(\theta_1)$. When $r = 1$ this prior is called the *first order* probability matching prior for posterior quantile, whereas when $r = 2$ it is referred as *second order* probability matching prior for posterior quantile. As Scricciolo (1999) remarks, Welch and Peers (1963) have shown that Jeffrey's prior is a second order probability matching prior for posterior quantile, which was also noticed by Datta and Mukerjee (2004) in terms of it satisfying certain partial differential equations. When nuisance parameters are present, therefore we consider a multivariate setting for our model parameters, the first order probability matching is not unique. Furthermore, Datta and Sweeting (2005) remark that in the case of multivariate parameters, the first order probability matching holds as there is an equivalence between the normal approximations in both Bayesian and frequentist contexts. Under certain partial differential equations, a second order probability matching can be obtained when we consider a single parameter of interest, let us say θ_1, and the rest of the d-dimensional model parameter vector $\boldsymbol{\theta}$ as nuisance parameters. Then, by denoting with $z_{\theta_1}^{1-\alpha}(\pi, \boldsymbol{X})$ the $1-\alpha$ marginal posterior quantile of θ_1 under the prior $\pi(\theta_1)$, we consider π a second order matching probability if $\Pr(\theta_1 \leq z_{\theta_1}^{1-\alpha}(\pi, \boldsymbol{X})) = 1 - \alpha + o(n^{-1})$ holds. In this multivariate framework, it is necessary to distinguish two approaches regarding the probability matching priors. The first is concerned with finding for each parameter of interest a particular second order prior; the resulting priors are called *simultaneous marginal* second order probability matching priors. The second method consists in obtaining priors by matching the corresponding frequentist and posterior joint cumulative distribution functions. Both approaches originated with the work of Datta (1996). Besides the quantile and distribution matching, other types of matching priors are possible such as when we consider highest posterior density regions, other credible regions like the inversion of likelihood ratio statistics and others. The highest posterior density regions have the smallest volume for a given credible level. As this type of priors do not constitute the subject of this thesis, a more

in-depth discussion is provided by Datta and Mukerjee (2004) and Datta and Sweeting (2005) and the references therein.

Fractional prior

This type of prior was introduced by O'Hagan (1997) as a way to solve the problems created by improper priors for *Bayes Factors* (BFs). As explained by Berger and Pericchi (1996), when improper priors are involved, the BF is defined up to a ratio of arbitrary constants. According to de Santis and Spezzaferri (1999), the BFs can be expressed as the ratio of posterior odds to the prior ones, thus suggesting the change in the odds favoured by the data. Before formally introducing the fractional prior, we have to establish the notion of *partial Bayes Factors* (PBFs) (de Santis and Spezzaferri, 1999). The idea that stands at the framework of PBFs is to split the sample of data $y(n)$ in a training set, denoted by $y(l)$, with the remaining $y(n-l)$ data being used to make the model comparisons, where n represents the sample size and $0 < l < n$ describes the size of the training data. The purpose of this split is to use the training data to compute proper posteriors $\pi(\cdot|y(l))$ starting from the improper prior $\pi^N(\cdot)$ and then use those proper posteriors as the priors when defining the PBFs. Thus, the partial Bayes Factor for a model M_j against a model M_k where the training sample has size l, that is $B_{jk}(l)$, is simply represented as

$$B_{jk}(l) = \frac{\int_{\Theta_j} f_j(y(n-l)|\theta_j)\pi_j(\theta_j|y(l))\,\mathrm{d}\theta_j}{\int_{\Theta_k} f_k(y(n-l)|\theta_k)\pi_k(\theta_k|y(l))\,\mathrm{d}\theta_k} = \frac{B^N_{jk}(y)}{B^N_{jk}(y(l))}, \tag{2.4}$$

where $B^N_{jk}(y)$ and $B^N_{jk}(y(l))$ are the BFs under the improper priors $\pi^N_j(\theta_j)$ and $\pi^N_k(\theta_k)$ when the full sample data and the training data are considered, respectively. The idea of O'Hagan (1995) bypasses the arbitrary choice of the training data $y(l)$ for the PBFs, by using a certain subunitary fractional power $b = l/n$ of the likelihood to solve the problems caused by improper priors to BFs. As such, the correction $B^N_{jk}(y(l))$ from equation (2.4) is replaced by

$B^b_{jk}(y)$, thus obtaining the *fractional Bayes Factor* (FBF) which is

$$B^{FBF}_{jk}(y) = \frac{B^N_{jk}(y)}{B^b_{jk}(y)},$$

where

$$B^b_{jk}(y) = \frac{\int_{\Theta_j} f^b_j(y|\theta_j)\pi^N_j(\theta_j)\,\mathrm{d}\theta_j}{\int_{\Theta_k} f^b_k(y|\theta_k)\pi^N_k(\theta_k)\,\mathrm{d}\theta_k}.$$

The choice of b can be either $b = l_0/n$, where l_0 is the minimal training sample size, that is the size for which $0 < \int_{\Theta_i} f_i(y(l_0)|\theta_i)\pi^N_i(\theta_i)\,\mathrm{d}\theta_i < \infty$ and no other subset of it can be found where the integrated likelihood is finite, or, as discussed in O'Hagan (1995), can be $b = \max\{l_0, \sqrt{n}\}/n$ or $b = \max\{l_0, \log(n)\}/n$.

The fractional prior (O'Hagan, 1997) represents the proper prior for which the Bayes Factor is asymptomatically equivalent for a sequence $\{b_n\}$ of fractional powers to the FBF under the improper prior.

An alternative to the fractional prior is represented by the intrinsic prior introduced by Berger and Pericchi (1996). Before providing an interpretation of this prior, we need to briefly describe the *intrinsic Bayes Factors* (IBFs). The IBFs are obtained from PBFs by averaging in a certain way across all possible L minimal training samples z_h, $h = 1, 2, \ldots L$. Amongst the averaging, some well-known examples are represented by the arithmetic (AIBF) and the geometric means (GIBF), characterised as:

$$B^{\mathrm{A}}_{jk}(y) = B^N_{jk}(y) \cdot \frac{\sum_{h=1}^{L} B^N_{kj}(z_h)}{L},$$

$$B^{\mathrm{G}}_{jk}(y) = B^N_{jk}(y) \cdot \left[\prod_{h=1}^{L} B^N_{kj}(z_h)\right]^{\frac{1}{L}}.$$

As outlined by de Santis and Spezzaferri (1999) and Consonni et al. (2018), the intrinsic prior is that proper prior for which the computed BFs are asymptotically equivalent to the IBFs.

2.2 Objective Methods for Discrete Parameters

In this section, we are reviewing some of the objective priors used for discrete parameters. The literature on the subject is geared more towards outlining context-specific fixes as summarised by Villa and Walker (2015b). Recently, a general framework to address setting priors for discrete parameter spaces was introduced by Villa and Walker (2015b). This framework is discussed in more detail in Section 2.3.

Jeffreys (1961) has proposed the following prior on the space of positive integers $\pi(n) \propto n^{-1}$, where $n \in \{1,2,3,\ldots\}$. Another prior on the positive integers based on information theoretical reasons, which we will cover next, was proposed by Rissanen (1983). Following Rissanen (1983) and Kass and Wasserman (1996), we start by assigning a binary string, called code word, to every positive integer as a bijective map. Another important assumption is that the resulting code is a prefix code, that is a code made from code words such that there are no code words which are the initial segments of other code words in the respective code. This assumption allows one to easily discriminate between the code words. Now, let us consider the initial knowledge about the positive integers expressed through the distribution $P = (P(1), P(2), \ldots)$ satisfying certain constraints. According to Rissanen (1983), P is called the "test" distribution and quantifies the initial knowledge about our positive integers. Furthermore, we have $P(i) < 1, \forall i$ and $P(i) \geq P(i+1), i > M,$ for some M. The first condition ensures that the distribution P is non-singular in the interval $[1, M]$, whereas the second condition can be intuitively described as saying that the bigger the integer is the less probable it is. There is also a third technical condition related to the information entropy, namely $H(P) = -\sum_{i=1}^{N} P(i)\log(P(i)) = \infty$. This third condition is needed to get a solution. We must note that even though the entropy is infinite, this does not lead to paradoxes, as every integer still has a finite code length. In correspondence with the distribution P, let us denote by $L = (L(1), L(2), \ldots)$ the sequence of the lengths for the code words. Then, the idea is to search for the code where the code lengths are the shortest. Formally, this means to

solve the following optimization problem:

$$\min_{L} \sup_{P} \lim_{N\to\infty} \frac{\sum_{i=1}^{N} P(i)L(i)}{-\sum_{i=1}^{N} P(i)\log(P(i))}. \tag{2.5}$$

Under certain regularity conditions, the solution to the optimization from (2.5) is $L_0(n) = \log_2(c) + \log_2(n) + \log_2(\log_2(n)) + \ldots$, where $c = 2.865064$. Following certain coding theory constraints, namely the Kraft–McMillan inequality for prefix codes, Rissanen (1983) has proposed the prior $\pi(n) \propto 2^{-L_0(n)}$, which is proper subject to the aforementioned constraints. The prior can be expressed as:

$$\pi(n) \propto \frac{1}{c} \cdot \frac{1}{n} \cdot \frac{1}{\log_2(n)} \cdot \frac{1}{\log_2(\log_2(n))} \cdot \ldots \cdot \frac{1}{\log_2(\log_2(\ldots \log_2(n) \ldots))}. \tag{2.6}$$

Note that the product from equation (2.6) has a finite number of terms as it contains just those factors for which the iterated logarithm is defined, that is for the positive arguments.

Another approach regarding objective priors for discrete parameters was introduced by Berger et al. (2012). It is known that for a finite discrete space, the standard reference theory will generate the discrete uniform prior on that space, whilst on a countably infinite space, it leads to the appearance of a non-constant normalisation factor which is to be avoided. As suggested by Berger et al. (2012) regarding finite discrete spaces, this would not constitute a problem when the parameter space does not contain any structure. When there is structure, the uniform prior is not desirable as it will become apparent from Example 1 which was originally provided by the authors. As such, Berger et al. (2012) recommend enclosing the discrete parameter problem into a continuous one and use procedures motivated by asymptotics. Among those procedures, we recall considering a consistent estimator for the required parameter or applying formal limiting operations on the data. These approaches lead to a continuity assumption for the parameter in cause. Another possible solution to deal

with structured discrete parameter spaces is the introduction of a continuous hierarchical hyperparameter as in the following example.

***Example* 1.** *Consider a random variable X distributed according to the* Hypergeometric(K,N,n) *where the finite population has size* $N \in \{0,1,2,\ldots\}$ *with exactly* $K \in \{0,1,\ldots,N\}$ *successes. Let k be the observed number of successes in* $n \in \{0,1,\ldots,N\}$ *draws without replacement. Then the probability of k successes from n draws without replacement is:*

$$\Pr(X = k|K,N,n) = \frac{\binom{N}{k} \cdot \binom{N-K}{n-k}}{\binom{N}{n}}.$$

Let us suppose that K is unknown, whereas N is known. Clearly, K has values in a finite discrete space, but embeds a certain structure through the hypergeometric distribution. It is known, that for large N compared to n, the random variable X is approximately distributed as Binomial$(n,K/N)$ *random variable. The objective prior for K should therefore be related in a certain way to the objective prior for the probability of success in the binomial, probability denoted by p, usually taken in the literature as the Jeffreys's prior, namely* $\pi(p) \sim \text{Beta}(0.5,0.5)$. *Berger et al. (2012) propose an objective prior that is compatible with the objective prior for p through a certain embedding strategy that involves the introduction of a continuous hyperparameter and then proceeding with the standard reference analysis. In the hypergeometric example from above, we may assume that* $K \sim$ Binomial(N,p) *with unknown p. Integrating out K, we have:*

$$\begin{aligned}\Pr(X = k|N,n,p) &= \sum_{K=0}^{N} \Pr(X = k|K,N,n) \cdot \Pr(K|N,p) \\ &= \binom{n}{k} p^k (1-p)^{n-k}. \end{aligned} \tag{2.7}$$

The algebraic expression from equation (2.7) *designates a* Binomial(n,p). *As Berger et al. (2012) remind us, the reference analysis for the unknown p parameter of a binomial simply*

yields the Jeffreys's prior, namely $\pi^{\mathrm{R}}(p) \sim \mathrm{Beta}(0.5, 0.5)$. *As such, integrating out this probability of success p using the previously mentioned reference prior, gives us the following reference prior for K:*

$$\pi^{\mathrm{R}}(K|N) = \frac{1}{\pi} \frac{\Gamma\left(K+\frac{1}{2}\right)\Gamma\left(N-K+\frac{1}{2}\right)}{\Gamma(K+1)\Gamma(N-K+1)}.$$

Basically, in the aforementioned example we have considered the following hierarchical model:

$$K|p \sim \mathrm{Binomial(N, p)}$$
$$p \sim \mathrm{Beta}(\alpha, \beta)$$

The choice of the $\mathrm{Beta}(\alpha, \beta)$ is a natural one, since this distribution is a conjugate to the binomial distribution. This means that if we marginalise over p, we simply obtain the $\mathrm{Beta-Binomial}$ distribution (Griffiths, 1973). That is $K \sim \mathrm{Beta-Binomial(N}, \alpha, \beta)$ with the probability mass given as

$$\Pr(K|N, \alpha, \beta) = \frac{\binom{N}{K} B(K+\alpha, N-K+\beta)}{B(\alpha, \beta)},$$

where $B(\alpha, \beta) = \dfrac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$ is the Beta function with parameters α and β. Note that

$$\binom{N}{K} = \frac{\Gamma(N+1)}{\Gamma(K+1)\Gamma(N-K+1)},$$

where Γ is the Gamma function. As in our designation of the hierarchical model, K depends on p, then an objective prior on K would definitely necessitate an objective prior on p. We know that Jeffreys's prior for p represents an objective choice. As such, if we substitute

$\alpha = \beta = 1/2$ which represents that objective prior for p, we obtain the reference prior mentioned in Example 1.

2.3 Objective Prior for Discrete Parameters

In this section, we are presenting the concepts that stand at the basis of our proposed methodologies from the next chapters, respectively the objective priors introduced by Villa and Walker (2015b) for discrete parameters, as well as their extension to model prior probabilities (Villa and Walker, 2015a). We will heavily utilise the latter methodology in our approach to change point analysis and GGMs.

To illustrate the idea, consider a probability distribution $f(x|m)$, where $m \in \mathbb{M}$ is a discrete parameter. Then, the prior $\pi(m)$ is obtained by objectively measuring what is lost if the value m is removed from the parameter space, and it is the true value. According to Berk (1966), if a model is misspecified, the posterior distribution asymptotically accumulates on the model which is the most similar to the true one, where the similarity is measured in terms of the KL divergence. Therefore, $D_{KL}(f(\cdot|m)\|f(\cdot|m'))$, where m' is the parameter characterising the nearest model to $f(x|m)$, represents the utility of keeping m. The objective prior is then obtained by linking the aforementioned utility, or more precisely the loss, via the self-information loss:

$$\pi(m) \propto \exp\left\{\min_{m' \neq m} D_{KL}(f(\cdot|m)\|f(\cdot|m'))\right\} - 1.$$

This self-information loss was outlined by Merhav and Feder (1998) in the universal prediction context. Consider the problem of making a prediction on an outcome, say $x_t \in \mathscr{X}$, after observing $\boldsymbol{x}_{t-1} = (x_1, x_2, \ldots, x_{t-1}) \in \mathscr{X}$. Consider the conditional probability $\Pr_t(x_t|\boldsymbol{x}_{t-1})$. Once x_t is known, the previously mentioned conditional probability assignment is evaluated through a certain loss function l, which should be monotonically decreasing

regarding the respective probability statement. Such a loss is the self-information loss, which states that for a certain event x, the loss incurred by a probability statement $P = \{\Pr(x), x \in \mathscr{X}\}$ associated to x is given by:

$$l(P,x) = -\log(\Pr(x)).$$

The self-information loss satisfies certain properties, besides the monotonicity one. Due to its logarithmic form, it is advantageous to work with it, as products of conditional probabilities are transformed to cumulative sums. This means that the self-information loss corresponding to the joint distribution of two independent probability statements is simply the aggregate of the self-information losses corresponding to each of the two individual statements. Another property is represented by the equivalence between the choice of the probability statement that minimises the self-information and the choice of the maximum likelihood estimator (MLE). Intuitively, self-information loss can be described as the measure of uncertainty we have about the occurrence of an event. As such, if an event occurs almost surely, then the uncertainty around it will be zero, therefore leading us to ascribe a value of zero to the self-information loss associated to the respective event. Clearly, as we are more unsure about the realisation of an event, this will lead to a higher surprise when it actually happens, thus ending with a higher self-information loss. In some contexts the self-information loss appears under the moniker of 'surprisal'. As another measure of uncertainty for the occurrence of an event is represented by the probability of its realisation, the link between the self-information loss and the respective probability statement is given by the aforementioned negative logarithmic form. We must also note that in the area of data compression, the self-information loss $-\log(\Pr(x))$ represents the ideal code length of x with respect to a probability statement $\Pr()$, as outlined by Merhav and Feder (1998).

Villa and Walker (2015a) have extended the concept behind the objective prior on discrete parameters as to allow them to define a prior on the space of models. All our subsequent

analysis in the later chapters is based upon this latter extension. To illustrate, let us consider k Bayesian models:

$$M_j = \{f_j(x|\theta_j), \pi_j(\theta_j)\}, \qquad j \in \{1, 2, \ldots, k\},$$

where $f_j(x|\theta_j)$ is the sampling density characterised by θ_j and $\pi_j(\theta_j)$ represents the prior on the model parameter.

Assuming the prior on the model parameter, $\pi_j(\theta_j)$, is proper, the model prior probability $\Pr(M_j)$ is proportional to the expected minimum KL divergence from M_j, where the expectation is considered with respect to $\pi_j(\theta_j)$. That is:

$$\Pr(M_j) \propto \exp\left\{\mathbb{E}_{\pi_j}\left[\inf_{\theta_i, i \neq j} D_{KL}(f_j(x|\theta_j) \| f_i(x|\theta_i))\right]\right\}, \qquad j = 1, \ldots, k. \tag{2.8}$$

Note that according to Villa and Walker (2015a), the quantities shown in the exponential from equation (2.8) represent the worth of model M_j weighted by the prior on the model parameters associated with M_j. By worth, the original authors simply mean what is lost if the model is removed from the list of models and it turns out to be the true model. This loss is simply the KL divergence between the model and the nearest one to it across the available options. The use of the prior distribution on the model parameters is due to the fact that these parameters are unknown so they need to be integrated out.

The model prior probabilities defined in equation (2.8) can be employed to derive the model posterior probabilities through:

$$\Pr(M_i|x) = \left[\sum_{j=1}^{k} \frac{\Pr(M_j)}{\Pr(M_i)} B_{ji}\right]^{-1}, \tag{2.9}$$

where B_{ji} is the Bayes factor between model M_j and model M_i, defined as

$$B_{ji} = \frac{\int f_j(x|\theta_j)\pi_j(\theta_j)\,\mathrm{d}\theta_j}{\int f_i(x|\theta_i)\pi_i(\theta_i)\,\mathrm{d}\theta_i},$$

with $i \neq j \in \{1,2,\ldots,k\}$.

2.4 Bayesian Analysis of Change Point Problems

There are several practical scenarios where it is inappropriate to assume that the distribution of the observations does not change. For example, financial datasets can exhibit alternate behaviours due to crisis periods. In this case it is sensible to assume changes in the underlying distribution. The change in the distribution can be either in the value of one or more of the parameters or, more in general, on the family of the distribution. In the latter case, for example, one may deem appropriate to consider a normal density for the stagnation periods, while a Student t, with relatively heavy tails, may be more suitable to represent observations in the more turbulent stages of a crisis. The task of identifying if, and when, one or more changes have occurred is not trivial and requires appropriate methods to avoid detection of a large number of changes or, at the opposite extreme, seeing no changes at all. The change point problem has been deeply studied from a Bayesian point of view. Chernoff and Zacks (1964) focused on the change in the means of normally distributed variables. Smith (1975) looked into the single change point problem when different knowledge of the parameters of the underlying distributions is available: all known, some of them known or none of them known. Smith (1975) focuses on the binomial and normal distributions. In Muliere and Scarsini (1985) the problem is tackled from a Bayesian nonparametric perspective. The authors consider Dirichlet processes with independent base measures as underlying distributions. In this framework, Petrone and Raftery (1997) have showed that the Dirichlet process prior could have a strong effect on the inference and may lead to wrong

conclusions in the case of a single change point. Raftery and Akman (1986) have approached the single change point problem in the context of a Poisson likelihood under both proper and improper priors for the model parameters. Carlin et al. (1992) build on the work of Raftery and Akman (1986) by considering a two level hierarchical model. Both papers illustrate the respective approaches by studying the well-known British coal-mining disaster dataset. In the context of multiple change points detection, Loschi and Cruz (2005) have provided a fully Bayesian treatment for the product partitions model of Barry and Hartigan (1992). Their application focused on stock exchange data. Stephens (1994) has extended the Gibbs sampler introduced by Carlin et al. (1992) in the change point literature to handle multiple change points. Hannart and Naveau (2009) have used Bayesian decision theory, in particular 0-1 cost functions, to estimate multiple changes in homoskedastic normally distributed observations. Schwaller and Robin (2017) extend the product partition model of Barry and Hartigan (1992) by adding a graphical structure which could capture the dependencies between multivariate observations. Fearnhead and Liu (2007) proposed a filtering algorithm for the sequential multiple change points detection problem in the case of piecewise regression models. Henderson and Matthews (1993) introduced a partial Bayesian approach which involves the use of a profile likelihood, where the aim is to detect multiple changes in the mean of Poisson distributions with an application to *haemolytic uraemic syndrome* (HUS) data. The same dataset was studied by Tian et al. (2009), who proposed a method which treats the change points as latent variables. Ko et al. (2015) have proposed an extension to the hidden Markov model of Chib (1998) by using a Dirichlet process prior on each row of the regime matrix. Their model is semiparametric, as the number of states is not specified in advance, but it grows according to the data size. Heard and Turcotte (2017) have proposed a new sequential Monte Carlo algorithm to infer multiple change points. Other contributions to the Bayesian change point literature are Harlé et al. (2016), Lai and Xing (2011), Martínez and Mena (2014) and Mira and Petrone (1996).

Whilst the literature covering change point analysis from a Bayesian perspective is vast when prior distributions are elicited, the documentation referring to analysis under minimal prior information is limited, see Moreno et al. (2005) and Girón et al. (2007). The former paper discusses the single change point problem in a model selection setting, whilst the latter paper, which is an extension of the former, tackles the multivariate change point problem in the context of linear regression models. Our work from Chapter 3 aims to contribute to the methodology for change point analysis under the assumption that the information about the number of change points and their location is minimal. First, we discuss the definition of an objective prior for change point location, both for single and multiple changes, assuming the number of changes is known a priori. Then, we define a prior on the number of change points via a model selection approach. Here, we assume that the change point coincides with one of the observations. As such, given $X_1, X_2, \ldots, X_n$ data points, the change point location is discrete.

2.5 Bayesian Analysis of Gaussian Graphical Models

New technologies allow the collection of large amounts of data up to a significant level of detail. To fully exploit the information in the data it is important that the possibly complex relationships among them are effectively captured and described. A statistical tool that allows one to exploit the power of graphs to represent such relationships among a, possibly large, number of variables, is a graphical model. Indeed, a graphical model can provide a geometrical representation of the dependencies among the variables with the immediacy that graphs exhibit. The use of this particular type of models is widespread within disciplines, including finance and economics (Giudici and Spelta (2016)), social sciences (McNally et al. (2015), Williams (2018)), speech recognition (Bilmes (2004), Bell and King (2007)) and biology (Wang et al. (2016)).

A sensible way of describing a graph is as a collection of two sets of objects: *vertices* and *edges* (Roverato, 2017). Vertices represent a finite set of elements, whereas the edges signify the existence of a link or interplay between pairs of those elements. In a diagram, the vertices are drawn as numerically labelled circles, while the edges can be represented by either a simple line or an arrow, symbolising the distinction between *undirected* and *directed* graphs, respectively. Formally, an edge is said to be undirected if the order in the pair of the connected vertices is not relevant; conversely, the edge is said to be directed and the order is represented by the direction of an arrow. Examples of both types of graphs can be seen in Figures 2.1 and 2.2.

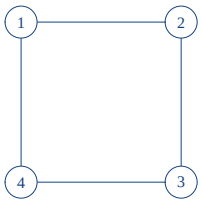


Fig. 2.1 A undirected graph with 4 vertices and 4 edges.

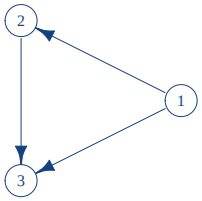


Fig. 2.2 A directed graph with 3 vertices and 3 arrows (edges).

An attractive feature of undirected graphs is *decomposability*, since it allows to divide a graph into subgraphs (graphs which are part of a larger graph). Decomposability can help with the computations and in the implementation of efficient inferential methods as subgraphs can be treated separately. To elaborate, a decomposable graph can be divided into smaller parts, called *cliques* and *separators*. A clique is a subgraph where all its vertices are connected to each other. When all the pairs between different vertices in a graph are joined together, we call the respective graph a *complete* graph. Clearly, a clique represents a

complete subgraph of a graph. When we refer to cliques across this work, we mean *maximal* cliques. A clique is maximal if it is not the subgraph of another clique in the graph. A separator has a more technical definition, but it can be intuitively illustrated as follows. Let us assume that a graph is formed by three subgraphs: A, B and C. Then B is a separator if the only way to move from a vertex in A to a vertex in C is through B. So a separator represents a subgraph in the graph, not necessarily complete which divides the graph in several subgraphs disconnected from each other. Note that in a decomposable graph, the separator must be complete. In contrast with cliques, when we specify a separator we actually mean a *minimal* separator, that is a separator which does not contain any other separator. In the Bayesian framework, the decomposability in cliques and separators allows to define priors which encode the statistical dependencies of a model. A more in-depth treatment of the graph notions described above is given in Chapter 4. An example of an undirected decomposable graph can be seen in Figure 2.3 with its decomposition in terms of cliques and separators seen in Figure 2.4.

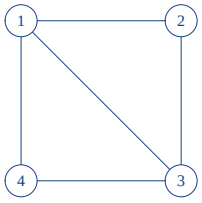


Fig. 2.3 A undirected decomposable graph with 4 vertices and 5 edges.

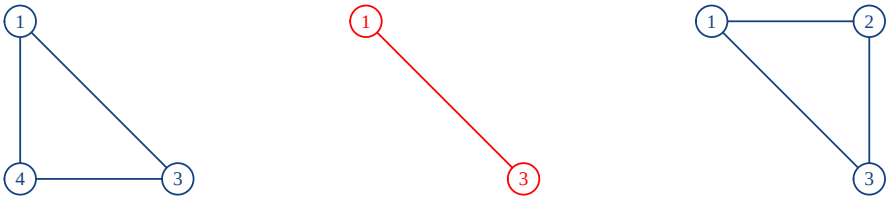


Fig. 2.4 The decomposition of the undirected graph from Figure 2.3 in two cliques (blue) and one separator (red).

