Variational Bayesian Learning and its Applications

by

Hui Zhao

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Doctor of Philosophy
in
Statistics

Waterloo, Ontario, Canada, 2013

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

This dissertation is devoted to studying a fast and analytic approximation method, called the *variational Bayesian* (*VB*) method, and aims to give insight into its general applicability and usefulness, and explore its applications to various real-world problems. This work has three main foci:

1. The general applicability and properties
2. Diagnostics for VB approximations
3. Variational applications

General applicability and properties

Generally, the variational inference has been developed in the context of the exponential family, for example, see Beal (2003) and Wainwright and Jordan (2008). Their solutions are open to further development. First, they usually consider the cases in the context of the conjugate exponential family. Second, the variational inferences are developed only with respect to natural parameters, which are often not the parameters of immediate interest. Moreover, the full factorization, which assumes all terms to be independent of one another, is the most commonly used scheme in the most of the variational applications.

We show that VB inferences can be extended to a more general situation. We propose a special parameterization for a parametric family, and also propose a factorization scheme with a more general dependency structure than is traditional in VB. Based on these new frameworks, we develop a variational formalism, in which VB has a fast implementation, and not be limited to the conjugate exponential setting. We also investigate its local convergence property, the effects of choosing improper priors, and the effects of different choices of factorization scheme.

Diagnostics for VB approximations

The essence of the VB method relies on making simplifying assumptions about the posterior dependence of a problem. By definition, the general posterior dependence structure is distorted. In addition, in the various applications, we observe that the posterior variances are often underestimated. MCMC methods are extensively used in the posterior simulations and approximations. We may consider using the posterior samples as benchmarks to calibrate the VB approximations from all aspects. In general, although MCMC methods make possible the computation of any posterior quantities, various practical issues are associated with their implementation. We aim to develop diagnostics test to assess VB approximations, and these methods are expected to be quick and easy to use, and to require no sophisticated tuning expertise.

We propose three methods to compute the actual posterior covariance matrix by only using the knowledge obtained from VB approximations: 1) To look at the joint posterior distribution and attempt to find an optimal affine transformation that links the VB and true posteriors; 2) Based on a marginal posterior density approximation proposed by Tierney et al. (1989), to work in specific low dimensional directions to estimate true posterior variances and correlations; 3) Based on a stepwise conditional approach, to construct and solve a set of system of equations that lead to estimates of the true posterior variances and correlations.

A key computation in the above methods is to calculate a univariate marginal or conditional variance. We propose a novel way to compute these quantities, called the VB Adjusted Independent Metropolis-Hastings (VBAIMH) method. It uses an independent Metropolis-Hastings (IMH) algorithm with proposal distributions configured by VB approximations. The variance of the target distribution is obtained by monitoring the acceptance rate of the converged chain.

Variational applications

One major question associated with the VB method is how well the approximations can work. We particularly study the mean structure approximations, and show how it is possible using VB approximations to approach model selection tasks such as determining the dimensionality of a model, or variable selection.

We also consider the variational application in Bayesian nonparametric modeling. The Dirichlet process (DP), introduced by Ferguson (1973), provides a means to specify a probability measure $P(dF)$ over the space of all (discrete) probability measures, and thus greater flexibility for modelling and inference. Variational Bayesian inference has recently been applied to DP-based problems, for example in Blei and Jordan (2006), who strictly speaking, use a mean-field method, rather than provide a full variational solution. Moreover, only cases, in which the base distribution and the likelihood function are in conjugate exponential families are considered.

We consider a non-conjugate setting for DP and present a full variational Bayesian solution, where the optimization is in terms of both the distributional family and the parameters of the approximating distribution. Our solution uses a truncated stick-breaking representation of the DP. We propose an empirical method to determine the number of distinct components in a finite dimensional DP. The posterior predictive distribution for DP is also often not available in a closed form. We show how to use the variational "tricks" to approximate this quantity.

As a concrete application study, we work through the VB method on regime-switching lognormal models and present solutions to quantify both the uncertainty in the parameters and model specification. Through a series numerical comparison studies with likelihood based methods and MCMC methods on the simulated and real data sets, we show that the VB method can recover exactly the model structure, gives the reasonable point estimates, and is very computationally efficient.

Acknowledgements

I would like to express my gratitude to my supervisor, Professor Paul Marriott for his supports in many aspects. I would like to express my gratitude to my thesis committee: Professor Mary Hardy, Professor Pengfei Li, Professor Paul McNicholas, and Professor Dinghai Xu for their advice and time.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

While, in principle, the calculation of the posterior distribution is mathematically straightforward, in practice the computation of many of its features, such as posterior densities, normalizing constants and posterior moments, is a major challenge in Bayesian analysis. Such computations typically involve high dimensional integrals which often have no analytical or tractable forms. The *variational Bayes (VB)* method was developed to generate tractable approximations to these quantities. This method provides analytic approximations to the posterior distribution by minimizing the Kullback-Leibler (KL) divergence from the approximations to the actual posterior, and has been demonstrated to be computationally very fast. This thesis is devoted to studying the VB method. Apart from VB there exist two commonly used approximation techniques to deal with these calculations: the Laplace approximation and sampling-based methods such as MCMC. However, as we shall see, all these approximation methods have their own strengths and limitations.

This work has three main foci, which we describe briefly here. First, currently, VB is mainly used in the context of the conjugate exponential family. The dissertation proposes a new class of parametric family which can exhibit a particular parameterization. This class is shown to have a fast VB implementation. We also propose a factorization scheme

with a more general dependency structure than is traditional in VB. Based on this new framework, variational inference can be extended to more general situations, and not be limited to the conjugate exponential setting. The thesis also aims to give insight into its general applicability and explore its theoretical properties.

Second, the essence of the VB method relies on making simplifying assumptions about the posterior dependence of a problem, trading off accuracy of the posterior dependence approximation with computational speed. Hence by definition, the general posterior dependence structure is distorted. The dissertation investigates how well the approximation represent the actual posterior distribution, especially their covariance structures. In particular, three methods are proposed to compute the actual posterior covariance matrix by only using the knowledge obtained from VB approximations and by using new algorithms whose computational speeds are of the order of the original VB algorithm.

Third, the dissertation is interested in various applications of VB in real-world examples. In particular, it investigates VB's mean structure approximations, and illustrate how it is possible, using VB approximations, to approach model selection tasks. For example the problem of finding the number of components in mixture models seems well suited for the VB approach. Moreover, the dissertation looks at the variational inference for the Bayesian nonparametric modelling.

The present chapter aims to provide a detailed literature review for VB. We first briefly review the Laplace approximation and MCMC methods and examine their strengths and limitations in applications in Section 1.1. We then elaborate the current development of VB and its various applications in different disciplines and area in Section 1.2 and 1.3. The main contributions of the dissertation are presented in Section 1.4 and 1.5. A technical introduction of the method is given in Chapter 2.

1.1 Posterior quantities

Bayesian analysis often requires the computation of posterior quantities. Typical examples include posterior distributions, the marginal likelihood, posterior predictive distributions, posterior moments, Bayesian credible intervals, Bayesian p-values, amongst others. In general, for sophisticated models these quantities are intractable since they involve high dimensional integrals. In practice, there are two types of approximation techniques which can be used: analytical or numerical.

One of the most important analytical approximations is the Laplace method (Kass and Raftery, 1995). It constructs a local normal approximation around a maximum *a posteriori* (MAP) estimate. This method is essentially based on the fact that in the large sample size limit, and given that some regularity conditions are satisfied, the posterior distribution converges to a normal distribution around its mode. Beal and Ghahramani (2003) list some limitations when applying this method: the Gaussian assumption can represent the posterior poorly for small sample sizes, and it is also poorly suited to bounded, constrained, or positive parameters. Further, the log-posterior may not be well approximated by a quadratic function for likelihoods with hidden variables, such as mixture models. For high dimensional integrals, when the dimension of the parameter vectors, d, may be thought of as increasing as the sample size n increases (Shun and McCullagh, 1995), the correction term, ordinarily $O(n^{-1})$ for fixed dimension, can become $O(1)$. This means, in practice, that the Laplace method may not work well for large, or even medium size, d problems. Even when the required regularity conditions hold, computing the inverse of Hessian matrix in high dimensional cases can also be burdensome (Beal and Ghahramani, 2003).

Sampling-based approximations, especially Markov chain Monte Carlo (MCMC) are a form of numerical approximation. MCMC methods provide a unifying framework within which many complex problems can be analyzed. These methods are straightforward and easy to implement without the requirements for sophisticated numerical analytic expertise

and possibly specialist software (Gelfand and Smith, 1990), and the requirements on the target distributions are quite minimal, which allows for settings where very little is known about them (Robert and Casella, 2010). Moreover, the MCMC perspective leads to efficient decompositions of high-dimensional problems into a sequence of smaller problems that are much easier to solve. Introductory material on MCMC methods can be found in Robert and Casella (2010); Liu (2008); Chen et al. (2000). Well-known methods include the Metropolis algorithm (Metropolis et al., 1953), the Gibbs sampler (Geman and Geman, 1984), the Metropolis-Hastings (M-H) algorithm (Hastings, 1970; Peskun, 1973) and its various varieties including, the random walk Metropolis-Hastings (Roberts et al., 1997), the independent Metropolis-Hastings algorithm (Tierney, 1994; Liu, 1996), the adaptive independent Metropolis-Hastings (Holden et al., 2009), the slice sampler (Neal, 2003), the reversible jump MCMC algorithm (Green, 1995), the birth-and-death MCMC algorithm (Stephens, 2000), the perfect sampling (Propp and Wilson, 1996). More detail can be found in Robert and Casella (2010); Chen et al. (2000); Liu (2008).

Although MCMC methods make possible the computation of posterior quantities, various practical issues are associated with their implementation. A primary concern is determining that the generated chain has in fact "converged". Although in theory, MCMC methods can be shown to have a geometric convergence rate for quite general classes of target distributions (Rosenthal, 1995; Roberts and Tweedie, 1999; Roberts and Rosenthal, 1998, 2004), in practice, many commonly used algorithms frequently fail to converge geometrically (Brooks, 1998). MCMC practitioners instead use convergence diagnostics to check if the sampled Markov chain appears to be stable for long enough runs (Roberts and Rosenthal, 2004). In the examples used in this proposal, for example, the M-H algorithm applied on the one-way random-effects model (Chapter 3), a Gibbs sampler on the Ising model (Chapter 2), and the reversible jump MCMC algorithm on a finite mixture model (Chapter 4), we had to use large runs to ensure the convergence of the chains. The computational cost can also be a concern, for example, posterior simulations for a 64×64 Ising model requires high storage

resources, and a reversible jump MCMC on a finite mixture model takes several hours to run. Other implementational issues include the difficulty of making a good choice of initial point, the implementation of MCMC algorithm in one long chain or several shorter chains in parallel, the tuning of M-H methods. Detailed discussions can be found in Brooks (1998).

This thesis is devoted to studying a fast and analytical approximation technique called the variational Bayesian method. This method tries to provide an approximation, which we call a *variational approximation*, of the posterior distribution by minimizing the Kullback-Leibler (KL) divergence from the approximations to the actual posterior, while constraining the VB solution to be "well-behaved". In this work we shall often use a well designed and calibrated MCMC method as a gold-standard to which we can compare the, hopefully much faster, VB solution. This is used, in particular, in the cases where we look at diagnostics for VB and for the application sections.

1.2 Variational methods

In Physics, variational free energy minimization (see Chapter 2) is a well-established principle (Feynman, 1972), and mean-field theory is a special application of this principle (Parisi, 1988). In the mean-field approximation, we minimize the free energy functional with respect to an approximating distribution of special forms; first the approximating distribution is *fully factorized*, which means random variables are assumed to be independent of each other, and secondly they are chosen from specified parametric families, with typically nice properties. The early applications of mean-field approximation can be found in the context of Bayesian neural networks (Hinton and Van Camp, 1993), belief networks (Saul et al., 1996), and factorial hidden Markov models (Ghahramani and Jordan, 1997). Barber and Bishop (1998) relaxed the first condition in the mean-field approximations, where the approximation is not necessary to be fully factorized.

Still in the context of Bayesian neural networks, MacKay (1995) considered the case where there are no restrictions on the distributional family of the approximations. This could be regarded as an early application of the general variational principle. Instead of using *variational calculus* (see Chapter 2), MacKay used a method which we call the *Kullback-Leibler distance decomposition*. Both methods will be reviewed in Chapter 2.

In the literature there also exists another approximating technique under the name of the variational method, using the principle of convex duality. In convex analysis (Rockafellar, 1997), a concave function $f(x)$ can be represented by $f(x) = \min_{\lambda}\{\lambda x - f^*(\lambda)\}$, where $f^*(\lambda)$ is called the conjugate function of $f(x)$. This implies that the function $f(x)$ is bounded above by a linear function. By using this principle, Jaakkola (1997), and Jaakkola and Jordan (1999) introduced a *variational method* to approximate the conditional distribution of a node in the context of graphical models. This approach is essentially a variational free energy minimization with a fixed functional form.

Although the method is originated in Physics, and is developed mainly in computer science, and widely applied in the domains such as machine learning, signal processing and the analysis of neuroimaging data, an early application of variational principle in the statistical literature can be found in Haff (1991), where a Bayes estimator of model parameters is obtained by minimizing a particular form of a Bayes risk function.

Theoretical properties of VB have been explored in recent years. For example, Neal and Hinton (1998) explained how the classical Expectation-Maximization (EM) algorithm (Dempster et al., 1977) can be considered in the framework of the variational method. The ordinary EM iteration alternates between an expectation (E) step, which evaluates the expected (complete) log-likelihood using the current estimate for the parameters, and a maximization (M) step, which computes parameters maximizing the expected log-likelihood found on the E step. The E step can be viewed as inferring the distribution over hidden variables from the variational Bayesian perspective, and the M step can be viewed as infer-

ring the distribution over model parameter, which is restricted to be a Dirac delta function. Wang and Titterington (2004, 2006) studied the convergence properties of the VB approximations. They also showed that for some special cases the approximating distributions will be asymptotically normally distributed.

Based on *stochastic complexity* (Watanabe, 2009), a different methodology is used to study theoretical properties of the VB method. Suppose that $y = \{y_i\}_{i=1}^n$ is a data set from the true distribution $f(y)$, and $p(y|\tau)$ is a hypothetic distribution of y with parameter τ. Stochastic complexity of the data (Watanabe, 2009) is defined as $F(Y) = -\log(Z(Y))$, where $Z(Y) = \int e^{-nH(\tau)}\pi(\tau)d\tau$, and $H(\tau) = \frac{1}{n}\sum_{i=1}^{n} \log \frac{f(y_i)}{p(y_i|\tau)}$. Essentially, it can be viewed as the negative log marginal likelihood. Hosino et al. (2005), Watanabe and Watanabe (2005), and Watanabe and Watanabe (2007) studied the asymptotic bounds for the expected stochastic complexity in the VB learning for some non-regular models (non-identifiable or whose Fisher information is not positive definite, Watanabe, 2009). They show that the expected stochastic complexity in the VB learning is smaller than those of regular models.

1.3 Application reviews

In this section we summarize some important applications of VB, and provide a flavour of how VB is broadly used.

As mentioned in the previous section, Bayesian *artificial neural networks* are an early application of the method where it is used in approximating the posterior distributions of weights in a *neural network*. Important references include Hinton and Van Camp (1993); Barber and Bishop (1998); MacKay (1995).

Another early application of the method can be found in the context of *independent component analysis (ICA)* (Ans et al., 1985), which is a method to extract the hidden independent components (sources) from a set of observations and has been broadly applied

in imaging processing (neuro/biomedical-imaging, natural images), signal processing, and the analysis of financial data. In short, an ICA problem can be formalised as $\mathbf{y}_t = \mathbf{A}\mathbf{s}_t + \epsilon_t$, where $\mathbf{y}_t$ is a L dimensional observable random vector at time t, and $\mathbf{s}_t$ is a I $(I \leq L)$ dimensional unobservable random vector (the source components) which are assumed to be independent of each other and at least one component is assumed non-Gaussian distributed (often modelled by a mixture of Gaussians with parameter θ). This setting can be found in Moulines et al. (1997); Attias (1999); Lappalainen (1999); Penny, and Roberts (2000). The term $\mathbf{A}$ is called the mixing matrix which is unknown, and ϵ_t represents Gaussian noise with zero mean and a unknown covariance matrix Λ. Lappalainen (1999); Miskin (2000); Attias (1999); Penny, and Roberts (2000); Lawrence and Bishop (2000) use the VB method to approximate the posteriors of $\mathbf{A}$, Λ, and θ.

Factor analysis is a method used to describe variability among observed variables in terms of low-dimensional unobserved variables called factors and is widely used in the social sciences, marketing, and other applied sciences. The model can be expressed as (Johnson and Wichern, 2002) $\mathbf{y}_t = \mathbf{A}\mathbf{s}_t + \mu + \epsilon_t$, where $\mathbf{y}_t$ is a d dimensional observable random vector, and $\mathbf{s}_t$ is a k $(k < d)$ dimensional unobservable random vector called, the common factors, $\mathbf{s}_t \sim \text{MVN}(\mathbf{0}, \mathbf{I})$, $\mathbf{A}$ is called a factor loading matrix (considered unknown), and μ is an unknown d dimensional constant vector, and ϵ_t represents Gaussian noise with zero mean and a diagonal covariance matrix Λ which is unknown. Moreover, $\mathbf{s}_t$ and ϵ_t are also assumed to be independent. Ghahramani and Beal (2000b); Zhao and Yu (2009) apply the VB method to approximate the marginal likelihood and the posterior distributions of Λ and $\mathbf{A}$.

Applications of VB in *directed graphical models or directed acyclic graphs (DAGs)* have been studied by many researchers. We classify these applications into two categories based on the implementation of VB. One uses the variational calculus, the other uses convex duality. The first type of application can be found in Attias (2000); Storkey (2000); Winn (2003);

Ghahramani and Beal (2000a); Beal and Ghahramani (2001, 2003, 2006). A common feature of these applications is that these researchers focus on a special class of DAGs where the nodes are split into a set of visible (observable) nodes, a set of hidden (unobservable) nodes and model parameters. VB is used to approximate the conditional distributions of hidden variables given the observed data and the conditional distributions of model parameters given the observed data. In particular, Storkey (2000) studies the dynamic tree model (Williams and Adams, 1999). Ghahramani and Beal (2000a); Beal and Ghahramani (2001, 2003, 2006) study conjugate-exponential belief networks (Ghahramani and Beal, 2000a). Winn (2003) also focuses on conjugate-exponential belief networks and summarized the VB learning in an algorithm called variational message passing and implemented the algorithm in a software package called VIBES. The second type of application can be found in Jaakkola and Jordan (1999), where they approximate the conditional distributions of the diseases given a set of observed findings in the QMR-DT database by using the principle of convex duality (The QMR-DT database is a large-scale bipartite belief network consisting of approximately 600 significant diseases and 4000 associated findings)

In *undirected graphical models*, applications can be found in McGrory et al. (2009) and Mohammad-Djafari and Ayasso (2009). McGrory et al. (2009) studies VB for a hidden Potts model. Mohammad-Djafari and Ayasso (2009) uses VB to approximate the normalizing constant in a Markov random field model. Other applications of the VB method also include, *principal component analysis* (Bishop, 1999), discrete hidden Markov models (MacKay, 1997), and finite mixture models (Corduneanu and Bishop, 2001).

In recent years more and more VB applications in conventional statistical problems have emerged. McGrory and Titterington (2007) use VB to calculate the deviance information criterion (DIC, Spiegelhalter et al. (2002)). An application in Poisson mixed models can be found in Hall et al. (2011). An application in generalized linear mixed models can be found in Ormerod and Wand (2012), while applications in parametric and nonparametric

regression models with missing data can be found in Faes et al. (2011).