A widely used statistical model for graphs is the *Gaussian Graphical Model* (GGM). Here, Gaussian means the multivariate distribution the data should follow. Furthermore, there is a direct equivalence between the zeros in the inverse of the covariance matrix associated with the aforementioned Gaussian distribution and the missing edges of the corresponding elements in the underlying graph structure. There are many useful reasons for assuming Normality. A remarkable one is that, among all distributions with same mean and same variance, the Normal assumption maximizes the entropy (Cover and Thomas, 2006). As a consequence, it imposes the least number of structural constraints beyond the first and second moments.

The literature around GGMs is vast, and it spans from frequentist to Bayesian approaches. Meinshausen and Bühlmann (2006) estimate the neighbourhood of vertices through the LASSO procedure (Tibshirani, 1996) and then put together those estimates to build the underlying graph. Of the same flavour as LASSO, Yuan and Lin (2007) have introduced a penalized likelihood method to estimate the concentration matrix, which for GGMs encodes the conditional independence. Friedman et al. (2008) have developed the graphical LASSO algorithm which is quite fast compared to other frequentist based algorithms. The above methods look at the regularization penalty being imposed on the concentration matrix. A method where the penalty is imposed to the inverse of the concentration matrix, the covariance matrix, is presented by Bien and Tibshirani (2011). Giudici and Green (1999) have applied the trans-dimensional reversible jump Markov chain Monte Carlo (RJMCMC) algorithm of Green (1995) to estimate the decomposable graphs that underlie the relationships in the data. This RJMCMC method was extended to estimate the structure in a case of multivariate lattice data by Dobra et al. (2011). Another trans-dimensional algorithm, this time based upon birth-death processes, was described by Mohammadi and Wit (2015). Jones et al. (2005) have reviewed the traditional MCMC (Markov chain Monte Carlo) methods used for graph search for both decomposable and non-decomposable cases when high-dimensional data is

considered and have proposed an alternative method to find high probability regions of the graph space. An MCMC method to estimate the normalising constant of the distribution which has its structure characterised by a non-decomposable graph has been proposed by Atay-Kayis and Massam (2005). Their idea was also used by Jones et al. (2005) when non-decomposable graphs were involved. For decomposable graphs, Carvalho and Scott (2009) have introduced a prior for the covariance matrix which helps to improve the accuracy in the graph search. In addition, they have also presented a graph prior which automatically guards against multiplicity.

The estimation methods in GGMs have been extensively studied in the literature for both directed (Friedman et al. (2000), Spirtes et al. (2000), Geiger and Heckerman (2002), Shojaie and Michailidis (2010), Stingo et al. (2010), Yajima et al. (2015), Consonni et al. (2017)) and undirected graphs (Dobra et al. (2004), Meinshausen and Bühlmann (2006), Yuan and Lin (2007), Banerjee et al. (2008), Friedman et al. (2008), Carvalho and Scott (2009), Kundu et al. (2019), Stingo and Marchetti (2015)).

We are tackling the GGM problem from the Bayesian perspective. In this approach there are two sources of randomness as discussed by Giudici and Green (1999). One is related to the multivariate distribution and the quantities that may parametrise it, the other has to do with the underlying graph G, equivalent to describing the conditional independence structure of the model under consideration. As such two kinds of priors are necessary: one related to the model parameters, Σ_G in our case, the other associated with the graph G. In Chapter 4, we propose a graph prior based on the loss-based method reviewed in Section 2.3. First, we revisit some of the graph priors encountered in the GGM literature. Then, we define our proposal and compare it with some of the aforementioned graph priors in simulated and real data studies.

2.6 Bayesian Analysis of Proper Binary Trees

This section outlines the idea of a tree in the context of Bayesian data analysis. We start by showing how trees arise as a structure which is extremely useful, especially in empirical studies (Linero, 2017). We then summarily review some of the literature around trees in this Bayesian framework.

Following the ideas from Chipman et al. (1998), Wu et al. (2007), Chipman et al. (2013), Linero (2017) and Chipman et al. (2010), let us consider n observations $(y_i, \boldsymbol{x}_i)$ where $\boldsymbol{x}_i = (x_1, x_2, \ldots, x_p)$ is the p-dimensional vector of predictors and the response y_i depends on $\boldsymbol{x}_i$. In the case of the BART (Bayesian Additive Regression Trees) methodology, the responses depend on the predictors in a linear fashion. The idea behind using a tree T, is that at each node of the tree the predictor space is split into non-overlapping regions according to one of the predictors and a threshold value. At each of the terminal nodes of the tree, called leaves, there is a parameter θ_l such that the responses corresponding to the predictors from that path through the tree are distributed according to $f(y|\theta_l)$, where $l = 1, \ldots, L_T$ are the leaves of the tree T. Clearly, the response values at each terminal node l are i.i.d. with distribution $f(y|\theta_l)$ and between terminal nodes of the same tree T, the response values are independent. Now let us denote with Θ the vector of parameters θ_l affiliated with the leaves of tree T. Obviously, for a different tree structure, the predictors and the corresponding responses will be split differently. As such, we may think of a particular tree T_k in the space of trees $\mathscr{T}$ as akin to a model in a model space. Therefore, we may apply the methodology of Villa and Walker (2015a). An example of a proper binary tree can be seen in Figure 2.5.

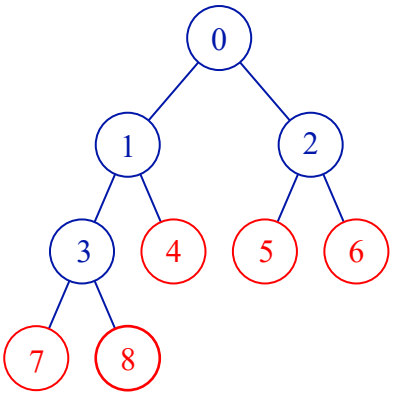


Fig. 2.5 A binary tree with 4 internal nodes (blue) and 5 terminal nodes (red). Note that this binary tree is proper as all internal nodes have exactly two children.

Now let us consider the following data generating process for a tree T_k to which we attribute the vector of parameters according to which the data is split, that is $\Theta_k = (\theta_1, \theta_2 \ldots, \theta_{L_{T_k}})$, where L_{T_k} are the leaves of T_k:

$$y_{lj} \overset{\text{i.i.d.}}{\sim} f(y_{lj}|\theta_l), \quad \forall l = 1,2,\ldots,L_{T_k} \quad \text{and} \quad \forall j = 1,2,\ldots,n_l.$$

Let h be the likelihood function of all n responses y_i denoted by $\boldsymbol{y} = (y_1, y_2, \ldots, y_n)$, that is:

$$h(\boldsymbol{y}|T_k, \Theta_k, \boldsymbol{X}) = \prod_{l=1}^{L_{T_k}} \prod_{j=1}^{n_l} f(y_{lj}|\theta_l),$$

where $\boldsymbol{X}$ is the $n \times p$ design matrix. Note that f designates a parametric family of distributions indexed by θ_l. These notations would be used again in Chapter 5, especially when defining the loss-based prior on trees.

The Bayesian approach to trees, in particular proper binary trees, starts with the works of Chipman et al. (1998) and Denison et al. (1998). The differences between the two papers consist in the tree priors they use, as well as the stochastic approaches utilised to explore the tree space. Chipman et al. (1998) consider a tree-generating process which also controls the shape of the tree through two hyperparameters for the prior, whereas Denison et al.

(1998) simply use a truncated Poisson distribution for the number of leaves. Regarding the stochastic algorithms employed, the former authors adopt a Metropolis-Hastings algorithm with a transition kernel which allows four possible moves (grow, prune, change, swap), while the latter apply the reversible jump Markov chain Monte Carlo algorithm of Green (1995). An extension to the model introduced by Chipman et al. (1998) was provided by Chipman et al. (2002) in the context of linear regression, which they called Bayesian treed models. Wu et al. (2007) have proposed a tree prior which takes into account both the tree size and the tree shape, while also adding a restructure step to the four aforementioned moves allowed in the previous tree searching algorithms. The aim of this additional step is to provide large changes in the tree structure whilst maintaining the number of leaves and the allocation of the observations to the subsequent leaves unchanged. Gramacy and Lee (2008) have developed a way to deal with nonstationary modelling by coupling a stationary Gaussian process with the partition provided by the tree structure. Chipman et al. (2010) have extended the single tree model of Chipman et al. (1998) and Chipman et al. (2002) in the case of linear regression to an ensemble of trees introducing the BART model which exhibits very good empirical performance. An addition to the tree and BART literatures was provided by Linero (2018) who advocated the use of a sparsity-inducing Dirichlet hyperprior when choosing the predictors that the split will be constructed around instead of the usual discrete uniform hyperprior. A review of trees in the Bayesian context is provided by Linero (2017). Some loss-based binary tree priors are introduced in Chapter 5, together with their justification, theoretical attributes and a potential application.

2.7 Outline of the Thesis

The outline of this thesis is as follows. In Chapter 3, we present our methodology for tackling change point problems, together with its application to simulated and real data. In particular, we define the loss-based prior on the change point locations. Then, we show how we can use

Bayesian model selection to determine the number of change points in a dataset. Chapter 4 is about describing and exploring our loss-based prior in the context of Gaussian graphical models. Notably, we test our method against other graph priors defined across the relevant literature when synthetic and real data scenarios are considered. In Chapter 5, we showcase some of the loss-based proposals for a tree prior. For each of the proposals we indicate the theoretical behaviours according to the choices of the tuning parameters. The chapter ends by outlining the plan for encapsulating one of the loss-based priors in a BART approach to a bike-sharing system. The last chapter (Chapter 6) summarises and concludes the main results of the thesis, as well as sketches a plan for future work.

3. Loss-based Prior applied to Change Point Problems

This chapter contains the methods we use to address change point problems. The first section introduces the loss-based prior on the positions of the change points, whilst the second section looks into finding the number of change points in a dataset as a model selection exercise. The third section is dedicated to validating our theoretical work through simulated and real datasets. The last section's purpose is to discuss our main contributions corresponding to this chapter. The body of this chapter has been taken from Hinoveanu et al. (2019).

3.1 Loss-based Prior on the Change Point Locations

This section is devoted to the derivation of the loss-based prior when the number of change points is known a priori. Specifically, let k be the number of change points and $m_1 < m_2 < \ldots < m_k$ their locations. We introduce the idea in the simple case where we assume that there is only one change point in the dataset (see Section 3.1.1). Then, we extend the results to the more general case where multiple change points are considered (see Section 3.1.2). Note that we assume that the change in the dataset occurs after the identified point. For instance, in the case of one change point, m implies that the actual change occurs from the X_{m+1} observation onwards.

A well-known objective prior for finite parameter spaces, in cases where there is no structure, is the uniform prior. As such, a natural choice for the prior on the change points location is the uniform, as discussed in Koop and Potter (2009). The corresponding loss-based prior is indeed the uniform, as shown below, which is a reassuring result as the objective prior for a specific parameter space, if it exists, should be unique.

3.1.1 Single Change Point

Let $\mathbf{X}^{(n)} = (X_1, \ldots, X_n)$ denote an n-dimensional vector of random variables, representing the random sample, and m be our single change point location, that is $m \in \{1, 2, \ldots, n-1\}$, such that

$$X_1, \ldots, X_m | \tilde{\theta}_1 \overset{\text{i.i.d.}}{\sim} f_1(\cdot|\tilde{\theta}_1)$$
$$X_{m+1}, \ldots, X_n | \tilde{\theta}_2 \overset{\text{i.i.d.}}{\sim} f_2(\cdot|\tilde{\theta}_2).$$

Note that we assume that there is a change point in the series, as such the space of m does not include the case $m = n$. In addition, we assume that $\tilde{\theta}_1 \neq \tilde{\theta}_2$ when $f_1 = f_2$. The sampling density for the vector of observations $\mathbf{x}^{(n)} = (x_1, \ldots, x_n)$ is:

$$f(\mathbf{x}^{(n)}|m, \tilde{\theta}_1, \tilde{\theta}_2) = \prod_{i=1}^{m} f_1(x_i|\tilde{\theta}_1) \prod_{i=m+1}^{n} f_2(x_i|\tilde{\theta}_2).$$

Let $m' \neq m$. Then, the KL divergence between the model parametrised by m and the one parametrised by m' is:

$$D_{KL}(f(\mathbf{x}^{(n)}|m, \tilde{\theta}_1, \tilde{\theta}_2) \| f(\mathbf{x}^{(n)}|m', \tilde{\theta}_1, \tilde{\theta}_2)) = \int f(\mathbf{x}^{(n)}|m, \tilde{\theta}_1, \tilde{\theta}_2) \log\left(\frac{f(\mathbf{x}^{(n)}|m, \tilde{\theta}_1, \tilde{\theta}_2)}{f(\mathbf{x}^{(n)}|m', \tilde{\theta}_1, \tilde{\theta}_2)}\right) \mathrm{d}\mathbf{x}^{(n)}.$$

Without loss of generality, consider $m < m'$. In this case, note that

$$\frac{f(\mathbf{x}^{(n)}|m,\tilde{\theta}_1,\tilde{\theta}_2)}{f(\mathbf{x}^{(n)}|m',\tilde{\theta}_1,\tilde{\theta}_2)} = \frac{\prod_{i=1}^{m} f_1(x_i|\tilde{\theta}_1)\prod_{i=m+1}^{m'} f_2(x_i|\tilde{\theta}_2)\prod_{i=m'+1}^{n} f_2(x_i|\tilde{\theta}_2)}{\prod_{i=1}^{m} f_1(x_i|\tilde{\theta}_1)\prod_{i=m+1}^{m'} f_1(x_i|\tilde{\theta}_1)\prod_{i=m'+1}^{n} f_2(x_i|\tilde{\theta}_2)} = \prod_{i=m+1}^{m'} \frac{f_2(x_i|\tilde{\theta}_2)}{f_1(x_i|\tilde{\theta}_1)}.$$

This leads to

$$\begin{aligned} D_{KL}(f(\mathbf{x}^{(n)}|m,\tilde{\theta}_1,\tilde{\theta}_2)\|f(\mathbf{x}^{(n)}|m',\tilde{\theta}_1,\tilde{\theta}_2)) &= \int f(\mathbf{x}^{(n)}|m,\tilde{\theta}_1,\tilde{\theta}_2)\left[\log\left(\prod_{i=m+1}^{m'}\frac{f_2(x_i|\tilde{\theta}_2)}{f_1(x_i|\tilde{\theta}_1)}\right)\right]\mathrm{d}\mathbf{x}^{(n)} \\ &= \int f(\mathbf{x}^{(n)}|m,\tilde{\theta}_1,\tilde{\theta}_2)\left[\sum_{i=m+1}^{m'}\log\left(\frac{f_2(x_i|\tilde{\theta}_2)}{f_1(x_i|\tilde{\theta}_1)}\right)\right]\mathrm{d}\mathbf{x}^{(n)} \\ &= \sum_{i=m+1}^{m'}\int f(\mathbf{x}^{(n)}|m,\tilde{\theta}_1,\tilde{\theta}_2)\left[\log\left(\frac{f_2(x_i|\tilde{\theta}_2)}{f_1(x_i|\tilde{\theta}_1)}\right)\right]\mathrm{d}\mathbf{x}^{(n)} \\ &= \sum_{i=m+1}^{m'}\left\{1^{n-1}\cdot\int f_2(x_i|\tilde{\theta}_2)\left[\log\left(\frac{f_2(x_i|\tilde{\theta}_2)}{f_1(x_i|\tilde{\theta}_1)}\right)\right]\mathrm{d}x_i\right\}. \end{aligned} \tag{3.1}$$

The 1^{n-1} factor from equation (3.1) is due to integrating with respect to all variables which are not indexed by i. Then, we obtain

$$\begin{aligned} &D_{KL}(f(\mathbf{x}^{(n)}|m,\tilde{\theta}_1,\tilde{\theta}_2)\|f(\mathbf{x}^{(n)}|m',\tilde{\theta}_1,\tilde{\theta}_2)) = \\ &\qquad\sum_{i=m+1}^{m'}\int f_2(x_i|\tilde{\theta}_2)\log\left(\frac{f_2(x_i|\tilde{\theta}_2)}{f_1(x_i|\tilde{\theta}_1)}\right)\mathrm{d}x_i. \end{aligned} \tag{3.2}$$

On the right hand side of equation (3.2), we can recognise the KL divergence from density f_2 to density f_1, thus getting:

$$\begin{aligned} &D_{KL}(f(\mathbf{x}^{(n)}|m,\tilde{\theta}_1,\tilde{\theta}_2)||f(\mathbf{x}^{(n)}|m',\tilde{\theta}_1,\tilde{\theta}_2)) = \\ &\qquad(m'-m)D_{KL}(f_2(\cdot|\tilde{\theta}_2)\|f_1(\cdot|\tilde{\theta}_1)). \end{aligned} \tag{3.3}$$

In a similar fashion, when $m > m'$, we have that:

$$D_{KL}(f(\mathbf{x}^{(n)}|m,\tilde{\theta}_1,\tilde{\theta}_2)\|f(\mathbf{x}^{(n)}|m',\tilde{\theta}_1,\tilde{\theta}_2)) = \\ (m-m')D_{KL}(f_1(\cdot|\tilde{\theta}_1)\|f_2(\cdot|\tilde{\theta}_2)). \tag{3.4}$$

In this single change point scenario, we can consider m' as a perturbation of the change point location m, that is $m' = m \pm l$ where $l \in \mathbb{N}^*$, such that $1 \leq m' < n$. Then, taking into account equations (3.3) and (3.4), the KL divergence becomes:

$$D_{KL}(f(\mathbf{x}^{(n)}|m,\tilde{\theta}_1,\tilde{\theta}_2)\|f(\mathbf{x}^{(n)}|m',\tilde{\theta}_1,\tilde{\theta}_2)) = \\ \begin{cases} l \cdot D_{KL}(f_2(\cdot|\tilde{\theta}_2)\|f_1(\cdot|\tilde{\theta}_1)), & \text{if } m < m' \\ \\ l \cdot D_{KL}(f_1(\cdot|\tilde{\theta}_1)\|f_2(\cdot|\tilde{\theta}_2)), & \text{if } m > m', \end{cases}$$

and

$$\begin{aligned} &\min_{m' \neq m}\left[D_{KL}(f(\mathbf{x}^{(\mathbf{n})}|m,\tilde{\theta}_1,\tilde{\theta}_2)\|f(\mathbf{x}^{(\mathbf{n})}|m',\tilde{\theta}_1,\tilde{\theta}_2))\right] = \\ &\quad = \min_{m' \neq m}\{l \cdot D_{KL}(f_2(\cdot|\tilde{\theta}_2)\|f_1(\cdot|\tilde{\theta}_1)), l \cdot D_{KL}(f_1(\cdot|\tilde{\theta}_1)\|f_2(\cdot|\tilde{\theta}_2))\} \\ &\quad = \min_{m' \neq m}\{D_{KL}(f_2(\cdot|\tilde{\theta}_2)\|f_1(\cdot|\tilde{\theta}_1)), D_{KL}(f_1(\cdot|\tilde{\theta}_1)\|f_2(\cdot|\tilde{\theta}_2))\} \cdot \underbrace{\min_{m' \neq m}\{l\}}_{1}. \end{aligned} \tag{3.5}$$

We observe that equation (3.5) is only a function of $\tilde{\theta}_1$ and $\tilde{\theta}_2$ and does not depend on m. Thus, $\pi(m) \propto 1$ and, therefore,

$$\pi(m) = \frac{1}{n-1}, \qquad m \in \{1,\ldots,n-1\}.$$

This prior was used, for instance, in an econometric context by Koop and Potter (2009) with the rationale of giving equal weight to every possible change point location.

3.1.2 Multivariate Change Point Problem

In this section, we address the change point problem in its generality by assuming that there are $1 \leq k < n$ change points. In particular, for the data $\mathbf{x}^{(n)} = (x_1,\ldots,x_n)$, we consider the following sampling distribution

$$f(\mathbf{x}^{(n)}|\boldsymbol{m},\tilde{\boldsymbol{\theta}}) = \prod_{i=1}^{m_1} f_1(x_i|\tilde{\theta}_1) \prod_{j=1}^{k-1} \prod_{i=m_j+1}^{m_{j+1}} f_{j+1}(x_i|\tilde{\theta}_{j+1}) \prod_{i=m_k+1}^{n} f_{k+1}(x_i|\tilde{\theta}_{k+1}), \tag{3.6}$$

where $\boldsymbol{m} = (m_1,\ldots,m_k)$, $1 \leq m_1 < m_2 < \ldots < m_k < n$, is the vector of the change point locations and $\tilde{\boldsymbol{\theta}} = (\tilde{\theta}_1,\ldots,\tilde{\theta}_k,\tilde{\theta}_{k+1})$ is the vector of the parameters of the underlying probability distributions. Schematically:

$$\begin{array}{llll}
X_1 & ,\ldots, & X_{m_1}|\tilde{\theta}_1 & \overset{\text{i.i.d.}}{\sim} f_1(\cdot|\tilde{\theta}_1) \\
X_{m_1+1} & ,\ldots, & X_{m_2}|\tilde{\theta}_2 & \overset{\text{i.i.d.}}{\sim} f_2(\cdot|\tilde{\theta}_2) \\
\vdots & ,\ldots, & \vdots & \vdots\ldots\vdots \\
X_{m_{k-1}+1} & ,\ldots, & X_{m_k}|\tilde{\theta}_k & \overset{\text{i.i.d.}}{\sim} f_k(\cdot|\tilde{\theta}_k) \\
X_{m_k+1} & ,\ldots, & X_n|\tilde{\theta}_{k+1} & \overset{\text{i.i.d.}}{\sim} f_{k+1}(\cdot|\tilde{\theta}_{k+1}).
\end{array}$$

If $f_1 = f_2 = \cdots = f_{k+1}$, then it is reasonable to assume that some of the θ's are different. Without loss of generality, we assume that $\tilde{\theta}_1 \neq \tilde{\theta}_2 \neq \cdots \neq \tilde{\theta}_k \neq \tilde{\theta}_{k+1}$. In a similar fashion to

the single change point case, we cannot assume $m_k = n$ since we require exactly k change points.

In this case, due to the multivariate nature of the vector $\boldsymbol{m} = (m_1, \ldots, m_k)$, the derivation of the loss-based prior is not as straightforward as in the one dimensional case. In fact, the derivation of the prior is based on heuristic considerations supported by Theorem 1 from below. In particular, we are able to prove an analogous of equations (3.3) and (3.4) when only one component is arbitrarily perturbed. Let us define the following functions:

$$d_j^{+1}(\tilde{\boldsymbol{\theta}}) = D_{KL}(f_{j+1}(\cdot|\tilde{\theta}_{j+1}) \| f_j(\cdot|\tilde{\theta}_j))$$
$$d_j^{-1}(\tilde{\boldsymbol{\theta}}) = D_{KL}(f_j(\cdot|\tilde{\theta}_j) \| f_{j+1}(\cdot|\tilde{\theta}_{j+1})),$$

where $j \in \{1, 2, \ldots, k\}$. The following Theorem (the proof of which is in Appendix A) is useful to understand the behaviour of the loss-based prior in the general case.

Theorem 1. *Let $f(\mathbf{x}^{(n)}|\boldsymbol{m}, \tilde{\boldsymbol{\theta}})$ be the sampling distribution defined in equation* (3.6) *and consider $j \in \{1, \ldots, k\}$. Let $\boldsymbol{m}'$ be such that $m'_i = m_i$ for $i \neq j$, and let the component m'_j be such that $m'_j \neq m_j$ and $m_{j-1} < m'_j < m_{j+1}$. Therefore,*

$$D_{KL}(f(\mathbf{x}^{(n)}|\boldsymbol{m}, \tilde{\boldsymbol{\theta}}) \| f(\mathbf{x}^{(n)}|\boldsymbol{m}', \tilde{\boldsymbol{\theta}}) = |m'_j - m_j| d_j^S(\tilde{\boldsymbol{\theta}}),$$

where $S = \operatorname{sgn}(m'_j - m_j)$.

Note that, Theorem 1 states that the minimum KL divergence is achieved when $m'_j = m_j + 1$ or $m'_j = m_j - 1$. This result is not surprising since the KL divergence measures the degree of similarity between two distributions. The smaller the perturbation caused by changes in one of the parameters is, the smaller the KL divergence between the two distributions is. Although Theorem 1 makes a partial statement about the multiple change points scenario, it provides a strong argument for supporting the uniform prior. Indeed, if

now we consider the general case of having k change points, it is straightforward to see that the KL divergence is minimised when only one of the components of the vector $\boldsymbol{m}$ is perturbed by (plus or minus) one unit. As such, the loss-based prior depends on the vector of parameters $\tilde{\boldsymbol{\theta}}$ only, as in the one-dimensional case, yielding the uniform prior for $\boldsymbol{m}$.

Therefore, the loss-based prior on the multivariate change point location is

$$\pi(\boldsymbol{m}) = \left\{ \binom{n-1}{k} \right\}^{-1}, \tag{3.7}$$

where $\boldsymbol{m} = (m_1, \ldots, m_k), 1 \leq m_1 < m_2 < \ldots < m_k < n$. The denominator in equation (3.7) has the above form because, for every number of k change points, we are interested in the number of k-subsets from a set of $n-1$ elements, which is $\binom{n-1}{k}$. The same prior was also derived in a different way by Girón et al. (2007).

Note that across this chapter, the $\pi(\boldsymbol{m})$ prior implicitly assumes that we know the number of change points k a priori. As such, in this context, $\pi(\boldsymbol{m})$ actually represents the $\pi(\boldsymbol{m}|k)$ distribution.

3.2 Loss-based Prior on the Number of Change Points

Here, we approach the change point analysis as a model selection problem. In particular, we define a prior on the space of models, where each model represents a certain number of change points (including the case of no change points). The method adopted to define the prior on the space of models is the one introduced in Villa and Walker (2015a).

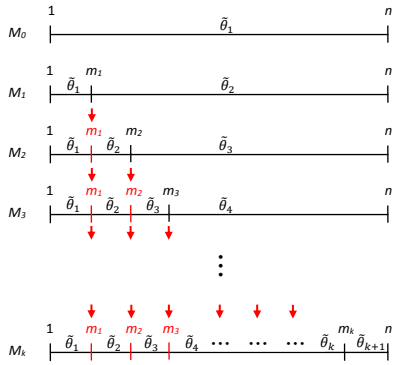


Fig. 3.1 Diagram showing the way we specify our models. The arrows indicate that the respective change point locations remain fixed from the previous model to the current one.

We proceed as follows. Assume we have to select from $k+1$ possible models. Let M_0 be the model with no change points, M_1 the model with one change point and so on. Generalising, model M_k corresponds to the model with k change points. The idea is that the current model encompasses the change point locations of the previous model. As an example, in model M_3 the first two change point locations will be the same as in the case of model M_2. To illustrate the way we envision our models, we have provided Figure 3.1. It has to be noted that the construction of the possible models from M_0 to M_k can be done in a different way to the one described here. Through this, we simply mean that of course the change point locations in a subsequent model can be uncoupled from the previous model like in Figure 3.2. As we can see in that figure, the single change point location from model M_1 is situated between the two change points from model M_2 in green. The only difference compared to the models defined in Figure 3.1 is the additional term $D_{KL}(f_1(\cdot|\tilde{\theta}_1)\|f_2(\cdot|\tilde{\theta}_2))$ being present when computing the quantity $D_{KL}(M_1\|M_2)$. Other extra terms would also appear when both change points from model M_2 would be placed before or after the location of the change point from M_1. Obviously, the approach to define the model priors stays unchanged. As we require a minimization of the quantity $D_{KL}(M_1\|M_2)$ when computing the model prior probabilities and because the KL divergence is non-negative, we prefer to construct models akin to the one shown in Figure 3.1 for the sake of notational and computational simplicity.

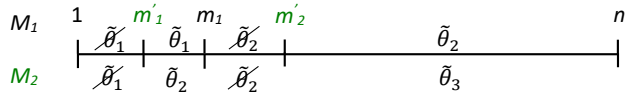


Fig. 3.2 Diagram showing a different way to specify the locations of the change points. The single change point location from model M_1 is situated between the two change points from model M_2 in green. The crossed parameters show the sections that do not contribute to the computation of $D_{KL}(M_1\|M_2)$.

Consistently with the notation used in Sections 2.3 and 3.1,

$$\theta_k = \begin{cases} \tilde{\theta}_1,\ldots,\tilde{\theta}_{k+1},m_1,\ldots,m_k & \text{if } k=1,\ldots,n-1 \\ \tilde{\theta}_1 & \text{if } k=0, \end{cases}$$

represents the vector of parameters of model M_k, where $\tilde{\theta}_1,\ldots,\tilde{\theta}_{k+1}$ are the model specific parameters and $m_1,\ldots,m_k$ are the change point locations, as in Figure 3.1.

Based on the way we have specified our models, which are in direct correspondence with the number of change points and their locations, we state Theorem 2 (the proof of which is in Appendix A).

Theorem 2. *Let*

$$D_{KL}(M_i\|M_j) = D_{KL}(f(\mathbf{x}^{(n)}|\theta_i)\|f(\mathbf{x}^{(n)}|\theta_j)).$$

For any $0 \le i < j \le k$ *integers, with* $k < n$*, and the convention* $m_{j+1} = n$*, we have the following:*

$$D_{KL}(M_i\|M_j) = \sum_{q=i+1}^{j} \left[(m_{q+1}-m_q)\cdot D_{KL}(f_{i+1}(\cdot|\tilde{\theta}_{i+1})\|f_{q+1}(\cdot|\tilde{\theta}_{q+1}))\right],$$

and

$$D_{KL}(M_j\|M_i) = \sum_{q=i+1}^{j} \left[(m_{q+1}-m_q)\cdot D_{KL}(f_{q+1}(\cdot|\tilde{\theta}_{q+1})\|f_{i+1}(\cdot|\tilde{\theta}_{i+1}))\right].$$

The result in Theorem 2 is useful when the model selection exercise is implemented. Indeed, the Villa and Walker (2015a) approach requires the computation of the KL divergences in Theorem 2. To recall their idea, let us consider k Bayesian models:

$$M_j = \{f_j(x|\theta_j), \pi_j(\theta_j)\}, \qquad j \in \{1,2,\ldots,k\},$$

where $f_j(x|\theta_j)$ is the sampling distribution parametrised by θ_j and $\pi_j(\theta_j)$ represents the prior on the model parameter (possibly vector of parameters) θ_j. Assuming the priors $\pi_j(\theta_j)$ are proper, the model prior probability $P(M_j)$ is proportional to the expected minimum KL divergence from M_j to M_i, with $i = 1,\ldots,k$ and $i \neq j$, where the expectation is considered with respect to $\pi_j(\theta_j)$. That is:

$$P(M_j) \propto \exp\left\{\mathbb{E}_{\pi_j}\left[\inf_{\theta_i, i\neq j} D_{KL}(f_j(x|\theta_j)\|f_i(x|\theta_i))\right]\right\}, \qquad j = 1,\ldots,k. \tag{3.8}$$

In other words, we assign a prior mass to model M_j which is proportional to the distance to the most similar model M_i $(i \neq j)$, in expectation. To illustrate, let us start by considering what is lost if model M_j is removed from the set of all the possible models and it is the true model. This loss is quantified by the KL divergence to the nearest model. The loss is then linked to the model prior probability via the self-information loss function (Merhav and Feder, 1998). The prior in (3.8) is then obtained by equating the two aforementioned losses.

Recalling equation (3.8), the objective model prior probabilities are then given by:

$$\Pr(M_j) \propto \exp\left\{\mathbb{E}_{\pi_j}\left[\inf_{\theta_i, i\neq j} D_{KL}(M_j \| M_i)\right]\right\}, \qquad j = 0, 1, \ldots, k. \tag{3.9}$$

For illustrative purposes, in Appendix A we derive the model prior probabilities to perform model selection among M_0, M_1 and M_2. Recall that the θ's represent notationally the change point locations and the parameters for the underlying distributions corresponding to the models which are indexed according to the number of change points. As such, taking into account the non-negativity of the KL divergence, in Appendix A we see that the overall minimum comes down to simply minimising over the individual segments in which the overall KL divergence between two models splits into according to Theorem 2. Afterwards, we simply compute the overall KL divergences between the required model and all other available models different than it and select the smallest one. This represents our loss in information associated to the respective model and it contributes to computing the corresponding model prior probability from equation (3.9).

It is easy to infer from equation (3.9) that model priors depend on the prior distribution assigned to the model parameters, that is on the level of uncertainty that we have about their true values. For the change point location, a sensible choice is the uniform prior which, as shown in Section 3.1, corresponds to the loss-based prior. For the model specific parameters, we have several options. If one wishes to pursue an objective analysis, intrinsic priors (Berger and Pericchi, 1996) may represent a viable solution since they are proper. Nonetheless, the method introduce by Villa and Walker (2015a) does not require, in principle, an objective choice as long as the priors are proper. Given that we use the latter approach, here we consider subjective priors for the model specific parameters.

Remark 1. In the case where the changes in the underlying sampling distribution are limited to the parameter values, the model prior probabilities defined in (3.9) follow the

uniform distribution. That is, $\Pr(M_j) \propto 1$. In the real data example illustrated in Section 3.3.2, we indeed consider a problem where the above case occurs.

Remark 2. As we assign a prior which depends on the number of change points, a legitimate question is how the dilution problem may affect our method, see George (2010). We would like to point out that the prior introduced in this chapter implicitly takes into account the numerosity of models with the same number of change points. Indeed, the methodology used in this work builds on Villa and Walker (2015a). In particular, the approach requires to assume a prior on the change point locations and, as highlighted above, the default choice in our methodology is the uniform, which takes into account for the dilution. Generally, the dilution property of a prior is concerned with not putting excessive mass on irrelevant models. A practical intuition of the dilution property can be seen from the following example taken from George (2010). Assume we are concerned with the problem of variable selection in linear models and the number of covariates is d. Let a single predictor be completely different from all others. Now, let us presume those $d-1$ predictors are approximately identical. Then, a prior which takes into account the dilution will put half of its mass on models that include the relevant predictor and half on models comprised of any subset of the other $d-1$ predictors. Clearly, the mass put on those $d-1$ explanatory variables needs to be diluted or spread uniformly within. We note that the methodology of Villa and Walker (2015a) has an inbuilt dilution property, as we are reducing the prior mass when identical models are compared, whilst putting more mass on models that are completely different. An effect of this property can be seen in Scenario 4 from Section 3.3 in regards to our loss-based prior.

Note that across this chapter, an important assumption which we have considered in regards to our methodology corresponds to the independence of the observations. In the case of time series which are dependent on past values, we would look at the works of Sandberg et al. (2001) who computed the KL divergence when Gaussian autoregressive

moving average (ARMA) processes were involved and Liu et al. (2013) who based their change point detection method for time series data on the Pearson divergence. For a Bayesian treatment of the change points when segmented ARMA models are concerned, we would like to mention the paper of Sadia et al. (2018).

3.2.1 A special case: selection between M_0 and M_1

Let us consider the case where we have to estimate whether there is or not a change point in a set of observations. This implies that we have to choose between model M_0 (i.e. no change point) and M_1 (i.e. one change point). Following our approach, we have:

$$\Pr(M_0) \propto \exp\left\{\mathbb{E}_{\pi_0}\left[\inf_{\tilde{\theta}_2} D_{KL}(f_1(\cdot|\tilde{\theta}_1)\|f_2(\cdot|\tilde{\theta}_2))\right]\right\},$$

and

$$\Pr(M_1) \propto \exp\left\{\mathbb{E}_{\pi_1}\left[(n-m_1)\cdot\inf_{\tilde{\theta}_1} D_{KL}(f_2(\cdot|\tilde{\theta}_2)\|f_1(\cdot|\tilde{\theta}_1))\right]\right\}. \tag{3.10}$$

Now, let us assume independence between the prior on the change point location and the prior on the parameters of the underlying sampling distributions, that is $\pi_1(m_1,\tilde{\theta}_1,\tilde{\theta}_2) = \pi_1(m_1)\pi_1(\tilde{\theta}_1,\tilde{\theta}_2)$. Let us further recall that, as per equation (3.7), $\pi_1(m_1) = 1/(n-1)$. As such, we observe that the model prior probability on M_1 becomes:

$$\Pr(M_1) \propto \exp\left\{\left(\frac{n}{2}\right)\mathbb{E}_{\pi_1(\tilde{\theta}_1,\tilde{\theta}_2)}\left[\inf_{\tilde{\theta}_1} D_{KL}(f_2(\cdot|\tilde{\theta}_2)\|f_1(\cdot|\tilde{\theta}_1))\right]\right\}. \tag{3.11}$$

We notice that the model prior probability for model M_1 is increasing when the sample size increases. This behaviour occurs whether there is or not a change point in the data. We propose to address the above problem by using a non-uniform prior for m_1. A reasonable

alternative, which works quite well in practice, would be the following shifted binomial as prior:

$$\pi_1(m_1) = \binom{n-2}{m_1-1}\left(\frac{n-1}{n}\right)^{m_1-1}\left(\frac{1}{n}\right)^{n-m_1-1}, 1 \leq m_1 \leq n-1. \tag{3.12}$$

To argument the choice of (3.12), we note that, as n increases, the probability mass will be more and more concentrated towards the upper end of the support. Therefore, from equations (3.10) and (3.12) follows:

$$\Pr(M_1) \propto \exp\left\{\left(\frac{2n-2}{n}\right)\mathbb{E}_{\pi_1(\tilde{\theta}_1,\tilde{\theta}_2)}\left[\inf_{\tilde{\theta}_1} D_{KL}(f_2(\cdot|\tilde{\theta}_2)\|f_1(\cdot|\tilde{\theta}_1))\right]\right\}.$$

For the more general case where we consider more than two models, the problem highlighted in equation (3.11) vanishes. As we observe in Appendix A, when we compute the KL divergence between any model different than the largest one and all the other models, there will be quantities that do not depend on the sample size. Because we minimise the divergence, it is often the case that those quantities will be chosen. In regards to the largest model, note that we actually compute the KL divergences between all the pairs of the underlying distributions. Since we require the minimal overall KL divergence, there will be elements where the individual divergences will be very small, but positive, thus suppressing the effect of the sample size.

3.3 Simulated and Real Data Analysis

This section outlines the behaviour of the prior in both simulated and real data studies. The first subsection is dedicated to analysing our methodology across four scenarios involving simulated data, as well as comparing the results we obtain with our prior against the method of Barry and Hartigan (1993). The second subsection concentrates on presenting the behaviour

for several real datasets, namely the number of disasters in the British coal mines between 1851-1962 and the absolute value of the daily logarithmic returns of the S&P500 index observed from the 14/01/2008 to the 31/12/2011.

3.3.1 Change Point Analysis on Simulated Data

In this section, we present the results of several simulation studies based on the methodologies discussed in Sections 3.1 and 3.2. We start with a scenario involving discrete distributions in the context of the one change point problem. We then show the results obtained when we consider continuous distributions for the case of two change points. The choice of the underlying sampling distributions is in line with Villa and Walker (2015a).

Single sample

Scenario 1. The first scenario concerns the choice between models M_0 and M_1. Specifically, for M_0 we have:

$$X_1, X_2, \ldots, X_n | p \overset{\text{i.i.d.}}{\sim} \text{Geometric}(p),$$

and for M_1 we have:

$$X_1, X_2, \ldots, X_{m_1} | p \overset{\text{i.i.d.}}{\sim} \text{Geometric}(p)$$
$$X_{m_1+1}, X_{m_1+2}, \ldots, X_n | \lambda \overset{\text{i.i.d.}}{\sim} \text{Poisson}(\lambda).$$

Let us denote with $f_1(\cdot|p)$ and $f_2(\cdot|\lambda)$ the probability mass functions of the Geometric and the Poisson distributions, respectively. The priors for the parameters of f_1 and f_2 are $p \sim \text{Beta}(a,b)$ and $\lambda \sim \text{Gamma}(c,d)$.

In the first simulation, we sample $n = 100$ observations from model M_0 with $p = 0.8$. To perform the change point analysis, we have chosen the following parameters for the priors on p and λ: $a = 2$, $b = 2$, $c = 3$ and $d = 1$. Applying the approach introduced in

Section 3.2, we obtain $\Pr(M_0) \propto 1.59$ and $\Pr(M_1) \propto 1.81$. These model priors yield the model posterior probabilities (refer to equation (2.9)) $\Pr(M_0|\mathbf{x}^{(n)}) = 0.92$ and $\Pr(M_1|\mathbf{x}^{(n)}) = 0.08$. As expected, the selection process strongly indicates the true model as M_0. Table 3.1 reports the above probabilities including other information, such as the appropriate Bayes factors.

The second simulation looked at the opposite setup, that is we sample $n = 100$ observations from M_1, with $p = 0.8$ and $\lambda = 3$. We have sampled 50 data points from the Geometric distribution and the remaining 50 data points from the Poisson distribution. In Figure 3.3, we have plotted the simulated sample, where it is legitimate to assume a change in the underlying distribution. Using the same prior parameters as above, we obtain $\Pr(M_0|\mathbf{x}^{(n)}) = 0.06$ and $\Pr(M_1|\mathbf{x}^{(n)}) = 0.94$. Again, the model selection process is assigning heavy posterior mass to the true model M_1. These results are further detailed in Table 3.1. Note that if we swap around the two distributions, the model prior probabilities will change, due to the asymmetry of the KL divergence (Cover and Thomas, 2006). As such, the ordering of the distributions definitely matters in terms of the nominal values, but the subsequent analysis leads to the same result as in the previous setup, namely the correct identification of the models given the two possible scenarios of a no change or a single change present in the data.

In the context of model misspecification, we keep the same modelling choices as above, namely the Geometric and Poisson distributions with the hyperparameters fixed at the values selected previously, but we consider the 100 sampled observations as following: for the M_0 case, all observations come from a Binomial$(10, 0.15)$ distribution, whilst when one change is present, we have the first 50 observations being sampled from Binomial$(10, 0.15)$ with the remaining 50 observations following a Binomial$(10, 0.25)$ distribution. As seen in Table 3.2, we are still able to correctly identify the true models, but with lower model posterior probabilities due to misspecification.

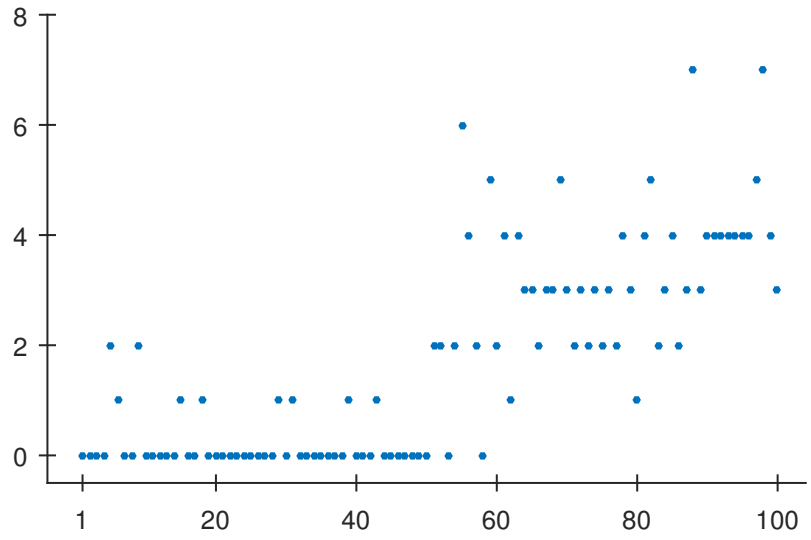


Fig. 3.3 Scatter plot of the data simulated from model M_1 in Scenario 1.

	True model	
	M_0	M_1
$\Pr(M_0)$	0.47	0.47
$\Pr(M_1)$	0.53	0.53
B_{01}	12.39	0.08
B_{10}	0.08	12.80
$\Pr(M_0\|\mathbf{x}^{(n)})$	0.92	0.06
$\Pr(M_1\|\mathbf{x}^{(n)})$	0.08	0.94

Table 3.1 Model priors, Bayes factors and model posterior probabilities for the change point analysis in Scenario 1. We considered samples from, respectively, model M_0 and model M_1.

Scenario 2. In this scenario we consider the case where we have to select among three models, that is model M_0:

$$X_1, X_2, \ldots, X_n | \lambda, \kappa \overset{\text{i.i.d.}}{\sim} \text{Weibull}(\lambda, \kappa),$$

model M_1:

$$X_1, X_2, \ldots, X_{m_1} | \lambda, \kappa \overset{\text{i.i.d.}}{\sim} \text{Weibull}(\lambda, \kappa)$$
$$X_{m_1+1}, X_{m_1+2}, \ldots, X_n | \mu, \tau \overset{\text{i.i.d.}}{\sim} \text{Log-normal}(\mu, \tau),$$

with $1 \leq m_1 \leq n-1$ being the location of the single change point, and model M_2:

$$X_1, X_2, \ldots, X_{m_1} | \lambda, \kappa \overset{\text{i.i.d.}}{\sim} \text{Weibull}(\lambda, \kappa)$$
$$X_{m_1+1}, X_{m_1+2}, \ldots, X_{m_2} | \mu, \tau \overset{\text{i.i.d.}}{\sim} \text{Log-normal}(\mu, \tau)$$
$$X_{m_2+1}, X_{m_2+2}, \ldots, X_n | \alpha, \beta \overset{\text{i.i.d.}}{\sim} \text{Gamma}(\alpha, \beta),$$

with $1 \leq m_1 < m_2 \leq n-1$ representing the locations of the two change points, such that m_1 corresponds exactly to the same location as in model M_1. These distributions were chosen for computational reasons, as the KL divergences between each pair of them was already derived by Villa and Walker (2015a). Analogously to the previous scenario, we sample from each model in turn and perform the selection to detect the number of change points.

	True model	
	M_0	M_1
$\Pr(M_0)$	0.47	0.47
$\Pr(M_1)$	0.53	0.53
B_{01}	1.36	0.55
B_{10}	0.74	1.82
$\Pr(M_0 \mid \mathbf{x}^{(n)})$	0.55	0.33
$\Pr(M_1 \mid \mathbf{x}^{(n)})$	0.45	0.67

Table 3.2 Model priors, Bayes factors and model posterior probabilities for the change point analysis in Scenario 1 when model misspecification is explored. We considered samples from, respectively, model M_0 and model M_1.

Let $f_1(\cdot|\lambda, \kappa), f_2(\cdot|\mu, \tau)$ and $f_3(\cdot|\alpha, \beta)$ represent the Weibull, Log-normal and Gamma densities, respectively, with $\tilde{\theta}_1 = (\lambda, \kappa)$, $\tilde{\theta}_2 = (\mu, \tau)$ and $\tilde{\theta}_3 = (\alpha, \beta)$. We assume a Normal

prior on μ and Gamma priors on all the other parameters as follows:

$$\lambda \sim \text{Gamma}(1.5,1) \qquad \kappa \sim \text{Gamma}(5,1) \qquad \mu \sim \text{Normal}(0.05,1),$$

$$\tau \sim \text{Gamma}(16,1) \qquad \alpha \sim \text{Gamma}(10,1) \qquad \beta \sim \text{Gamma}(0.2,0.1).$$

In the first exercise, we have simulated $n = 100$ observations from model M_0, where we have set $\lambda = 1.5$ and $\kappa = 5$. We obtain the following model priors: $\Pr(M_0) \propto 1.09$, $\Pr(M_1) \propto 1.60$ and $\Pr(M_2) \propto 1.37$, yielding the posteriors $\Pr(M_0|\mathbf{x}^{(n)}) = 0.96$, $\Pr(M_1|\mathbf{x}^{(n)}) = 0.04$ and $\Pr(M_2|\mathbf{x}^{(n)}) = 0.00$. We then see that the approach assigns high mass to the true model M_0. Table 3.3 reports the above probabilities and the corresponding Bayes factors. The second

	True model		
	M_0	M_1	M_2
$\Pr(M_0)$	0.27	0.27	0.27
$\Pr(M_1)$	0.39	0.39	0.39
$\Pr(M_2)$	0.34	0.34	0.34
B_{01}	36.55	3.24×10^{-4}	4.65×10^{-40}
B_{02}	1.84×10^{3}	0.02	1.27×10^{-45}
B_{12}	50.44	55	2.72×10^{-6}
$\Pr(M_0|\mathbf{x}^{(n)})$	0.96	0.00	0.00
$\Pr(M_1|\mathbf{x}^{(n)})$	0.04	0.98	0.00
$\Pr(M_2|\mathbf{x}^{(n)})$	0.00	0.02	1.00

Table 3.3 Model priors, Bayes factors and model posterior probabilities for the change point analysis in Scenario 2. We considered samples from, respectively, model M_0, model M_1 and model M_2.

simulation was performed by sampling 50 observations from a Weibull with parameter values as in the previous exercise, and the remaining 50 observations from a Log-normal density with location parameter $\mu = 0.05$ and scale parameter $\tau = 16$. The data is displayed in Figure 3.4.

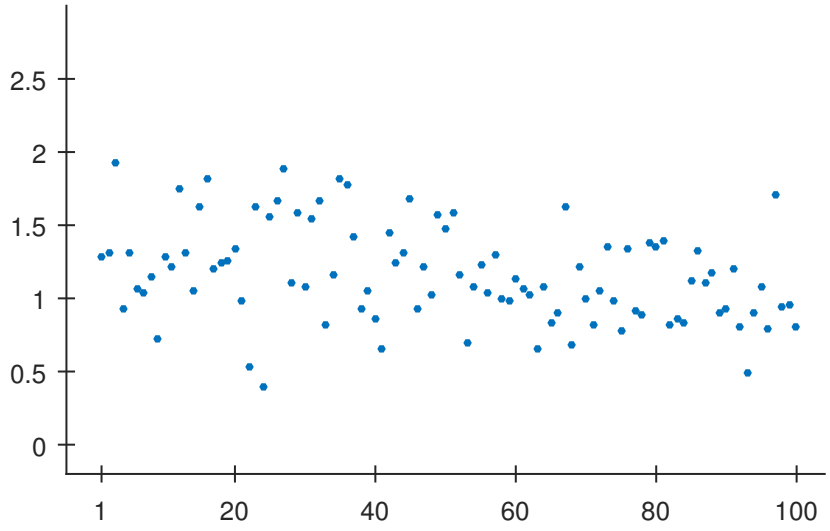


Fig. 3.4 Scatter plot of the observations simulated from model M_1 in Scenario 2.

The model posterior probabilities are $\Pr(M_0|\mathbf{x}^{(n)}) = 0.00$, $\Pr(M_1|\mathbf{x}^{(n)}) = 0.98$ and $\Pr(M_2|\mathbf{x}^{(n)}) = 0.02$, which are reported in Table 3.3. In this case as well, we see that the model selection procedure indicates M_1 as the true model, as expected.

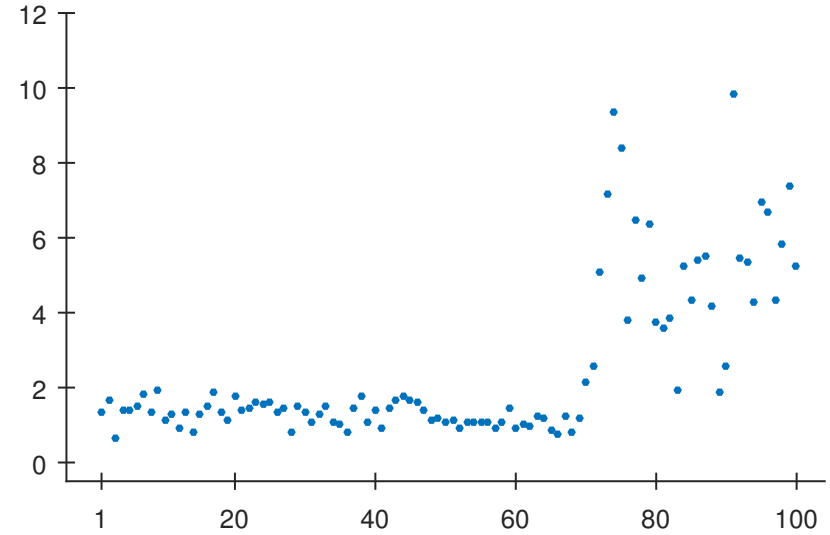


Fig. 3.5 Scatter plot of the observations simulated from model M_2 in Scenario 2.

Finally, for the third simulation exercise we sample 50 and 20 data points from, respectively, a Weibull and a Log-normal with parameter values as defined above. The last 30 observations are sampled from a Gamma distribution with parameters $\alpha = 10$ and $\beta = 2$. The sampled data is displayed in Figure 3.5. From Table 3.3, we note that the posterior distribution on the model space accumulates on the true model M_2.

Frequentist Analysis

In this section, we perform a frequentist analysis of the performance of the proposed prior by drawing repeated samples from different scenarios. In particular, we look at a two change points problem where the sampling distributions are Student-t with different degrees of freedom. In this scenario, we perform the analysis with 60 repeated samples generated by different densities with the same mean values.

Then, we repeat the analysis of Scenario 2 by selecting 100 samples for $n = 500$ and $n = 1500$. We consider different sampling distributions with the same mean and variance. In this scenario, where we added the further constraint of the equal variance, it is interesting to note that the change in distribution is captured when we increase the sample size, meaning that we learn more about the true sampling distributions.

We also compare the performances of the loss-based prior with the uniform prior when we analyse the scenario with different sampling distributions, namely Weibull/Log-normal/Gamma. It is interesting to note that the uniform prior is unable to capture the change in distribution even for a large sample size. On the contrary, the loss-based prior is able to detect the number of change points when $n = 1500$. Furthermore, for $n = 500$, even though both priors are not able to detect the change points most of the times, the loss-based prior has a higher frequency of success when compared to the uniform prior.