1.4 Our examples

In this thesis we apply the VB method to a variety of models. Here we describe these applications, and in which chapter they occur.

In Chapter 2, we use a *Markov random field (MRF)* model to illustrate the concept of model simplification in Example 2.1. *A Weibull model* in Example 2.2 and *Logistic regression models* in Example 2.3 are used in studying the functional forms of VB approximations.

In Chapter 3, a *Normal random sample with unknown mean and variance* is used in Section 3.3 to illustrate the computational advantages of VB. *The Ising model*, *linear regression models*, and a *finite mixture of normals* are used in Section 3.3.2 to explore the mean structure approximations under VB. The *finite mixture of normals* is also used to study the local convergence property of VB in Section 3.4.1. The *one-way random-effects model* and *finite mixture models* are used in Example 3.5 and Example 3.6 to demonstrate the effects of choosing improper priors. They are also use in Example 3.7 and Example 3.8 to show the effects of choosing different factorization schemes.

In Chapter 4, a multivariate normal, *a normal random sample with unknown mean and variance* and *finite mixtures of normals* are used to demonstrate the proposed diagnostics methods. *The Dirichlet process* (Chapter 5) and *regime-switching lognormal models* (Chapter 6) are studied in detail.

1.5 Main contribution

This thesis aims to give insight into the general applicability of the VB method, explore its theoretical properties, and also apply it to a wide variety of problems.

Generally, there are several limitations in the current development of variational inference. First, it is usual to consider only conjugate exponential settings. Second, variational inferences are developed mostly with respect to natural parameters. Moreover, a full factorization, which assume complete independence, imposes a too strong constraint on the approximating distribution. We propose a particularly attractive parameterization for a parametric family, and also propose a factorization scheme with a more general dependency structure than is traditional in VB. These new frameworks provide greater modelling flexibility, and are not limited to the conjugate exponential setting.

We investigate some properties of VB, for example: its local convergence property, the effects of choosing different priors, and the effects of choosing different factorization scheme.

The essence of VB relies on making simplifying assumptions about the posterior dependence of a problem. By definition, posterior covariance structures are distorted. We propose three diagnostics methods to compute the actual posterior covariance matrix by looking at joint, marginal and conditional aspects of VB approximations. We propose a novel way, called the VB Adjusted Independent Metropolis-Hastings (VBAIMH) method, to compute the important quantities.

We also study the mean structure of VB approximations, and show how it is possible using VB approximations to approach model selection tasks such as determining the dimensionality of a model, or variable selection.

We also consider the variational application approach in Bayesian nonparametric modelling, especially for the Dirichlet process (DP). As an alternative to MCMC methods, which has been extensively studied in DP based-models, this work presents a full variational solution for DP with non-conjugate settings. We propose an empirical method to determine the number of distinct components in a truncated DP. We show how to use the variational techniques to approximate the posterior predictive distribution.

As a concrete application study, we work through the VB method on regime-switching

lognormal models and present solutions to quantify both the uncertainty in the parameters and model specification. We show that VB can recover exactly the model structure, gives the reasonable point estimates, and is very computationally efficient.

We also explore the asymptotic property of VB approximations.

Chapter 2

The variational Bayesian method

This chapter serves two purposes. First, it is intended to provide the reader with a detailed introduction to the *Variational Bayes (VB)* method and the techniques involved in its implementation. Second, its introduces a new framework for the VB method to overcome some limitations in its current development.

2.1 Motivation

Complex posterior dependence across model parameters is the main reason for the difficulty of calculation of posterior quantities. The VB approximation makes simplifying assumptions regarding this aspect of the problem with a full factorization into independent terms being the most commonly used scheme in the literature. For example, most of the applications reviewed in Chapter 1 use this form of factorization. However, a full factorization might impose too strong constraints on the approximating distribution. In many situations, we hope to capture the posterior dependence structure, or at least partial dependence structure, while having tractable approximations.

Generally, variational inference has been mainly investigated in the context of the exponential family. For example, Beal (2003) and Wainwright and Jordan (2008) provide a general variational formalism, called Variational Bayesian EM (VBEM), for the conjugate exponential family with latent variables. There are several limitations with these developments. First, they mainly consider the cases assuming conjugate priors. Second, the variational inferences are developed only with respect to natural parameters, which are often not the parameters of immediate interests.

To overcome the limitations in the current development, we propose a conditional factorization scheme which allows more general dependence structures than is traditional in VB. Further, we propose a special parameterization for a parametric family, called the parameter separation parameterization, which includes the distributions from both exponential family and non-exponential family, and whose prior distributions are not limited to the conjugate settings. The new framework provides greater modelling flexibility, and leads to greater approximation accuracy. Based on this framework, we develop the variational formalism, to which the VBEM is a special case.

The present chapter consists of two parts. We first provide the reader with a detailed introduction to the VB method. Section 2.2 introduces the concept of model simplification, Section 2.3 introduces two important ideas in the variational method, the Kullback-Leibler divergence and a lower bound of the log marginal likelihood, which are both used to describe the discrepancy between a posterior distribution and its VB approximation, Section 2.4 then describes two techniques, the Kullback-Leibler divergence decomposition and variational calculus, which are used to obtain the functional forms of the VB approximations. Our new development is presented in Section 2.5, including: Section 2.5.1 discusses a conditional factorization; Section 2.5.2 introduces the new *parameter separation parameterization*.

2.2 Simplification

Broadly speaking, the main reason for the intractability associated with posterior quantities derives from a complex posterior dependence structure. If we can reduce this complexity, we might expect the possibility of tractable posterior integrals. A very simple motivating example is that simplification can be achieved by assuming posterior independence between certain random variables. This idea is illustrated by an example from image processing.

Example 2.1 Markov random fields (MRFs). *In image analysis (Li, 2009), a site in a lattice represents an image pixel. The interrelationship between a set of sites is modelled by a neighborhood system. For example, Figure 2.1 (a) represents a second-order neighborhood system described in the form of an undirected graph where neighboring sites are connected by edges. Let* X_{ij} *be the random variable which indicates the grey level for the* ij^{th} *pixel and let* $X = \{X_{ij}\}_{i=1,j=1}^{n,m}$. *If we assume that the set of random variables* X *satisfies the Markov property, then by the Hammersley-Clifford theorem (Hammersley and Clifford, 1971), the joint probability of* X *can be expressed as,*

$$p(X = x) = \frac{1}{Z(\beta)} e^{-\beta \varepsilon(x)}, \tag{2.1}$$

where $\varepsilon(x) = \sum_{c \in \mathcal{C}} \Phi_c(x)$, *and* $\Phi_c(x)$ *is a potential function (Li, 2009) defined over a clique* c, *which is a subgraph such that any two vertices are connected by one and only one edge (Diestel, 2010). Here* $\mathcal{C}$ *is the set of all possible cliques with respect to a neighborhood system.*

One way to simplify the model given in Figure 2.1 (a) is to assume conditional independence between every two nodes connected by a diagonal edge, which results in Figure 2.1 (b) which is referred to as the first-order neighborhood system (Li, 2009). We could assume an even simpler model where every random variable is independent of all of other random variables, which results in Figure 2.1 (c). In this example, the joint distribution

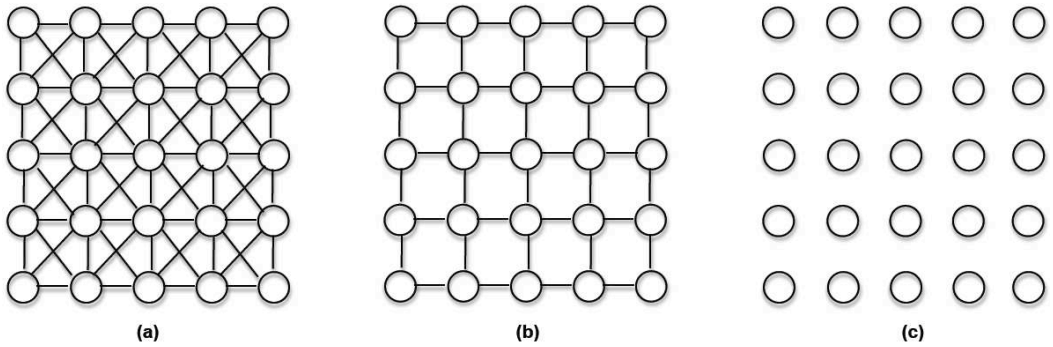


Figure 2.1: A graphical illustration of the model simplification (a) a second-order neighborhood system, (b) a first-order neighborhood system, and (c) a full factorization.

of the set of random variables X factors into a product of conditional distributions, whose probability distribution can be written as $q(\mathbf{x}) = \prod_i q_i(x_i)$. Generally speaking, we refer to such a simplifying procedure as a *factorization*. In the *VB* method, we consider $q(\mathbf{x})$ as an approximation of $p(\mathbf{x})$ given in (2.1). This setting is similar to the mean-field approximation (Parisi, 1988), where we explicitly specify a particular parametric family to $q(\mathbf{x})$. In VB we implicitly "choose" $q(\mathbf{x})$ by some criteria, that will be discussed in next section.

2.3 Two criteria

The immediate gain from a simplified model is the possibility of tractable integrals, since a set of low dimensional integrals might be expected to be more tractable than a single high dimensional one. Of course there will be loss due to the simplification, and criteria are required to assess the approximation accuracy.

The *Kullback–Leibler (KL)* divergence, defined in (2.2) below, is a measure of the discrepancy between two distributions. It is worth noting although it is intuitively interpreted as a "distance", the KL divergence is not a true distance, since the divergence from a probability measure P to a probability measure Q is typically not equal to that from Q to P (Critchley

et al., 1994).

Suppose, in a Bayesian inference problem, that we use $q(\boldsymbol{\tau})$ to approximate the posterior $p(\boldsymbol{\tau}|y)$, where y is the data and $\boldsymbol{\tau} = \{\tau_1, \cdots, \tau_p\}$ the model parameter vector. The KL divergence between them is defined as,

$$\text{KL } [q(\boldsymbol{\tau})||p(\boldsymbol{\tau}|\mathbf{y})] = \int q(\boldsymbol{\tau}) \log \frac{q(\boldsymbol{\tau})}{p(\boldsymbol{\tau}|\mathbf{y})} d\boldsymbol{\tau}, \tag{2.2}$$

provided the integral exists. We want to balance two things, having the discrepancy between p and q small while keeping q tractable. Hence we want to seek $q(\boldsymbol{\tau})$ which minimizes (2.2) while keeping $q(\tau)$ in analytically tractable form. First, note that the evaluation of (2.2) requires $p(\boldsymbol{\tau}|\mathbf{y})$, which may be unavailable since in the general Bayesian problem its normalizing constant is one of the main intractable integrals. However, we note that

$$\begin{aligned} \text{KL } [q(\boldsymbol{\tau})||p(\boldsymbol{\tau}|\mathbf{y})] &= \int q(\boldsymbol{\tau}) \log \frac{q(\boldsymbol{\tau})}{p(\boldsymbol{\tau}|\mathbf{y})p(\mathbf{y})} d\boldsymbol{\tau} + \log p(\mathbf{y}) \\ &= -\int q(\boldsymbol{\tau}) \log \frac{p(\boldsymbol{\tau}, \mathbf{y})}{q(\boldsymbol{\tau})} d\boldsymbol{\tau} + \log p(\mathbf{y}). \end{aligned} \tag{2.3}$$

Thus minimizing (2.2) is equivalent to maximizing the first term of the right-hand side of (2.20). The key computational point is that often the term $p(\boldsymbol{\tau}, \mathbf{y})$ is available even when the full posterior $\frac{p(\tau,y)}{\int p(\tau,y)d\tau}$ is not.

Definition 2.1 (HZ)[1] *Let $p(\boldsymbol{\tau}, \mathbf{y})$ be the joint distribution of data $\mathbf{y}$ and model parameter $\boldsymbol{\tau}$, and $q(\boldsymbol{\tau})$ be a probability density function over $\boldsymbol{\tau}$. A real valued* function F *of q is defined as*

$$F(q) = \int q(\boldsymbol{\tau}) \log \frac{p(\boldsymbol{\tau}, \mathbf{y})}{q(\boldsymbol{\tau})} d\boldsymbol{\tau}, \tag{2.4}$$

[1]For readers' convenience, we will give the source for the referred definitions, theorems, and lemmas in this dissertation, and use HZ (the initial of the dissertation's author) to indicate the contribution of this work.

when the integral exists. The negative of the function F is known as the variational free energy. *The log marginal likelihood is defined as* $\log p(\mathbf{y}) = \log \int p(\boldsymbol{\tau}, \mathbf{y}) d\boldsymbol{\tau}$.

It is worth noting that the marginal likelihood has long been used for Bayesian model checking and Bayesian model selection problems (Berger, 1985). The following lemma shows that there exists a close relationship between the log marginal likelihood and the function F.

Lemma 2.1 (HZ) *Given any approximating distribution q, the log marginal likelihood is bounded below by $F(q)$.*

Proof: This is a direct application of Jensen's inequality (Jensen, 1906) and the concavity of the log function, or alternatively the positivity of the KL divergence and the identity (2.20).

2.4 An optimization problem

Definition 2.2 (HZ) *Let*

$$\hat{q} = \arg\max_{q \in Q} F(q),$$

where Q is a set of probability density functions over the parameter space. Then $\hat{q}$ is called the variational approximation *or* variational posterior distribution*, and parameters of $\hat{q}$ (such as its mean, variance, etc), are called* variational parameters.

Some of the power of Definition 2.2 comes when we assume that all elements of Q have tractable posterior quantities. In that case all variational parameters will then also be tractable when the optimization can be achieved.

Two techniques can be used to obtain $\hat{q}(\boldsymbol{\tau})$. For simplicity and clarity let us assume, at least initially, that Q is the set of all densities which factorise as

$$q(\boldsymbol{\tau}) = \prod_{i=1}^{d} q_i(\tau_i).$$

Note then, here we have the property of tractability across Q as described above.

The function F is then given by

$$F\big(q\big) = \int \prod_{i=1}^{d} q_i(\tau_i) \log \frac{p(\boldsymbol{\tau}, \mathbf{y})}{\prod_{i=1}^{d} q_i(\tau_i)} d\boldsymbol{\tau}, \tag{2.5}$$

We want to maximize (2.5) over $q_i(\tau_i), i = 1, \cdots, d$. By the method of Lagrange multipliers[2], we define the Lagrangian as

$$\tilde{F}\big(q_1, \cdots, q_d\big) = F\big(q\big) + \sum_{i=1}^{d} \lambda_i \left(\int q_i(\tau_i) d\tau_i - 1 \right). \tag{2.6}$$

We take the functional derivative (Gelfand and Fomin, 1964) of (2.6) with respect to each $q_i(\tau_i)$ and equate the derivative to zero, obtaining, for $i = 1, \cdots, d$

$$\begin{aligned} \frac{\partial \tilde{F}(q(\boldsymbol{\tau}))}{\partial q_i(\tau_i)} &= -\left(\log q_i(\tau_i) + 1\right) + \int \prod_{j \neq i}^{d} \hat{q}_j(\tau_j) \log p(\boldsymbol{\tau}, \mathbf{y}) \ \mathrm{d}\boldsymbol{\tau}_{\backslash \boldsymbol{i}} + \lambda_i = 0 \\ \Rightarrow \quad \hat{q}_i(\tau_i) &= \frac{1}{Z_i} \exp\left[\int \prod_{j \neq i}^{d} \hat{q}_j(\tau_j) \log p(\boldsymbol{\tau}, \mathbf{y}) \ \mathrm{d}\boldsymbol{\tau}_{\backslash \boldsymbol{i}} \right], \end{aligned} \tag{2.7}$$

where $\boldsymbol{\tau}_{\backslash \boldsymbol{i}}$ denotes the complement of τ_i in $\boldsymbol{\tau}$, and Z_i is a normalization term. It is worth

[2]Note that the positivity constraints of q_i require Karush-Kuhn-Tucker (KKT) multipliers (Kuhn and Tucker, 1951; Karush, 1939) in a general nonlinear programming problem. However, in our problem the solution using KKT multipliers for the constraint with the form of $q_i > 0$ is in fact the same as the one given by Lagrange multipliers. Therefore, for simplicity we use only the Lagrange multipliers.

noting that the evaluation of the resultant integral in (2.7) requires all other $\hat{q}_j$ for $j \neq i$, and also the corresponding integral is not guaranteed to exist. We will discuss these problems in detail in Section 2.5.1. The above derivation uses well-established results from the calculus of variations (Gelfand and Fomin, 1964). The terminology *"variational"* Bayesian method comes from the roots of this technique. Another method, which we call the *Kullback–Leibler divergence decomposition*, can be used to obtain the same result. We can rewrite (2.20) for any given i as

$$\begin{aligned}
\mathrm{KL}\left[q(\boldsymbol{\tau})||p(\boldsymbol{\tau}|\mathbf{y})\right] &= \int q_i(\tau_i)\log q_i(\tau_i)\ \mathrm{d}\tau_i + \sum_{j\neq i}\int q_j(\tau_j)\log q_j(\tau_j)\ \mathrm{d}\tau_j + \log p(\mathbf{y}) \\
&\quad - \int \log\left(\frac{1}{Z_i}\exp\left[\int\prod_{j\neq i}^{d} q_j(\tau_j)\log p(\boldsymbol{\tau},\mathbf{y})\ \mathrm{d}\boldsymbol{\tau}_{\backslash \boldsymbol{i}}\right]\right) q_i(\tau_i)\ \mathrm{d}\tau_i - \log Z_i \\
&= \mathrm{KL}\left(q_i(\tau_i)||\frac{1}{Z_i}\exp\left[\int\prod_{j\neq i}^{d} q_j(\tau_j)\log p(\boldsymbol{\tau},\mathbf{y})\ \mathrm{d}\boldsymbol{\tau}_{\backslash \boldsymbol{i}}\right]\right) + K, \qquad (2.8)
\end{aligned}$$

where $K = \sum_{j\neq i}\int q_j(\tau_j)\log q_j(\tau_j)\ \mathrm{d}\tau_j - \log Z_i + \log p(\mathbf{y})$.

The first term on the right hand side of (2.8) is the only term which depends on $q_i(\tau_i)$. Then, the minimum value of KL $[q(\boldsymbol{\tau})||p(\boldsymbol{\tau}|\mathbf{y})]$ is achieved when the first term of the right-hand side of (2.8) equals to zero. Thus, we obtained

$$\hat{q}_i(\tau_i) = \frac{1}{Z_i}\exp\left[\int\prod_{j\neq i}^{d}\hat{q}_j(\tau_j)\log p(\boldsymbol{\tau},\mathbf{y})\ \mathrm{d}\boldsymbol{\tau}_{\backslash \boldsymbol{i}}\right].$$

2.5 Variational Bayesian marginal approximations

A full factorization into independent terms is the most commonly used scheme in the literature. For example, most of the applications reviewed in Chapter 1 use this form of factorization. We propose a factorization scheme in which more complex dependence struc-

tures can be used, while keeping tractability. This will provide greater modelling flexibility and approximation accuracy by the VB method.