Scenario 3. In this scenario, we consider the case where the sampling distributions belong to the same family, that is Student-t, where the true model has two change points. In particular,

let $f_1(\cdot|\nu_1), f_2(\cdot|\nu_2)$ and $f_3(\cdot|\nu_3)$ represent the densities of three standard t distributions, respectively. We assume that ν_1, ν_2 and ν_3 are positive integers strictly greater than one so as to have defined mean for each density. Note that this allows us to compare distributions of the same family with equal mean. The priors assigned to the number of degrees of freedom assume a parameter space of positive integers strictly larger than 1. As such, we define them as follows:

$$\nu_1 \sim 2 + \text{Poisson}(30) \qquad \nu_2 \sim 2 + \text{Poisson}(3) \qquad \nu_3 \sim 2 + \text{Poisson}(8).$$

In this experiment, we consider 60 repeated samples, each of size $n = 300$ and with the following structure:

- $X_1, \ldots, X_{100}$ from a Student-t distribution with $\nu_1 = 30$,
- $X_{101}, \ldots, X_{200}$ from a Student-t distribution with $\nu_2 = 3$,
- $X_{201}, \ldots, X_{300}$ from a Student-t distribution with $\nu_3 = 8$.

Table 3.4 reports the frequentist results of the simulation study. First, note that $P(M_1) = P(M_2) = P(M_3) = 1/3$ as per Remark 1 in Section 3.2. For all the simulated samples, the loss-based prior yields a posterior with the highest probability assigned to the true model M_2. We also note that the above posterior is on average 0.75 with a variance 0.02, making the inferential procedure extremely accurate.

	Mean posterior	Variance posterior	Freq. true model
$\Pr(M_0\|\mathbf{x}^{(n)})$	0.01	3.84×10^{-4}	0/60
$\Pr(M_1\|\mathbf{x}^{(n)})$	0.24	0.0160	0/60
$\Pr(M_2\|\mathbf{x}^{(n)})$	0.75	0.0190	60/60

Table 3.4 Average model posterior probabilities, variance and frequency of true model for the Scenario 3 simulation exercise.

Scenario 4. In this scenario, we perform repeated sampling from the setup described in Scenario 2 above, where the true model has two change points. In particular, we draw 100 samples with $n = 500$ and $n = 1500$. For $n = 500$, the loss-based prior probabilities are $P(M_0) = 0.18$, $P(M_1) = 0.16$ and $P(M_2) = 0.66$. For $n = 1500$, the loss-based prior probabilities are $P(M_0) = 0.015$, $P(M_1) = 0.014$ and $P(M_2) = 0.971$. The simulation results are reported, respectively, in Table 3.5 and in Table 3.6. The two change point locations for $n = 500$ are at the 171^{st} and 341^{st} observations. For $n = 1500$, the first change point is the 501^{st} observation, while the second is at the 1001^{st} observation. We note that there is a sensible improvement in detecting the true model, using the loss-based prior, when the sample size increases. In particular, we move from 30% to 96%.

To compare the loss-based prior with the uniform prior we have run the simulation on the same data samples used above. The results for $n = 500$ and $n = 1500$ are in Table 3.7 and in Table 3.8, respectively. Although we can observe an improvement when the sample size increases, the uniform prior does not lead to a clear detection of the true model for both sample sizes.

	Mean posterior	Variance posterior	Freq. true model
$\Pr(M_0\|\mathbf{x}^{(n)})$	9.88×10^{-4}	2.60×10^{-5}	0/100
$\Pr(M_1\|\mathbf{x}^{(n)})$	0.63	0.0749	70/100
$\Pr(M_2\|\mathbf{x}^{(n)})$	0.37	0.0745	30/100

Table 3.5 Average model posterior probabilities, variance and frequency of true model for the Scenario 4 simulation exercise with $n = 500$ and the loss-based prior.

	Mean posterior	Variance posterior	Freq. true model
$\Pr(M_0\|\mathbf{x}^{(n)})$	1.33×10^{-13}	1.76×10^{-24}	0/100
$\Pr(M_1\|\mathbf{x}^{(n)})$	0.08	0.0200	4/100
$\Pr(M_2\|\mathbf{x}^{(n)})$	0.92	0.0200	96/100

Table 3.6 Average model posterior probabilities, variance and frequency of true model for the Scenario 4 simulation exercise with $n = 1500$ and the loss-based prior.

	Mean posterior	Variance posterior	Freq. true model
$\Pr(M_0\|\mathbf{x}^{(n)})$	16×10^{-4}	7.15×10^{-5}	0/100
$\Pr(M_1\|\mathbf{x}^{(n)})$	0.82	0.0447	91/100
$\Pr(M_2\|\mathbf{x}^{(n)})$	0.18	0.0443	9/100

Table 3.7 Average model posterior probabilities, variance and frequency of true model for the Scenario 4 simulation exercise with $n = 500$ and the uniform prior.

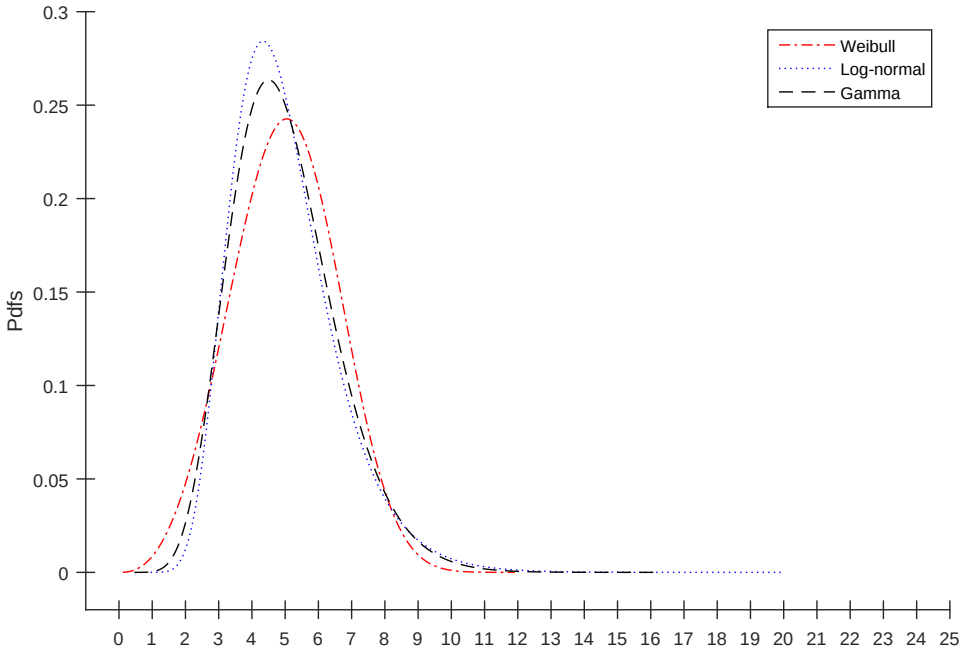


Fig. 3.6 The densities of Weibull(λ, κ), Log-normal(μ, τ) and Gamma(α, β) with the same mean (equal to 5) and the same variance (equal to 2.5).

	Mean posterior	Variance posterior	Freq. true model
$\Pr(M_0\|\mathbf{x}^{(n)})$	8.64×10^{-12}	7.45×10^{-21}	0/100
$\Pr(M_1\|\mathbf{x}^{(n)})$	0.501	0.1356	49/100
$\Pr(M_2\|\mathbf{x}^{(n)})$	0.499	0.1356	51/100

Table 3.8 Average model posterior probabilities, variance and frequency of true model for the Scenario 4 simulation exercise with $n = 1500$ and the uniform prior.

Finally, we conclude this section with a remark. One may wonder why the change point detection requires an increase in the sample size, and the reply can be inferred from Figure 3.6, which displays the density functions of the distributions employed in this scenario. As it can be observed, the densities are quite similar, which is not surprising since these distributions have the same means and the same variances. The above similarity can also be appreciated in terms of Hellinger distance, see Table 3.9. In other words, from Figure 3.6 we can see that the main differences in the underlying distributions are in the tail areas. It is therefore necessary to have a relatively large number of observations in order to be able to discern differences in the densities, because in this case only we would have a sufficient representation of the whole distribution.

	Hellinger distances		
	Weibull(λ,κ)	Log-normal(μ,τ)	Gamma(α,β)
Weibull(λ,κ)		0.1411996	0.09718282
Log-normal(μ,τ)			0.04899711

Table 3.9 Hellinger distances between all the pairs formed from a Weibull(λ,κ), Log-normal(μ,τ) and Gamma(α,β). The six hyperparameters are such that the distributions have the same mean=5 and same variance=2.5.

Comparison to Barry and Hartigan's method

In this section we perform a comparison of the proposed change point method to the one described in Barry and Hartigan (1993). The simulation study is performed by considering

three different scenarios: we simulate data, assumed to be normally distributed, and which exhibits, respectively, one, two and three change points.

The proposal of Barry and Hartigan (1993) is based on a product partition approach. In particular, product models on partitions represent a framework for Bayesian inference on change points. The authors highlight that, even if the initial probability model for partitions and parameters is not a product model, under specific conditions it represents a suitable approximation for the analysis. The product partition method is based upon splitting the data into *contiguous blocks* where the parametric model is the same. This splitting is called a *partition*. Furthermore, the respective partition is treated as a random variable and its distribution is based upon the *cohesions* within the constituent blocks. According to Müller et al. (2011), a cohesion for a block is a non-negative function which measures how close together the elements of that block are. Essentially in the method of Barry and Hartigan (1993) regarding normal data, we are interested in the differences between the means of those blocks. Let us assume that a n sized normally distributed sample is split into b blocks. Then, let us have a partition $\rho = (i_0, i_1, \ldots, i_b)$, where $0 = i_0 < i_1 < \ldots < i_b = n$, such that the observations in the block $i_0 + 1, i_0 + 2, \ldots, i_1$ denoted by $i_0 i_1$ come from the same parametric model, the observations from block $i_1 i_2$ come from a different parametric model and so on. The distribution of ρ is $f(\rho) = K c_{i_0 i_1} c_{i_1 i_2} \cdot \ldots \cdot c_{i_{b-1} i_b}$, where c_{ij} is the prior cohesion in the block ij and K is the normalising constant. Following Yao (1984), these prior cohesions are $c_{ij} = (1-\iota)^{j-i-1}\iota$ for $j < n$ and $c_{ij} = (1-\iota)^{j-i-1}$ for $j = n$, where $\iota \in [0,1]$ is the probability for the existence of a change at each individual element from block ij. Note that the respective prior cohesions suggest a discrete renewal process for the change points with independent and geometrically distributed inter-arrival times. This observation stands at the basis of the subsequent analysis provided by Barry and Hartigan (1993).

To make the results comparable, we assume normality as the Barry and Hartigan (1993) method is particularly showcased based on this assumption. As described in detail below,

we consider for each scenario normal distributions with variance 1 (assumed as known) and differences in the mean (at each change point) of, respectively, 1, 2.5 and 3. In addition, in each scenario we consider as a possible model the no-change point model. The prior distribution for the means is a normal with zero mean and large variance (i.e. 10^6).

To perform the simulations, for the Barry and Hartigan (1993) method, we employ the `R` package `bcp`, developed by Erdman and Emerson (2007), and assume a change point when the posterior probability is at least 0.5. All simulations have a burnin of 10000, with a total number of iterations of 100000.

One change point. We consider the following model for the case with one change point:

$$X_1, X_2, \ldots, X_{m_1} | \mu_{11} \overset{\text{i.i.d.}}{\sim} \text{Normal}(\mu_{11}, 1)$$
$$X_{m_1+1}, X_{m_1+2}, \ldots, X_n | \mu_{21} \overset{\text{i.i.d.}}{\sim} \text{Normal}(\mu_{21}, 1)$$

Model M_0 corresponds to no changes in the mean of the data. We set $\mu_{21} = \mu_{11} + \Delta_1$ with $\Delta_1 \in \{0, 1, 2.5, 3\}$ and $\mu_{11} = 0$. In Table 3.10, we see the frequency of identifying the true model amongst 100 repeated samples for different sampling scenarios.

n	Frequency of identifying the true model							
	$\Delta_1 = 0 \quad (M_0)$		$\Delta_1 = 1 \quad (M_1)$		$\Delta_1 = 2.5 \quad (M_1)$		$\Delta_1 = 3 \quad (M_1)$	
	Our method	bcp	Our method	bcp	Our method	bcp	Our method	bcp
100	100/100	95/100	43/100	17/100	100/100	86/100	100/100	94/100
250	100/100	99/100	100/100	7/100	100/100	86/100	100/100	99/100
500	100/100	100/100	100/100	7/100	100/100	91/100	100/100	98/100

Table 3.10 Frequency of identifying the true model (the one within the nearby parentheses to the Δ_1 values) amongst 100 repeated samples for different sampling scenarios. The change point location is in $m_1 = 70, 175, 350$ for, respectively, $n = 100, 250, 500$.

Two change points We consider the following model for the case with two change points:

$$X_1, X_2, \ldots, X_{m_1} | \mu_{12} \overset{\text{i.i.d.}}{\sim} \text{Normal}(\mu_{12}, 1)$$

$$X_{m_1+1}, X_{m_1+2}, \ldots, X_{m_2} | \mu_{22} \overset{\text{i.i.d.}}{\sim} \text{Normal}(\mu_{22}, 1)$$

$$X_{m_2+1}, X_{m_2+2}, \ldots, X_n | \mu_{32} \overset{\text{i.i.d.}}{\sim} \text{Normal}(\mu_{32}, 1)$$

As before, model M_0 corresponds to no changes in the mean. In the simulations we set $\mu_{22} = \mu_{12} + \Delta_2$ and $\mu_{32} = \mu_{12}$ with $\Delta_2 \in \{0, 1, 2.5, 3\}$ and $\mu_{12} = 0$. In Table 3.11, we see the frequency of identifying the true model amongst 100 repeated samples for different sampling scenarios.

n	Frequency of identifying the true model							
	$\Delta_2 = 0 \quad (M_0)$		$\Delta_2 = 1 \quad (M_2)$		$\Delta_2 = 2.5 \quad (M_2)$		$\Delta_2 = 3 \quad (M_2)$	
	Our method	bcp	Our method	bcp	Our method	bcp	Our method	bcp
100	100/100	96/100	3/100	9/100	100/100	67/100	100/100	89/100
250	100/100	100/100	86/100	5/100	100/100	78/100	100/100	90/100
500	100/100	98/100	100/100	1/100	100/100	76/100	100/100	90/100

Table 3.11 Frequency of identifying the true model (the one within the nearby parentheses to the Δ_2 values) amongst 100 repeated samples for different sampling scenarios. The location of the first change point is $m_1 = 30, 75, 150$, respectively, for $n = 100, 250, 500$, and for the second change point is $m_2 = 70, 175, 350$.

Three change points Finally, we consider the following model for the case with three change points:

$$X_1, X_2, \ldots, X_{m_1} | \mu_{13} \overset{\text{i.i.d.}}{\sim} \text{Normal}(\mu_{13}, 1)$$
$$X_{m_1+1}, X_{m_1+2}, \ldots, X_{m_2} | \mu_{23} \overset{\text{i.i.d.}}{\sim} \text{Normal}(\mu_{23}, 1)$$
$$X_{m_2+1}, X_{m_2+2}, \ldots, X_{m_3} | \mu_{33} \overset{\text{i.i.d.}}{\sim} \text{Normal}(\mu_{33}, 1)$$
$$X_{m_3+1}, X_{m_3+2}, \ldots, X_n | \mu_{43} \overset{\text{i.i.d.}}{\sim} \text{Normal}(\mu_{43}, 1)$$

We set $\mu_{23} = \mu_{13} + \Delta_3$, $\mu_{33} = \mu_{13} + 2\Delta_3$ and $\mu_{43} = \mu_{13} + 3\Delta_3$ with $\Delta_3 \in \{0, 1, 2.5, 3\}$ and $\mu_{13} = 0$. In Table 3.12, we see the frequency of identifying the true model amongst 100 repeated samples for different sampling scenarios.

n	Frequency of identifying the true model							
	$\Delta_3 = 0 \quad (M_0)$		$\Delta_3 = 1 \quad (M_3)$		$\Delta_3 = 2.5 \quad (M_3)$		$\Delta_3 = 3 \quad (M_3)$	
	Our method	bcp	Our method	bcp	Our method	bcp	Our method	bcp
100	100/100	97/100	0/100	1/100	100/100	65/100	100/100	88/100
250	100/100	98/100	30/100	0/100	100/100	69/100	100/100	88/100
500	100/100	100/100	100/100	0/100	100/100	73/100	100/100	80/100

Table 3.12 Frequency of identifying the true model (the one within the nearby parentheses to the Δ_3 values) amongst 100 repeated samples for different sampling scenarios. The location of the first change point is $m_1 = 25, 62, 125$ for, respectively, $n = 100, 250, 500$; the location of the second change point is $m_2 = 50, 125, 250$ and the location of the third change point is $m_3 = 75, 188, 375$.

By looking at the above tables, we note the following. In general, both methods improve the detection of the change points as n increases, which is an expected result as the information about change points in the sample increases. For the cases where $\Delta = 2.5, 3$, our method

appears to have a better performance than the one in Barry and Hartigan (1993). This is more obvious for the smaller Δ. Furthermore, the proposed approach seems to select the model with the true number of change points when this number increases. A noteworthy aspect is that the Barry and Hartigan (1993) method diminishes its performance as n increases when the difference between the means is relatively small (i.e. $\Delta = 1$). A possible explanation is due to a degenerate behaviour of the product partition model; however, we did not investigate further as it does not impact the performance of our method.

3.3.2 Change Point Analysis on Real Data

In this section, we illustrate the proposed approach applied to real data. We first consider a well known dataset which has been extensively studied in the literature of the change point analysis, that is the British coal-mining disaster data (Carlin et al., 1992). The second set of data we consider refers to the daily returns of the S&P 500 index observed over a period of four years.

British Coal-Mining Disaster Data

The British coal-mining disaster data consists of the yearly number of disasters for the British coal miners over the period 1851-1962. Here, a disaster represents an event where at least 10 persons died. It is believed that the change in the working conditions, and in particular, the enhancement of the security measures, led to a decrease in the number of disasters. This calls for a model which can take into account a change in the underlying distribution around a certain observed year. With the proposed methodology we wish to detect if the assumption is appropriate. In particular, if a model with one change point is more suitable to represent the data than a model where no changes in the sampling distribution are assumed. Figure 3.7 shows the number of disasters per year in the British coal-mining industry from 1851 to 1962. As in Chib (1998), we assume a Poisson sampling distribution with a possible change

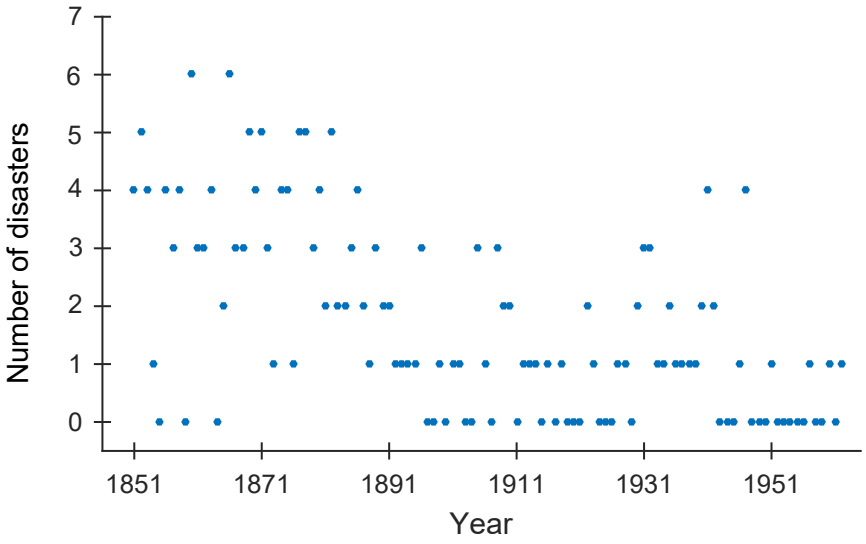


Fig. 3.7 Scatter plot of the British coal-mining disaster data.

in the parameter value. That is

$$X_1, X_2, \ldots, X_m | \phi_1 \overset{\text{i.i.d.}}{\sim} \text{Poisson}(\phi_1)$$
$$X_{m+1}, X_{m+2}, \ldots, X_n | \phi_2 \overset{\text{i.i.d.}}{\sim} \text{Poisson}(\phi_2),$$

where m is the unknown location of the single change point, such that $1 \leq m \leq n$, and a Gamma$(2,1)$ is assumed for ϕ_1 and ϕ_2. The case $m = n$ corresponds to the scenario with no change point, that is model M_0. The case $m < n$ assumes one change point, that is model M_1.

Let $f_1(\cdot|\phi_1)$ and $f_2(\cdot|\phi_2)$ be the Poisson distributions with parameters ϕ_1 and ϕ_2, respectively. Then, the analysis is performed by selecting between model M_0, that is when the sampling distribution is f_1, and model M_1, where the sampling distribution is f_1 up to a certain $m < n$ and f_2 from $m+1$ to n.

As highlighted in Remark 1 from Section 3.2, the prior on the model space is the discrete uniform distribution, that is $\Pr(M_0) = \Pr(M_1) = 0.5$. The proposed model selection approach

leads to the Bayes factors $B_{01} = 1.61 \times 10^{-13}$ and $B_{10} = 6.20 \times 10^{12}$, where it is obvious that the odds are strongly in favour of model M_1. Indeed, we have $\Pr(M_1|\mathbf{x}^{(n)}) \approx 1$.

Daily S&P 500 Absolute Log-Return Data

The second real data analysis aims to detect the number of change points in the absolute value of the daily logarithmic returns of the S&P 500 index observed from the 14/01/2008 to the 31/12/2011 (see Figure 3.8). Note that in this analysis we do not provide information on the locations where the changes occur. As underlying sampling distributions we consider the Weibull and the Log-normal (Yu, 2001), and the models among which we select are as follows. M_0 is a Weibull(λ, κ), M_1 is formed by a Weibull(λ, κ) and a Log-normal(μ_1, τ_1) and, finally, M_2 is formed by a Weibull(λ, κ), a Log-normal(μ_1, τ_1) and a Log-normal(μ_2, τ_2). An interesting particularity of this problem is that we will consider a scenario where the changes are in the underlying distribution as well as in the parameter values of the same distribution. As suggested in Section 4.1.3 of Kass and Raftery (1995), due to the large sample size of the dataset, we could approximate the Bayes factor by using the Schwartz criterion. Therefore, in this case the specification of the priors for the parameters of the underlying distributions is not necessary. From the results in Table 3.13, we see that the model indicated by the proposed approach is M_2. In other words, there is very strong indication that there are two change points in the dataset. From Table 3.13, we note that the priors on models M_1 and M_2 assigned by the proposed method are the same. This is not surprising as the only difference between the two models is an additional Log-normal distribution with different parameter values. Note that applying the methodology for the case with maximum two changes as outlined in Appendix A and taking into account Remark 1 provides the reasoning behind the fact that the model prior probabilities for M_1 and M_2 are the same. Recall that Remark 1 points out that when the changes are related just to the parameter values

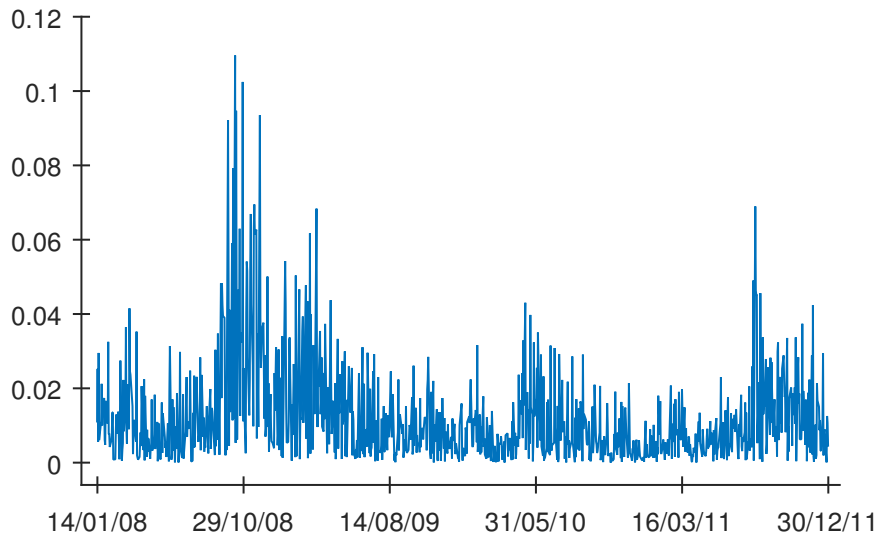


Fig. 3.8 Absolute daily log-returns of the S&P 500 index from 14/01/08 to 30/12/11.

of the underlying distributions and not to the nature of those distributions themselves, the KL divergence between the corresponding distributions is actually 0.

3.4 Discussion

Bayesian inference in change point problems under the assumption of insufficient prior information has not been deeply explored in the past, as the limited literature on the matter shows.

We contribute to the area by deriving an objective prior distribution to detect change point locations, when the number of change points is known a priori. As a change point location can be interpreted as a discrete parameter, we apply recent results in the literature (Villa and Walker, 2015b) to make inference. The resulting prior distribution, which is the discrete uniform distribution, is not new in the literature (Girón et al., 2007), and therefore can be considered as a validation of the proposed approach.

$\Pr(M_0)$	0.36
$\Pr(M_1)$	0.32
$\Pr(M_2)$	0.32
B_{01}	7.72×10^{18}
B_{02}	3.30×10^{-3}
B_{12}	4.28×10^{-22}
$\Pr(M_0\|\mathbf{x}^{(n)})$	0.00
$\Pr(M_1\|\mathbf{x}^{(n)})$	0.00
$\Pr(M_2\|\mathbf{x}^{(n)})$	1.00

Table 3.13 Model prior, Bayes factor and model posterior probabilities for the S&P 500 change point analysis.

A second major contribution is in defining an objective prior on the number of change points, which has been approached by considering the problem as a model selection exercise. The results of the proposed method on both simulated and real data, show the strength of the approach in estimating the number of change points in a series of observations. A point to note is the generality of the scenarios considered. Indeed, we consider situations where the change is in the value of the parameter(s) of the underlying sampling distribution, or in the distribution itself. For the simulation study we have compared the proposed method with an existing Bayesian approach for the detection of change points (Barry and Hartigan, 1993). Of particular interest is the last real data analysis (S&P 500 index), where we consider a scenario where we have both types of changes, that is the distribution for the first change point and on the parameters of the distribution for the second.

The aim of this work was to set up a novel approach to address change point problems. In particular, we have selected prior densities for the parameters of the models to reflect a scenario of equal knowledge, in the sense that model priors are close to represent a uniform distribution. Two remarks are necessary here. First, in the case prior information about the true value of the parameters is available, and one wishes to exploit it, the prior densities will need to reflect it and, obviously, the model prior will be impacted by the choice. Second, in

applications it is recommended that some sensitivity analysis is performed, so as to investigate if and how the choice of the parameter densities affects the selection process.

4. Loss-based Prior applied to Gaussian Graphical Models

This chapter describes the methodology we use in the context of *Gaussian graphical models* (GGMs). The first section outlines some of the graph priors found across the specific literature, out of which three will be used as a comparison for our graph prior introduced in the subsequent section. The second section contains the description of our graph prior. Simulated and real data analyses are performed in the third section, whilst the last section concludes this chapter with a discussion concerning our main contributions to the GGM literature. The contents of this chapter have been taken from Hinoveanu et al. (2018).

4.1 Graph Priors for Gaussian Graphical Models

We mentioned that graphical models help when modelling complex data. As the name suggests for GGMs, the data is assumed to be sampled from a multivariate Gaussian distribution. Let $\mathbf{X} = (X_1, X_2, \ldots, X_p)^{\mathbf{T}}$ be a p-dimensional random vector which follows a multivariate Gaussian distribution, that is

$$\mathbf{X} \sim N_p(\mathbf{0}, \Sigma_G),$$

where $\mathbf{0} \in \mathbb{R}^p$ is a p-dimensional column vector of zero means and $\Sigma_G \in \mathbb{R}^{p \times p}$ is the positive-definite covariance matrix. Let $\mathbf{x} = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n)^{\mathbf{T}}$ be the $n \times p$ matrix of observations,

where $\mathbf{x}_i$, for $i = 1, \ldots, n$, is a p-dimensional realisation from the multivariate Gaussian distribution. The link between the assumed sampling distribution and the graph is specified by completing a positive-definite matrix with respect to an undirected graph (Atay-Kayis and Massam, 2005; Giudici and Green, 1999; Roverato and Whittaker, 1998). For an arbitrary positive-definite matrix Γ and an undirected graph G, Σ_G is the unique positive-definite matrix completion of Γ with respect to G. This means that for the pairs of vertices which share an edge, the corresponding entries of Σ_G are the same as Γ. The entries for the missing edges are set to be 0 in the concentration (precision) matrix, that is Σ_G^{-1}. Therefore, we have a link between the multivariate sampling distribution and the graph structure represented by the zeros of the concentration matrix Σ_G^{-1}. In the GGMs framework, the dimension p of the multivariate Gaussian distribution also represents the number of vertices in the undirected graph G. As our sampling distribution is Gaussian, the concentration matrix has a clear interpretation. The entries of the concentration matrix encode the conditional independence structure of the distribution (Lauritzen, 1996). As such, if and only if the $(i, j)^{\text{th}}$ element of the concentration matrix is 0, the random variables X_i and X_j are conditionally independent given all other variables in the matrix (pairwise Markov property); or, equivalently, given their neighbours (local Markov property). The previous statement is based upon the idea that in a GGM the global, local and pairwise Markov properties are equivalent. For more details about these properties, we refer the reader to Lauritzen (1996).

Following Lauritzen (1996), a graph G is represented by the pair $G = (V, E)$ with V a finite set of vertices and E a subset of $V \times V$ of ordered pairs of distinct edges. Throughout the present chapter we will consider $V = \{1, 2, \ldots, p\}$, where p is a strictly positive integer. In the GGMs setting, p represents the dimension of the multivariate Normal distribution. In this chapter we consider undirected graphs with no loops and without multiple edges between pairs of distinct vertices.

Now, let us provide the following descriptions:

- vertices connected by an edge are called *neighbours* or *adjacent*
- a sequence of distinct vertices $i_0 = i, \ldots, i_n = j$, where the pair $(i_{l-1}, i_l) \in E, \forall\, l = 1, 2, \ldots, n$, is called a *path* of length n from vertex i to vertex j (in Figure 2.3, for example, $4, 1, 2$ represents one path of length 2 from vertex 4 to vertex 2, whereas $4, 1, 3, 2$ is an alternative path of length 3 between the same vertices)
- a subset of V is an $(i, j)-$separator when all the paths from i to j go through the respective subset. Subset $C \subseteq V$ separates A from B if C is a $(i, j)-$*separator* $\forall\, i \in A, j \in B$ (in Figure 2.3, $\{1, 3\}$ is a separator and it splits the graph in two subgraphs as seen in Figure 2.4)
- a graph where $(i, j) \in E, \forall\, i, j \in V$ is called a *complete* graph
- a subgraph represents a subset of V such that the edge set is restricted to those edges that have both endpoints in the respective subset. We call a maximal complete subgraph a *clique* (in Figure 2.4 we can see the two cliques that the undirected graph from Figure 2.3 is separated into, that is the subgraphs $\{1, 3, 4\}$ and $\{1, 2, 3\}$)
- the decomposition of an undirected graph is a triple (A, C, B) where $V = A \cup C \cup B$ for disjoint sets A, C and B such that C separates A from B and C is complete. Therefore, the graph is decomposed in the subgraphs $G_{A \cup C}$ and $G_{B \cup C}$
- a decomposable graph can be broken up into cliques and separators
- for a non-decomposable graph there will be subgraphs which cannot be decomposed further and are not complete

An example of a non-decomposable graph is in Figure 2.1, while if we swap the arrows for lines in Figure 2.2, thus transforming the directed graph into an undirected one, we observe a decomposable graph. A decomposition can be seen in Figure 2.4.

Note that through prior on a graph we simply mean the prior distribution on the number of edges of that graph where the support of the distribution is just the number of possible edges in a graph with a fixed p number of vertices.

Assuming G decomposable, Giudici and Green (1999) discuss the following prior on G:

$$\pi(G) = d^{-1},$$

where d is the number of decomposable graphs on a specific vertex set V. If we consider unrestricted graphs, the above prior is the uniform prior on the graph space and has the form:

$$\pi^{\mathrm{UP}}(G) = \frac{1}{2^{\binom{|V|}{2}}}.$$

where $|V|$ is the number of vertices in the graph. A criticism in using a uniform prior is that it assigns more mass to medium size graphs compared to, for example, the empty graph or the full graph.

To address the problem, Jones et al. (2005) set independent Bernoulli trials on the edge inclusions, such that the prior probability is $\phi = 2/(|V|-1)$ leading to the expected number of edges equal to $|V|$. Thus, the prior on G is:

$$\pi(G) \propto \phi^k \cdot (1-\phi)^{m-k}, \tag{4.1}$$

where $0 \leq k \leq m$ is the number of edges in the graph G and $m = \binom{|V|}{2}$ represents the maximum number of edges possible in that respective graph. Clearly, a ϕ close to zero would encourage sparser graphs, while for $\phi \to 1$, more mass will be put on complex graphs.

Carvalho and Scott (2009) recommend a fully Bayesian approach, where ϕ should be inferred from the data. As such, they assume that $\phi \sim \text{Beta}(a,b)$, leading to:

$$\pi(G) \propto \frac{\beta\,(a+k, b+m-k)}{\beta(a,b)}. \tag{4.2}$$

By setting $a = b = 1$ (equivalent to setting a uniform prior on ϕ) in equation (4.2), they obtain the prior on G as:

$$\pi^{\text{CS}}(G) \propto \frac{1}{(m+1)} \binom{m}{k}^{-1}. \tag{4.3}$$

A property of the prior in equation (4.3) is that it corrects for multiplicity. That is, as more noise vertices are added to the true graph, the number of false positives (edges which are erroneously included in the graph) remains constant. Furthermore, this prior was used in the context of Bayesian variable selection as outlined by Scott and Berger (2010). We note that its presence in that context had the same effect as in the graph framework, namely the multiplicity adjustment argument.

A somewhat similar form of the prior in equation (4.3) was derived by Armstrong et al. (2009). Their prior, called the *sized based prior*, uses the $A_{p,k}$ parameter representing the number of decomposable graphs instead of the combinatorial coefficient in the formula from above. The value of $A_{p,k}$ is estimated using an MCMC scheme and a recurrence relationship with graphs that have up to 5 vertices. The recurrence as codified by the Lemmas 1 and 2 provided by the original authors is $A_{p,k} = \binom{m}{k} - F_{p,k}$ where $F_{p,k}$ describes the number of non-decomposable graphs satisfying the following conditions:

(a) for $p \geq 0$ then $F_{p,0} = F_{p,1} = F_{p,m} = 0$

(b) for $p \geq 2$ then $F_{p,2} = F_{p,m-1} = 0$

(c) for $p \geq 3$ then $F_{p,3} = 0$

(d) for $p \geq 4$ then $F_{p,4} = F_{p,m-2} = 3 \cdot \binom{p}{4}$

(e) for $p \geq 5$ then $F_{p,5} = 12 \cdot \binom{p}{5} + 3 \cdot (m-6) \cdot \binom{p}{4}$

(f) for $p \geq 6$ then $A_{p,k}$ is based upon an initial estimate and a MCMC sampling scheme

The enumerating of the non-decomposable graphs present amongst all graphs with 5 vertices (conditions (a)-(e)) is based upon identifying all possible chordless 4-cycles and 5-cycles. As outlined by Lauritzen (1996), a *n-cycle* is simply a path of length n where the starting and the end points are the same vertex. When describing a path, a *chord* represents an edge between non-consecutive vertices. Figure 2.3 depicts a 4-cycle which has the chord $(1,3)$, whereas the 4-cycle from Figure 2.1 is chordless.

4.2 A Loss-based Prior for Gaussian Graphical Models

In this section, we present a prior based on a methodology that involves loss functions (Villa and Walker, 2015a).

We follow the insight provided by Villa and Lee (2019), where the method has been applied to variable selection in linear regression models, by adding an additional loss component to account for model complexity. We designed the penalty term to penalize complex graphs, meaning graphs with a relatively large number of edges. For instance, this is in line with the approach suggested by Cowell et al. (2007). Therefore, for a given number of vertices p with a maximum number of edges m, our prior has the form:

$$\pi(G) \propto \exp\left\{ \underbrace{\mathbb{E}_\pi\left[\inf_{\Sigma_{G'},} D_{KL}(f(\mathbf{x}|\mathbf{0},\Sigma_G)||f(\mathbf{x}|\mathbf{0},\Sigma_{G'}))\right]}_{\text{loss due to information}} \right.$$
$$\left. \underbrace{-h\left[(1-c)|G| + c\log\binom{m}{|G|}\right]}_{\text{loss due to graph complexity}} \right\}, \tag{4.4}$$

with $h \in [0, +\infty)$ and $c \in [0, 1]$. The component of the prior that penalizes for complexity takes into account the number of the edges of the graph, $|G|$, as well as the number of graphs with the same number of edges, $\binom{m}{|G|}$. The former can be interpreted as an *absolute* complexity of the graph, whilst the latter is weighing the complexity of the graph relatively to all the graphs with the same number of edges (i.e. *relative* complexity). Note that the last one is considered in the log-scale to mitigate the exponential behaviour of the binomial coefficient for large m. This makes the two terms approximately on the same order of magnitude. The two components are mixed by means of c, while h represents the constant up to which a loss function is defined. Noting that the KL divergence in (4.4) is minimized for $\Sigma_G = \Sigma_{G'}$, as such is zero, the prior will have the form:

$$\pi(G) \propto \exp\left\{-h\left[(1-c)|G| + c\log\binom{m}{|G|}\right]\right\}. \tag{4.5}$$

The tuning parameter h allows to set the prior in order to control the sparsity of the graph. In particular, for $h \to \infty$, the prior in equation (4.5) will decrease quickly to zero, assigning most of the mass to simple graphs. On the other hand, small values of h result in a prior where its mass is more evenly distributed over the whole space of graphs. In fact, if we set $h = 0$ the prior in (4.5) will become $\pi(G) \propto 1$, that is the uniform prior. An interesting feature of the prior in (4.5) is that it has, as particular cases, other well-known priors, besides the uniform prior. By setting, $c = 1$ and $h = 1$ we recover the prior in equation (4.3) proposed by Carvalho and Scott (2009).

If we set $c = 0$ we obtain

$$\pi(G) \propto \exp\{-h|G|\},$$

which resembles the prior of Villa and Lee (2019), introduced in the context of linear regression.

Let $M(G)$ represent the set of symmetric positive-definite matrices constrained by G, which means there is an equivalence between the zeroes of the concentration matrix Σ_G^{-1} and the missing edges from graph G. The function $f(\mathbf{x}|\Sigma_G, G)$ denotes the multivariate Gaussian sampling distribution with covariance matrix Σ_G. Then, the graph posterior probability is:

$$\pi(G|\mathbf{x}) \propto \pi(G) \int_{\Sigma_G \in M(G)} f(\mathbf{x}|\Sigma_G, G)\pi(\Sigma_G|G)\,\mathrm{d}\Sigma_G.$$

Although our prior is suitable for both decomposable and non-decomposable graphs, here we focus on the former class of graphs so that we can compare the performance of our prior to other priors available in the literature.

Regarding the marginal likelihood, we are using the hyper-inverse Wishart g-prior of Carvalho and Scott (2009) as the prior for the constrained covariance matrix Σ_G. This prior arises as the implied fractional prior of the covariance matrix (O'Hagan, 1997) for the following noninformative prior (see Geisser and Cornfield (1963), Sun and Berger (2007)), whose form was purposely selected to maintain conjugacy:

$$\pi_N(\Sigma|G) \propto \frac{\prod_{C\in\mathscr{C}} \det(\Sigma_C)^{-|C|}}{\prod_{S\in\mathscr{S}} \det(\Sigma_S)^{-|S|}}.$$

Here, $\mathscr{C}$ and $\mathscr{S}$ represent the clique and separator sets for graph G, respectively. Furthermore, the hyper-inverse Wishart g-prior is a conjugate prior for the multivariate Gaussian distribution. As such, the marginal likelihood can be expressed in closed form as (see Carvalho and Scott (2009)):

$$f(\mathbf{x}|G) = (2\pi)^{-np/2} \frac{H_G(gn, g\mathbf{x}^T\mathbf{x})}{H_G(n, \mathbf{x}^T\mathbf{x})},$$

with $H_G(b,D)$ denoting the normalising constant of the hyper-inverse Wishart distribution with degrees of freedom parameter $b \in \mathbb{R}^+$ and scale matrix $D \in M(G)$. For a decomposable graph, $H_G(b,D)$ can be expressed as a ratio of products over the cliques and separators (see Dawid and Lauritzen (1993), Atay-Kayis and Massam (2005), Carvalho and Scott (2009)),

that is

$$H_G(b,D) = \frac{\prod_{C\in\mathscr{C}} \det\left(\frac{1}{2}D_C\right)^{\frac{b+|C|-1}{2}} \Gamma_{|C|}\left(\frac{b+|C|-1}{2}\right)^{-1}}{\prod_{S\in\mathscr{S}} \det\left(\frac{1}{2}D_S\right)^{\frac{b+|S|-1}{2}} \Gamma_{|S|}\left(\frac{b+|S|-1}{2}\right)^{-1}},$$

where

$$\Gamma_a(x) = \pi^{\frac{a(a-1)}{4}} \prod_{j=1}^{a} \Gamma(x+(1-j)/2)$$

represents the multivariate gamma function.

As recommended by Carvalho and Scott (2009), in all our further analyses we set $g = 1/n$. To explore the graph space we have used the *feature-inclusion stochastic search* (FINCS) algorithm of Scott and Carvalho (2008). FINCS is a serial procedure which utilises three types of moves: local, resampling and global. The local moves depend on updated estimates of the posterior edge inclusion probabilities. Resampling of one of the previously visited models is done in proportion to the their posterior probabilities. The global moves allow us to explore those regions that would not be accessible in a finite number of local steps and with the help of the local moves, they try to address the multimodality of the problem at hand. Clearly, FINCS is not an MCMC scheme, but a hybrid algorithm designed to explore a collection of likely graphs.

As suggested by Scott and Carvalho (2008), for small-to-moderate-sized graphs like the graph with 25 vertices used by them in their simulation study from Section 4, the convergence of FINCS is quite fast irrespective of the starting graph. The necessity of the global move becomes apparent when the true graph has a lot of vertices. Then, a version of FINCS with only local moves and resampling steps would get trapped in the local hills, a behaviour which was also observed with the standard Metropolis-Hastings. Moreover, taking into account the enormity of the graph space to be explored, even a global variant of FINCS would depend on the starting graph. Here, the original authors have used an initial estimated

graph based on conditional regressions which is the default setting in the FINCS algorithm. Furthermore, Scott and Carvalho (2008) recommend a mixture of 80% to 90% local moves with the remainder used for global moves. Out of those local moves, 10% to 15% should be dedicated to the resampling step. This is why in all our considered simulated and real data analyses, we have used the default setting of the original authors from Section 4 of their paper, namely a global version of the FINCS algorithm with a resampling step every 10 iterations and a global move used every 50 iterations. A more detailed outline of the algorithm is provided in Appendix B.

4.3 Simulated and Real Data Analysis

In this section, we are showing the behaviour of the prior in equation (4.5) in both simulated and real data scenarios. When we focus on decomposable graphs, the inference is made by implementing the FINCS algorithm, whilst when we utilise the main algorithm of Mohammadi and Wit (2019), we look at unrestricted graphs.

For the analyses, on simulated and real data, we compare four priors on G. Namely, the Carvalho and Scott prior (CS prior), the uniform prior (UP prior) and the proposed prior with two different settings: in the first we have $h = 1$ and $c = 0$ (VL prior) and for the second we have $h = 1$ and $c = 0.5$ (MP prior). Thus:

$$\pi^{\text{VL}}(G) \propto \exp\{-|G|\} \qquad \text{and} \qquad \pi^{\text{MP}}(G) \propto \exp\left\{-\left[\frac{1}{2}|G| + \frac{1}{2}\log\binom{m}{|G|}\right]\right\}.$$

The above choices of the two priors have been dictated by the following reasons. The VL prior allows to highlight the choice of a prior that penalises for the *absolute* graph complexity without including any prior information on the rate of penalisation (controllable by setting h). The choice of the MP prior is driven by the motivation of understanding how equal weights for the two types of the considered penalties, i.e. *absolute* versus *relative*, interplay.

4.3.1 Simulated Data Example

The first simulation study has been taken from Carvalho and Scott (2009). We start from a graph with 10 vertices and 20 edges, which is represented in Figure 4.1. We have then added 5 and 40 noise vertices for, respectively, the first and the second simulation. These noise vertices represent vertices unconnected to each other or with the 10 vertices graph. The data has been simulated from a zero mean multivariate normal distribution with the covariance matrix designed to represent the dependencies of the above graphs. In both cases the sample size was of $n = 50$ observations. That is, we have sampled 50 realisations for a $p = 15$ vertices graph and a $p = 50$ vertices graph, where each graph contains just the edges shown in Figure 4.1, through the `R` package BDgraph of Mohammadi and Wit (2019).

For the simulated data, we are using a single covariance matrix for each of the two cases. More precisely, to simulate the data we have used the `bdgraph.sim()` function with the following arguments: the adjacency matrix was given respectively by one of the two graph structures described previously and the *G*-Wishart prior was the default one. We have run FINCS for 5 million iterations and set a global move every 50 iterations; the resampling step was considered at every 10 iterations. During the FINCS search, we have saved the best 1000 graphs. That is, we have used the default setting of FINCS which outputs in descending order the first 1000 most likely graphs according to the associated posterior probabilities (Carvalho and Scott, 2009). The estimated edge posterior inclusion probabilities were computed as

$$\hat{q}_{ij} = \frac{\sum_{r=1}^{t} \mathbb{1}_{(i,j)\in G_r} f(\mathbf{x}|G_r)\pi(G_r)}{\sum_{r=1}^{t} f(\mathbf{x}|G_r)\pi(G_r)},$$

with t being the number of uniquely discovered graphs in terms of the log-score amongst all our iterations, and reported in Table 4.1, for the case $p = 15$, and in Table 4.2, for the case $p = 50$.

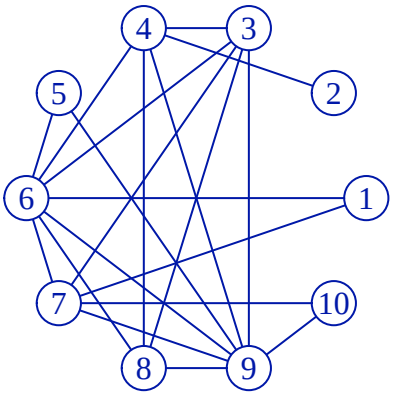


Fig. 4.1 The 10 vertices graph we have used in our simulation study.

Edge	Noise Vertices: 5 (p=15)			
	CS prior	VL prior ($c = 0$)	MP prior ($c = 0.5$)	UP prior
(1,6)	0.167	0.234	0.216	0.158
(1,7)	0.916	0.981	0.960	0.997
(2,4)	0.079	0.173	0.126	0.184
(3,4)	0.014	0.017	0.018	0.321
(3,6)	0.961	0.994	0.987	0.999
(3,7)	0.198	0.355	0.282	0.311
(3,8)	0.997	1.000	0.999	1.000
(3,9)	0.013	0.012	0.013	0.025
(4,6)	0.023	0.025	0.027	0.366
(4,8)	0.005	0.003	0.005	0.006
(4,9)	0.493	0.877	0.721	0.984
(5,6)	0.007	0.003	0.005	0.007
(5,9)	0.698	0.958	0.878	0.994
(6,7)	0.014	0.014	0.015	0.013
(6,8)	0.005	0.009	0.007	0.018

(6,9)	0.011	0.013	0.011	0.297
(7,9)	0.213	0.153	0.179	0.097
(7,10)	1.000	1.000	1.000	1.000
(8,9)	0.006	0.007	0.007	0.015
(9,10)	0.785	0.874	0.834	0.962
FPs:	0	1	0	2

Table 4.1 The estimated edge posterior inclusion probabilities together with the remaining false positive flags (FPs) when the number of noise vertices is 5.

Edge	Noise Vertices: 40 (p=50)			
	CS prior	VL prior ($c = 0$)	MP prior ($c = 0.5$)	UP prior
(1,6)	1.000	1.000	1.000	1.000
(1,7)	1.000	1.000	1.000	1.000
(2,4)	0.454	0.996	0.753	1.000
(3,4)	0.002	0.003	0.003	0.120
(3,6)	0.000	0.000	0.000	0.000
(3,7)	0.000	0.000	0.000	0.000
(3,8)	0.999	1.000	1.000	1.000
(3,9)	0.001	0.001	0.001	0.006
(4,6)	0.000	0.000	0.000	0.000
(4,8)	1.000	1.000	1.000	1.000
(4,9)	0.089	0.001	0.016	0.002
(5,6)	0.000	0.000	0.000	0.001
(5,9)	1.000	1.000	1.000	1.000
(6,7)	1.000	1.000	1.000	1.000
(6,8)	0.000	0.000	0.000	0.001

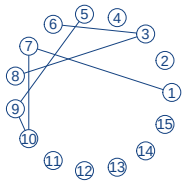
(6,9)	0.991	1.000	1.000	1.000
(7,9)	0.992	1.000	1.000	1.000
(7,10)	0.000	0.000	0.000	0.001
(8,9)	0.912	1.000	0.985	1.000
(9,10)	1.000	1.000	1.000	1.000
FPs:	0	11	2	41

Table 4.2 The estimated edge posterior inclusion probabilities together with the remaining false positive flags (FPs) when the number of noise vertices is 40.

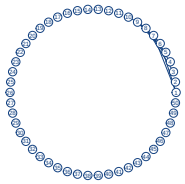
In Table B.1 from Appendix B, we can see the computed Pearson correlation matrix corresponding to the sample used in Table 4.1, whereas in Table B.2 we have that correlation matrix for the sample we have dealt with in the case of Table 4.2. Note that the estimated edge posterior inclusion probabilities for $(8,9)$ in Table 4.1 are very low in comparison to the ones from Table 4.2. A possible explanation is that due to the fact that we randomly sample the data necessary for our simulation study through the aforementioned `bdgraph.sim()`, the initial information in the generated random sample considered in Table 4.1 about a possible connection between variables 8 and 9 is very faint in contrast to the sample used in Table 4.2. Taking into account Tables B.1 and B.2, we indeed see that this is the case, as the correlation coefficients for $(8,9)$ are 0.14 and 0.98, respectively. Furthermore, the sample size is quite small at 50.

In terms of false positive flags (FPs), we see an increase for the VL and UP priors when moving from 5 to 40 noise vertices; although of different sizes. In fact, the VL prior moves from 1 to 11 false positives, while the UP prior moves from 2 to 41. For the MP prior, that is when we mix the VL and the CS prior with equal weights, the increase in FPs is marginal. To compare the inferential results of the priors we consider the median probability graphs, that is the graphs composed by all the edges with a posterior inclusion probability of at

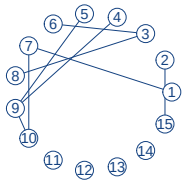
least 0.5 (Carvalho and Scott, 2009). In both cases the priors yield to similar graphs, with the exception of edge (4, 9) for the experiment with $p = 15$ and (2, 4) for the experiment with $p = 50$. The above edges are not included in the graph derived by using the CS prior, although the posterior inclusion probability is close to 0.5 (0.49 and 0.45, respectively). The estimated median graphs under the CS, VL, MP and UP priors can be seen in Figure 4.2, where the left column represents the case where 5 noise vertices were added, whereas the right column designates the case for the insertion of 40 noise vertices.



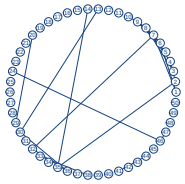
(a) CS with 6 edges for 5 noise vertices



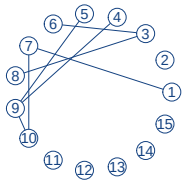
(b) CS with 10 edges for 40 noise vertices



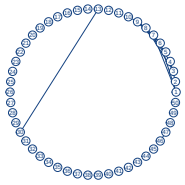
(c) VL with 8 edges for 5 noise vertices



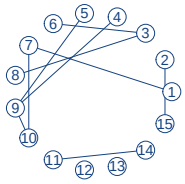
(d) VL with 22 edges for 40 noise vertices



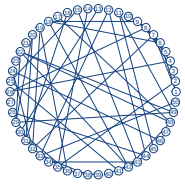
(e) MP with 7 edges for 5 noise vertices



(f) MP with 13 edges for 40 noise vertices



(g) UP with 9 edges for 5 noise vertices



(h) UP with 52 edges for 40 noise vertices

Fig. 4.2 The corresponding posterior graphs and their sizes where the estimated edge posterior inclusion probability is greater than 0.5 obtained under the four priors (CS, VL, MP and UP) in the case of inserting 5 noise vertices (left column) or 40 noise vertices (right column).

In the second simulation exercise, we study the performance of the proposed prior when initial information about the number of edges is available (and one wishes to reflect this in the prior). The results are compared to the ones obtained by using the Bernoulli prior (see equation (4.1)) implemented in the BDgraph package. We consider both the case of accurate prior information as well as the case where the prior information about the true number of edges is not accurate. We have considered the scenarios with $n = 50$ and $n = 100$, and we have repeated the analysis for 250 randomly generated samples. Computational details are that we have employed the BDgraph package appropriately modified to allow the implementation of our prior, and we have run 200000 iterations with a burn-in of 100000.