2.5.1 A conditional factorization

Suppose $\boldsymbol{\tau}_I$ is a d dimension parameter vector, indexed by $I = \{1, \cdots, d\}$. When there is no chance of confusion we will write this as $\boldsymbol{\tau}$. We consider a VB approximation for the posterior $p(\boldsymbol{\tau}|y)$, which is factorized as,

$$q(\boldsymbol{\tau}) = \prod_i^K q(\boldsymbol{\tau}_{\mathcal{F}_i}) = \prod_i^K q(\boldsymbol{\tau}_{\mathcal{C}_i}|\boldsymbol{\tau}_{\mathcal{P}_i}) q(\boldsymbol{\tau}_{\mathcal{P}_i}), \tag{2.9}$$

where $\{\mathcal{F}_i\}_{i=1}^K$ is a partition of the index set I, for $K \leq d$, $\mathcal{F}_i = \mathcal{C}_i \cup \mathcal{P}_i$ and $\mathcal{C}_i \neq \emptyset$ for $i = 1, \cdots, K$. If the set $\mathcal{P}_i$ is an empty set, then $q(\boldsymbol{\tau}_{\mathcal{C}_i}|\boldsymbol{\tau}_{\mathcal{P}_i})$ denotes the unconditional density $q(\boldsymbol{\tau}_{\mathcal{C}_i})$. The generic functional forms of $q(\boldsymbol{\tau}_{\mathcal{C}_i}|\boldsymbol{\tau}_{\mathcal{P}_i})$ and $q(\boldsymbol{\tau}_{\mathcal{P}_i})$ are given as follows.

Theorem 2.1 (Variational Bayesian marginal approximations, HZ) *Let $p(\mathbf{y}, \boldsymbol{\tau})$ be the joint distribution of data $\mathbf{y}$ and model parameter $\boldsymbol{\tau}$. The variational approximations $q(\boldsymbol{\tau}_{\mathcal{C}_i}|\boldsymbol{\tau}_{\mathcal{P}_i})$ and $q(\boldsymbol{\tau}_{\mathcal{P}_i})$ in the decomposition (2.9) are given by*

$$q(\boldsymbol{\tau}_{\mathcal{C}_i}|\boldsymbol{\tau}_{\mathcal{P}_i}) \propto \exp\left(E_{q(\boldsymbol{\tau}_{\setminus(\mathcal{C}_i \cup \mathcal{P}_i)})}[\log p(\mathbf{y}, \boldsymbol{\tau})]\right), \tag{2.10}$$

$$q(\boldsymbol{\tau}_{\mathcal{P}_i}) \propto \exp\left(-E_{q(\boldsymbol{\tau}_{\mathcal{C}_i}|\boldsymbol{\tau}_{\mathcal{P}_i})}[\log q(\boldsymbol{\tau}_{\mathcal{C}_i}|\boldsymbol{\tau}_{\mathcal{P}_i})]\right) \exp\left(E_{q(\boldsymbol{\tau}_{\setminus\mathcal{P}_i})}[\log p(\mathbf{y}, \boldsymbol{\tau})]\right), \tag{2.11}$$

where $q(\boldsymbol{\tau}_{\setminus(\mathcal{C}_i \cup \mathcal{P}_i)})$ and $q(\boldsymbol{\tau}_{\setminus\mathcal{P}_i})$ are defined as,

$$q(\boldsymbol{\tau}_{\setminus(\mathcal{C}_i \cup \mathcal{P}_i)}) = \prod_{j \neq i} q(\boldsymbol{\tau}_{\mathcal{C}_j}|\boldsymbol{\tau}_{\mathcal{P}_j}) q(\boldsymbol{\tau}_{\mathcal{P}_j}), \quad q(\boldsymbol{\tau}_{\setminus\mathcal{P}_i}) = q(\boldsymbol{\tau}_{\mathcal{C}_i}|\boldsymbol{\tau}_{\mathcal{P}_i}) \prod_{j \neq i} q(\boldsymbol{\tau}_{\mathcal{C}_j}|\boldsymbol{\tau}_{\mathcal{P}_j}) q(\boldsymbol{\tau}_{\mathcal{P}_j}).$$

Theorem 2.1 can be proved either by using the calculus of variations or by the Kullback-Leibler divergence decomposition. The Appendix provides the proof by using the Kullback-

Leibler divergence decomposition.

For Theorem 2.1, we have the following comments.

First, expressions (2.10) and (2.11) imply that the evaluation of $q(\boldsymbol{\tau}_{\mathcal{C}_i}|\boldsymbol{\tau}_{\mathcal{P}_i})$ and $q(\boldsymbol{\tau}_{\mathcal{P}_i})$ requires knowledge of all other $q(\boldsymbol{\tau}_{\mathcal{C}_j}|\boldsymbol{\tau}_{\mathcal{P}_j})$ and $q(\boldsymbol{\tau}_{\mathcal{P}_j})$ for $j \neq i$. This structure leads to considering an iterative updating scheme, described in Algorithm 1, where the convergence of the function F is used to define the stopping rule.

Algorithm 1 Variational Bayesian algorithm

Initialize $q^{(0)}(\boldsymbol{\tau}_{\mathcal{C}_i}|\boldsymbol{\tau}_{\mathcal{P}_i})$ and $q^{(0)}(\boldsymbol{\tau}_{\mathcal{P}_i})$ at step 0, for $i = 1, \cdots, K$
Evaluate $F\big(q^{(0)}(\boldsymbol{\tau})\big)$
$t \Leftarrow 1$
while $F\big(q^{(t)}(\boldsymbol{\tau})\big)$ increases **do**
  for $i = 1$ to K **do**
  1. Infer $q^{(t)}(\boldsymbol{\tau}_{\mathcal{C}_i}|\boldsymbol{\tau}_{\mathcal{P}_i})$ at step t by

$$q^{(t)}(\boldsymbol{\tau}_{\mathcal{C}_i}|\boldsymbol{\tau}_{\mathcal{P}_i}) \;\propto\; \exp\Big(E_{q(\boldsymbol{\tau}_{\setminus(\mathcal{C}_i\cup\mathcal{P}_i)})}[\log p(\mathbf{y},\boldsymbol{\tau})]\Big)$$

  where $q(\boldsymbol{\tau}_{\setminus(\mathcal{C}_i\cup\mathcal{P}_i)})$ is given by

$$q(\boldsymbol{\tau}_{\setminus(\mathcal{C}_i\cup\mathcal{P}_i)}) = \prod_{j=1}^{i-1} q^{(t)}(\boldsymbol{\tau}_{\mathcal{C}_j}|\boldsymbol{\tau}_{\mathcal{P}_j}) q^{(t)}(\boldsymbol{\tau}_{\mathcal{P}_j}) \prod_{j=i+1}^{K} q^{(t-1)}(\boldsymbol{\tau}_{\mathcal{C}_j}|\boldsymbol{\tau}_{\mathcal{P}_j}) q^{(t-1)}(\boldsymbol{\tau}_{\mathcal{P}_j})$$

  2. Infer $q^{(t)}(\boldsymbol{\tau}_{\mathcal{P}_i})$ at step t by

$$q^{(t)}(\boldsymbol{\tau}_{\mathcal{P}_i}) \;\propto\; \exp\Big(E_{q^{(t)}(\boldsymbol{\tau}_{\mathcal{C}_i}|\boldsymbol{\tau}_{\mathcal{P}_i})}[\log q^{(t)}(\boldsymbol{\tau}_{\mathcal{C}_i}|\boldsymbol{\tau}_{\mathcal{P}_i})]\Big) \exp\Big(E_{q(\boldsymbol{\tau}_{\setminus\mathcal{P}_i})}[\log p(\mathbf{y},\boldsymbol{\tau})]\Big)$$

  where $q(\boldsymbol{\tau}_{\setminus\mathcal{P}_i})$ is given by,

$$q(\boldsymbol{\tau}_{\setminus\mathcal{P}_i}) = q^{(t)}(\boldsymbol{\tau}_{\mathcal{C}_i}|\boldsymbol{\tau}_{\mathcal{P}_i}) \prod_{j=1}^{i-1} q^{(t)}(\boldsymbol{\tau}_{\mathcal{C}_j}|\boldsymbol{\tau}_{\mathcal{P}_j}) q^{(t)}(\boldsymbol{\tau}_{\mathcal{P}_j}) \prod_{j=i+1}^{K} q^{(t-1)}(\boldsymbol{\tau}_{\mathcal{C}_j}|\boldsymbol{\tau}_{\mathcal{P}_j}) q^{(t-1)}(\boldsymbol{\tau}_{\mathcal{P}_j})$$

  end for
  $t \Leftarrow t + 1$
end while

Second, this algorithm will be particularly useful when all the relevant integrals have simple closed forms. Ways in which this can work include choosing exponential families or by using conjugate priors.

Third, if we assume that the prior $\pi(\boldsymbol{\tau})$ can be factorized in the same form as (2.9), then we have the following result.

Corollary 2.1 (Variational Bayesian marginal approximations with factorized priors, HZ) *Let $\pi(\boldsymbol{\tau})$ be the prior distribution which can be written in the form $\prod_{i=1}^{K} \pi(\boldsymbol{\tau}_{\mathcal{C}_i}|\boldsymbol{\tau}_{\mathcal{P}_i})\pi(\boldsymbol{\tau}_{\mathcal{P}_i})$. Then, (2.10) and (2.11) in Theorem 2.1 are given by*

$$q(\boldsymbol{\tau}_{\mathcal{C}_i}|\boldsymbol{\tau}_{\mathcal{P}_i}) \propto \pi(\boldsymbol{\tau}_{\mathcal{C}_i}|\boldsymbol{\tau}_{\mathcal{P}_i}) \exp\left(E_{q(\boldsymbol{\tau}_{\setminus \mathcal{C}_i \cup \mathcal{P}_i})}\left[\log\left(p(\mathbf{y}|\boldsymbol{\tau})\right)\right]\right) \tag{2.12}$$

$$q(\boldsymbol{\tau}_{\mathcal{P}_i}) \propto \pi(\boldsymbol{\tau}_{\mathcal{P}_i}) \exp\left(E_{q(\boldsymbol{\tau}_{\mathcal{C}_i}|\boldsymbol{\tau}_{\mathcal{P}_i})}\left[\log \frac{\pi(\boldsymbol{\tau}_{\mathcal{C}_i}|\boldsymbol{\tau}_{\mathcal{P}_i})}{q(\boldsymbol{\tau}_{\mathcal{C}_i}|\boldsymbol{\tau}_{\mathcal{P}_i})}\right]\right) \exp\left(E_{q(\boldsymbol{\tau}_{\setminus \mathcal{P}_i})}\left[\log\left(p(\mathbf{y}|\boldsymbol{\tau})\right)\right]\right) \tag{2.13}$$

Proof: These are direct applications of the linearity of expectation.

It may be of interest to note that the results of (2.12) and (2.13) exhibit a pattern,

variational distribution $\propto$ *prior distribution* $\times$ *"expected likelihood function".*

Fourth, as discussed above, although Equation (2.10) and (2.11) give generic functional forms, it is not guaranteed that either $E_{q(\boldsymbol{\tau}_{\setminus(\mathcal{C}_i \cup \mathcal{P}_i)})}[\log p(\mathbf{y}, \boldsymbol{\tau})]$ or $E_{q(\boldsymbol{\tau}_{\setminus \mathcal{P}_i})}[\log p(\mathbf{y}, \boldsymbol{\tau})]$ have closed-form representations as is illustrated in the following two examples.

Example 2.2 Weibull model. *In a reliability study, we may assume the data y_i is modelled by a Weibull distribution with a density function*

$$f(y|\alpha, \beta) \propto \alpha\beta y^{\alpha-1} e^{-\beta y^{\alpha}}, \quad \textit{where } y > 0, \alpha > 0, \beta > 0$$

In a Bayesian setting, we can choose the prior distribution as $\pi(\alpha, \beta|\gamma, \eta) \propto e^{-\alpha}\beta^{\gamma-1}e^{-\eta\beta}$, *where* γ *and* η *are hyper-parameters. The VB approximations are given by*

$$q(\beta) \propto \beta^{\gamma-1}\exp(-\eta\beta)\beta^n \exp\left(-\beta\sum_{i=1}^{n}\int q(\alpha)x_i^\alpha d\alpha\right),$$
$$q(\alpha) \propto \exp(-\alpha)\alpha^n \exp\left(\alpha\sum_{i=1}^{n}\log x_i - \sum_{i=1}^{n}\big(\int q(\beta)\beta d\beta\big)x_i^\alpha\right).$$

We can see that $q(\alpha)$ *is not in a form which gives easy evaluation of* $\int q(\alpha)x_i^\alpha d\alpha$. *Thus, although* $q(\beta)$ *is a Gamma density function, it is hard to compute its shape parameter which involves* $\int q(\alpha)x_i^\alpha d\alpha$.

Example 2.3 Logistic regression models. *Suppose the observed data* $Y_i \sim Bin(m_i, \theta_i)$, *for* $i = 1, \ldots, n$, *and there exist the associated explanatory variables* $x_i = \{x_{i1}, \ldots, x_{ip}\}$. *The parameters* θ_i *can be modelled by*

$$\log\left(\frac{\theta_i}{1-\theta_i}\right) = x_i^T\beta.$$

In a Bayesian setting, we chose $\pi(\beta) = \prod_i \pi(\beta_i)$ *as the prior. Thus, the posterior of* β *is given by*

$$p(\beta|y) \propto \prod_i^n \left[e^{x_i^T\beta}\right]^{y_i}\left[\frac{1}{1+e^{x_i^T\beta}}\right]^{m_i}\pi(\beta).$$

Under a full factorization, the VB approximation of $q(\beta_j)$ *is given by*

$$q(\beta_j) \propto \pi(\beta_j)\exp\left(\sum_{i=1}^{n} y_i x_{ij}\beta_j - \sum_{i=1}^{n} m_i \int\prod_{k\neq j} q(\beta_k)\log(1+e^{x_i^T\beta})d\beta_{\setminus j}\right),$$

and again tyically the term of $\int\prod_{k\neq j} q(\beta_k)\log(1+e^{x_i^T\beta})d\beta_{\setminus j}$ *has no closed-form.*

2.5.2 Parameter separation parameterization

Generally, variational inference has been mainly developed in the context of the exponential family. For example, Beal (2003) and Wainwright and Jordan (2008) provide a general variational formalism for the conjugate exponential family. There are several limitations with these developments. First, they mainly consider the cases assuming conjugate priors. Second, the variational inferences are developed only with respect to natural parameters, which are often not the parameters of immediate interests. In the present section, we show that VB inferences can be extended to a more general situation.

We consider a particular form of a parameterization for a parametric family, which we call the *parameter separation parameterization*. We will show that with this parameterization the distributional families of VB approximations are not changed during the iterative updates, and the convergence of variational parameters can be used as the stopping rule of Algorithm 1.

Definition 2.3 (HZ) *A parametric family* $\{P_{\boldsymbol{\tau}} : \boldsymbol{\tau} \in \mathbf{R}^d\}$ *is said to have a parameter separation parameterization if and only if the logarithm of its* density function *can be written as*

$$\log f(x) = h(x) + \sum_{c=1}^{C} \left(\prod_{i=1}^{d} g_{c,i}(\tau_i, x) \right), \tag{2.14}$$

where C *is a positive integer, and* h *and* $g_{c,i}$ *are real-valued functions.*

Example 2.4 *The density function of a normal distribution is given as,*

$$f(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}$$

The logarithm of this density function is given as,

$$\log f(x;\mu,\sigma^2) = \left(\underbrace{-\frac{1}{2}\log 2\pi}_{h(x)}\right) + \left(\overbrace{-\log\sigma}^{g_{11}(\sigma^2,x)}\ \overbrace{1}^{g_{12}(\mu,x)}\right) + \left(\underbrace{-\frac{1}{2\sigma^2}}_{g_{21}(\sigma^2,x)}\ \underbrace{(x-\mu)^2}_{g_{22}(\mu,x)}\right)$$

Many distributions can be written in the form of (2.14), for example, Gamma, Erlang, Inverse Gaussian, Normal, Log-normal, Logit-normal, Pareto, Weibull, Laplace, von Mises, and Multinomial (The form of the parameter separation parameterization for these distributions are given in the Appendix). These include both *exponential family* and non-*exponential family* examples.

Lemma 2.2 (HZ) *Suppose* $\mathbf{y} = \{y_t\}_{t=1}^n$ *are i.i.d. with a density function* $p(y;\boldsymbol{\tau})$ *having a parameter separation parameterization, where* $\boldsymbol{\tau} \in \mathbf{R}^p$*. For any given parameter subset of* $\boldsymbol{\theta} \subseteq \boldsymbol{\tau}$*, the likelihood function can be written as*

$$p(\mathbf{y}|\boldsymbol{\tau}) \;=\; \exp\left(\sum_t^n \left(h(y_t) + \sum_{c=1}^{C_\theta} g_c(\boldsymbol{\theta}, y_t)K_c + J\right)\right), \tag{2.15}$$

where K_c *and* J *are constant with respect to* $\boldsymbol{\theta}$*, and* C_θ *is the number of terms in the Expansion (2.14) whose* $g_{c,i}$ *function is not a constant function with respect to* $\boldsymbol{\theta}$*.*

The following result shows how the variational approximation behaves when we work on these particular parameterization.

Theorem 2.2 (Variational Bayesian marginal approximations with parameter separation parameterization, HZ) *Suppose* $\mathbf{y} = \{y_t\}_{t=1}^n$ *are i.i.d. with a density function* $p(y;\boldsymbol{\tau})$ *having a parameter separation parameterization, where* $\boldsymbol{\tau} \in \mathbf{R}^d$*. Then,*

(i) *For any given* $\boldsymbol{\tau}_{\mathcal{F}_i} = \boldsymbol{\tau}_{\mathcal{C}_i} \cup \boldsymbol{\tau}_{\mathcal{P}_i}$ *defined in the decomposition (2.9), the likelihood function*

can be written as

$$p(\mathbf{y}|\boldsymbol{\tau}) \;=\; \exp\left(\sum_{t}^{n}\left(h(y_t)+\sum_{c=1}^{C_{F_i}} g_c(\boldsymbol{\tau}_{\mathcal{F}_i},y_t)K_{F_i,c}+J_{F_i}\right)\right), \tag{2.16}$$

where $K_{F_i,c}$ *and* J_{F_i} *are constant with respect to* $\boldsymbol{\tau}_{\mathcal{F}_i}$. *The variation distribution* $q(\boldsymbol{\tau}_{\mathcal{C}_i}|\boldsymbol{\tau}_{\mathcal{P}_i})$ *is given by*

$$q(\boldsymbol{\tau}_{\mathcal{C}_i}|\boldsymbol{\tau}_{\mathcal{P}_i}) \;\propto\; \pi(\boldsymbol{\tau}_{\mathcal{C}_i}|\boldsymbol{\tau}_{\mathcal{P}_i})\exp\left(\sum_{t}^{n}\left(h(y_t)+\sum_{c=1}^{C_{P_i}} g_c(\boldsymbol{\tau}_{\mathcal{F}_i},y_t)K^*_{F_i,c}+J^*_{F_i}\right)\right), \tag{2.17}$$

where $K^*_{F_i,c}=E_{q(\boldsymbol{\tau}_{\setminus\mathcal{C}_i\cup\mathcal{P}_i})}[K_{F_i,c}]$ *and* $J^*_{F_i}=E_{q(\boldsymbol{\tau}_{\setminus\mathcal{C}_i\cup\mathcal{F}_i})}[J_{F_i}]$.

(*ii*) *For any given* $\boldsymbol{\tau}_{\mathcal{P}_i}$, *defined in the decomposition (2.9), the likelihood function can be written as*

$$p(\mathbf{y}|\boldsymbol{\tau}) \;=\; \exp\left(\sum_{t}^{n}\left(h(y_t)+\sum_{c=1}^{C_{P_i}} g_c(\boldsymbol{\tau}_{\mathcal{P}_i},y_t)K_{P_i,c}+J_{P_i}\right)\right), \tag{2.18}$$

where $K_{P_i,c}$ *and* J_{P_i} *are constant with respect to* $\boldsymbol{\tau}_{\mathcal{P}_i}$. *The variation distribution* $q(\boldsymbol{\tau}_{\mathcal{P}_i})$ *is given by*

$$\begin{aligned} q(\boldsymbol{\tau}_{\mathcal{P}_i}) \;\propto\; & \pi(\boldsymbol{\tau}_{\mathcal{P}_i})\exp\left(E_{q(\boldsymbol{\tau}_{\mathcal{C}_i}|\boldsymbol{\tau}_{\mathcal{P}_i})}\left[\log\frac{\pi(\boldsymbol{\tau}_{\mathcal{C}_i}|\boldsymbol{\tau}_{\mathcal{P}_i})}{q(\boldsymbol{\tau}_{\mathcal{C}_i}|\boldsymbol{\tau}_{\mathcal{P}_i})}\right]\right) \\ & \exp\left(\sum_{t}^{n}\left(h(y_t)+\sum_{c=1}^{C_{F_i}} g_c(\boldsymbol{\tau}_{\mathcal{P}_i},y_t)K^*_{P_i,c}+J^*_{P_i}\right)\right) \end{aligned} \tag{2.19}$$

Where $K^*_{P_i,c}=E_{q(\boldsymbol{\tau}_{\setminus\mathcal{P}_i})}[K_{P_i,c}]$ *and* $J^*_{P_i}=E_{q(\boldsymbol{\tau}_{\setminus\mathcal{P}_i})}[J_{P_i}]$.

Proof: The results of (2.16) and (2.18) are direct applications of Lemma 2.2. The results of (2.17) and (2.19) are direct applications of the linearity of expectation.

Theorem 2.2 implies two important properties of the VB method when it is applied to a parametric family with a parameter separation parameterization. First, the distributional families of VB approximations are not changed during the iterations of Algorithm 1. The result of (2.16) gives a particular form of the likelihood function with respect to $\boldsymbol{\tau}_{\mathcal{C}_i}$ and $\boldsymbol{\tau}_{\mathcal{P}_i}$. The result of (2.17) gives the functional form of the variational distribution of $q(\boldsymbol{\tau}_{\mathcal{C}_i}|\boldsymbol{\tau}_{\mathcal{P}_i})$. Comparing these two results, we see that these two expressions share the same set of functions of $\{g_c(\boldsymbol{\tau}_{\mathcal{F}_i}, y_i)\}_{c=1}^{C_i}$. Up to prior distributions, the difference between these two results only lies in the constant terms of $\{K_{F_i,c}\}_{c=1}^{C_i}$ and $\{K^*_{F_i,c}\}_{c=1}^{C_i}$, and J_{F_i} and $J^*_{F_i}$.

Second, Theorem 2.2 implies that in practice we can use the convergence of variational parameters as the stopping rule in Algorithm 1. This is because that at each iteration, the dependence of $q(\boldsymbol{\tau}_{\mathcal{C}_i}|\boldsymbol{\tau}_{\mathcal{P}_i})$ on other $q(\boldsymbol{\tau}_{\mathcal{C}_j}|\boldsymbol{\tau}_{\mathcal{P}_j})$ and $q(\boldsymbol{\tau}_{\mathcal{P}_j})$ for $j \neq i$ is only through the constants of $\{K^*_{F_i,c}\}_{c=1}^{C_i}$, which partially form the variational parameters. Therefore, iterative updates to the whole approximating distributions are in fact turned to be the updates to the variational parameters. The convergence of variational parameters in this thesis is measured by specifying a threshold for the norm of the difference between the values of the variational parameters in the last two iterations. We usually choose a threshold for example as 10^{-5} or even small.

Due to the linearity property of expectation, Theorem 2.2 is easily extended to a hierarchical setting, as long as at each layer or stage, the parametric family has a parameter separation parameterization.

The variational Bayesian EM (VBEM) for conjugate-exponential models, discussed in Beal (2003) and Wainwright and Jordan (2008), can be shown to be a special case of Theorem 2.2. Suppose the latent variable x and observable variable y are sampled from the natural exponential family with parameter θ, having a conjugate prior to the complete likelihood. We assume the variational approximation of $q(x, \theta)$ to $p(x, \theta|y)$ is factorized as $q(x, \theta) = q(x)q(\theta)$. Plugging $q(x)$ and $q(\theta)$ into Equation (2.17), with the likelihood and the prior in the form

of the exponential family, we will obtain the formula for VBEM, given in Beal (2003) and Wainwright and Jordan (2008).

2.6 Discussion

In this chapter, we first provide the reader with a detailed introduction to the VB method and the techniques involved in its implementation. We then discuss the limitations in the current development of the VB method. We show that VB inferences can be extended to a more general situation by proposing a special parameterization for a parametric family, the parameter separation parameterization, and a factorization scheme with a more general dependency structure. Based on this new framework, we develop a variational formalism, in which VB has particularly nice properties. This chapter mainly focuses on the mathematical derivations. In the next chapter, we will see how we can use this development into real-world problems.

2.7 Appendix

2.7.1 Proof of Theorem 2.1

Proof: : The Kullback-Leibler divergence from $q(\tau)$ to $p(\tau|y)$ can be written as

$$KL(q(\tau)||p(\tau|y)) = \log p(y) - \int q(\tau) \log \frac{p(\tau, y)}{q(\tau)} d\tau. \tag{2.20}$$

Plugging (2.9) into (2.20) and re-arrange the terms with respect to $q(\tau_{\mathcal{C}_i}|\tau_{\mathcal{P}_i})$, we can obtain the following expression:

$$\begin{aligned}\mathrm{KL}(q(\tau)||p(\tau|y)) =& E_{q(\tau_{\mathcal{P}_i})}\left[\mathrm{KL}(q(\tau_{\mathcal{C}_i}|\tau_{\mathcal{P}_i})||\frac{1}{Z}\exp(E_{q(\tau_{\setminus\mathcal{C}_i\cup\mathcal{P}_i})}[\log p(y,\tau)]))\right] \\ &+\log p(y)+K \end{aligned} \tag{2.21}$$

where Z is a normalization constant, and K is a constant with respect to $q(\tau_{\mathcal{C}_i}|\tau_{\mathcal{P}_i})$. The first term on the right hand side of (2.21) is the only term which depends on $q(\tau_{\mathcal{C}_i}|\tau_{\mathcal{P}_i})$. Then, the minimum value of $\mathrm{KL}[q(\tau)||p(\tau|y)]$ is achieved when the first term of the right-hand side of (2.21) equals to zero. Thus, we obtained

$$q(\tau_{\mathcal{C}_i}|\tau_{\mathcal{P}_i}) = \frac{1}{Z}\exp\left(E_{q(\tau_{\setminus(\mathcal{C}_i\cup\mathcal{P}_i)})}[\log p(y,\tau)]\right).$$

Similar to (2.11).

2.7.2 Examples of the parameter separation parameterization

Gamma distribution:

$$\begin{aligned} f(x;\alpha,\beta) &= x^{\alpha-1}\frac{e^{\left(-\frac{x}{\beta}\right)}}{\Gamma(\alpha)\,\beta^{\alpha}};\alpha>0,\beta>0,x>0 \\ \log f(x;\alpha,\beta) &= -\log\Gamma(\alpha)-\beta\log\alpha+(\alpha-1)\log x-\frac{x}{\beta} \end{aligned}$$

Erlang distribution:

$$\begin{aligned} f(x;\alpha,\beta) &= \frac{\beta^{\alpha}x^{\alpha-1}e^{-\beta x}}{(\alpha-1)!};\alpha\in\mathbb{N},\beta>0,x\geq 0 \\ \log f(x;\alpha,\beta) &= \alpha\log\beta-\log(\alpha-1)!+(\alpha-1)\log x-\beta x \end{aligned}$$

Inverse Gaussian:

$$\begin{aligned} f(x;\mu,\lambda) &= \left(\frac{\lambda}{2\pi x^3}\right)^{1/2} \exp\left\{\frac{-\lambda(x-\mu)^2}{2\mu^2 x}\right\} ; \mu>0, \lambda>0, x>0 \\ \log f(x;\mu,\lambda) &= \frac{1}{2}\frac{\lambda}{2\pi x^3} - \frac{-\lambda(x-\mu)^2}{2\mu^2 x} \end{aligned}$$

Normal distribution:

$$\begin{aligned} f(x;\mu,\sigma^2) &= \tfrac{1}{\sqrt{2\pi\sigma^2}}\, e^{-\frac{(x-\mu)^2}{2\sigma^2}} ; \mu \in \mathbf{R}, \sigma^2>0, x \in \mathbf{R} \\ \log f(x;\mu,\sigma^2) &= -\frac{1}{2}(\log 2\pi + \log\sigma^2) - \frac{(x-\mu)^2}{2\sigma^2} \end{aligned}$$

Log-normal distribution:

$$\begin{aligned} f(x;\mu,\sigma^2) &= \frac{1}{x\sigma\sqrt{2\pi}} \exp\left\{-\frac{(\ln x-\mu)^2}{2\sigma^2}\right\} ; \mu \in \mathbf{R}, \sigma^2>0, x>0 \\ \log f(x;\mu,\sigma^2) &= -\log x - \frac{1}{2}(\log 2\pi + \log\sigma^2) - \frac{(\ln x-\mu)^2}{2\sigma^2} \end{aligned}$$

Logit-normal distribution:

$$\begin{aligned} f(x;\mu,\sigma^2) &= \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{(\mathrm{logit}(x)-\mu)^2}{2\sigma^2}\right\} \frac{1}{x(1-x)} ; \mu \in \mathbf{R}, \sigma^2>0, x \in (0,1) \\ \log f(x;\mu,\sigma^2) &= -\log x(1-x) - \frac{1}{2}(\log 2\pi + \log\sigma^2) - \frac{(\mathrm{logit}(x)-\mu)^2}{2\sigma^2} \end{aligned}$$

where $\mathrm{logit}(x) = \log\left(\frac{x}{1-x}\right)$.

Pareto distribution:

$$\begin{aligned} f(x;\alpha,\beta) &= \frac{\alpha\,\beta^\alpha}{x^{\alpha+1}}; \alpha>0, x>\beta>0, \\ \log f(x;\mu,\sigma^2) &= \log\alpha + \alpha\log\beta - (\alpha+1)\log x \end{aligned}$$

Weibull distribution:

$$\begin{aligned} f(x;\alpha,\beta) &= \alpha\beta x^{\alpha-1}e^{-\beta x^{\alpha}}\ ;\alpha>0,\beta>0,x>0 \\ \log f(x;\alpha,\beta) &= \log\alpha+\log\beta+(\alpha-1)\log x-\beta x^{\alpha} \end{aligned}$$

Laplace distribution:

$$\begin{aligned} f(x;\mu,\beta) &= \frac{1}{2\,\beta}\exp\left(-\frac{|x-\mu|}{\beta}\right);\ \mu\in\mathbf{R},\beta>0,x><\in\mathbf{R} \\ \log f(x;\mu,\beta) &= -\log 2\beta-\frac{|x-\mu|}{\beta} \end{aligned}$$

von Mises distribution:

$$\begin{aligned} f(x;\mu,\kappa) &= \frac{e^{\kappa\cos(x-\mu)}}{2\pi I_0(\kappa)};\ \mu\in\mathbf{R},\kappa>0,x\in[0,2\pi) \\ \log f(x;\mu,\kappa) &= -\log 2\pi I_0(\kappa)-\kappa\cos(x-\mu), \end{aligned}$$

where $I_0(\kappa)$ is a modified Bessel function of order 0.

Multinomial distribution:

$$\begin{aligned} f(x_1,\ldots,x_k;n,p_1,\ldots,p_k) &= \frac{n!}{x_1!\cdots x_k!}p_1^{x_1}\cdots p_k^{x_k}; \\ &\quad p_i>0,\sum_i p_i=1,X_i\in\{0,\ldots,n\},\sum_i X_i=n \\ \log f(x_1,\ldots,x_k;n,p_1,\ldots,p_k) &= \log\frac{n!}{x_1!\cdots x_k!}+\sum_i x_k\log p_k \end{aligned}$$

Chapter 3

Properties of the VB Method

In Chapter 2 we introduced the working principle of the VB method, describe its implementation, and proposed a new general formalism for the variational solutions. In this chapter we aim to provide the reader with concrete examples and show how to apply the developments in Chapter 2 to these real problems. In addition, this chapter is intended to give insight into the general applicability and properties of the VB method.

3.1 Motivation

One major question associated with the VB method is how accurate are the approximations in a given situation? We are also interested in the advantages VB possesses over other approximation methods, as well its disadvantages. The present chapter applies the variational method, especially the developments from Chapter 2, to a variety of models and problems. Through these numerical studies we aim to obtain insights into its general applicability and properties. The present chapter investigates the following models: a Normal random sample with unknown mean and variance, one-way random-effects model, the Ising model, linear regression models, and finite mixture models. This set of models was selected to give a broad

range of problems which occur in applied statistics, but of course it is not comprehensive in any sense.

For these models, we study the mean structure approximations using VB, and show how it is possible using VB approximations to approach model selection tasks such as determining the dimensionality of a model, or variable selection. For different models, we apply different strategies. In the linear regression model, the regression setup was embedded in a hierarchical framework where latent variables are used to identify subset choices. In the finite mixture of normals model, the number of components are determined by the number of components whose mixing proportions are non-zero.

The present chapter also looks at some problems with the VB method. In particular, we show VB might only converge to a local optimal point, and we also look at the effects of choosing improper priors, and the effects of different choices of factorization scheme.

The chapter is organized as follows. In Section 3.2, we work through three models with the complete variational solutions. In Section 3.3 we discuss the advantages of the VB method, and look at its mean structure approximations. In Section 3.4, we discuss potential problems.

3.2 Applications

In this section, we will work through three models: a normal random sample with unknown mean and variance; a one-way random-effects model; a finite mixture model. These models are well-established, but important and widely applicable. We first describe these models, and then provide the variational solutions on them. The VB derivations are given step by step, and we hope the ideas discussed in Chapter 2 are brought in gradually. These applications will be repeated using as demonstrations and examples throughout the rest of the dissertation.

3.2.1 A normal random sample

Suppose that a set of observed data $y = \{y_i\}_{i=1}^n$ are assumed to be independently and identically normal distributed with unknown mean μ and variance σ^2. We use $L(y|\mu, \sigma^2)$ to denote the likelihood function, and use $\pi(\mu)$ and $\pi(\sigma^2)$ to denote the priors. We choose $\pi(\mu) = N(\mu; \gamma, \eta^2)$ and $\pi(\sigma^2) = \mathrm{IG}(\sigma^2; \alpha, \beta)$ as the prior distributions, where γ, η^2, α, and β are hyper-parameters. This setting is referred to as a semi-conjugate prior (Gelman et al., 1995). It is worth noting that the priors of $\pi(\mu)$ and $\pi(\sigma^2)$ are not conjugate to the likelihood function in this setting.

In the VB learning, suppose that we choose a full factorization of $q(\mu, \sigma^2) = q(\mu)q(\sigma^2)$. Thus the function F of $q(\mu, \sigma^2)$ is written as

$$F\big(q\big) = \int q(\mu)q(\sigma^2) \log \frac{L(y|\mu, \sigma^2)\pi(\mu)\pi(\sigma^2)}{q(\mu)q(\sigma^2)} d\mu d\sigma^2. \tag{3.1}$$

The corresponding Lagrangian is given by

$$\begin{aligned} \tilde{F}\big(q(\mu), q(\sigma^2)\big) \;=\; & \int q(\mu)q(\sigma^2) \log \frac{L(y|\mu, \sigma^2)\pi(\mu)\pi(\sigma^2)}{q(\mu)q(\sigma^2)} d\mu d\sigma^2 \\ & + \lambda_1 \left(\int q(\mu) d\mu - 1 \right) + \lambda_2 \left(\int q(\sigma^2) d\sigma^2 - 1 \right) \end{aligned} \tag{3.2}$$

We then take the functional derivative of (3.2) with respect to $q(\mu)$ and equate to zero, obtaining

$$\begin{aligned} & \frac{\partial \tilde{F}(q(\mu), q(\sigma^2))}{\partial q(\mu)} = \log \pi(\mu) - (\log q(\mu) + 1) + \int q(\sigma^2) \log L(y|\mu, \sigma^2) d\sigma^2 + \lambda_1 = 0 \\ \Rightarrow \quad & \hat{q}(\mu) = \frac{1}{Z_\mu} \pi(\mu) \exp \left[\int \hat{q}(\sigma^2) \log L(y|\mu, \sigma^2) d\sigma^2 \right], \end{aligned} \tag{3.3}$$

where Z_μ is the normalizing constant. The evaluation of (3.3) gives

$$\begin{aligned}\hat{q}(\mu) &= \frac{1}{Z_\mu}\exp\left(-\frac{1}{2\eta^2}(\mu-\gamma)^2\right)\exp\left(-\frac{n}{2}(\mu-\bar{y})^2 E_{\hat{q}(\sigma^2)}[\frac{1}{\sigma^2}]\right),\\ &= N(\gamma_{vb},\eta^2_{vb}),\quad \gamma_{vb}=\frac{\frac{\gamma}{\eta^2}+n\bar{y}E_{\hat{q}(\sigma^2)}[\frac{1}{\sigma^2}]}{\frac{1}{\eta^2}+nE_{\hat{q}(\sigma^2)}[\frac{1}{\sigma^2}]},\quad \eta^2_{vb}=\frac{1}{\frac{1}{\eta^2}+nE_{\hat{q}(\sigma^2)}[\frac{1}{\sigma^2}]},\end{aligned} \tag{3.4}$$

where $\bar{y}$ is the sample mean. Similarly, we can obtain $\hat{q}(\sigma^2)$ as

$$\hat{q}(\sigma^2)=IG(\alpha_{vb},\beta_{vb}),\quad \alpha_{vb}=\alpha+\frac{n}{2},\quad \beta_{vb}=\beta+\frac{S^2}{2}+\frac{n}{2}E_{\hat{q}(\mu)}[(\mu-\bar{y})^2], \tag{3.5}$$

where S^2 is the total sum of squares of y. Note that computing $\hat{q}(\mu)$ requires $\hat{q}(\sigma^2)$ and vice versa. Thus, iteration is needed to find a possible solution. A general algorithm for the iterative update is discussed in Chapter 2.

This model will be used in this work in the following ways: in Section 3.3 to show the computational advantages of VB; in Section 4.3.2 as a demonstration of our new diagnostics methods.