First, we have simulated from a graph with 6 vertices and 3 edges, and assumed that the prior information about the expected number of edges was correct. To have prior distributions with mean of 3, we have set $h = 0.28$ and $c = 0.11$ for the MP prior and chose a probability of success of 0.2 for the Bernoulli prior. We have compared the two priors by considering the average size of the posterior graphs over the 250 samples. Table 4.3 shows the statistics of the simulation study, including the 99% bootstrap confidence interval based on one million replicates. We note that the MP prior outperforms the Bernoulli prior as the confidence intervals contain the true graph size for both $n = 50$ and $n = 100$. For the second case, we have sampled from a graph with 6 vertices and 5 edges assuming that the prior information about the true graph size is as before (i.e. 3 edges). If we keep the MP prior with the same setting as above, the 99% confidence intervals are $(4.04, 4.56)$ and $(4.25, 4.76)$ for, respectively, $n = 50$ and $n = 100$. However, the MP prior allows to set h and c to have the same prior mean as above and a larger variance. In Table 4.4 we report the frequentist summaries for MP prior with a variance of 35.5 and the Bernoulli prior. The MP's variance of 35.5 is attained for $h = 1.36$ and $c = 0.93$. This variance is larger than in the case of accurate prior information (that variance was 11) as we would like to include more uncertainty about the prior number of edges, because now we are unsure about the respective prior knowledge. The variance of the Bernoulli prior is 2.4 which is simply the formula for the variance of the Bernoulli distribution where the number of trials was set to the number of vertices and the probability of success was 0.2. These parameters in the Bernoulli prior lead to a prior mean of 3. We note, in the case of inaccurate prior information, that the confidence intervals for the MP prior contain the true number of edges. Although there is a discrepancy in terms of variance between the MP prior and the Bernoulli prior, this shows a higher versatility of the MP prior as it allows to control two pieces of prior information (mean and variance) by the choice of the parameters h and c.

Prior	$n = 50$		$n = 100$	
	Average Size	99% Confidence Interval	Average Size	99% Confidence Interval
MP	2.81	(2.61, 3.02)	2.96	(2.78, 3.14)
Bernoulli	1.88	(1.74, 2.04)	2.22	(2.08, 2.36)

Table 4.3 Frequentist summaries for the MP prior and the Bernoulli prior when prior information is accurate.

Prior	$n = 50$		$n = 100$	
	Average Size	99% Confidence Interval	Average Size	99% Confidence Interval
MP	5.47	(4.64, 6.34)	4.50	(3.98, 5.08)
Bernoulli	3.36	(3.13, 3.60)	3.82	(3.61, 4.03)

Table 4.4 Frequentist summaries for the MP prior and the Bernoulli prior when prior information is not accurate.

4.3.2 Real Data Examples

In this section we illustrate our prior in real data scenarios. We compare the performance of our prior with the other priors considered in the previous section. We have selected three datasets, encompassing different sizes, both in terms of variables and in terms of number of observations. The results, obtained with the same settings for the FINCS algorithm as implemented for simulation studies during Section 4.3.1, are presented in the next subsections. For comparison purposes, edges have been selected as part of the estimated graph if their posterior inclusion probability was at least 0.5 (median probability graph).

Let us recall that the estimated edge posterior inclusion probabilities are computed using model averaging over the uniquely discovered graphs in terms of the log-score. Unfortunately, due to the exponential nature of the graph space for a fixed p (there are $2^{\binom{p}{2}}$ graphs), these posterior inclusion probabilities are just a search heuristic as Scott and Carvalho (2008) also

outlined. They converge to the true edge posterior probabilities, when we could talk about finding an exact graph, in very particular cases, namely when the true graph is very sparse (its number of edges is 5 or 6), the number of vertices is very small (p is 4 or 5) or when we would be able to visit all the graphs in a graph space in the provided number of iterations.

The Multivariate Flow Cytometry Dataset

Sachs et al. (2005) consider flow cytometry measurements for 11 phosphorylated proteins and phospholipids across a total number of 7466 observations. The 11 proteins considered have the following nomenclature: Raf, Mek, Plcg, PIP2, PIP3, Erk, Akt, PKA, PKC, P38, Jnk. The purpose of their study was to infer a Bayesian network to reveal possible connections between enzymes. We have centred the data and the key results are reported in Table 4.5 and Table 4.6.

The most sparse graph was produced using the VL prior, and the included edges are listed in Table 4.5. In Table 4.6 , we can see the edges that were omitted for the VL prior, but included for the others. The most complex graph is selected under the CS prior, where 5 extra edges are added, while the MP and the UP priors include, respectively, 1 and 2 edges more than the VL prior. To note, edge $(1,8)$, which is included by all the priors except the VL prior, has a posterior inclusion probability for the latter prior relatively close to 0.5, suggesting that it is likely to be the sole relevant difference among the priors. For the remaining edges in Table 4.6, a more conservative threshold (e.g. set at 0.7) would have excluded them from all the graphs. For the included edges (Table 4.5), there is strong agreement among the priors, as the posterior inclusion probabilities are all quite close to one. The estimated graphs identified by FINCS under the four aforementioned priors can be seen in Figure 4.3.

Index	Edge	CS prior	VL prior	MP prior	UP prior
1	(1,2)	1.000	1.000	1.000	1.000

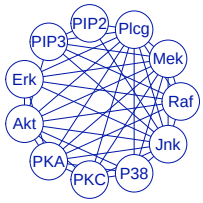
2	(1,3)	1.000	1.000	1.000	1.000
3	(1,6)	1.000	1.000	1.000	1.000
4	(1,7)	1.000	1.000	1.000	1.000
5	(1,11)	0.999	0.999	0.999	0.999
6	(2,3)	1.000	1.000	1.000	1.000
7	(2,6)	1.000	1.000	1.000	1.000
8	(2,7)	1.000	1.000	1.000	1.000
9	(2,8)	0.999	0.997	0.998	0.999
10	(2,10)	0.892	0.932	0.907	0.904
11	(2,11)	0.999	1.000	0.999	0.999
12	(3,4)	1.000	1.000	1.000	1.000
13	(3,5)	1.000	1.000	1.000	1.000
14	(3,6)	1.000	1.000	1.000	1.000
15	(3,7)	1.000	1.000	1.000	1.000
16	(3,8)	1.000	1.000	1.000	1.000
17	(3,9)	0.978	0.910	0.952	0.957
18	(3,10)	0.999	0.983	0.996	0.997
19	(3,11)	1.000	1.000	1.000	1.000
20	(4,5)	1.000	1.000	1.000	1.000
21	(5,7)	1.000	1.000	1.000	1.000
22	(5,11)	0.947	0.938	0.924	0.923
23	(6,7)	1.000	1.000	1.000	1.000
24	(6,8)	1.000	1.000	1.000	1.000
25	(6,11)	1.000	1.000	1.000	1.000
26	(7,8)	1.000	1.000	1.000	1.000
27	(7,9)	1.000	1.000	1.000	1.000

28	(7,10)	1.000	1.000	1.000	1.000
29	(7,11)	1.000	1.000	1.000	1.000
30	(8,9)	1.000	1.000	1.000	1.000
31	(8,10)	1.000	1.000	1.000	1.000
32	(8,11)	1.000	1.000	1.000	1.000
33	(9,10)	1.000	1.000	1.000	1.000
34	(9,11)	1.000	1.000	1.000	1.000
35	(10,11)	1.000	0.999	1.000	1.000

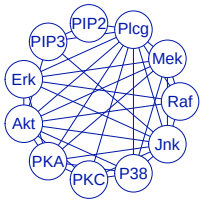
Table 4.5 Edges with a posterior inclusion probability of at least 0.5 for all four priors considered.

Index	Edge	CS prior	VL prior	MP prior	UP prior
1	(1,5)	0.550	0.043	0.182	0.216
2	(1,8)	0.832	0.436	0.644	0.677
3	(2,5)	0.561	0.046	0.190	0.224
4	(2,9)	0.656	0.322	0.480	0.507
5	(4,11)	0.528	0.197	0.338	0.363

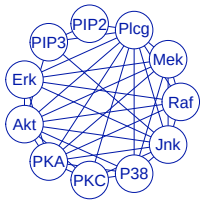
Table 4.6 Edges with a posterior inclusion probability smaller than 0.5 under the VL prior, but with a value larger than 0.5 under at least one of the other three priors.



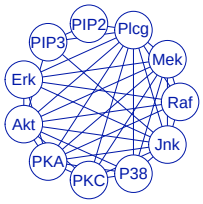
(a) CS with 40 edges



(b) VL with 35 edges



(c) MP with 36 edges



(d) UP with 37 edges

Fig. 4.3 The corresponding posterior graphs and their sizes where the estimated edge posterior inclusion probability is greater than 0.5 obtained under the four priors (CS, VL, MP and UP) through the FINCS algorithm for the flow cytometry dataset.

The PTSD Symptoms for Earthquake Survivors in Wenchuan, China Dataset

This dataset (McNally et al., 2015) represents the measurement of 17 symptoms associated with PTSD (Post-traumatic stress disorder) reported by 362 survivors of an earthquake from the Wenchuan county in the Sichuan province, China. Each of the participants indicated through a ordinal scale from 1 to 5 how affected they were by every single one of the 17 PTSD symptoms, where 1 signifies not being bothered by the symptom at hand, whereas 5 corresponds to an extreme response to the same symptom. All participants have lost at

least one child in the respective earthquake. The data is available with the `R` package APR (Mair, 2015). Amongst those 362 answers, in 18 cases, there was missing information associated with one or several symptoms. These cases were discarded, leaving a final sample of 344 participants, and the data was centred. In Table 4.7, we provide the mapping between the numeric identifiers for the variables and the corresponding PTSD symptoms and their meaning as given by McNally et al. (2015). These numeric identifiers will be used in the subsequent tables and figures shown in this subsection.

Numeric Identifier	PTSD symptom
1	"*intrusion* = intrusive memories, thoughts, or images of the trauma"
2	"*dreams* = traumatic dreams"
3	"*flash* = flashbacks"
4	"*upset* = feeling upset in response to reminders of trauma"
5	"*physior* = physiological reactivity to reminders of the trauma"
6	"*avoidth* = avoidance of thoughts or feelings about the trauma"
7	"*avoidact* = avoidance of activities or situations reminiscent of the trauma"
8	"*amnesia* = having trouble remembering parts of the traumatic experience"
9	"*lossint* = loss of interest in previously enjoyed activities"
10	"*distant* = feeling distant or cut off from people"
11	"*numb* = feeling emotionally numb"
12	"*future* = feeling that your future will be cut short"
13	"*sleep* = difficulty falling or staying asleep"
14	"*anger* = feeling irritable or having angry outbursts"
15	"*concen* = difficulty concentrating"

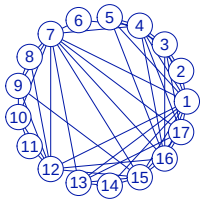
16	"*hyper* =hyper-vigilant or watchfull or super alert"
17	"*startle* = feeling easily startled or jumpy"

Table 4.7 The mapping between the numeric identifiers for the variables and the corresponding PTSD symptoms and their meaning as provided by McNally et al. (2015).

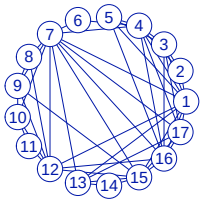
The sparser graph is identified under the MP prior and it contains 44 edges. With exception of edge $(13, 16)$, the remaining 43 edges were also included in the other three priors. Table 4.8 reports the 8 edges not included under all four priors. The estimated graphs identified by FINCS under the four aforementioned priors can be seen in Figure 4.4.

Index	Edge	CS prior	VL prior	MP prior	UP prior
1	(1,14)	0.608	0.492	0.413	0.763
2	(1,17)	1.000	1.000	0.456	1.000
3	(2,4)	0.513	0.512	0.385	0.463
4	(3,17)	0.528	0.531	0.246	0.634
5	(4,17)	0.994	0.969	0.442	0.998
6	(7,17)	0.908	0.895	0.414	0.999
7	(9,11)	0.495	0.405	0.431	0.663
8	(13,16)	0.027	0.019	0.562	0.045

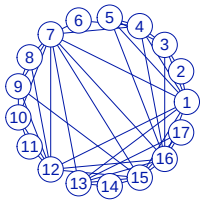
Table 4.8 Edges with a posterior inclusion probability larger than 0.5 for one to three of the four considered priors.



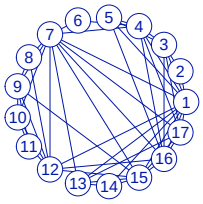
(a) CS with 49 edges



(b) VL with 48 edges



(c) MP with 44 edges



(d) UP with 49 edges

Fig. 4.4 The corresponding posterior graphs and their sizes where the estimated edge posterior inclusion probability is greater than 0.5 obtained under the four priors (CS, VL, MP and UP) through the FINCS algorithm for the PTSD symptoms dataset.

In addition, we have modified the Gaussian copula part (Mohammadi et al., 2017) of the BDgraph package so that our prior can be utilised. The modifications simply consist of adding the parameters h and c to the `bdgraph()` function in `R` and modifying the `C++` code which deals with the prior computations to accommodate for our prior. No alterations were done to the Gaussian copula part of the main algorithm. Amongst the computations we had to rework, an important part is represented by the ratios between the priors on the proposed new graph and the current one, as seen in the work of Mohammadi and Wit (2015). In our

case, when the proposed graph includes a new edge, the ratio is

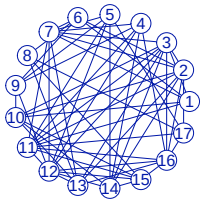
$$\exp\left\{h\left[c\left(\log\left(\frac{m-|G_{proposed}|+1}{|G_{proposed}|}\right)-1\right)+1\right]\right\},$$

whereas when the proposed graph contains one less edge than the current one, the ratio becomes

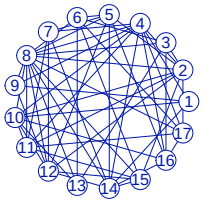
$$\exp\left\{h\left[c\left(\log\left(\frac{|G_{proposed}|+1}{m-|G_{proposed}|}\right)+1\right)-1\right]\right\},$$

with $|G_{proposed}|$ representing the number of edges in the proposed graph.

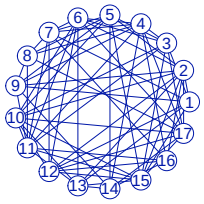
We have compared the maximum posterior graphs which were found under the CS, VL, UP and a specific case of the mixture prior where the parameters were set to $h = 1$ and $c = 0.5$ when 2000000 iterations were used with the first 1000000 discarded. We have also included the Bernoulli prior (denoted with BD) where the probability of edge inclusion was set to 0.004 so that a priori it had the same mean as the VL prior, namely 0.582. The maximum posterior probability graph found under the Bernoulli prior has a size of 59 and can be seen in Figure 4.5. Furthermore, as we can observe in Figure 4.5, under the CS prior, the posterior graph had a size of 59, whilst for the VL prior, the posterior had a size of 58. When we use the mixture prior, the size of the maximum posterior probability graph is 68. The most dense graph is obtained under the UP which had a size of 71. The number of common edges between the BD and various graphs is as following: 20 edges with the CS, 29 edges with the VL, 27 with the MP and 26 with the UP. Clearly, under this particular running of the bdmcmc (birth death Markov chain Monte Carlo) algorithm, the sparser graph is found under the VL prior, but the other maximum posterior probability graphs corresponding to the BD and CS priors are close to its size. The only exception is given by the uniform prior which leads to a denser graph.



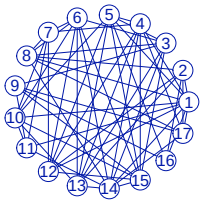
(a) BD with 59 edges



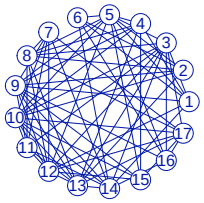
(b) VL with 58 edges



(c) MP with 68 edges



(d) CS with 59 edges



(e) UP with 71 edges

Fig. 4.5 The corresponding maximum posterior probability graph sizes under the four priors (CS, VL, MP and UP) together with the Bernoulli prior (BD) for the PTSD symptoms data.

Note that in all the graphs from Figures 4.4 and 4.5, besides the graphs from Figures 4.5b and 4.5e , the edge $(6, 7)$ is present. This edge signifies a strong interaction between the *avoidth* and *avoidact* PTSD symptoms. This link was also noticed by McNally et al. (2015). Its possible omission in the two graphs could be due to the fact that in Figure 4.5 we simply show the maximum posterior graphs and not estimated graphs based on model averaging.

The Breast Cancer Dataset

Hess et al. (2006) have collected gene expression data for 133 patients which had breast cancer. This dataset was also analysed by Ambroise et al. (2009) and made available through the `R` package SIMONE (Statistical Inference for MOdular NEtworks) developed by one of the authors. There are 26 genes considered in the study. The dataset is split in two groups, one pertaining to the pathological complete response (pCR) to the chemotherapy treatment started after surgery, whereas the other corresponds to the disease still being present in the patients (not-pCR). First, we have looked at the not-pCR cases which was recorded for 99 patients. The remaining 34 patients had a positive response to the treatment (the pCR case). The data has been centred.

The estimated graphs are reported in Figure 4.6, where we have shown the results under each prior, that is CS, VL, MP and UP; furthermore, the analysis has been performed on each group separately. Comparing the performance of the priors, we note that the CS prior and the MP prior give relatively sparse graphs for both groups, 21 and 23 edges for the pCR, and 25 and 26 for the not-pCR, respectively. The VL prior yields to slightly larger graphs (28 and 39) while the UP is the prior resulting in the most complex graphs (42 and 46). If we compare, within each prior, the obtained graphs for the two groups, we consistently notice that the graphs for the pCR group are sparser than the graphs for the not-pCR group. As the sample size for the not-pCR group is larger than the size for the pCR group, it may be that

the more complex posterior graphs are a result of the higher amount of information from the observations.

Index	Edge	CS prior	VL prior	MP prior	UP prior
1	(1,14)	0.098	0.100	0.138	0.824
2	(1,15)	0.759	0.819	0.742	0.143
3	(2,8)	0.056	0.314	0.127	0.870
4	(4,6)	0.109	0.129	0.099	0.622
5	(4,7)	0.461	0.343	0.475	0.970
6	(4,8)	0.176	0.889	0.165	0.313
7	(4,11)	0.019	0.879	0.012	0.000
8	(4,13)	0.003	0.643	0.003	0.000
9	(4,15)	0.320	0.200	0.344	0.850
10	(4,17)	0.000	0.887	0.001	0.005
11	(4,19)	0.002	0.661	0.003	0.000
12	(6,9)	0.160	0.568	0.162	0.998
13	(6,15)	0.365	0.705	0.372	1.000
14	(6,26)	0.003	0.480	0.004	0.999
15	(7,8)	0.000	0.001	0.000	0.965
16	(7,11)	0.001	0.801	0.001	0.000
17	(7,15)	0.000	0.000	0.000	0.912
18	(7,16)	0.024	0.092	0.046	0.521
19	(7,17)	0.000	1.000	0.001	1.000
20	(7,23)	0.002	0.005	0.004	0.956
21	(8,12)	0.000	0.000	0.000	0.606
22	(8,23)	0.058	0.005	0.024	0.865
23	(9,15)	0.878	0.479	0.894	0.034

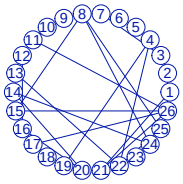
24	(9,26)	0.213	0.770	0.296	1.000
25	(11,13)	0.145	0.041	0.201	0.551
26	(11,14)	0.560	0.041	0.636	0.931
27	(11,17)	0.364	0.968	0.408	0.472
28	(11,19)	0.000	0.849	0.000	0.000
29	(12,17)	0.000	0.741	0.000	0.951
30	(12,24)	0.002	0.872	0.001	0.985
31	(13,14)	0.291	0.237	0.604	0.979
32	(14,20)	0.003	0.013	0.009	0.752
33	(17,19)	0.003	0.998	0.001	0.006
34	(17,23)	0.018	0.065	0.007	0.583
35	(17,25)	0.036	0.980	0.075	0.999

Table 4.9 Posterior inclusion probabilities not included under all the four compared priors for the not-pCR case.

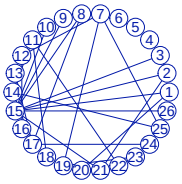
Index	Edge	CS prior	VL prior	MP prior	UP prior
1	(2,9)	0.001	0.008	0.004	0.538
2	(2,10)	0.001	0.001	0.001	0.985
3	(5,16)	0.190	0.522	0.333	0.994
4	(5,17)	0.309	0.750	0.510	0.996
5	(6,16)	0.011	0.016	0.011	0.759
6	(6,17)	0.111	0.775	0.464	0.983
7	(8,10)	0.000	0.000	0.000	1.000
8	(8,15)	0.621	0.028	0.561	1.000
9	(8,16)	0.001	0.995	0.017	1.000

10	(8,20)	0.001	0.011	0.001	0.720
11	(8,25)	0.953	0.969	0.972	0.251
12	(8,26)	0.241	0.998	0.389	1.000
13	(9,26)	0.004	0.021	0.010	0.601
14	(10,15)	0.000	0.000	0.000	0.999
15	(10,16)	0.001	0.641	0.001	1.000
16	(10,18)	0.000	0.010	0.000	0.996
17	(10,21)	0.009	0.003	0.007	0.987
18	(10,26)	0.001	0.005	0.002	1.000
19	(11,16)	0.000	0.984	0.002	0.010
20	(11,18)	0.056	0.980	0.045	0.001
21	(14,20)	0.652	0.008	0.660	0.972
22	(15,16)	0.000	0.004	0.000	0.999
23	(15,26)	0.991	0.993	0.994	0.013
24	(16,17)	0.000	0.055	0.000	0.963
25	(16,25)	0.012	0.037	0.007	0.785
26	(16,26)	0.000	0.995	0.007	1.000
27	(17,22)	0.000	0.000	0.000	0.741
28	(17,25)	0.000	0.000	0.000	0.751
29	(18,26)	0.362	0.021	0.508	0.062
30	(20,24)	0.001	0.000	0.003	0.972

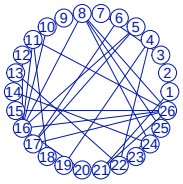
Table 4.10 Posterior inclusion probabilities not included under all the four compared priors for the pCR case.



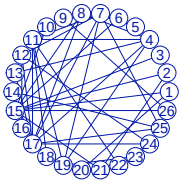
(a) CS with 21 edges (pCR)



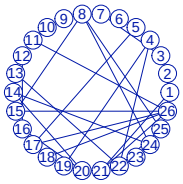
(b) CS with 25 edges (not-pCR)



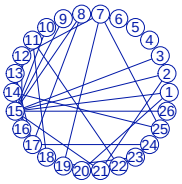
(c) VL with 28 edges (pCR)



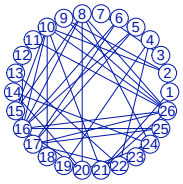
(d) VL with 39 edges (not-pCR)



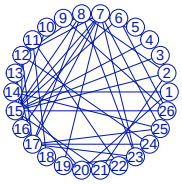
(e) MP with 23 edges (pCR)



(f) MP with 26 edges (not-pCR)



(g) UP with 42 edges (pCR)



(h) UP with 46 edges (not-pCR)

Fig. 4.6 The corresponding posterior graphs and their sizes where the estimated edge posterior inclusion probability is greater than 0.5 obtained under the four priors (CS, VL, MP and UP) for the two groups.The first column corresponds to the pCR group, whilst the second column contains the identified posterior graphs for the not-pCR group.

4.4 Discussion

In the present work, we have illustrated a novel prior on the space of graphs in the context of GGMs. The prior is derived using a loss with two components: one relative to the informational content of the graph and one related to its complexity. The results were obtained by implementing the FINCS algorithm and comparisons were made with two alternative weakly informative priors: the uniform prior and the prior advocated by Carvalho and Scott (2009), both of which can be seen as a particular case of the proposed prior.

We would like to provide some general remarks about setting the parameters h and c for the proposed prior. There are several ways to approach the issue:

- one could set h and c to reflect subjective prior information. See the example where we compare the proposed prior to the Bernoulli prior in Section 4.3.1.
- an alternative choice is to set $c = 0$ so that the prior will reduce to the global loss component only. Here, the parameter h can be either set according to some prior information or in a default manner (see Villa and Lee (2019)).
- the third choice is to set $c = 1$ and $h = 1$ and obtain the prior of Carvalho and Scott (2009). This would be the choice if one is interested in multiplicity correction.
- finally, one could fix $h = 1$ and then set c so to have a desired balance between the global and the local losses due to complexity. We have suggested that a default choice is for $c = 0.5$. In this scenario as well, given that the prior will depend on the total number of edges, there is correction for multiplicity.

Simulation studies, performed under a non-informative assumption, show that the best configuration of the proposed prior is when equal weight is given to absolute and relative complexity. In fact, the results are similar to the CS prior. In another simulation study we

have analysed the performance of the prior when prior information is available and it is reflected in the construction of the prior distribution. We have noticed favourable evidence, in particular when compared to the Bernoulli prior used by Mohammadi and Wit (2019). Here, we show that the dependence of our prior on the two tuning parameters allows to better include initial information when it is not limited to one piece only (i.e. expected number of edges).

Finally, we have illustrated the prior for three real datasets of different dimensionality and size. The proposed prior, in terms of sparsity, yields results in line with the CS prior, with the clear better performance for the first dataset (Flow Cytometry dataset).

5. An Extension of the Loss-based Methodology to Proper Binary Trees

This chapter shows how to extend the prior introduced in Chapter 4 to a different discrete structure, namely the binary tree. We start by re-establishing how do we use a tree in the context of data analysis. We then formally define our loss-based prior in the case of proper binary trees.

Let us recall that for a tree T_k with L_{T_k} leaves, the joint probability is simply:

$$h(\boldsymbol{y}|T_k,\Theta_k,\boldsymbol{X}) = \prod_{l=1}^{L_{T_k}}\prod_{j=1}^{n_l} f(y_{lj}|\theta_l), \quad \forall l = 1,2,\ldots,L_{T_k} \quad \text{and} \quad \forall j = 1,2,\ldots,n_l, \tag{5.1}$$

where $\boldsymbol{X}$ is the $n \times p$ design matrix and Θ_k is simply the collection of parameters associated with tree T_k.

Now if we consider a tree as a model and apply the methodology of Villa and Walker (2015a), we have:

$$\pi(T_k) \propto \exp\left\{\mathbb{E}_{\pi_k}\left[\inf_{\Theta_{k'},k'\neq k} D_{KL}(h(\boldsymbol{y}|T_k,\Theta_k,\boldsymbol{X})||h(\boldsymbol{y}|T_{k'},\Theta_{k'},\boldsymbol{X}))\right]\right\}, \tag{5.2}$$

where π_k denotes the priors on the tree parameters Θ_k. We know that the y_i responses are independent between terminal nodes and the joint distribution of them can be written as in equation (5.1). For this particular form of the prior, let us consider the next toy example. Let

us assume that the data generating process follows a Bernoulli distribution, that is

$$y_{lj} \overset{\text{i.i.d.}}{\sim} \text{Bernoulli}(\theta_l), \quad \forall l = 1,2,\ldots,L_{T_k} \quad \text{and} \quad \forall j = 1,2,\ldots,n_l.$$

Based on the additive property of the KL divergence when independent distributions are considered, the joint divergence from equation (5.2) is simply the sum of individual KL divergences between Bernoulli distributed variables. As each individual divergence is non-negative, the infinimum for the joint divergence is attained when each individual KL is minimised. The minimal KL divergence between two Bernoulli distributions is reached when the parameters of the two distributions are the same and this minimum is then 0. As such, we obtain $\pi(T_k) \propto 1$. Note that $T_k \in \mathscr{T}$, where $\mathscr{T}$ is a countably infinite discrete structure, which usually has an upper bound in practice (see Denison et al. (1998)). Therefore, it is necessary to consider model complexity when defining the prior on T_k.