3.2.2 The one-way random-effects model

In the one-way random-effects model, we consider J independent experiments, with experiment j estimating the parameter θ_j from n_j independent normally distributed data points, y_{ij}, with a common unknown error variance σ^2. We define y_j as $y_j=(y_{1j},\cdots,y_{n_jj})$. Parameters θ_j are assumed independently drawn from a normal distribution with mean μ and variance τ^2. The parameters of μ, τ^2 and σ^2 are further treated as random variables. This

model is given by

$$
\begin{aligned}
&y_{ij}|\theta_j, \sigma^2 \sim N(\theta_j, \sigma^2), \\
&\theta_j|\mu, \tau^2 \sim N(\mu, \tau^2), \\
&\sigma^2 \sim \pi(\sigma^2), \mu \sim \pi(\mu), \tau^2 \sim \pi(\tau^2) \text{ for } i = 1, \ldots, n_j; j = 1, \ldots, J,
\end{aligned} \tag{3.6}
$$

where $\pi(\sigma^2)$, $\pi(\mu)$, and $\pi(\tau^2)$ are prior distributions. To have these priors providing little influence on the posterior distributions, we assign non-informative uniform priors for μ, $\log(\sigma^2)$, and τ^2. If we were to assign a uniform prior distribution for $\log(\tau^2)$, the posterior distribution would be improper. Thus, we get the prior distribution for μ, $\log(\sigma^2)$, and τ^2 is given by $\pi(\sigma^2, \mu, \tau^2) \propto \frac{1}{\sigma^2}$. The joint distribution of $y, \theta_1, \cdots, \theta_J, \sigma^2, \mu, \tau^2$ is given as follows:

$$
p(y, \theta, \sigma^2, \mu, \tau^2) = \prod_{j=1}^{J} \prod_{i=1}^{n_j} \phi(y_{ij}; \theta_j, \sigma^2) \prod_{j=1}^{J} \phi(\theta_j; \mu, \tau^2) \frac{1}{\sigma^2}, \tag{3.7}
$$

where $\phi(.)$ denotes the normal density function. We denote $q(\theta, \sigma^2, \mu, \tau^2)$ as the VB approximation for the posterior distribution of $p(\theta, \sigma^2, \mu, \tau^2|y)$. We assume q has a full factorization:

$$
q(\theta, \sigma^2, \mu, \tau^2) = \prod_{j=1}^{J} q(\theta_j) q(\sigma^2) q(\mu) q(\tau^2). \tag{3.8}
$$

It is straightforward to check that the distributions at each stage of model (3.7) all have a parameter separate parameterization, and then Theorem 2.2 can be used. By plugging (3.7)

into (2.17) or (2.19), we can obtain the following results:

$$
\begin{aligned}
q(\theta_j) &= N(g_j, k_j^2);\ g_j = -\frac{\frac{cn_j}{d}\bar{y}_{+j} + \frac{a}{b}e}{\frac{a}{b} + \frac{cn_j}{d}}, k_j^2 = \frac{1}{\frac{a}{b} + \frac{cn_j}{d}} \quad \text{for } j = 1, \cdots, J \\
\sigma^2 &= IG(c, d);\ c = \frac{n}{2}, d = \frac{1}{2}\left(\sum_{j=1}^{J}\sum_{i=1}^{n_j}(y_{ij} - g_j)^2 + \sum_{j=1}^{J} n_j k_j^2\right) \\
q(\mu) &= N(e, f^2);\ e = \frac{1}{J}\sum_{j=1}^{J} g_j, f^2 = \frac{b}{aJ} \\
q(\tau^2) &= IG(a, b);\ a = \frac{J}{2}, b = \frac{1}{2}\sum_{j=1}^{J}((g_j - e)^2 + F^2)
\end{aligned} \tag{3.9}
$$

where IG denotes the gamma distribution. The VB algorithm requires an iterative updates on the parameters of a, b, c, d, e, f^2, g_j, k_j^2 till they converge.

This model will be used in this work in the following ways: in Section 3.4.2 to demonstrate the effects of choosing improper priors; in Section 3.4.3 to show the effects of choosing different factorization schemes; in Chapter 5 for the Dirichlet process prior.

3.2.3 Finite mixture models

The density of a finite, K-component mixture of normals can be expressed as

$$
f(x_i|\Psi) = \sum_{j=1}^{K} \pi_j \phi(x_i|\mu_j, \sigma_j^2), \tag{3.10}
$$

where $\pi_j > 0, \sum_{j=1}^{K} \pi_j = 1$, ϕ is the normal density function, $\Psi = (\pi, \mu_1, \sigma_1^2, \cdots, \mu_K, \sigma_K^2)$. By introducing a latent indicator variable $z_i = (z_{i1}, \cdots, z_{iK}) \sim \text{Multinomial}(1, \pi)$, (3.10) can be written as

$$
f(x_i, z_i|\Psi) = \prod_{j=1}^{K} \{\pi_j \phi(x|\mu_j, \sigma_j^2)\}^{z_{ij}}. \tag{3.11}
$$

Given a data set $x = \{x_i\}_{i=1}^n$, the posterior distribution is given by

$$p(\Psi, z|x) \propto \prod_{i=1}^{n} f(x_i, z_i|\Psi)p(\Psi), \tag{3.12}$$

where $p(\Psi)$ is the prior distribution. Based on the conjugacy considerations, we choose the priors as follows:

$$\begin{aligned} p(\pi) &= \mathrm{Dir}(\pi_1, \cdots, \pi_K; a_1, \cdots, a_K),\ a_i = \frac{a_0}{K}, a_0 > 0, \\ p(\mu_j, \sigma_j^2) &= \mathrm{N}\left(\mu_j|\sigma_j^2; c_j, \frac{\sigma_j^2}{d_j^2}\right)\mathrm{IG}(\sigma_j^2; e_j, f_j) \end{aligned} \tag{3.13}$$

where a_0, c_j, d_j^2, e_j, f_j are hyper-parameters.

We choose a VB approximation of (3.12) which has a factorization of the form

$$q(\Psi, z) = q(\pi)\prod_{j=1}^{K} q(\mu_j|\sigma_j^2)q(\sigma_j^2)\prod_{i=1}^{n} q(z_i).$$

The logarithm of the density function of $f(x_i, z_i|\Psi)$ given in (3.11) can be written as

$$\log f(x_i, z_i|\Psi) = \sum_{j=1}^{K} z_{ij}(\log \pi_j + \log \phi(x|\mu_j, \sigma_j^2)).$$

In addition, the logarithm of the density function of $p(\pi)$ given in (3.13) can be written as

$$\log p(\pi) \propto \sum_{i=1}^{K}\left(\frac{a_0}{K} - 1\right)\log \pi_i.$$

Then, it is straightforward to check that $f(x_i, z_i|\Psi)$, $p(\pi)$, and $p(\mu_j, \sigma_j^2)$ all have a parameter seperation parameterization; and an application of Theorem 2.2 yields the following

results,

$$
\begin{aligned}
\hat{q}(z) &= \prod_{i=1}^{n} \text{multinomial}(z_i; 1, q_{i1}, \cdots, q_{iK}), \\
q_{ij} &= \frac{\exp\left(\psi(a_j') - \psi(\sum_{j=1}^{K} a_j') - \frac{1}{2}(\log f_j' - \psi(e_j')) - \frac{1}{2}(c_j' - x_i')^2 \frac{e_j'}{f_j'} - \frac{1}{2}\frac{1}{d_j^2 + \sum_{i=1}^{n} q_{ij}}\right)}{\sum_{j=1}^{K} \exp\left(\psi(a_j') - \psi(\sum_{j=1}^{K} a_j') - \frac{1}{2}(\log f_j' - \psi(e_j')) - \frac{1}{2}(c_j' - x_i')^2 \frac{e_j'}{f_j'} - \frac{1}{2}\frac{1}{d_j^2 + \sum_{i=1}^{n} q_{ij}}\right)}, \\
\hat{q}(\pi) &= \text{Dir}(a_1', ..., a_K') \; ; a_j' = a_j + \sum_{i=1}^{n} q_{ij}, \\
\hat{q}(\mu_j | \sigma_j^2) &= \text{N}\left(c_j', d_j'^2\right), c_j' = \frac{d_j^2 c_j + \sum_{i=1}^{n} x_i q_{ij}}{d_j^2 + \sum_{i=1}^{n} q_{ij}}, d_j'^2 = \frac{\sigma_j^2}{d_j^2 + \sum_{i=1}^{n} q_{ij}}, \\
\hat{q}(\sigma_j^2) &= \text{IG}\left(e_j', f_j'\right), e_j' = e_j + \frac{\sum_{i=1}^{n} q_{ij}}{2}, f_j' = f_j + \frac{d_j^2}{2}(c_j' - c_j)^2 + \frac{1}{2}\sum_{i=1}^{n} q_{ij}(c_j' - x_i)^2. \quad (3.14)
\end{aligned}
$$

VB again gives simple and conjugate approximations, and we can have the following observations on these variational parameters. The term $\sum_{i=1}^{n} q_{ij}$ can be interpreted as the number of times the state j being visited in the process. As $n \to \infty$, c_j' converges to a sample average, and the ratio of $\frac{f_j'}{e_j'}$ converges to a sample variance.

This model will be used in this work in the following ways: in Section 3.4.1 to study the local convergence property of VB; in Section 3.4.3 to show the effects of choosing different factorization schemes; in Section 3.3.2 to demonstrate how to use the VB to approach the model selection problem; in Section 4.3.3 as a demonstration in the diagnostics methods.

3.3 The advantages

One major question associated with the VB method is how well the approximations work. In this section, we particularly look at its approximation of the mean structure. In addition, we compare its performance with MCMC methods through numerical studies.

3.3.1 Tractability

A crucial computational advantage of the VB method lies in that it often provides analytically tractable marginal approximations. We demonstrate this by a simple example.

Let us revisit the model of a normal random sample with unknown mean and variance. The exact conditional posterior distributions of $p(\mu|\sigma^2, y)$ and the marginal posterior distributions of $p(\sigma^2|y)$ are given by

$$\begin{aligned} p(\mu|\sigma^2, y) &= N(\gamma_p, \eta_p^2), \quad \gamma_p = \frac{\frac{\gamma}{\eta^2} + \frac{n\bar{y}}{\sigma^2}}{\frac{1}{\eta^2} + \frac{n}{\sigma^2}}, \quad \eta_p^2 = \frac{1}{\frac{1}{\eta^2} + \frac{n}{\sigma^2}}, \\ p(\sigma^2|y) &\propto \eta_p \phi(\gamma_p; \gamma, \eta^2) \varphi(\sigma^2; \alpha, \beta) \prod_{i=1}^{n} \phi(y_i; \gamma_p, \sigma^2), \end{aligned}$$

where ϕ denotes the normal density function and φ denotes the inverse-gamma density function. Despite the apparent simplicity of this model, the marginal posterior distribution of $p(\sigma^2|y)$ has no simple conjugate form, and this density function is not easy to use, for example computing moments. The marginal posterior distribution of $p(\mu|y)$ is also not easy to evaluate. However, we can see that their variational approximation of $q(\mu)$ in (3.4) and $q(\sigma^2)$ in (3.5) give very simple and conjugate forms.

For the one-way random-effects model, the finite mixtures of normal model, and most of the models we present in this thesis, VB gives simple marginal distributions. One natural question may be raised: how well these marginal approximations work? In the next section, we will, in particular, look at the mean structure approximations by the VB method.

3.3.2 Mean structure approximations

Posterior means are commonly used point estimators in the Bayesian analysis, which minimize the mean square error (MSE). In this section we are interested in variational approximations for posterior means. We will study three different models; Ising, linear regression,

and finite mixtures of normals. In the Bayesian setting, they present sufficiently complex dependence structures where we can apply the variational principle. MCMC methods are well-established with these models. Through a series of numerical studies, we will compare VB with MCMC methods. These comparative studies will give the reader an initial experience of how VB approximations can work. In addition, we will show how it is possible using VB approximations to approach model selection tasks such as determining the dimensionality of a model, or variable selection.

The Ising model

We continue looking at the Example 2.1 and consider a special case of MRFs, an Ising model. Graphically, an Ising model can be depicted by Figure 2.1 (b), where each $x_i \in \chi_i = \{-1, +1\}$, and $\varepsilon(x) = -\sum_i \sum_{j\sim i} x_i x_j$, where $j \sim i$ represents the neighbors of i but not including i. In VB, we assume that the joint distribution $p(x)$ is approximated by $q(x)$ which satisfies $q(x) = \prod_i q(x_i)$. The KL divergence is given by

$$\begin{aligned} KL(q(x)||p(x)) &= \sum_i \int_{\chi_i} q(x_i) \log q(x_i) dq(x_i) \\ &\quad - \int_{\chi} \prod_i q(x_i)(-\beta\varepsilon(x) - \log z(\beta)) dq(x). \end{aligned} \tag{3.15}$$

Thus, $q(x_i)$ which minimizes (3.15), has the form,

$$q(x_i) = \frac{\exp\left\{\beta \sum_{j\sim i} x_i E_{\hat{q}(x_j)}[x_j]\right\}}{\sum_{x_i} \exp\left\{\beta \sum_{j\sim i} x_i E_{\hat{q}(x_j)}[x_j]\right\}},$$

where we can see the expected values of the neighbors of i are used in the approximation. This result is same as the mean-field approximation (Parisi, 1988). Again, each $q(x_i)$ requires its neighboring $q(x_j)$ for $j \neq i$, hence a iterative update is needed.

In our experiments, we consider a simple Ising model of size 64×64 with the value of $\beta = 0.1$. Since the marginal distribution of $p(x_i)$ is intractable, we use a Gibbs sampler to obtain samples from the true $p(x_i)$ and report the last sample. Meanwhile we draw samples from the VB approximations directly. An examination of the graphical pattern presented by these two sets of samples could provide us an intuitive comparison of $q(x)$ and $p(x)$.

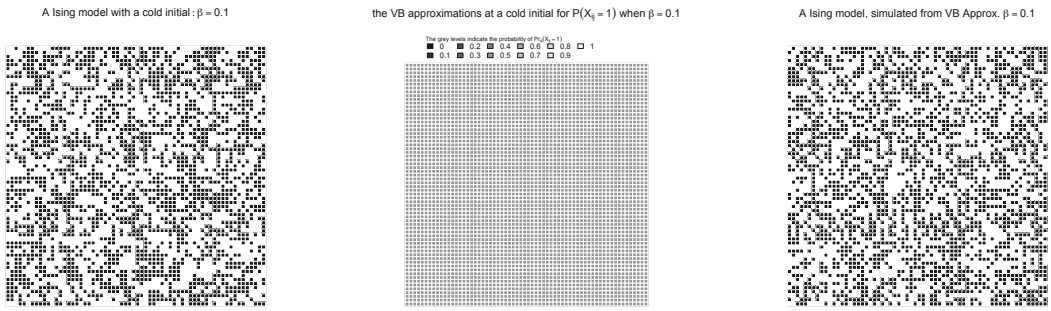
We start the Gibbs sampler by setting the initial value of each x_i randomly and independently to be -1 or $+1$, referred to as a hot initial condition. A chain of size 10^6 is generated. The program (in R) takes about 90 seconds on a MacBook computer (2 GHz Intel Core 2 Duo processor, 1 GB 667 MHz SDRAM, and Mac OS X 10.4.11). Figure 6.2(a) shows the last sample simulation. In contrast, VB approximations take about 10 iterations (about 1.5 seconds) to converge. The convergence here refers to the convergence of the variational parameters, which is measured by specifying a threshold for the norm of the difference between the values of the variational parameters in the last two iterations. We usually can choose a threshold fox example as 10^{-4} Figure 6.2(b) displays the probability of x_i being in state $+1$ approximated by VB through a grey level on each site; black means $Pr(x_i = 1) = 0$, and white means $Pr(x_i = 1) = 1$. Numerically the VB approximation on each x_i reports $Pr(x_i = 1) = 0.5 \pm 0.0001$. Figure 6.2(c) shows samples drawn from the VB approximations. We can see that the VB samples shown in Figure 6.2(a) and MCMC samples shown in Figure 6.2(c) exhibit similar patterns, where each x_i is nearly randomly arranged.

Regression models

We consider the canonical regression setup,

$$Y|X, \beta, \sigma^2 \sim MVN(X\beta, \sigma^2 I_n), \tag{3.16}$$

where Y is $n \times 1$, X is $n \times p$, $\beta = (\beta_1, \cdots, \beta_p)$ and σ^2 is a scalar.



(a) The MCMC outputs (b) The VB approx. (c) The VB simulations

Figure 3.1: An Ising model with $\beta = 0.1$ and a hot initial

There exists extensive literature on finding strategies to approach the variable selection problem in linear regression analysis. From a Bayesian perspective, one method is to consider that each component of β is modelled as having come from one of two normal distributions with different variances. By introducing a latent variable γ_i with a Bernoulli(w_i) distribution, we have

$$\beta_i|\gamma_i \sim (1-\gamma_i)N(0,\tau_i^2) + \gamma_i N(0, c_i^2\tau_i^2), \tag{3.17}$$

where c_i and τ_i are hyper-parameters, and τ_i^2 is chosen to be small and c_i^2 large. When $\gamma_i = 0$, β_i is typically "close" to 0, while if $\gamma_i = 1$, β_i would be more dispersed. Thus γ_i can be regarded as an indicator of whether x_i should be included in the model or not. This setting can be found in the work of Mitchell and Beauchamp (1988); George and McCulloch (1993); Chipman (1996); Brown et al. (1998); George and McCulloch (1997); Ntzoufras et al. (2000). Given the data y, x and choices of c_i and τ_i, we want to obtain the marginal posterior of $p(\gamma|y)$ and use it to decide about the model structure.

We choose the following prior distributions; $\pi(\gamma) = \prod_i^P w_i^{\gamma_i}(1-w_i)^{1-\gamma_i}$ and $\pi(\sigma^2) = \text{IG}(a,b)$, where $\{w_i\}_{i=1}^P$, a, and b are hyper-parameters. Based on (3.16) and (3.17), the

posterior distribution is given by

$$p(\beta, \gamma, \sigma^2|y) \propto f(y|\beta, \sigma^2) f(\beta|\gamma)\pi(\gamma)\pi(\sigma^2). \tag{3.18}$$

We consider a VB approximation with factorization $q(\beta, \gamma, \sigma^2) = \prod_{i=1}^{p} q(\beta_i) \prod_{i=1}^{p} q(\gamma_i) q(\sigma^2)$. The logarithm of the density function of $f(y|\beta, \sigma^2)$ is given by

$$\log f(y|\beta, \sigma^2) \propto -n \log \sigma - \frac{1}{2\sigma^2}(Y - X\beta)^T (Y - X\beta),$$

which is a polynomial for each β_i. In addition, the logarithm of the density for $\beta_i|\gamma_i$ in (3.17) can be written as

$$(1 - \gamma_i) \log \phi(\beta_i; 0, \tau_i^2) + \gamma_i \log \phi(\beta_i; 0, c_i^2 \tau_i^2).$$

Then the conditions for parameter seperation parameterization are satisfied. Applying Theorem (2.2), we obtain,

$$\begin{aligned}
\hat{q}(\beta_i) &= \mathrm{N}(\mu_i, \nu_i^2), \quad \mu_i = \frac{e_i}{l_i(\frac{1-w_i^{'}}{\tau_i^2} + \frac{w_i^{'}}{c_i^2\tau_i^2} + \frac{1}{l_i})}, \nu_i^2 = \frac{1}{\frac{1-w_i^{'}}{\tau_i^2} + \frac{w_i^{'}}{c_i^2\tau_i^2} + \frac{1}{l_i}}, i = 1, \cdots, p, \\
&\quad e_i = \frac{\sum_{t=1}^{n} y_t x_{ti} - \sum_{t=1}^{n} x_{ti} \sum_{j \neq i}^{P} x_{tj}\mu_j}{\sum_{t=1}^{n} x_{ti}^2}, l_i = \frac{1}{\frac{a'}{b'}\sum_{t=1}^{n} x_{ti}^2} \\
\hat{q}(\sigma^2) &= \mathrm{IG}(a', b'), \quad a' = a + \frac{n}{2}, b' = b + \frac{y^T y - 2\mu^T x^T y + \mathrm{tr}(x\nu x^T + x\mu\mu^T x^T)}{2}, \\
&\quad \mu = (\mu_1, \cdots, \mu_P), \nu = \mathrm{diag}(\nu_1^2, \cdots, \nu_P^2) \\
\hat{q}(\gamma_i) &= \mathrm{Bernoulli}(w_i^{'}), \quad w_i^{'} = \frac{e^{k_{2i}} w_i}{e^{k_{2i}} w_i + e^{k_{1i}}(1 - w_i)}, i = 1, \cdots, p, \\
&\quad k_{1i} = -\log c_i\tau_i - \frac{1}{2c_i^2\tau_i^2}(\mu_i^2 + \nu_i^2), k_{2i} = -\log \tau_i - \frac{1}{\tau_i^2}(\mu_i^2 + \nu_i^2).
\end{aligned}$$

We see VB approximations are simple and well-recogonised distribution, in contrast, the exact marginal posteriors of $p(\gamma|y)$, $p(\beta|y)$, and $p(\sigma^2|y)$ have no closed forms. We will use

$q(\gamma_i)$ to make inference on γ_i.

We consider two sets of simulated data and one real data set, which are used in George and McCulloch (1993), where they used Gibbs samplers to obtain the posterior samples. George and McCulloch proposed that after the chain has reached approximate stationarity, the values of γ_i corresponding to the most promising subsets of x_i will appear with the highest frequency, because it is just those values which have largest probability under $p(\gamma|y)$. Thus a tabulation of the high-frequency values of γ_i can be used to identify the corresponding subsets of predictors as potentially promising.

Example 3.1 *Suppose there are $p = 5$ explanatory variables, and they follow $X_{5\times 1} \sim MVN(0, I_5)$. We first generate 60 samples from $X_{5\times 1}$, and then generate the dependent variable according to the model*

$$y = x_4 + 1.2x_5 + \epsilon, \tag{3.19}$$

where $\epsilon \sim N(0, 1)$. We choose the values of the hyper-parameters of $a = 2$, $b = 1$ based on the fact of $E_{\pi(\sigma^2)}[\sigma^2] = \frac{b}{a-1}$, and set $w_i = 0.5$ for which $\pi(\gamma)$ turns to be a uniform distribution on γ. The choices of τ_i and c_i were discussed in detail in George and McCulloch (1993) where the general principle is that "if $\beta_i \sim N(0, \tau_i^2)$, then β_i can be approximately replaced by 0 since $|\beta_i| \leq 3\tau_i$", and if $\beta_i \sim N(0, c_i^2\tau_i^2)$, "one would want to choose c_i large enough to give support to values of β_i that are substantively different from 0, but not so large that unrealistic values of β_i are supported". Here, we set $\tau_i = 0.1$, $c_i = 30$, for $i = 1, \cdots, p$.

In experiments, we generate 200 data sets from model (3.19). For each data set, we randomly generate initial values to start VB, and the VB method converges less than 10 iterations, and gives almost exactly same results. Table 3.1 gives the average value of the probability of $\gamma_i = 1$ under the VB approximation for these 200 data sets, and the numbers in parenthesis are the standard deviations. It clearly shows that VB suggests that x_4 and x_5

should be included in the final model.

Table 3.1: The VB approximation of $q(\gamma)$ in Example 3.1

	β_1	β_2	β_3	β_4	β_5
$q(\gamma_i = 1)$	0.047(0.0053)	0.048(0.0083)	0.047(0.0079)	1.000(0)	1.000(0)

Table 3.2: High Frequency Models in Example 3.1 (George and McCulloch, 1993)

Model predictors	5	4, 5	2, 5	2, 4, 5
Frequencies	0.258	0.242	0.070	0.055

Table 3.2 (cited from George and McCulloch (1993)) gives the four highest frequency values of γ estimated by Gibbs samplers. The two most frequent models are $y = f(x_5)$ and $y = f(x_4, x_5)$. We can see that the predictor of x_4 are often excluded, and there are non-negligible probabilities for the models which include x_2. Thus we may conclude that VB has found the correct structure

Example 3.2 *This example is similar to Example 3.1, except that x_3 is replaced by $x_3 = x_5 + 0.15\kappa, \kappa \sim N(0,1)$, yielding a very high correlation between x_3 and x_5 (the correlation coefficient is about 0.98). The response variable was generated according to $y = x_4 + 1.2x_5 + \epsilon, \epsilon \sim N(0,1)$. We still generate 200 data sets from this model. For each data set, we randomly generate initial values to start VB, and the VB method converges within 10 iterations. Table 3.3 shows the result for one of the 200 data sets. For some of the data sets, VB will pick up x_3, while for the others, VB will choose x_5. The different choice on x_3 or x_5 may be mainly due to the strong proxy of x_3 for x_5. However, the positive aspect is that VB can always identify x_4. Table 3.4 gives the most frequent models suggested by Gibbs samplers. The Gibbs sampler seems to be more diluted by this proxy. The predictor of x_4 can not be identified in the first two most frequent models.*

Table 3.3: The VB approximation of $q(\gamma)$ in Example 3.2

	β_1	β_2	β_3	β_4	β_5
$q(\gamma_i = 1)$	< 0.05	< 0.05	1.000	1.000	< 0.05

Table 3.4: High Frequency Models in Example 3.2 (George and McCulloch, 1993)

Model predictors	3	5	3,4	4, 5
Frequencies	0.146	0.123	0.098	0.086

Example 3.3 *The Hald data (Draper and Smith, 1981), which has been used by various authors to illustrate variable selection procedures, consists of* $n = 13$ *observations on a dependent variable* Y *(which is the heat evolved during a chemical reaction) and* $p = 4$ *independent variables. The inclusion of* x_1 *and* x_2 *is favored by conventional selection procedures (Draper and Smith, 1981). We set* $w_i = 0.5$, $a = 2$, $b = 1$, *and* $\tau_i = 0.6$, $c_i = 60$, *for* $i = 1, \cdots, p$. *The VB method converges after 253 iterations and only favors* x_1 *being included (Table 3.5). This result slightly differs from the conventional procedures since* x_2 *is not favored by VB. The most frequent models suggested by the Gibbs sampler is tabulated in Table 3.6*

Table 3.5: The VB approximation of $q(\gamma)$ for the Hald data

	x_1	x_2	x_3	x_4
$q(\gamma_i = 1)$	0.92061674	0.09733090	0.03404244	0.02277744

Table 3.6: High Frequency Models for the Hald data (George and McCulloch, 1993)

Model predictors	None	1	4	3	2	1,2	1,3	1,4
Frequencies	0.44	0.25	0.07	0.06	0.05	0.03	0.03	0.03

Finite mixture models

The variational solution for a mixture of normals model has been derived in (3.14) in Section 3.2.3. Again, VB gives simple and well-recogonised distributions for the marginal or conditional marginal approximations. Making inference on the number of components is a fundamental and challenging problem in the application of finite mixture models. A large number of statistical methods have been proposed and investigated over the past a few decades. In this section, we propose an empirical means to estimate the number of components, which uses the marginal approximation for the mixing proportions. We compare its performance with the sampling based method, in particular, the reversible jump MCMC (RJMCMC), Green (1995).

Example 3.4 *Since $q(\pi)$ is a Dirichlet (see (3.14)), then we have $E[\pi_j] = \frac{a'_j}{\sum_{i=1}^{K} a'_j} = \frac{a_j+\sum_{i=1}^{n} q_{ij}}{a_0+n}$, where $\sum_{i=1}^{n} q_{ij}$ can be interpreted as the number of observations from component j. We usually choose a_j to be small with respect to sample size n, then the zero value of $\sum_{i=1}^{n} q_{ij}$ will result in $E[\pi_j]$ close to zero. We refer the component with $E[\pi_j]$ close to zero as an empty component. Our approach to the order selection problem is selecting the number of the non-empty components.*

In the numerical studies, we consider a three component mixtures of normals, and use a simulated data. We generate a sample of size 400 (Figure 3.2) from the model of $f(x) = 0.3\phi(x; 0, 1^2) + 0.4\phi(x; 2, 0.7^2) + 0.3\phi(x; 4.5, 0.8^2)$. For the VB method, we will assume no knowledge about the exact number of components, and, in fact, mis-specify it as 10. This allows redundancy and potentially over-fitting. The hyper-parameters are chosen as follows: we set $a_0 = 0.0001$; we divide the data to 10 groups according an ascending order, and then set c_j equal to the mean of each group, and $e_j = 2$, f_j equal to the variance of each group (see Section 3.2.3 for the definition of a_0, c_j, e_j, and f_j). The VB algorithm is given in Algorithm 2. The variational parameters are updated in the order of a'_j first, then c'_j, e'_j, and f'_j last. To start the iterations, we set $\sum_{i=1}^{n} q_{ij} = \frac{400}{10} = 40$ for each j, and $\sum_{i=1}^{n} x_i q_{ij} = c \times \sum_{i=1}^{n} q_{ij}$,

where c is sampled from $N(c_j, f_j)$.

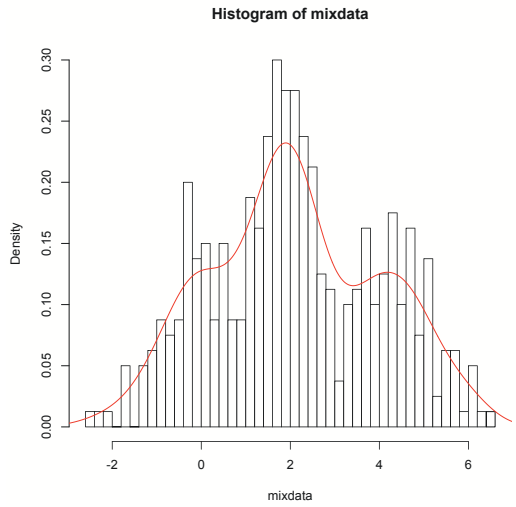


Figure 3.2: Simulated data for a 3 component normal mixture model

Algorithm 2 Variational Bayes algorithm for the mixtures of normal model

Initialization
while $a_j^{'}$, $c_j^{'}$, $d_j'^2$, $e_j^{'}$, and $f_j^{'}$ do not converge **do**
According to the formula given in (3.14), computing $a_j^{'}$, $c_j^{'}$, $d_j'^2$, $e_j^{'}$, and $f_j^{'}$ respectively.
end while

The convergence of the variational parameters is determined by specifying a threshold for the Euclidean distance between the parameter estimates in the last two iterations. For this particular data set (Figure 3.2), we set the threshold to be 10^{-4}*, it takes 283 iterations for VB to converge. Table 3.7 gives the posterior means under the VB approximations for all model parameters. We can see that, except for* $E[\pi_1]$*,* $E[\pi_7]$ *and* $E[\pi_{10}]$*, the expected value of mixing proportions for the the rest of the components are very close to zero. This implies that VB suggests 3 components for this data set. In fact the number of 1.25e-08 exactly equals* $\frac{a_j}{n+a_0}$*, where* $a_j = 0.00001$*,* $a_0 = 0.0001$ *and* $n = 400$*. Moreover, we can see from Table 3.7 the mean values of* μ_j *and* σ_j^2 *of component of 1, 7, 10, are reasonably close to the actual*

model parameters.

Table 3.7: The numerical results for the VB approximations in Example 3.3

$E[\pi_1]$	$E[\mu_1]$	$E[\sigma_1^2]$	0.297	-0.064	0.778
$E[\pi_2]$	$E[\mu_2]$	$E[\sigma_2^2]$	1.25e-08	-0.046	0.038
$E[\pi_3]$	$E[\mu_3]$	$E[\sigma_3^2]$	1.25e-08	0.765	0.065
$E[\pi_4]$	$E[\mu_4]$	$E[\sigma_4^2]$	1.25e-08	1.428	0.020
$E[\pi_5]$	$E[\mu_5]$	$E[\sigma_5^2]$	1.25e-08	1.813	0.011
$E[\pi_6]$	$E[\mu_6]$	$E[\sigma_6^2]$	1.25e-08	2.171	0.009
$E[\pi_7]$	$E[\mu_7]$	$E[\sigma_7^2]$	0.386	1.957	0.284
$E[\pi_8]$	$E[\mu_8]$	$E[\sigma_8^2]$	1.25e-08	3.609	0.055
$E[\pi_9]$	$E[\mu_9]$	$E[\sigma_9^2]$	1.25e-08	4.384	0.043
$E[\pi_{10}]$	$E[\mu_{10}]$	$E[\sigma_{10}^2]$	0.318	4.353	0.783

In RJMCMC method, the order of K is treated as a random variable. We use a discrete uniform$(1, \cdots, K_{max})$ with $K_{max} = 15$ as a prior. Each RJMCMC iteration contains an update on $\{\pi_j, \mu_j, \sigma_j^2\}_{j=1}^K$ and $\{z_i\}_{i=1}^n$ with an ordering constraint on μ_j, a splitting or combining move, and a birth or death move. The pseudo-code for RJMCMC is given in Algorithm 3. It obversely shows that in contrast to the simple arithmetic operations in the VB algorithm (Algorithm 2), each iteration in the RJMCMC algorithm demands non-trivial computational efforts. Furthermore, to ensure that the generated chain is indeed converged, we may have to run a large number of iterations for RJMCMC, for example, 1×10^6 iterations. However, it only takes 283 iterations for VB to converge. Therefore, the computational advantage of the VB method is clear.

For RJMCMC we run 2×10^6 iterations for this data set. The acceptance rates of splitting, combining, birth, and death are 0.1083, 0.1659, 0.0877, *and* 0.9055 *respectively. Figure 3.3(a) shows that the MCMC algorithm mixes well over K. Histogram (Figure 3.3(b)) shows there are several competing explanations of the data. The model with 3 components has the largest posterior probability (Table 3.8). However, there are also large posterior probabilities for other models, for example,* 2,4,5, *and* 6 *components. This can be explained as the effects*

Algorithm 3 RJMCMC algorithm for the mixtures of normal model

Initialization.

while $t <$ the iteration number **do**

1. Given the current value of K, generating samples for μ_j, $j = 1, \cdots, K$, and checking the ordering constraints on μ_j. If ordering constraints are not satisfied, discarding μ_j and using the values in previous iteration.
2. Given μ_j, generating samples for π_j,σ_j^2, for $j = 1, \cdots, K$, and $\{z_i\}_{i=1}^n$
3. Splitting one mixture component into two, or combining two into one
   (a) making a random choice between splitting or combining
   (b) For splitting
      i. Randomly choosing one component
      ii. Creating two new components based on the selected component. Checking the adjacent condition for the newly created components. If not satisfied, then the splitting is rejected
      iii. Reallocating all those observations y_i to the new components
      iv. Computing μ_i^*, σ_i^{*2}, π_i^*, $i = 1, 2$ for the new components.
      v. Computing the acceptance probability, and rejecting or accepting the splitting
   (c) For combining
      i. Randomly choosing two adjacent components
      ii. Merging two components to form a new one, and reallocating all those observations y_i to the new component
      iii. Computing μ^*, σ_i^{*2}, π^* for the new component
      iv. Computing the acceptance probability, and rejecting or accepting the combining
4. The birth or death of an empty component
   (a) making a random choice between birth or death
   (b) For birth
      i. Creating a new component with corresponding μ^*, σ^{*2}, π^*
      ii. Rescaling the existing π_j
      iii. Computing the acceptance probability, and rejecting or accepting the birth
   (c) For death
      i. Randomly choosing an empty component
      ii. Rescaling the existing π_j
      iii. Computing the acceptance probability, and rejecting or accepting the death

$t \Leftarrow t + 1$

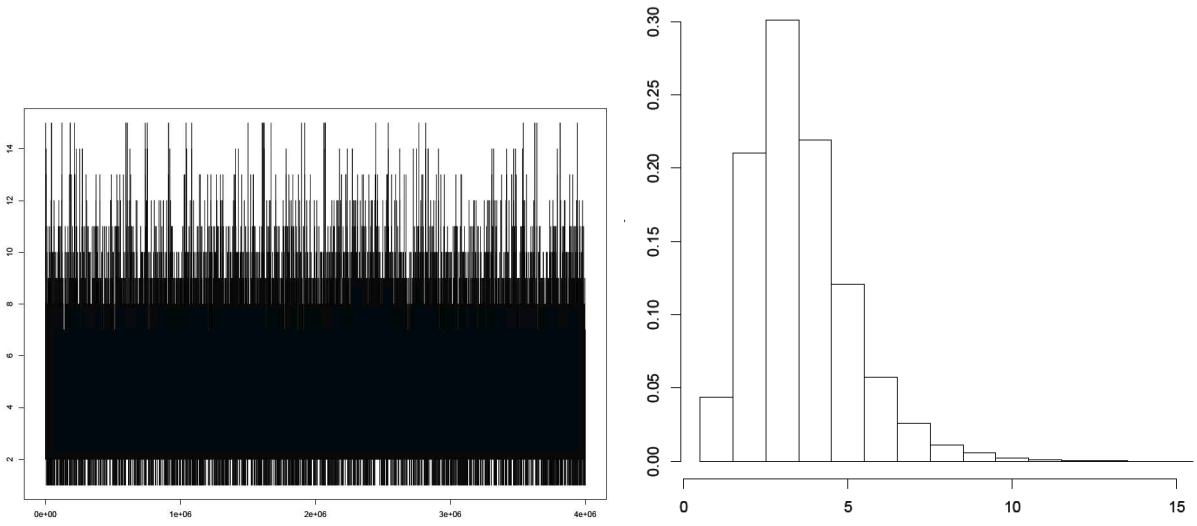
end while

of splitting moves and birth moves, mainly due to splitting moves since the acceptance rate of splitting is about 11%. *As the adjacent condition (see Algorithm 3), the newly created components would surround the original split component, and they would tend to close to each other in terms of their means.*

Given $K = 3$, *trace plots (Figure 3.4(a), 3.5(a), 3.6(a)) shows good mixing patterns within a fixed number of K. Histograms (Figure 3.4(b), 3.5(b)) still shows evidence of label switching, and therefore the moments computed by those samples might loses interpretability.*

Table 3.8: The posterior distribution of K given by the RJMCMC method in Example 3.3

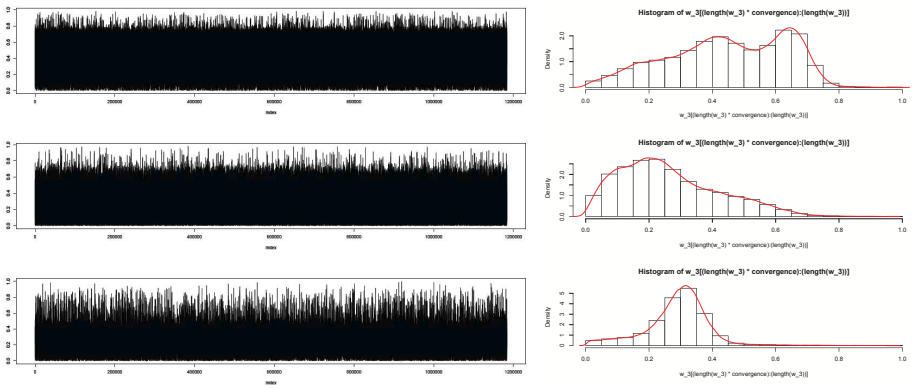
$K =$	1	2	3	4	5	6	7	8
$p(K\|y)$	0.044	0.211	0.301	0.219	0.121	0.057	0.026	0.011
$K =$	9	10	11	12	13	14	15	
$p(K\|y)$	0.006	0.002	0.001	< 0.001	< 0.001	< 0.001	< 0.001	



(a) A trace plot of K

(b) Histogram of K, last 10% samples

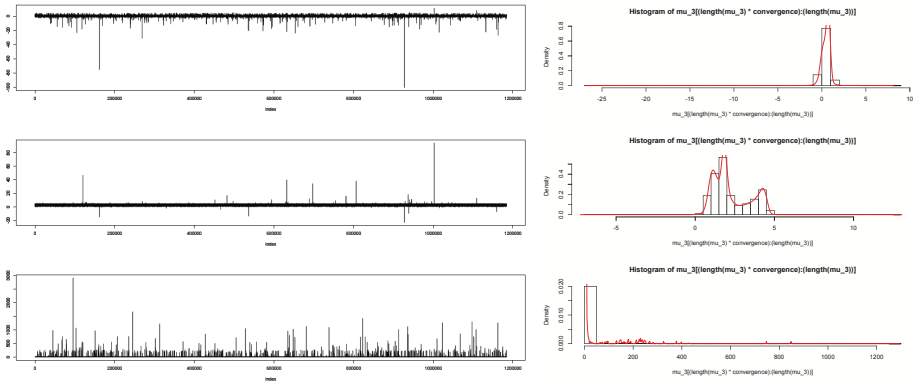
Figure 3.3: Samples of K generated by the RJMCMC in Example 3.3



(a) Trace plots of π_j, K=3

(b) Histograms of π_j, K=3, last 10% samples

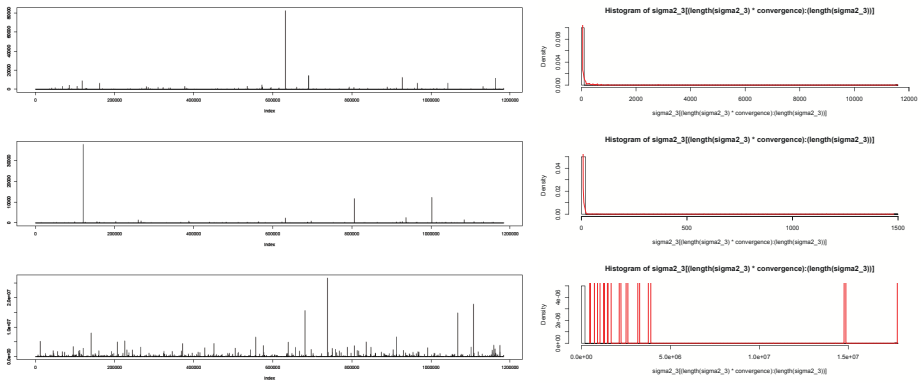
Figure 3.4: Samples of π_j generated by the RJMCMC method, when $K = 3$, in Example 3.3



(a) Trace plots of μ_j, K=3

(b) Histograms of μ_j, K=3, last 10% samples

Figure 3.5: Samples of μ_j generated by the RJMCMC method, when $K = 3$, in Example 3.3



(a) Trace plots of σ_j^2, K=3

(b) Histograms of σ_j^2, K=3, last 10% samples

Figure 3.6: Samples of σ_j^2 generated by the RJMCMC method, when $K = 3$, in Example 3.3

3.4 Some properties

In previous section, we have seen that VB can provide a good posterior mean structure approximation, and we show how we can use the posterior mean to make inference on model structures. In this section, we intend to discuss more its theoretical properties and give insight into its general applicability.