5.1 Loss-based Prior on Binary Tree Structures

As seen in the Bayesian literature built around trees from Section 2.6, the types of trees we are interested are proper binary trees. These are the trees where the internal nodes always have two children. We know that the number of proper binary trees that have L_T leaves is simply C_{L_T-1} (Shaun et al., 2016), where $C_n = \dfrac{1}{n+1}\binom{2n}{n}$ is the n^{th} *Catalan number*. Through an analogy with the prior defined in Chapter 4, the loss-based tree prior has two components, namely a loss due to information and another related to the tree structure complexity, that is:

$$\pi(T_k) \propto \exp\{\text{Loss}_\text{I} + \text{Loss}_\text{C}\},$$

where Loss_I is the loss due to information (seen in equation (5.2) for tree structures) and Loss_C is the loss due to complexity.

As we can embed a given tree into another more complex tree, the loss due to information will be 0. Recall from Chapters 2, 3 and 4, that the loss in information is represented by the minimal KL divergence between the required model and all other available models. This minimization is realised with respect to the model parameters corresponding to those other models. So, it is clear that we may find a more complex tree that through a certain fixing of its parameters (they are approximately the same to the ones of the required model) simply reduces to the tree structure we would like to put a prior on. In essence, for every tree we may find another more developed tree which contains the respective tree and can be reduced to it through a specific way of choosing its tree parameters. This leads to the loss of information being 0. For a similar argument, we can refer to the work of Grazian et al. (2018) in the context of finite mixture models. As such, we can focus on the loss due to the tree structure complexity. Here, there are several choices depending on the parameters that a modeller wants to focus on. If we define the tree complexity as simply the number of leaves in the particular tree, a tree prior could essentially be akin to the prior utilised by Villa and Lee (2019) in the case of linear regression models, that is:

$$\pi^{\mathrm{VL}}(T_k) \propto \exp\{-\xi|L_{T_k}|\},$$

where $|L_{T_k}|$ represents the number of leaves corresponding to tree T_k and $\xi \in [0, +\infty)$. This prior, which we denote with VL, puts more mass on simpler trees according to the ξ parameter. A large value of ξ strongly emphasizes a priori trees with a small number of leaves, which could be useful in the BART (Bayesian Additive Regression Trees) methodology. A ξ close to zero is equivalent to a flat tree prior.

An alternative choice to the prior of Villa and Lee (2019) would be one that also penalizes the depth of the trees. Letting D_{T_k} represent the depth of tree T_k, we could define a tree prior

(designated by the VLD moniker) as:

$$\pi^{\text{VLD}}(T_k) \propto \exp\{-[\xi|L_{T_k}| + \omega|D_{T_k}|]\},$$

where ξ and $|L_{T_k}|$ were defined previously, whilst $|D_{T_k}|$ is the depth of the tree and $\omega \in [0,+\infty)$. By introducing the depth element and through the ω parameter we would be able to control how much mass is put on balanced shallow trees. Clearly a large ω would imply that this kind of trees would be favoured, whereas with ω close zero this prior's behaviour would be reduced to the case of the VL one. For a ξ close to zero, the penalty term would depend just on the depth element. As outlined previously, with a larger ω, we will heavily hinder the development of deep trees which could be especially useful in a BART setting.

Another choice for the penalty corresponding to the tree complexity could be a form similar to the one used for GGMs. Let us recall that the number of proper binary trees is related to the number of leaves of the tree through the Catalan numbers. Similarly to the argument presented in Chapter 4, the complexity of a tree could be split into absolute and relative parts. The absolute complexity essentially corresponds to just the number of leaves for the tree. The relative complexity is related to the number of trees with a particular number of leaves, essentially emphasizing the weight of a tree in a particular tree class. By a tree class we simply mean all the trees that have the same number of leaves. As such, we define the following prior which we denote with HLV:

$$\pi^{\text{HLV}}(T_k) \propto \exp\left\{-\xi\left[(1-\delta)|L_{T_k}| + \delta \log C_{L_{T_k}-1}\right]\right\},$$

with $\xi \in [0,+\infty)$ and $\delta \in [0,1]$. The δ parameter interpolates between the absolute and relative complexities. Clearly a δ close to 0, reduces the behaviour of the HLV prior to the VL one, whilst a δ close to 1 indicates that the relative complexity element is the most dominant in the prior's behaviour.

5.2 Discussion and Future Work

In this chapter, we have outlined what is the likelihood when discrete structures called binary trees are involved and also introduced the loss-based prior on those respective structures. We gave a reasoning for one element of the loss-based prior being 0, whilst also proposing various forms for the term which corresponds to the binary tree complexity.

An avenue for future work in this chapter is represented by the behaviour of one of the proposed loss-based priors under simulated and real data, together with a comparison with other tree priors mentioned in the literature. In particular, we would like to use our prior, possibly the one that is based on absolute and relative tree complexity, under the BART methodology for the classification of the existence or not of bikes in a city's docking stations at a certain time interval. This knowledge could allow a better redistribution of the bikes across the stations. Classical decision trees and random forests have been used to aid in modelling of the bike distributions across various stations in a bike-sharing system (BSS) by Yang et al. (2016). There are several cities for which BSS data is readily available from online sources. The benefit of the study will be twofold. On one side is the proposal of a new tree prior, followed by a Bayesian application of binary trees to predict the number of bikes that leave a station in a specific period. As our analysis will be based on BART, we could include various covariates like weather, season or other factors into the discussion.

6. Conclusion and Future Work

This thesis introduces priors based on the works of Villa and Walker (2015a) and Villa and Walker (2015b) in the context of heterogeneous data, namely change points and Gaussian graphical models (GGMs).

Chapter 2 contains the literature review part of the thesis. The first two sections review some of the objective priors utilised across the respective literature. The third section outlines the main ideas of Villa and Walker (2015b) together with their extension to model prior probabilities (Villa and Walker, 2015a). The next three sections are dedicated to discussing some of the influential papers related to the change point, GGMs and binary trees frameworks, respectively. Furthermore, these three sections also help identifying what are the methodologies that we would like our proposed techniques be compared with into the subsequent chapters.

Our first contribution is provided in Chapter 3. As a change point location can be interpreted as a discrete parameter, we may apply the methodology of Villa and Walker (2015b). This leads us to the discrete uniform prior on the change point locations which was also derived in a different way by Girón et al. (2007). Then, we may envision a Bayesian model selection exercise to detect the number of change points in various datasets by utilising the prior on the change point locations in the model prior probability context of Villa and Walker (2015a). We would like to emphasize that the simulated and real data we looked at during this chapter contained changes that could concern either the parameters of a distribution or the distributional family altogether.

Chapter 4 contains our second contribution, namely the proposal of a graph prior inspired by the approach undertaken by Villa and Lee (2019) in the case of linear regression models. In that chapter, we provide a reasoning for the existence of our proposed graph prior in the respective form in terms of absolute and relative model complexity penalties. We then show its behaviour relative to other graph priors from the GGM literature in terms of the estimated posterior edge inclusion probabilities when the FINCS algorithm is utilised or the maximum a posteriori estimated graph when the main algorithm from the BDgraph package is used. As the behaviour of our proposed prior is controlled through two parameters, in simulation studies we may observe that our prior performs best regarding the control of false positives when equal importance is given to both the absolute and relative complexity penalty parts of the prior.

The last contribution is showcased in Chapter 5. Motivated by the graph prior, we have looked at another discrete structure represented by proper binary trees and proposed several tree priors. Their justification follows somehow the arguments presented during Chapter 4, that is we argue that the loss due to information is 0. As such, the loss due to tree complexity impacts our proposed loss-based tree prior the most. This loss due to tree complexity may have different forms, thus encouraging different behaviours. A possible line for future work is represented by taking one of the proposed tree priors and comparing its behaviour with alternatives from the literature in terms of simulated and real data analysis. This future line of development could prove especially fruitful in the case of the BART methodology, as there is a need of tree priors which yield sparser trees as outlined by Chipman et al. (2010). In particular, we would like to use one of the proposed loss-based tree priors in a BART setting to model the bike distributions in a bike-sharing system.

References

Ambroise, C., Chiquet, J., and Matias, C. (2009). Inferring Sparse Gaussian Graphical Models with Latent Structure. *Electronic Journal of Statistics*, 3:205–238.

Armstrong, H., Carter, C. K., Wong, K. F. K., and Kohn, R. (2009). Bayesian Covariance Matrix Estimation using a Mixture of Decomposable Graphical Models. *Statistics and Computing*, 19(3):303–316.

Atay-Kayis, A. and Massam, H. (2005). A Monte Carlo Method for Computing the Marginal Likelihood in Nondecomposable Gaussian Graphical Models. *Biometrika*, 92(2):317–335.

Banerjee, O., El Ghaoui, L., and d'Aspremont, A. (2008). Model Selection Through Sparse Maximum Likelihood Estimation for Multivariate Gaussian or Binary Data. *The Journal of Machine Learning Research*, 9:485–516.

Barry, D. and Hartigan, J. A. (1992). Product Partition Models for Change Point Problems. *The Annals of Statistics*, 20(1):260–279.

Barry, D. and Hartigan, J. A. (1993). A Bayesian Analysis for Change Point Problems. *Journal of the American Statistical Association*, 88(421):309–319.

Bell, P. and King, S. (2007). Sparse Gaussian Graphical Models for Speech Recognition. In *INTERSPEECH 2007, 8th Annual Conference of the International Speech Communication Association, Antwerp, Belgium, August 27-31, 2007*, pages 2113–2116.

Berger, J. and Bernardo, J. (1992a). On the Development of Reference Priors. In Bernardo, J.M.., Berger, J.O.., Dawid, A.P.., and Smith, A.F.M.., editors, *Bayesian Statistics 4*, pages 35–60. Oxford University Press, London.

Berger, J. O. and Bernardo, J. M. (1992b). Ordered Group Reference Priors with Application to the Multinomial Problem. *Biometrika*, 79(1):25–37.

Berger, J. O., Bernardo, J. M., and Sun, D. (2009). The Formal Definition of Reference Priors. *The Annals of Statistics*, 37(2):905–938.

Berger, J. O., Bernardo, J. M., and Sun, D. (2012). Objective Priors for Discrete Parameter Spaces. *Journal of the American Statistical Association*, 107(498):636–648.

Berger, J. O. and Pericchi, L. R. (1996). The Intrinsic Bayes Factor for Model Selection and Prediction. *Journal of the American Statistical Association*, 91(433):109–122.

Berk, R. H. (1966). Limiting Behavior of Posterior Distributions when the Model is Incorrect. *The Annals of Mathematical Statistics*, 37(1):51–58.

Bernardo, J. M. (1979). Reference Posterior Distributions for Bayesian Inference. *Journal of the Royal Statistical Society. Series B (Methodological)*, 41(2):113–147.

Bernardo, J. M. (2005). Reference analysis. In Dey, D. and Rao, C., editors, *Bayesian Thinking*, volume 25 of *Handbook of Statistics*, pages 17 – 90. Elsevier.

Bien, J. and Tibshirani, R. J. (2011). Sparse Estimation of a Covariance Matrix. *Biometrika*, 98(4):807–820.

Bilmes, J. A. (2004). Graphical Models and Automatic Speech Recognition. In Johnson, M., Khudanpur, S. P., Ostendorf, M., and Rosenfeld, R., editors, *Mathematical Foundations of Speech and Language Processing*, pages 191–245, New York, NY. Springer New York.

Carlin, B. P., Gelfand, A. E., and Smith, A. F. M. (1992). Hierarchical Bayesian Analysis of Changepoint Problems. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 41(2):389–405.

Carvalho, C. M. and Scott, J. G. (2009). Objective Bayesian Model Selection in Gaussian Graphical Models. *Biometrika*, 96(3):497–512.

Chernoff, H. and Zacks, S. (1964). Estimating the Current Mean of a Normal Distribution which is Subjected to Changes in Time. *The Annals of Mathematical Statistics*, 35(3):999–1018.

Chib, S. (1998). Estimation and Comparison of Multiple Change-point Models. *Journal of Econometrics*, 86(2):221–241.

Chipman, H., George, E. I., Gramacy, R. B., and McCulloch, R. (2013). Bayesian Treed Response Surface Models. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 3(4):298–305.

Chipman, H. A., George, E. I., and McCulloch, R. E. (1998). Bayesian CART Model Search. *Journal of the American Statistical Association*, 93(443):935–948.

Chipman, H. A., George, E. I., and McCulloch, R. E. (2002). Bayesian Treed Models. *Machine Learning*, 48(1):299–320.

Chipman, H. A., George, E. I., and McCulloch, R. E. (2010). BART: BAYESIAN ADDITIVE REGRESSION TREES. *The Annals of Applied Statistics*, 4(1):266–298.

Consonni, G., Fouskakis, D., Liseo, B., and Ntzoufras, I. (2018). Prior Distributions for Objective Bayesian Analysis. *Bayesian Analysis*, 13(2):627–679.

Consonni, G., La Rocca, L., and Peluso, S. (2017). Objective Bayes Covariate-Adjusted Sparse Graphical Model Selection. *Scandinavian Journal of Statistics*, 44(3):741–764.

Cover, T. M. and Thomas, J. A. (2006). *Elements of Information Theory (Wiley Series in Telecommunications and Signal Processing)*. Wiley-Interscience, New York, NY, USA.

Cowell, R. G., Dawid, A. P., Lauritzen, S. L., and Spiegelhalter, D. J. (2007). *Probabilistic Networks and Expert Systems: Exact Computational Methods for Bayesian Networks*. Springer Publishing Company, Incorporated, 1st edition.

Datta, G. S. (1996). On Priors Providing Frequentist Validity of Bayesian Inference for Multiple Parametric Functions. *Biometrika*, 83(2):287–298.

Datta, G. S. and Mukerjee, R. (2004). *Probability Matching Priors: Higher Order Asymptotics*, volume 178. Springer-Verlag New York.

Datta, G. S. and Sweeting, T. J. (2005). Probability Matching Priors. In Dey, D. K. and Rao, C., editors, *Handbook of Statistics:Bayesian Thinking, Modeling and Computation*, volume 25, pages 91–114. Elsevier B.V.

Dawid, A. P. and Lauritzen, S. L. (1993). Hyper Markov Laws in the Statistical Analysis of Decomposable Graphical Models. *The Annals of Statistics*, 21(3):1272–1317.

de Finetti, B. (1937). La Prévision: Ses Lois Logiques, Ses Sources Subjectives. *Annales de l'Institut Henri Poincaré*, 17:1–68.

de Santis, F. and Spezzaferri, F. (1999). Methods for Default and Robust Bayesian Model Comparison: The Fractional Bayes Factor Approach. *International Statistical Review / Revue Internationale de Statistique*, 67(3):267–286.

Denison, D. G. T., Mallick, B. K., and Smith, A. F. M. (1998). A Bayesian CART Algorithm. *Biometrika*, 85(2):363–377.

Dobra, A., Hans, C., Jones, B., Nevins, J. R., Yao, G., and West, M. (2004). Sparse Graphical Models for Exploring Gene Expression Data. *Journal of Multivariate Analysis*, 90(1):196–212. Special Issue on Multivariate Methods in Genomic Data Analysis.

Dobra, A., Lenkoski, A., and Rodriguez, A. (2011). Bayesian Inference for General Gaussian Graphical Models With Application to Multivariate Lattice Data. *Journal of the American Statistical Association*, 106(496):1418–1433.

Erdman, C. and Emerson, J. (2007). bcp: An R Package for Performing a Bayesian Analysis of Change Point Problems. *Journal of Statistical Software*, 23(3):1–13.

Fearnhead, P. and Liu, Z. (2007). On-line Inference for Multiple Changepoint Problems. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(4):589–605.

Friedman, J., Hastie, T., and Tibshirani, R. (2008). Sparse Inverse Covariance Estimation with the Graphical Lasso. *Biostatistics*, 9(3):432–441.

Friedman, N., Linial, M., Nachman, I., and Pe'er, D. (2000). Using Bayesian Networks to Analyze Expression Data. *Journal of Computational Biology*, 7(3-4):601–620. PMID: 11108481.

Geiger, D. and Heckerman, D. (2002). Parameter Priors for Directed Acyclic Graphical Models and the Characterization of Several Probability Distributions. *The Annals of Statistics*, 30(5):1412–1440.

Geisser, S. and Cornfield, J. (1963). Posterior Distributions for Multivariate Normal Parameters. *Journal of the Royal Statistical Society. Series B (Methodological)*, 25(2):368–376.

George, E. I. (2010). Dilution Priors: Compensating for Model Space Redundancy. In Berger, J. O., Cai, T. T., and Johnstone, I. M., editors, *Borrowing Strength: Theory Powering Applications – A Festschrift for Lawrence D. Brown*, volume 6 of *Collections*, pages 158–165. Institute of Mathematical Statistics, Beachwood, Ohio, USA.

Ghosh, M. (2011). Objective Priors: An Introduction for Frequentists. *Statistical Science*, 26(2):187–202.

Girón, F. J., Moreno, E., and Casella, G. (2007). Objective Bayesian Analysis of Multiple Changepoints for Linear Models (with discussion). In Bernardo, J.M., Bayarri, M.J., Berger, J.O., Dawid, A.P., Heckerman, D., Smith, A.F.M., and West, M., editors, *Bayesian Statistics 8*, pages 227–252. Oxford University Press, London.

Giudici, P. and Green, P. (1999). Decomposable Graphical Gaussian Model Determination. *Biometrika*, 86(4):785–801.

Giudici, P. and Spelta, A. (2016). Graphical Network Models for International Financial Flows. *Journal of Business & Economic Statistics*, 34(1):128–138.

Goldstein, M. (2006). Subjective Bayesian Analysis: Principles and Practice. *Bayesian Analysis*, 1(3):403–420.

Gramacy, R. B. and Lee, H. K. H. (2008). Bayesian Treed Gaussian Process Models With an Application to Computer Modeling. *Journal of the American Statistical Association*, 103(483):1119–1130.

Grazian, C., Villa, C., and Liseo, B. (2018). On a Loss-based Prior for the Number of Components in Mixture Models. *ArXiv e-prints*.

Green, P. J. (1995). Reversible Jump Markov Chain Monte Carlo Computation and Bayesian Model Determination. *Biometrika*, 82(4):711–732.

Griffiths, D. A. (1973). Maximum Likelihood Estimation for the Beta-Binomial Distribution and an Application to the Household Distribution of the Total Number of Cases of a Disease. *Biometrics*, 29(4):637–648.

Hannart, A. and Naveau, P. (2009). Bayesian Multiple Change Points and Segmentation: Application to Homogenization of Climatic Series. *Water Resources Research*, 45(10):W10444.

Harlé, F., Chatelain, F., Gouy-Pailler, C., and Achard, S. (2016). Bayesian Model for Multiple Change-Points Detection in Multivariate Time Series. *IEEE Transactions on Signal Processing*, 64(16):4351–4362.

Heard, N. A. and Turcotte, M. J. M. (2017). Adaptive Sequential Monte Carlo for Multiple Changepoint Analysis. *Journal of Computational and Graphical Statistics*, 26(2):414–423.

Henderson, R. and Matthews, J. N. S. (1993). An Investigation of Changepoints in the Annual Number of Cases of Haemolytic Uraemic Syndrome. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 42(3):461–471.

Hess, K. R., Anderson, K., Symmans, W. F., Valero, V., Ibrahim, N., Mejia, J. A., Booser, D., Theriault, R. L., Buzdar, A. U., Dempsey, P. J., Rouzier, R., Sneige, N., Ross, J. S., Vidaurre, T., Gómez, H. L., Hortobagyi, G. N., and Pusztai, L. (2006). Pharmacogenomic Predictor of Sensitivity to Preoperative Chemotherapy With Paclitaxel and Fluorouracil, Doxorubicin, and Cyclophosphamide in Breast Cancer. *Journal of Clinical Oncology*, 24(26):4236–4244.

Hinoveanu, L. C., Leisen, F., and Villa, C. (2018). A Loss-based Prior for Gaussian Graphical Models. *ArXiv e-prints*.

Hinoveanu, L. C., Leisen, F., and Villa, C. (2019). Bayesian Loss-based Approach to Change Point Analysis. *Computational Statistics & Data Analysis*, 129:61–78.

Jeffreys, H. (1961). *Theory of Probability*. Clarendon Press Oxford, 3rd edition.

Jones, B., Carvalho, C., Dobra, A., Hans, C., Carter, C., and West, M. (2005). Experiments in Stochastic Computation for High-Dimensional Graphical Models. *Statistical Science*, 20(4):388–400.

Kass, R. E. and Raftery, A. E. (1995). Bayes Factors. *Journal of the American Statistical Association*, 90(430):773–795.

Kass, R. E. and Wasserman, L. (1996). The Selection of Prior Distributions by Formal Rules. *Journal of the American Statistical Association*, 91(435):1343–1370.

Ko, S. I. M., Chong, T. T. L., and Ghosh, P. (2015). Dirichlet Process Hidden Markov Multiple Change-point Model. *Bayesian Analysis*, 10(2):275–296.

Koop, G. and Potter, S. M. (2009). Prior Elicitation in Multiple Change-point Models. *International Economic Review*, 50(3):751–772.

Kullback, S. and Leibler, R. A. (1951). On Information and Sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86.

Kundu, S., Mallick, B. K., and Baladandayuthapani, V. (2019). Efficient Bayesian Regularization for Graphical Model Selection. *Bayesian Analysis*, 14(2):449–476.

Lai, T. L. and Xing, H. (2011). A Simple Bayesian Approach to Multiple Change-Points. *Statistica Sinica*, 21(2):539–569.

Lauritzen, S. L. (1996). *Graphical Models*. Clarendon Press, Oxford.

Lindley, D. (1972). *Bayesian Statistics: A Review*. Society for Industrial and Applied Mathematics.

Lindley, D. V. (1956). On a Measure of the Information Provided by an Experiment. *The Annals of Mathematical Statistics*, 27(4):986–1005.

Linero, A. R. (2017). A Review of Tree-based Bayesian Methods. *Communications for Statistical Applications and Methods*, 24(6):543–559.

Linero, A. R. (2018). Bayesian Regression Trees for High-Dimensional Prediction and Variable Selection. *Journal of the American Statistical Association*, 113(522):626–636.

Liu, S., Yamada, M., Collier, N., and Sugiyama, M. (2013). Change-point Detection in Time-series Data by Relative Density-ratio Estimation. *Neural Networks*, 43:72 – 83.

Loschi, R.H. and Cruz, F.R.B. (2005). Extension to the Product Partition Model: Computing the Probability of a Change. *Computational Statistics & Data Analysis*, 48(2):255–268.

Mair, P. (2015). *APR: Applied Psychometrics With R*. R package version 0.0-6/r205.

Martínez, A. F. and Mena, R. H. (2014). On a Nonparametric Change Point Detection Model in Markovian Regimes. *Bayesian Analysis*, 9(4):823–858.

McNally, R. J., Robinaugh, D. J., Wu, G. W. Y., Wang, L., Deserno, M. K., and Borsboom, D. (2015). Mental Disorders as Causal Systems: A Network Approach to Posttraumatic Stress Disorder. *Clinical Psychological Science*, 3(6):836–849.

Meinshausen, N. and Bühlmann, P. (2006). High-Dimensional Graphs and Variable Selection with the Lasso. *The Annals of Statistics*, 34(3):1436–1462.

Merhav, N. and Feder, M. (1998). Universal Prediction. *IEEE Transactions on Information Theory*, 44(6):2124–2147.

Mira, A. and Petrone, S. (1996). Bayesian Hierarchical Nonparametric Inference for Change-Point Problems. In Bernardo, J.M.., Berger, J.O.., Dawid, A.P.., and Smith, A.F.M.., editors, *Bayesian Statistics 5*, pages 693–703. Oxford University Press, London.

Mohammadi, A., Abegaz, F., van den Heuvel, E., and Wit, E. C. (2017). Bayesian Modelling of Dupuytren Disease by using Gaussian Copula Graphical Models. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 66(3):629–645.

Mohammadi, A. and Wit, E. C. (2015). Bayesian Structure Learning in Sparse Gaussian Graphical Models. *Bayesian Analysis*, 10(1):109–138.

Mohammadi, A. and Wit, E. C. (2019). BDgraph: An R Package for Bayesian Structure Learning in Graphical Models. *Journal of Statistical Software*, 89(3):1–30.

Moreno, E., Casella, G., and Garcia-Ferrer, A. (2005). An Objective Bayesian Analysis of the Change Point Problem. *Stochastic Environmental Research and Risk Assessment*, 19(3):191–204.

Muliere, P. and Scarsini, M. (1985). Change-point Problems: A Bayesian Nonparametric Approach. *Aplikace matematiky*, 30(6):397–402.

Müller, P., Quintana, F., and Rosner, G. L. (2011). A Product Partition Model With Regression on Covariates. *Journal of Computational and Graphical Statistics*, 20(1):260–278.

O'Hagan, A. (1995). Fractional Bayes Factors for Model Comparison. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):99–138.

O'Hagan, A. (1997). Properties of Intrinsic and Fractional Bayes Factors. *Test*, 6(1):101–118.

Petrone, S. and Raftery, A. E. (1997). A Note on the Dirichlet Process Prior in Bayesian Nonparametric Inference with Partial Exchangeability. *Statistics & Probability Letters*, 36(1):69–83.

Raftery, A. E. and Akman, V. E. (1986). Bayesian Analysis of a Poisson Process with a Change-point. *Biometrika*, 73(1):85–89.

Ramsey, F. P. (1926). Truth and Probability. In Braithwaite, R. B., editor, *The Foundations of Mathematics and other Logical Essays*, chapter 7, pages 156–198. McMaster University Archive for the History of Economic Thought.

Rissanen, J. (1983). A Universal Prior for Integers and Estimation by Minimum Description Length. *The Annals of Statistics*, 11(2):416–431.

Robert, C. P. (2007). *The Bayesian Choice:From Decision-Theoretic Foundations to Computational Implementation*. Springer Science+Business Media, LLC, 233 Spring Street, New York, NY 10013, USA, 2nd edition.

Roverato, A. (2017). *Graphical Models for Categorical Data*. SemStat Elements. Cambridge University Press.

Roverato, A. and Whittaker, J. (1998). The Isserlis Matrix and its Application to Non-decomposable Graphical Gaussian Models. *Biometrika*, 85(3):711–725.

Sachs, K., Perez, O., Pe'er, D., Lauffenburger, D. A., and Nolan, G. P. (2005). Causal Protein-Signaling Networks Derived from Multiparameter Single-Cell Data. *Science*, 308(5721):523–529.

Sadia, F., Boyd, S., and Keith, J. M. (2018). Bayesian Change-point Modeling with Segmented ARMA Model. *PLOS ONE*, 13(12):1–23.

Sandberg, I., Lo, J., Fancourt, C., Principe, J., Katagiri, S., and Haykin, S. (2001). *Nonlinear Dynamical Systems: Feedforward Neural Network Perspectives*. Adaptive and Cognitive Dynamic Systems: Signal Processing, Learning, Communications and Control. Wiley.

Schwaller, L. and Robin, S. (2017). Exact Bayesian Inference for Off-line Change-point Detection in Tree-structured Graphical Models. *Statistics and Computing*, 27(5):1331–1345.