3.4.1 Local optimization

The convergence of the VB algorithm may only be to a local maximum. We illustrate this local convergence property by an example of a two component mixtures of normals problem, for which the variances and mixing proportion of each component are assumed known. The parameters of interests are the means of each component. For a demonstration purpose, we have the following model,

$$f(x) = 0.3\phi(x; \mu_1, 0.6) + 0.7\phi(x; \mu_2, 1.5), \tag{3.20}$$

where μ_1 and μ_2 are parameters of interest. Suppose that 100 samples are generated from (3.20) with $\mu_1 = -2$ and $\mu_2 = 4$ and this data set is shown in Fig 3.7.

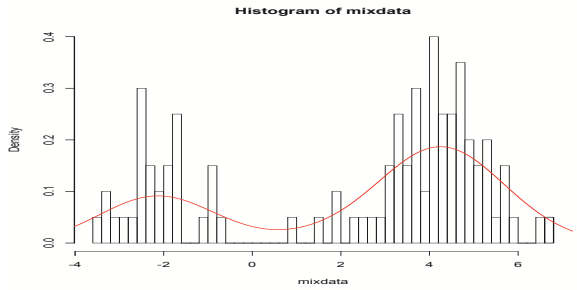


Figure 3.7: The simulated data of a two component mixtures of normals model.

We choose $N(c_j, d_j^2)$ as the priors for μ_j, where c_j and d_j^2 are hyper-parameters, for $j = 1, 2$. As we have already derived in Section 3.2.3, the VB approximations of the posterior distributions of μ_j are given by, $q_j(\mu_j) = N(c_j', d_j'^2)$. The function F is given by

$$F(q_1, q_2) = -\frac{1}{2}\sum_{j=1}^{2}\frac{(c_j' - c_j)^2}{d_j'^2} + \sum_{i=1}^{100}\log\left(\sum_{j=1}^{2} e^{-\frac{(x_i - c_j')^2}{2\sigma_j^2}}\right),$$

which is a function of c_j', $j = 1, 2$.

When choosing different starting points on the plane of (c_1', c_2'), for example, $(1, -4)$, $(9, -5.4)$, $(7.5, 9)$, $(2, 8.5)$, $(-4, -4)$, the function F (Figure 3.8) converges to one of two maxima, -180.92 at $(4.23, -1.58)$ and -112.00 at $(-2.16, 4.18)$.

3.4.2 Improper priors

Improper priors have been studied by many researchers (Jeffreys, 1961; Hartigan, 1964; Berger, 1985; Box et al., 1978; Bernardo and Smith, 1994). When using them, one must

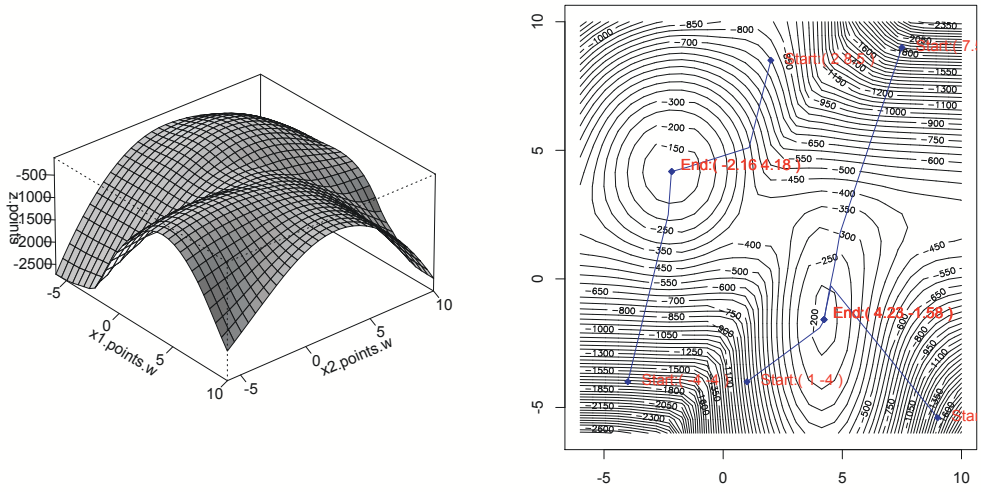


Figure 3.8: The left panel is the function of (c_1', c_2'). The right panel shows the different convergence paths starting from different points.

always check the posteriors are in fact proper. The following two cases show an improper prior can lead to an improper VB approximation.

Example 3.5 (Improper VB posteriors). *Section 3.2.2 provides the variational solution for the one-way random-effects model, where the VB approximation of* $q(\tau^2)$ *is given by*

$$q(\tau^2) \quad = \quad \frac{\pi(\tau^2)(\tau^2)^{-\frac{J}{2}} e^{-\frac{b}{\tau^2}}}{\int \pi(\tau^2)(\tau^2)^{-\frac{J}{2}} e^{-\frac{b}{\tau^2}} d\tau^2}$$

where $\pi(\tau^2)$ *is the prior distribution on* τ^2 *and* $b > 0$*. If we assume there are only two experiments, that is* $J = 2$*, and choose* $\pi(\tau^2) \propto (\tau^2)^2$*, then* $q(\tau^2)$ *is an improper density function.*

Example 3.6 (Failure to converge). *We still consider the one-way random-effects model, and apply this model to a real data set, the coagulation time data (Box et al., 1978) (Table 3.9), which contains coagulation time in seconds for blood drawn from 24 subjects randomly*

allocated to four different diets. As given in (3.9), under a full factorization, the VB approx-

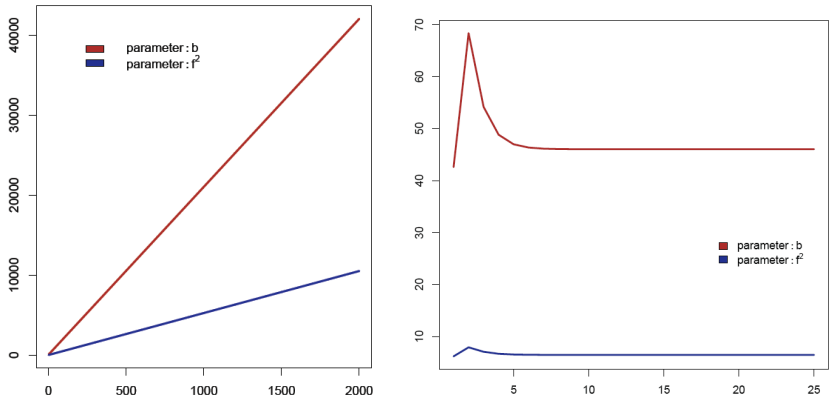
Table 3.9: The coagulation time data (Box, Hunter and Hunter, 1978)

Diet	Measurements
A	$62, 60, 63, 59$
B	$63, 67, 71, 64, 65, 66$
C	$68, 66, 71, 67, 68, 68$
D	$56, 62, 60, 61, 63, 64, 63, 59$

imations for $p(\tau^2|y)$ and $p(\mu|y)$ are given by

$$q(\tau^2) = IG(a, b), \qquad a = \frac{J}{2} - 1, b = \frac{1}{2}\sum_{j=1}^{J}((g_j - e)^2 + f^2)$$

$$q(\mu) = N(e, f^2), \qquad e = \frac{\sum_i^J g_j}{J}, f^2 = \frac{b}{aJ}$$



(a) Using improper priors (b) Using informative priors

Figure 3.9: The convergence of parameters b and f^2

Figure 3.9(a) shows the values of variational parameter of b and f^2 during the 2000 iterations, and the lack of convergence is obvious. We think the improper prior on μ causes the convergence failure. To see this, we use an informative prior for μ and keep non-informative

priors for τ^2 *and* $\log(\sigma^2)$*,*

$$\pi(\mu, \tau^2, \sigma^2) \propto \phi(\mu; s, t^2)\frac{1}{\sigma^2},$$

where is a normal density function. Thus, the VB approximation of $q(\tau^2)$ *and* $q(\mu)$ *are given by*

$$q(\tau^2) = IG(a, b), \qquad a = \frac{J}{2} - 1, b = \frac{1}{2}\sum_{j=1}^{J}((g_j - e)^2 + f^2)$$

$$q_(\mu) = N(e, f^2), \qquad e = \frac{\sum_i^J g_j}{J}, f^2 = \frac{1}{\frac{1}{t^2} + \frac{aJ}{b}}.$$

Due to the term $\frac{1}{t^2}$ *in* f^2*, both values of* b *and* f^2 *are converged after 18 iterations (see Figure 3.9(b)).*

3.4.3 Factorization effects

On one hand, factorization makes possible tractable integrals. However this means that the approximation is missing the actual posterior dependence, and there is a cost to this. We call this a *factorization effect*. Consider then the following two examples

Example 3.7 (Example 3.5 continued). *The convergence failure of variational parameters of* b *and* f^2 *in* $q(\tau^2)$ *and* $q(\mu)$ *in Example 3.5 can be interpreted alternatively as a factorization effect caused by using a full factorization. To see this, still using improper priors, a conditional factorization gives*

$$q(\tau^2) = IG(a, b); \qquad a = \frac{J}{2}, b = \frac{1}{2}\sum_{j=1}^{J}((g_j - e)^2 + k_j^2)$$

$$q(\mu|\tau^2) = N(e, f^2); \qquad e = \frac{1}{J}\sum_{j=1}^{J} g_j, f^2 = \frac{\tau^2}{J}$$

The variational parameter b converges after 18 iterations (Figure 3.10) .

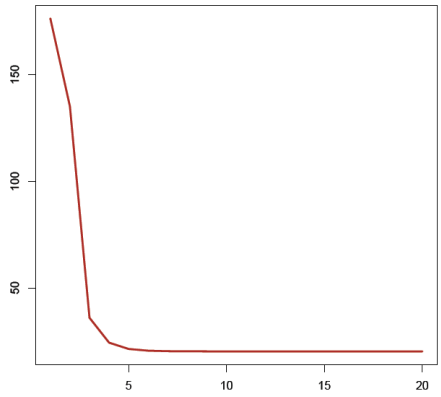


Figure 3.10: The convergence of parameters b

Example 3.8 Dependency. *An obvious problem in VB approximations is some of dependency intrinsic in posterior distributions would be lost due to factorization constraints. Let's revisit the finite mixture model introduced in Section 3.2.3, and consider a two component mixture of normals model,*

$$f(x|p, \mu_1, \sigma_1^2, \mu_2, \sigma_2^2) = p\phi(x; \mu_1, \sigma_1^2) + (1-p)\phi(x; \mu_2, \sigma_2^2)$$

where $p, \mu_1, \sigma_1^2, \mu_2, \sigma_2^2$ *are parameters of interest. We simulate a sample of size 200 from the model* $f(x) = 0.4\phi(x; 1, 1) + 0.6\phi(x; 3.5, 0.5^2)$*, and the priors are chosen as,*

$$\pi(p) = Beta(a, b), \quad \pi(\mu_i, \sigma_i^2) = N(\mu_i; c_i, \sigma_i^2/d_i^2)IG(\sigma_i^2; e_i, f_i), \text{ for } i = 1, 2,$$

where $a, b, c_i, d_i^2, e_i, f_i$ *are hyper-parameters. The posterior is given by*

$$p(p, \mu_1, \sigma_1^2, \mu_2, \sigma_2^2|x) \propto \pi(p) \prod_{i=1}^{2} \pi(\mu_i, \sigma_i^2) \prod_{j=1}^{n} f(x_j|p, \mu_1, \sigma_1^2, \mu_2, \sigma_2^2). \tag{3.21}$$

The correlation matrix is given in Table 3.10. It is obvious that under a full factorization the VB approximation of $q(p, \mu_1, \sigma_1^2, \mu_2, \sigma_2^2) = q(p)q(\mu_1)q(\sigma_1^2)q(\mu_2)q(\sigma_2^2)$ *can not capture these correlation structures.*

Table 3.10: The correlation matrix of simulated data, a two component mixtures of normals

	p	μ_1	σ_1^2	μ_2	σ_2^2
p	1	0.83	0.81	0.69	-0.60
μ_1	0.83	1	0.94	0.84	-0.66
σ_1^2	0.81	0.94	1	0.73	-0.76
μ_2	0.69	0.84	0.73	1	-0.50
σ_2^2	-0.60	-0.66	-0.76	-0.50	1

The loss of the correlation structure in VB approximations is an inevitable problem. It will be an interesting topic if we can recover the true posterior covariance by only using VB approximation. We propose three methods to achieve this goal. More details can be found in Chapter 4

3.5 Discussion

The present chapter applies variational methods to a variety of models, including: a normal random sample with unknown mean and variance, the one-way random-effects model, the Ising model, linear regression models, and finite mixture models. Through a series of numerical studies, we show that VB can provide good mean structure approximations; we also show how it is possible, using VB approximations, to approach model selection tasks such as determining the dimensionality of a model, or variable selection.

The chapter also explores general applicability and properties of VB. We show that the convergence of VB may only be to a local maximum. Improper priors might result

in improper approximations or convergence failures. Although factorization makes possible tractable integrals, it comes with a cost. We show factorization might result in convergence failures, and by definition, the dependence structure is distorted.

We are particularly interested in the last problem stated above. That is, how well the approximations can represent the true posteriors, especially the covariance structure, and whether we can estimate their actual covariance matrix by using only the information provided by the VB approximations. The next chapter will investigate these questions.

Chapter 4

Diagnostics for the VB Method

In Chapter 3 we see that the VB method has shown itself to be a powerful approximation method in many applications. This Chapter describes some diagnostics methods which can assess how good the VB approximation to the true posterior is, particularly with regards to its covariance structure. The methods proposed also allow us to generate simple corrections when the approximation error is large. It looks at joint, marginal and conditional aspects of the approximate posterior and shows how to apply these techniques in both simulated and real data examples.

4.1 Motivation

The present chapter looks at diagnostics tests to evaluate the quality of VB solutions. As is typical with diagnostic testing in statistics - think of diagnostic testing in regression analysis for example - we look at *necessary* conditions for adequacy. A VB solution may be inadequate from a number of perspectives. Here we list some of the most important. Firstly, by definition the posterior covariance structure is distorted, so both posterior variances and correlations can be wrong. Secondly, VB convergence is only local, hence it might miss other

'better' solutions. In particular it might focus on a single mode of a multimodal solution. Thirdly, there may be errors in higher order posterior moments, such as skewness or kurtosis. A given diagnostic test is designed to detect a particular kind of error, and the key idea of the present chapter is to make available a number of computationally fast diagnostics, with a computation time of the order of the VB itself, which target particularly common forms of inadequacy, specifically the first on the list above.

The essence of the VB method relies on making simplifying assumptions about the posterior dependence of a problem. This results in a high dimension integral being decomposed into a set of low dimensional ones which may be expected to be more tractable. The applications and examples shown in Chapter 3 have demonstrated that the VB method is very computational efficiency with respect to sampling based methods, for example, reversible jump MCMC (Green, 1995) may require many millions of iterations to obtain posterior samples for a finite mixture of Normals, however, the VB approximations can need only a few hundred. Moreover, in Section 3.3 we study the variational approximations to posterior mean structures, and show that VB can give good approximations and is good at finding overall structural features – such as the number of components in a mixture, and the variable selection in regression problem.

However through various numerical studies we observe that sometimes the posterior variance can be underestimated, and by definition the general posterior dependence structure is distorted. This underestimation of the variance has also been reported by other researchers, for example, Bishop (2006); Rue et al. (2009). This motivates the work in this Chapter to develop diagnostics to see how well the VB approximations represent the actual posterior distributions, and to some extent to provide corrections when these errors are large. We emphasize that these tests are only designed for these forms of error and may not detect errors of different kind.

We propose three diagnostics methods which only use the information obtained from VB

approximations. The first method looks at the the joint posterior distribution and attempts to find an optimal affine transformation which links the VB and true posteriors. The second method is based on a marginal posterior density approximation technique proposed by Tierney et al. (1989). Here we work in specific low dimensional directions to estimate true posterior variances and correlations. The third method based on a stepwise conditional approach allowing us to construct and solve a set of system of equations which lead to estimates of the true posterior variances and correlations.

We also proposes a novel method to calculate the variance of a marginal or conditional distributions of the posterior. This method uses an independent Metropolis-Hastings algorithm with the proposal kernel being given by the VB approximations. Instead of using the sample moments, the variance of the target distribution is computed by reading the acceptance probability of the generated MCMC chain.

This chapter is organized as follows. Section 2 presents the proposed methods in details. The applications of the methods on the simulated data and real-world data are provided in Section 3. Conclusions and discussions are available in section 4.

4.2 The three diagnostic methods

We consider the posterior distribution of a $p-$dimensional vector parameter $\theta = (\theta_1, \cdots, \theta_p)$, with density function $p(\theta|x)$ where x is an independent and identically distributed random sample. We denote the VB approximation by $q(\theta)$. We denote the true posterior mean by $\mu = (\mu_1, \cdots, \mu_p)$, the covariance matrix by Σ with variance of $\sigma_i^2, i = 1, \cdots, p$ and correlation coefficients by $\{\rho_{ij}\}$.

4.2.1 Optimal affine transformations of joint distributions

We denote the random vector associated with a VB approximation by η. We search for an optimal affine transformation of η (denoted by $A\eta + B$,where A is a $p \times p$ matrix and B is a $p \times 1$ vector) to get close to θ, the random vector associated with the true posterior. To find the values of A and B, we generate an independent random sample of size n from the VB approximations, denoted as $\{\eta_i\}_{i=1}^n$. The values of A and B are obtained by maximizing a likelihood function

$$Lik(A, B) := \prod_{i=1}^{n} (p(\theta_i|y; A, B)|\det(A)|), \tag{4.1}$$

where $\theta_i = A\eta_i + B$, and $\det(A)$ is the determinant of A and the corresponding estimates are denoted by $\hat{A}$ and $\hat{B}$. Sampling from $q(\eta)$ is typically straightforward since $q(\eta)$ usually has a factorization form of $q(\eta) = \prod_i^p q_i(\eta_i)$, and $q_i(\eta_i)$ often have a well-known distributional form. The maximization of (4.1) with respect to A and B is possible because it does not require the unknown normalizing constant of the posterior, $p(\theta|x)$.