Scott, J. G. and Berger, J. O. (2010). Bayes and Empirical-Bayes Multiplicity Adjustment in the Variable-selection Problem. *The Annals of Statistics*, 38(5):2587–2619.

Scott, J. G. and Carvalho, C. M. (2008). Feature-Inclusion Stochastic Search for Gaussian Graphical Models. *Journal of Computational and Graphical Statistics*, 17(4):790–808.

Scricciolo, C. (1999). Probability Matching Priors: A Review. *Statistical Methods & Applications*, 8(1):83–100.

Shannon, C. E. (1948). A Mathematical Theory of Communication. *Bell System Technical Journal*, 27(3):379–423.

Shaun, B., Tom, F., and Frank, S. (2016). *Algebra, Logic And Combinatorics*. LTCC Advanced Mathematics Series. World Scientific Publishing Company.

Shojaie, A. and Michailidis, G. (2010). Penalized Principal Component Regression on Graphs for Analysis of Subnetworks. In Lafferty, J. D., Williams, C. K. I., Shawe-Taylor, J., Zemel, R. S., and Culotta, A., editors, *Advances in Neural Information Processing Systems 23*, pages 2155–2163. Curran Associates, Inc.

Smith, A. F. M. (1975). A Bayesian Approach to Inference about a Change-point in a Sequence of Random Variables. *Biometrika*, 62(2):407–416.

Spirtes, P., Glymour, C., and Scheines, R. (2000). *Causation, Prediction, and Search*. MIT press, 2nd edition.

Stephens, D. A. (1994). Bayesian Retrospective Multiple-Changepoint Identification. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 43(1):159–178.

Stingo, F. and Marchetti, G. M. (2015). Efficient Local Updates for Undirected Graphical Models. *Statistics and Computing*, 25(1):159–171.

Stingo, F. C., Chen, Y. A., Vannucci, M., Barrier, M., and Mirkes, P. E. (2010). A Bayesian Graphical Modeling Approach to MicroRNA Regulatory Network Inference. *The Annals of Applied Statistics*, 4(4):2024–2048.

Sun, D. and Berger, J. O. (2007). Objective Bayesian Analysis for the Multivariate Normal Model. In Bernardo, J.M., Bayarri, M.J., Berger, J.O., Dawid, A.P., Heckerman, D., Smith, A.F.M., and West, M., editors, *Bayesian Statistics 8*, pages 525–563. Oxford University Press, London.

Tian, G.-L., Ng, K. W., Li, K.-C., and Tan, M. (2009). Non-iterative Sampling-based Bayesian Methods for Identifying Changepoints in the Sequence of Cases of Haemolytic Uraemic Syndrome. *Computational Statistics & Data Analysis*, 53(9):3314–3323.

Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288.

Villa, C. and Lee, J. E. (2019). A Loss-based Prior for Variable Selection in Linear Regression Methods. *Bayesian Analysis*. Forthcoming.

Villa, C. and Walker, S. (2015a). An Objective Bayesian Criterion to Determine Model Prior Probabilities. *Scandinavian Journal of Statistics*, 42(4):947–966.

Villa, C. and Walker, S. G. (2015b). An Objective Approach to Prior Mass Functions for Discrete Parameter Spaces. *Journal of the American Statistical Association*, 110(511):1072–1082.

Wang, T., Ren, Z., Ding, Y., Fang, Z., Sun, Z., MacDonald, M. L., Sweet, R. A., Wang, J., and Chen, W. (2016). FastGGM: An Efficient Algorithm for the Inference of Gaussian Graphical Model in Biological Networks. *PLOS Computational Biology*, 12(2):e1004755.

Welch, B. L. and Peers, H. W. (1963). On Formulae for Confidence Points Based on Integrals of Weighted Likelihoods. *Journal of the Royal Statistical Society. Series B (Methodological)*, 25(2):318–329.

Williams, D. R. (2018). Bayesian Inference for Gaussian Graphical Models: Structure Learning, Explanation, and Prediction. *PsyArXiv e-prints*.

Wu, Y., Tjelmeland, H., and West, M. (2007). Bayesian CART: Prior Specification and Posterior Simulation. *Journal of Computational and Graphical Statistics*, 16(1):44–66.

Yajima, M., Telesca, D., Ji, Y., and Müller, P. (2015). Detecting Differential Patterns of Interaction in Molecular Pathways. *Biostatistics*, 16(2):240–251.

Yang, Z., Hu, J., Shu, Y., Cheng, P., Chen, J., and Moscibroda, T. (2016). Mobility Modeling and Prediction in Bike-Sharing Systems. In *ACM International Conference on Mobile Systems, Applications, and Services (MobiSys)*.

Yao, Y.-C. (1984). Estimation of a Noisy Discrete-Time Step Function: Bayes and Empirical Bayes Approaches. *The Annals of Statistics*, 12(4):1434–1447.

Yu, J. (2001). Chapter 6 - Testing for a Finite Variance in Stock Return Distributions. In Knight, J. and Satchell, S. E., editors, *Return Distributions in Finance (Quantitative Finance)*, pages 143–164. Butterworth-Heinemann, Oxford.

Yuan, M. and Lin, Y. (2007). Model Selection and Estimation in the Gaussian Graphical Model. *Biometrika*, 94(1):19–35.

A. Appendix to the Change Point Chapter

Model prior probabilities to select among models M_0, M_1 and M_2

Here, we show how model prior probabilities can be derived for the relatively simple case of selecting among scenarios with no change points (M_0), one change point (M_1) or two change points (M_2). First, by applying the result in Theorem 2, we derive the KL divergences between any two models. That is:

- the prior probability for model M_0 depends on the following quantities:

$$\begin{aligned} D_{KL}(M_0 \| M_1) =& (n - m_1) \cdot D_{KL}(f_1(\cdot|\tilde{\theta}_1) \| f_2(\cdot|\tilde{\theta}_2)) \\ D_{KL}(M_0 \| M_2) =& (m_2 - m_1) \cdot D_{KL}(f_1(\cdot|\tilde{\theta}_1) \| f_2(\cdot|\tilde{\theta}_2)) \\ &+ (n - m_2) \cdot D_{KL}(f_1(\cdot|\tilde{\theta}_1) \| f_3(\cdot|\tilde{\theta}_3)) \end{aligned}$$

- the prior probability for model M_1 depends on the following quantities:

$$\begin{aligned} D_{KL}(M_1 \| M_2) =& (n - m_2) \cdot D_{KL}(f_2(\cdot|\tilde{\theta}_2) \| f_3(\cdot|\tilde{\theta}_3)) \\ D_{KL}(M_1 \| M_0) =& (n - m_1) \cdot D_{KL}(f_2(\cdot|\tilde{\theta}_2) \| f_1(\cdot|\tilde{\theta}_1)) \end{aligned}$$

- the prior probability for model M_2 depends on the following quantities:

$$\begin{aligned} D_{KL}(M_2\|M_1) =& (n-m_2)\cdot D_{KL}(f_3(\cdot|\tilde{\theta}_3)\|f_2(\cdot|\tilde{\theta}_2)) \\ D_{KL}(M_2\|M_0) =& (m_2-m_1)\cdot D_{KL}(f_2(\cdot|\tilde{\theta}_2)\|f_1(\cdot|\tilde{\theta}_1)) \\ &+ (n-m_2)\cdot D_{KL}(f_3(\cdot|\tilde{\theta}_3)\|f_1(\cdot|\tilde{\theta}_1)) \end{aligned}$$

The next step is to derive the minimum KL divergence computed at each model:

- for model M_0:

$$\begin{aligned} \inf_{\theta_1} D_{KL}(M_0\|M_1) &= \underbrace{\left[\inf_{m_1\neq n}(n-m_1)\right]}_{1}\cdot\left[\inf_{\tilde{\theta}_2} D_{KL}(f_1(\cdot|\tilde{\theta}_1)\|f_2(\cdot|\tilde{\theta}_2))\right] \\ &= \inf_{\tilde{\theta}_2} D_{KL}(f_1(\cdot|\tilde{\theta}_1)\|f_2(\cdot|\tilde{\theta}_2)) \\ \inf_{\theta_2} D_{KL}(M_0\|M_2) &= \underbrace{\left[\inf_{m_1\neq m_2}(m_2-m_1)\right]}_{1}\cdot\left[\inf_{\tilde{\theta}_2} D_{KL}(f_1(\cdot|\tilde{\theta}_1)\|f_2(\cdot|\tilde{\theta}_2))\right] \\ &\quad + \underbrace{\left[\inf_{m_2\neq n}(n-m_2)\right]}_{1}\cdot\left[\inf_{\tilde{\theta}_3} D_{KL}(f_1(\cdot|\tilde{\theta}_1)\|f_3(\cdot|\tilde{\theta}_3))\right] \\ &= \inf_{\tilde{\theta}_2} D_{KL}(f_1(\cdot|\tilde{\theta}_1)\|f_2(\cdot|\tilde{\theta}_2)) + \inf_{\tilde{\theta}_3} D_{KL}(f_1(\cdot|\tilde{\theta}_1)\|f_3(\cdot|\tilde{\theta}_3)) \end{aligned}$$

- for model M_1:

$$\begin{aligned} \inf_{\theta_2} D_{KL}(M_1\|M_2) &= \underbrace{\left[\inf_{m_2\neq n}(n-m_2)\right]}_{1}\cdot\left[\inf_{\tilde{\theta}_3} D_{KL}(f_2(\cdot|\tilde{\theta}_2)\|f_3(\cdot|\tilde{\theta}_3))\right] \\ &= \inf_{\tilde{\theta}_3} D_{KL}(f_2(\cdot|\tilde{\theta}_2)\|f_3(\cdot|\tilde{\theta}_3)) \\ \inf_{\theta_0=\tilde{\theta}_1} D_{KL}(M_1\|M_0) &= (n-m_1)\cdot\inf_{\tilde{\theta}_1} D_{KL}(f_2(\cdot|\tilde{\theta}_2)\|f_1(\cdot|\tilde{\theta}_1)) \end{aligned}$$

- for model M_2:

$$\begin{aligned}\inf_{\theta_1} D_{KL}(M_2\|M_1) =& (n-m_2)\cdot \inf_{\tilde{\theta}_2} D_{KL}(f_3(\cdot|\tilde{\theta}_3)\|f_2(\cdot|\tilde{\theta}_2))\\ \inf_{\theta_0=\tilde{\theta}_1} D_{KL}(M_2\|M_0) =& (m_2-m_1)\cdot \inf_{\tilde{\theta}_1} D_{KL}(f_2(\cdot|\tilde{\theta}_2)\|f_1(\cdot|\tilde{\theta}_1))\\ &+ (n-m_2)\cdot \inf_{\tilde{\theta}_1} D_{KL}(f_3(\cdot|\tilde{\theta}_3)\|f_1(\cdot|\tilde{\theta}_1))\end{aligned}$$

Therefore, the model prior probabilities can be computed through equation (3.9), so that:

- the model prior probability $\Pr(M_0)$ is proportional to the exponential of the minimum between:

$$\left\{\mathbb{E}_{\pi_0}\left[\inf_{\tilde{\theta}_2} D_{KL}(f_1(\cdot|\tilde{\theta}_1)\|f_2(\cdot|\tilde{\theta}_2))\right], \mathbb{E}_{\pi_0}\left[\inf_{\tilde{\theta}_2} D_{KL}(f_1(\cdot|\tilde{\theta}_1)\|f_2(\cdot|\tilde{\theta}_2)) + \inf_{\tilde{\theta}_3} D_{KL}(f_1(\cdot|\tilde{\theta}_1)\|f_3(\cdot|\tilde{\theta}_3))\right]\right\}$$

- the model prior probability $\Pr(M_1)$ is proportional to the exponential of the minimum between:

$$\left\{\mathbb{E}_{\pi_1}\left[\inf_{\tilde{\theta}_3} D_{KL}(f_2(\cdot|\tilde{\theta}_2)\|f_3(\cdot|\tilde{\theta}_3))\right], \mathbb{E}_{\pi_1}\left[(n-m_1)\cdot \inf_{\tilde{\theta}_1} D_{KL}(f_2(\cdot|\tilde{\theta}_2)\|f_1(\cdot|\tilde{\theta}_1))\right]\right\}$$

- the model prior probability $\Pr(M_2)$ is proportional to the exponential of the minimum between:

$$\left\{ \mathbb{E}_{\pi_2}\left[(n-m_2) \cdot \inf_{\tilde{\theta}_2} D_{KL}(f_3(\cdot|\tilde{\theta}_3) \| f_2(\cdot|\tilde{\theta}_2)) \right], \right.$$
$$\mathbb{E}_{\pi_2}\left[(m_2-m_1) \cdot \inf_{\tilde{\theta}_1} D_{KL}(f_2(\cdot|\tilde{\theta}_2) \| f_1(\cdot|\tilde{\theta}_1)) + (n-m_2) \right.$$
$$\left. \left. \cdot \inf_{\tilde{\theta}_1} D_{KL}(f_3(\cdot|\tilde{\theta}_3) \| f_1(\cdot|\tilde{\theta}_1)) \right] \right\}$$

Proofs

Proof of Theorem 1

We distinguish two cases: $S = +1$ and $S = -1$. When $S = +1$, equivalent to $m_j < m'_j$:

$$
\begin{aligned}
&D_{KL}(f(\mathbf{x}^{(n)}|\boldsymbol{m},\tilde{\boldsymbol{\theta}})\|f(\mathbf{x}^{(n)}|\boldsymbol{m}',\tilde{\boldsymbol{\theta}})) = \int f(\mathbf{x}^{(n)}|\boldsymbol{m},\tilde{\boldsymbol{\theta}}) \cdot \ln\left(\frac{f(\mathbf{x}^{(n)}|\boldsymbol{m},\tilde{\boldsymbol{\theta}})}{f(\mathbf{x}^{(n)}|\boldsymbol{m}',\tilde{\boldsymbol{\theta}})}\right) \mathrm{d}\mathbf{x}^{(n)} \\
=&\int f(\mathbf{x}^{(n)}|\boldsymbol{m},\tilde{\boldsymbol{\theta}}) \cdot \left[\sum_{i=m_j+1}^{m'_j} \ln\left(\frac{f_{j+1}(x_i|\tilde{\theta}_{j+1})}{f_j(x_i|\tilde{\theta}_j)}\right)\right] \mathrm{d}\mathbf{x}^{(n)} \\
=&\sum_{i=m_j+1}^{m'_j} \int f(\mathbf{x}^{(n)}|\boldsymbol{m},\tilde{\boldsymbol{\theta}}) \cdot \left[\ln\left(\frac{f_{j+1}(x_i|\tilde{\theta}_{j+1})}{f_j(x_i|\tilde{\theta}_j)}\right)\right] \mathrm{d}\mathbf{x}^{(n)} \\
=&\sum_{i=m_j+1}^{m'_j} \left\{1^{n-1} \cdot \int f_{j+1}(x_i|\tilde{\theta}_{j+1}) \cdot \left[\ln\left(\frac{f_{j+1}(x_i|\tilde{\theta}_{j+1})}{f_j(x_i|\tilde{\theta}_j)}\right)\right] \mathrm{d}x_i\right\} \\
=&\sum_{i=m_j+1}^{m'_j} D_{KL}(f_{j+1}(x_i|\tilde{\theta}_{j+1})\|f_j(x_i|\tilde{\theta}_j)) \\
=&(m'_j - m_j) \cdot D_{KL}(f_{j+1}(\cdot|\tilde{\theta}_{j+1})\|f_j(\cdot|\tilde{\theta}_j)) \\
=&(m'_j - m_j) \cdot d_j^{+1}(\tilde{\boldsymbol{\theta}}).
\end{aligned} \tag{A.1}
$$

When $S = -1$, equivalent to $m_j > m'_j$, in a similar fashion, we get

$$
D_{KL}(f(\mathbf{x}^{(n)}|\boldsymbol{m},\tilde{\boldsymbol{\theta}})\|f(\mathbf{x}^{(n)}|\boldsymbol{m}',\tilde{\boldsymbol{\theta}})) = (m_j - m'_j) \cdot d_j^{-1}(\tilde{\boldsymbol{\theta}}) \tag{A.2}
$$

From equations (A.1) and (A.2), we get the result in Theorem 1.

Proof of Theorem 2

We recall that the model parameter θ_i is the vector $(m_1, m_2, \ldots, m_i, \tilde{\theta}_1, \tilde{\theta}_2, \ldots, \tilde{\theta}_{i+1})$, where $i = 0, 1, \ldots, k$. Here, $\tilde{\theta}_1, \tilde{\theta}_2, \ldots, \tilde{\theta}_{i+1}$ represent the parameters of the underlying sampling

distributions considered under model M_i and $m_1, m_2, \ldots, m_i$ are the respective i change point locations. In this setting,

$$f(\mathbf{x}^{(n)}|\theta_i) = \prod_{r=1}^{m_1} f_1(x_r|\tilde{\theta}_1) \prod_{t=1}^{i-1} \prod_{r=m_t+1}^{m_{t+1}} f_{t+1}(x_r|\tilde{\theta}_{t+1}) \prod_{r=m_i+1}^{n} f_{i+1}(x_r|\tilde{\theta}_{i+1}) \tag{A.3}$$

We proceed to the computation of $D_{KL}(M_i\|M_j)$, that is the KL divergence introduced in Section 3.2. Similarly to the proof of Theorem 1, we obtain the following result.

$$\begin{aligned} D_{KL}(M_i\|M_j) = &\sum_{r=m_{i+1}+1}^{m_{i+2}} \int f(\mathbf{x}^{(n)}|\theta_i) \ln\left(\frac{f_{i+1}(x_r|\tilde{\theta}_{i+1})}{f_{i+2}(x_r|\tilde{\theta}_{i+2})}\right) \mathrm{d}\mathbf{x}^{(n)} \\ &+ \sum_{r=m_{i+2}+1}^{m_{i+3}} \int f(\mathbf{x}^{(n)}|\theta_i) \ln\left(\frac{f_{i+1}(x_r|\tilde{\theta}_{i+1})}{f_{i+3}(x_r|\tilde{\theta}_{i+3})}\right) \mathrm{d}\mathbf{x}^{(n)} + \\ &\ldots + \sum_{r=m_j+1}^{n} \int f(\mathbf{x}^{(n)}|\theta_i) \ln\left(\frac{f_{i+1}(x_r|\tilde{\theta}_{i+1})}{f_{j+1}(x_r|\tilde{\theta}_{j+1})}\right) \mathrm{d}\mathbf{x}^{(n)}. \end{aligned}$$

Given equation (A.3), if we integrate out the variables not involved in the logarithms, we obtain

$$\begin{aligned} D_{KL}(M_i\|M_j) = &(m_{i+2} - m_{i+1}) \cdot D_{KL}(f_{i+1}(\cdot|\tilde{\theta}_{i+1})\|f_{i+2}(\cdot|\tilde{\theta}_{i+2})) \\ &+ (m_{i+3} - m_{i+2}) \cdot D_{KL}(f_{i+1}(\cdot|\tilde{\theta}_{i+1})\|f_{i+3}(\cdot|\tilde{\theta}_{i+3})) + \\ &\ldots + (n - m_j) \cdot D_{KL}(f_{i+1}(\cdot|\tilde{\theta}_{i+1})\|f_{j+1}(\cdot|\tilde{\theta}_{j+1})). \end{aligned}$$

In a similar fashion, it can be shown that

$$\begin{aligned} D_{KL}(M_j\|M_i) = &(m_{i+2} - m_{i+1}) \cdot D_{KL}(f_{i+2}(\cdot|\tilde{\theta}_{i+2})\|f_{i+1}(\cdot|\tilde{\theta}_{i+1})) \\ &+ (m_{i+3} - m_{i+2}) \cdot D_{KL}(f_{i+3}(\cdot|\tilde{\theta}_{i+3})\|f_{i+1}(\cdot|\tilde{\theta}_{i+1})) + \\ &\ldots + (n - m_j) \cdot D_{KL}(f_{j+1}(\cdot|\tilde{\theta}_{j+1})\|f_{i+1}(\cdot|\tilde{\theta}_{i+1})) \end{aligned}$$

B. Appendix to the Gaussian Graphical Models Chapter

FINCS algorithm

The FINCS algorithm is schematically outlined below.

Given the data and some parameters do the following steps:

Step 1 Initialize a graph based on the triangular regression done on the data. In this context, triangular regression simply means that we take each column of the data matrix and we regress it, through the ordinary least squares method, on the submatrix formed from the remaining columns to the respective column's right. As such, the number of columns will become smaller and smaller as we progress towards the last column. For each considered column starting from the leftmost one, we take each estimated regression coefficient and we compare its absolute value to three times its standard error. If the absolute value is at least as big as three times the standard error, then the respective regression coefficient is different than 0 and the corresponding edge is present in the underlying graph, otherwise we consider the respective covariate as not influencing the response thus the analogous edge does not appear in the graph.

Step 2 Loop over the iterations in a serial manner:

1 At a certain number of iterations do a global move through a randomized median triangulation pair. Starting from a random median graph, we add or delete an edge such that decomposability is maintained and the log score is improved

2 At a certain number of iterations we resample one of the previous saved local graphs

3 Do a local move by deleting or adding an edge that maintains decomposability. When an edge is added, it is done in proportion to the estimated posterior probability of inclusion $\hat{q}_{ij}$ for edge (i, j), whereas when there is a deletion, the edge is affected in inverse proportion to the estimated inclusion probabilities

4 Save the local graph in a finite resampling list and remove those graphs that do not improve the log score.

According to Scott and Carvalho (2008), a *randomized median triangulation pair* represents a pair of decomposable graphs chosen in a certain way from the median graph G_N which will often be non-decomposable. One of the pair members will be the minimal decomposable supergraph $G^+ \supset G_N$, whilst the other will be the maximal decomposable subgraph $G^- \subset G_N$. Based on the posterior probabilities, we choose one of G^+ or G^- as our current generated graph at the respective iteration step. This randomised median triangulation pair allows the exploration of new regions in the decomposable graph space.

Pearson correlation matrices

Variables \ Variables	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	1.00	0.23	-0.36	0.20	-0.08	-0.37	-0.51	-0.16	0.43	0.42	-0.12	0.02	-0.24	-0.08	0.09
2	0.23	1.00	0.00	0.32	0.03	0.13	-0.10	-0.03	0.10	0.02	0.12	0.15	0.06	-0.05	-0.36
3	-0.36	0.00	1.00	-0.19	0.11	0.51	0.37	0.58	-0.22	-0.24	-0.01	-0.05	0.20	0.10	0.06
4	0.20	0.32	-0.19	1.00	-0.01	0.23	-0.29	-0.10	0.40	0.18	0.03	0.23	0.10	-0.09	-0.04
5	-0.08	0.03	0.11	-0.01	1.00	0.15	-0.01	-0.08	-0.43	-0.11	0.14	-0.09	-0.00	0.20	-0.21
6	-0.37	0.13	0.51	0.23	0.15	1.00	0.24	0.26	-0.22	-0.22	0.11	-0.01	0.21	0.10	-0.03
7	-0.51	-0.10	0.37	-0.29	-0.01	0.24	1.00	0.22	-0.52	-0.82	0.05	0.07	0.14	0.19	-0.31
8	-0.16	-0.03	0.58	-0.10	-0.08	0.26	0.22	1.00	0.14	-0.13	-0.04	0.14	0.17	0.13	-0.04
9	0.43	0.10	-0.22	0.40	-0.43	-0.22	-0.52	0.14	1.00	0.56	-0.17	-0.04	0.14	-0.08	0.27
10	0.42	0.02	-0.24	0.18	-0.11	-0.22	-0.82	-0.13	0.56	1.00	-0.23	-0.03	0.05	-0.13	0.30
11	-0.12	0.12	-0.01	0.03	0.14	0.11	0.05	-0.04	-0.17	-0.23	1.00	0.22	-0.12	-0.35	0.06
12	0.02	0.15	-0.05	0.23	-0.09	-0.01	0.07	0.14	-0.04	-0.03	0.22	1.00	0.08	-0.08	-0.21
13	-0.24	0.06	0.20	0.10	-0.00	0.21	0.14	0.17	0.14	0.05	-0.12	0.08	1.00	-0.01	0.12
14	-0.08	-0.05	0.10	-0.09	0.20	0.10	0.19	0.13	-0.08	-0.13	-0.35	-0.08	-0.01	1.00	-0.17
15	0.09	-0.36	0.06	-0.04	-0.21	-0.03	-0.31	-0.04	0.27	0.30	0.06	-0.21	0.12	-0.17	1.00

Table B.1 Pearson correlation matrix computed for the sample used in Table 4.1.

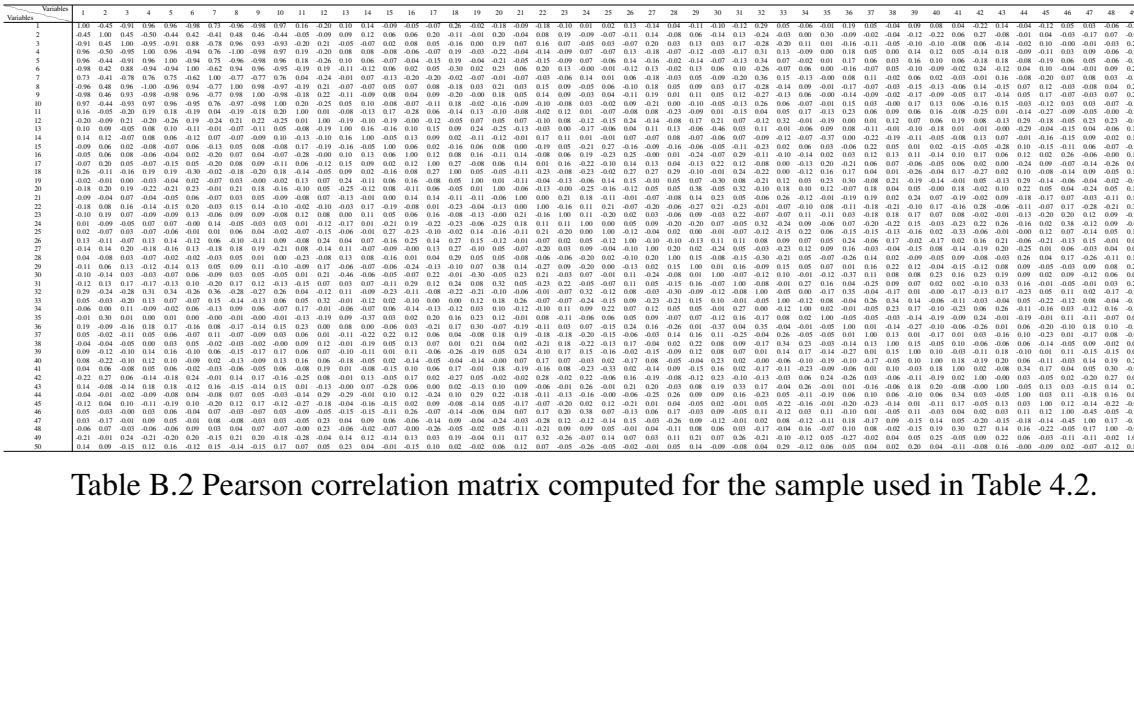


Table B.2 Pearson correlation matrix computed for the sample used in Table 4.2.