For small or medium dimensional problems we only restrict the transformation matrix A to be general lower triangular, with the positive diagonal elements for identification reasons. For more complex problems sparser classes of matrices can be used, trading off the accuracy of the test with speed.

4.2.2 Marginal approximations

This method considers a projection of the vector parameter θ in a direction α, denoted as $\alpha^T\theta$. The variance of $\alpha^T\theta$ is given by $\alpha^T\Sigma\alpha$, which is a function of $\{\sigma_i^2\}_{i=1}^p$ and $\{\rho_{ij}\}$. If we have the projections in different directions, we can obtain a system of equations which can be easily solved to obtain the values of $\{\sigma_i^2\}_{i=1}^p$ and $\{\rho_{ij}\}$.

The key computation of this method is to calculate the value of the marginal variance. In order to be computationally efficient and exploit the VB solution we propose the following new method.

Suppose $p(\theta)$ is a target distribution and q is a proposal distribution. The independent Metropolis-Hastings (IMH) algorithm will produce a transition from $\theta^{(t)}$ to $\theta^{(t+1)}$ described in Algorithm 4. Theorem 4.1, proved in Appendix, establishes a connection between *the expected acceptance probability*, or the *expected acceptance rate (EAR)* and the closeness of the target distribution $p(\theta)$ and the proposal distribution q measured in Kullback-Leibler (KL) divergence.

Theorem 4.1 *If there exists a constant M such that $p(\theta) < Mq(\theta)$, then $KL(p||q) < \log(M)$ and the* expected acceptance rate (EAR) *is at least $\frac{1}{M}$ when the chain is stationary.*

Heuristically Theorem 4.1 states the closer the target and the proposal, the higher the EAR. It is obvious that when $p(\theta)$ and q are identical, the optimal acceptance rate equals to one. This result is different from other types of Metropolis-Hastings algorithms. For examples, for random-walk Metropolis-Hastings algorithm the optimal acceptance rate is close to 0.234 (Roberts et al., 1997); for Metropolis adjusted Langevin algorithms an overall acceptance rate is close to 0.574 (Roberts and Rosenthal, 1998).

Algorithm 4 Independent Metropolis-Hastings (IMH) algorithm

Given $\theta^{(t)}$

Step 1 propose $\eta^t \sim q$.

Step 2 Accept

$$\theta^{(t+1)} = \begin{cases} \eta^t & \text{with prob. } \min\left\{\frac{p(\eta^t)q(\theta^{(t)})}{p(\theta^{(t)})q(\eta^t)}, 1\right\}; \\ \theta^{(t)} & \text{otherwise.} \end{cases}$$

Motivated by this general result, first consider a special case in which the target distribution is a univariate normal with mean of μ and variance of σ_t^2 and the proposal distribution

is a normal with the same mean and variance σ_p^2 (assume $\sigma_p^2 > \sigma_t^2$). It can be shown that the EAR is monotone decreasing as the proposal variance of σ_p^2 increases. Conversely, it says that given a fixed value of proposal variance of σ_p^2, the value of the target variance σ_p^2 is one-to-one correspondence to the value of EAR. This implies by monitoring the acceptance probability, we can obtain the value of the target variance. A table of expected acceptance rate versus the value of target variance is given in the Appendix.

After this motivation let us consider the method in practice. Consider two basics facts: firstly posteriors approach to normality when sample size is large, and secondly VB provides good mean structure approximations. Hence we propose a new method to compute the target variance. We call it a VB Adjusted Independent Metropolis-Hastings method (VBAIMH). The variance of the target distribution is obtained by checking the acceptance rate for a standard normal kernel centred at the VB mean, being used as the proposal. In fact, the idea of using acceptance rates to compute the target variances can be further extended to using acceptance rates as a key diagnostic to how close the VB distribution is to the true posterior. More discussion can be seen shortly.

The new approach above has several advantages. First it does not require any particular tuning tricks to run the IMH algorithm. We only need the posterior mean values produced by the VB approximation to configure the proposal kernel. This is a significant advantage over other MCMC methods, for example, the ordinary Independent Metropolis-Hastings, or random walk Metropolis-Hastings, in which the implementation issues are the major concerns.

Secondly, this VB kernel allows the MCMC chain to locate the regions of high posterior probability more efficiently, since the proposal kernel is around the posterior mode, at least locally; then we only need a short chain to compute acceptance rates. This is another significant advantage over the other MCMC methods, where the computational cost can be a big concern. While the IMH algorithm is well known to perform poorly in high dimensions

(Geweke, 1989), in this method we are sampling from a univariate distribution defined by the projections.

Third, when an acceptance rate is low, the generated sample may not represent the target distribution well, and the moments calculated upon these samples can be wrong. However, a low acceptance rate on its own reflects a significant discrepancy between the target and the proposal.

It is worth noting that when the projection is along the direction of a single parameter of θ_i: that is, for example, we set $\alpha = (1, 0, \cdots, 0)$, the VBAIMH can provide a possible mechanism to calibrate the EAR as a diagnostic tool to measure inaccuracy in the marginal approximations of VB, by using $q(\theta_i)$ directly as the proposal distribution. When the acceptance rate is low, it clearly indicates the approximation will be inaccurate. Thus it gives us two uses: firstly a diagnostics tool in the general case and secondly it is possible to give a correction to the VB approximation. More discussion can be found in Section 4.4.

The above idea can be further extended to more general situations, where the diagnostics are targeted to more specific errors. For example, if only a subset of parameters is of immediate concern, which is of particular usefulness in the high dimensional problems.

To run an IMH algorithm we also need to know the density function of $\alpha^T\theta$, at least proportional to a normalizing constant. Tierney et al. (1989) proposed an elegant marginal approximation of this posterior distribution.

Suppose the parameter of interest is $\omega = g(\theta)$, where g is a continuous real-valued function on R^p. The posterior distribution of $p(\omega|x)$ can be approximated as follows

$$p(\omega|x) \propto \frac{\hat{p}(\omega|x)}{|R_\omega|^{1/2}(b_\omega^T R_\omega^{-1} b_\omega)^{1/2}}, \tag{4.2}$$

where

$$\begin{aligned}\hat{p}(\omega|x) &= \sup_{\theta:g(\theta)=\omega} p(\theta|x), b_\omega = \frac{\partial g(\theta)}{\partial \theta}\Big|_{\theta=\theta_\omega}, \\ R_\omega &= \frac{\partial^2 \log p(\theta|x)}{\partial \theta \theta^T}\Big|_{\theta=\theta_\omega},\end{aligned}$$

and θ_ω conditionally maximizes $p(\theta|x)$ with respect to θ for each given ω.

4.2.3 The conditional stepwise method

The unknown quantities in a true posterior covariance matrix Σ are $\{\sigma_i^2\}_{i=1}^p$ and $\{\rho_{ij}\}$ and the difference between these and the VB versions which we are using as our diagnostics. The stepwise method looks at conditional distributions constructed from the true posterior and compares them to ones based on the VB solution. Algorithm 5 gives a description on the proposed method. The three steps can be explained as follows.

Step 1 uses a linear transformation of $Y = Q\theta$ to scale the variances of $\{\sigma_i^2\}_{i=1}^p$ to be the ratios of $\{\sigma_i^2\}_{i=1}^p$ over their variational estimations, and uses a linear transformation of $Z = MY$ to further scale these ratios to be one, which only leaves $\{\rho_{ij}\}$ in Σ to be found.

Step 2 finds a series of conditional bivariate random vector U_{ij}, to which the eigenvalues of their covariance matrix can be computed numerically after a rotation.

Step 3 constructs a system of equations of $f_k(\{\rho_{ij}\})$ by linking the analytical expression of the correlation coefficient for the conditional bivariate U_{ij}, obtained based on the posterior normality (when sample size is large), to their numerical values of r_k, obtained by using the relationship between eigenvalues, and variances and correlation coefficients in a bivariate covariance matrix. The values of $\{\sigma_i^2\}_{i=1}^p$ can be obtained by reversing Step 1.

The key computations in Algorithm 5 involve computing the values of a univariate conditional or marginal variances, that is $m_i^2, i = 1, \cdots, p$ in Step 1 and $\lambda_{k,1}^2$ and $\lambda_{k,2}^2$,

Algorithm 5 The stepwise method

Step 1

- Define $Y = Q\theta$ and $\mu^s = Q\mu = (\mu_1^s, \cdots, \mu_p^s)$, where $Q = \left(\frac{1}{var_{q_i}(\theta_i)}\right)$ is a diagonal matrix. Denote $Y_i|Y_{-i}$ as the conditional Y_i conditioning on $Y_j = \mu_j^s, j \neq i$.
- Denote $m_i^2 = var(Y_i|Y_{-i})$. Obtain $m_i^2, i = 1, \cdots, p$, numerically.
- Define $Z = MY$, and $\mu^{ss} = M\mu^s = (\mu_1^{ss}, \cdots, \mu_p^{ss})$, where $M = \left(\frac{1}{m_i}\right)$ is a diagonal matrix. .

Step 2

- Denote $U_{ij} = Z_{ij}|Z_{-ij}, i \neq j$ as the conditional bivariate (Z_i, Z_j) conditioning on $Z_t = \mu_t^{ss}, t \neq i, j$.
- Let $R = \begin{pmatrix} \cos(\frac{\pi}{4}) & -\sin(\frac{\pi}{4}) \\ \sin(\frac{\pi}{4}) & \cos(\frac{\pi}{4}) \end{pmatrix}$. Define $V_k = (V_{k,1}, V_{k,2}) = RU_{ij}$, where $k = 1, \cdots, \frac{p(p-1)}{2}$ for all the pair of $i \neq j$. Denote $\lambda_{k,1}^2 = var(V_{k,1})$ and $\lambda_{k,2}^2 = var(V_{k,2})$.
- Obtain $\lambda_{k,1}^2$ and $\lambda_{k,2}^2$, $k = 1, \cdots, \frac{p(p-1)}{2}$, numerically.

Step 3

- Based on the posterior normality assumption, compute the correlation coefficient for the conditional bivariate U_{ij}, and denote it as $f_k(\{\rho_{ij}\})$, where $k = 1, \cdots, \frac{p(p-1)}{2}$, for all the pair of $i \neq j$.
- Compute $r_k = \left(\frac{\lambda_{k,1}^2}{\lambda_{k,2}^2} - 1\right) / \left(\frac{\lambda_{k,1}^2}{\lambda_{k,2}^2} + 1\right), k = 1, \cdots, \frac{p(p-1)}{2}$. Solve the system of equations of $f_k(\{\rho_{ij}\}) = r_k$ to obtain the value of $\{\rho_{ij}\}$.
- Based on the posterior normality assumption, compute the conditional variance of $var(Y_i|Y_{-i})$, and denote it as $g_i(\sigma_i^2)$, where $i = 1, \cdots, p$. Solve the equation $g_i(\sigma_i^2) = m_i^2$ to obtain the value of σ_i^2.

$k = 1, \cdots, \frac{p(p-1)}{2}$ in Step 2. These values can be computed by the VBAIMH method. The definition of r_k in Step 3 derives from the following fact. For a bivariate distribution, suppose the variances are σ_1^2 and σ_2^2 and correlation is ρ. The eigenvalues of covariance matrix are given as $\lambda = \frac{(\sigma_1^2+\sigma_2^2) \pm \sqrt{(\sigma_1^2-\sigma_2^2)^2+4\rho^2\sigma_1^2\sigma_2^2}}{2}$. When $\sigma_1^2 = \sigma_2^2$, the eigenvalues are given by $\lambda_1 = (1+\rho)\sigma^2$, and $\lambda_2 = (1-\rho)\sigma^2$. Then it is easy to show that

$$\rho = \left(\frac{\lambda_1}{\lambda_2} - 1\right) / \left(\frac{\lambda_1}{\lambda_2} + 1\right). \tag{4.3}$$

To illustrate the stepwise method, we give a three-dimension example, where the derivation is given step by step, and the ideas of the method is brought in gradually. The interested readers can find the example in the Appendix.

4.3 Numerical studies

We will work through four models with simulated or real datasets to demonstrate the proposed methods. For each model we compute its variational approximation, obtained by minimizing the the Kullback-Leibler (KL) divergence. The distributional families of these approximations range widely; for example in the cases considered here they are Normal, t, Beta, Inverse Gamma, and Dirichlet.

We start with a very basic illustrative example: a large sample multivariate normal case with simulated data. The second example looks at a Normal random sample with unknown mean and variance with a real data set. In this case posterior normality is not assumed, showing that normality is not needed for the methods to have power. We thirdly consider a two-component mixture of Normals model. In addition, in Chapter 6, we apply the methods to the regime-switching lognormal model, which can be considered a high-dimensional case with six interest parameters, and 528 latent nuisance parameters. These models present a wide range of complex dependence structures, and MCMC methods have been intensively

studied with them; these models will provide good testimony for the proposed methods. For the regime-switching lognormal model, we used the real data set of the TSX monthly total return index in the period from January 1956 to December 1999, which contains 528 observations in total, see for discussion Hardy (2001, 2002); Hartman and Heaton (2011).

4.3.1 Multivariate normal distributions

We consider a 3-dimension vector parameter $\theta = (\theta_1, \theta_2, \theta_3)$, and the posterior distribution of θ and its VB approximation are all assumed to be a multivariate Normal distribution, and these two normal distributions have the same mean values. For the illustration purposes we will arbitrarily choose the values of the covariance matrices for the true posterior and the VB approximation. The following is an example; the actual variances are chosen to be 0.1^2, 1.3^2, and 4^2 with correlation 0.51 between θ_1 and θ_2, 0.37 between θ_1 and θ_3, and -0.3. In the VB approximation, the variances are assumed to be $\frac{0.1^2}{2.2}$, $\frac{1.3^2}{5.1}$, and $\frac{4^2}{6.9}$, and all correlation are assumed 0.

Our goal is that given the covariance of VB approximations and the density function of the posterior distribution up to a normalizing constant we will compute the true covariance structure, more precisely, to find the values of correlation coefficients of 0.51, 0.37, and -0.3 and the ratios of the posterior variances versus the VB variances, 2.2, 5.1, and 6.9.

We first apply the affine transformation method. A sample of size of 600 is generated from the VB distribution. We restrict the transformation matrix A to be a lower triangular matrix with positive diagonal elements. There are no constraints on the three parameters in the translation vector B. Maximizing the posterior probability over these 9 parameters can be done with the Newton's method or standard search methods. The resulted $\hat{A}$ and $\hat{B}$ are

given by

$$\hat{A} = \begin{pmatrix} 1.527 & 0.000 & 0.000 \\ 10.498 & 2.007 & 0.000 \\ 23.601 & -3.627 & 1.918 \end{pmatrix}; \hat{B} = \begin{pmatrix} -0.004 \\ -1.1089 \\ 1.758 \end{pmatrix}$$

Given $\hat{A}$ and $\hat{B}$, the estimated $\hat{\Sigma}_p$ can be computed by $\hat{A}\ \Sigma_v\ \hat{A}^T$. Table 4.1 gives the the actual and estimated variance ratios and correlation coefficients.

Table 4.1: 3 diagnostics methods for a multivariate Normal

	Variance ratios			Correlation coeff.		
	θ_1	θ_2	θ_3	ρ_1	ρ_2	ρ_3
The true values	2.2	5.1	6.9	0.51	0.37	-0.30
Affine transformations	2.33	5.54	6.65	0.52	0.41	-0.24
Marginal approximations	2.54	4.80	7.04	0.56	0.32	-0.30
Stepwise method	2.21	5.04	6.72	0.50	0.40	-0.28

Second, we use the method using marginal approximations. It requires 6 projections. We denote a projection direction as $\alpha = (\alpha_1, \alpha_2, \alpha_3)$, and the marginal variance along the projection direction as l. Thus, we can obtain a polynomial equation involving σ_i^2 and ρ_{ij} given by

$$\alpha_1^2\sigma_1^2 + \alpha_2^2\sigma_2^2 + \alpha_3^2\sigma_3^2 + 2\alpha_1\alpha_2\rho_{12}\sigma_1^2\sigma_2^2 + 2\alpha_1\alpha_3\rho_{13}\sigma_1^2\sigma_3^2 + 2\alpha_2\alpha_3\rho_{23}\sigma_2^2\sigma_3^2 = l$$

For each direction we simulate a sample of size 6000, and we record the proportion of the samples, which are accepted in each VBAIHM iteration for the last 3000 samples. We use this proportion as an empirical acceptance rate, and then obtain the values of l (Table 4.2) by checking the EAR table. Solving the 6 polynomial equations, we obtained the following values; $\sigma_1 = 1.59$, $\sigma_2 = 2.19$, $\sigma_3 = 2.65$, $\rho_{12} = 0.56$, $\rho_{13} = 0.32$, $\rho_{23} = -0.30$. The variance ratios are given in the third column in Table 4.1.

Table 4.2: The marginal approximation method for a multivariate normal

direction	Acc. rate	EAR reading: l_i
$\frac{1}{3}(1,1,1)$	0.790	1.94
$\frac{1}{3}(1,-1,1)$	0.805	1.86
$\frac{1}{3}(1,1,-1)$	0.764	2.12
$\frac{1}{3}(1,-1,-1)$	0.764	0.47
$\frac{1}{3}(1,0.5,1)$	0.865	1.52
$\frac{1}{3}(0.5,1.5,1)$	0.791	1.94

Finally, we work through the conditional stepwise method. The notation used here follow that given in Algorithm 5. For Step 1 and 2, we simulate a sample of size 6000 and use the last 50% sample points to calculate the acceptance rate, and the conditional variance is obtained from EAR table readings. All numerical results are given in Table 4.3. Solving the polynomial equations obtains the values for $\rho_{12} = 0.50$, $\rho_{13} = 0.40$ and $\rho_{23} = -0.28$ and the variance ratios. These values are given in the fourth row in Table 4.1.

Table 4.3: The stepwise method for a multivariate Normal

	Marginal variance	Acceptance rate	EAR table readings	Eigenvalue ratio
Step 1	m_1^2	0.989	0.97	-
	m_2^2	0.730	2.42	-
	m_3^2	0.620	3.6	-
Step 2	$\lambda_{1,1}^2, \lambda_{1,2}^2$	0.832, 0.651	0.588, 3.220	-
	$\lambda_{2,1}^2, \lambda_{2,2}^2$	0.848, 0.681	0.602, 2.800	-
	$\lambda_{3,1}^2, \lambda_{3,2}^2$	0.718, 0.852	2.520, 0.625	-
Step 3	r_1	-	-	0.691
	r_2	-	-	0.646
	r_3	-	-	-0.603

4.3.2 Normal random sample

In this example, we consider a real dataset which contains 1033 records of weights for some Major League Baseball (MLB) Players (Onge et al., 2008). Figure 4.1(a) is a histogram of the dataset. This plot suggests that it may be reasonable to model the data by a normal distribution with the mean μ and variance σ^2. We are interested in making inferences on μ and σ^2. In a Bayesian setting, we consider the priors as $\mu \sim N(\gamma, \eta^2)$ and $\sigma^2 \sim \text{IG}(\alpha, \beta)$, where IG denotes the inverse Gamma distribution. This setting has been discussed in Section 3.2.1. The joint posterior distribution for μ and σ^2 is given by

$$p(\mu, \sigma^2|y) \sim \left(\frac{1}{\sigma^2}\right)^{-(\frac{n}{2}+\alpha+1)} \exp\left(-\frac{1}{\sigma^2}\left(\frac{S^2}{2}+\beta+\frac{n(\mu-\bar{y})}{2}\right)-\frac{\mu-\gamma}{2\eta^2}\right),$$

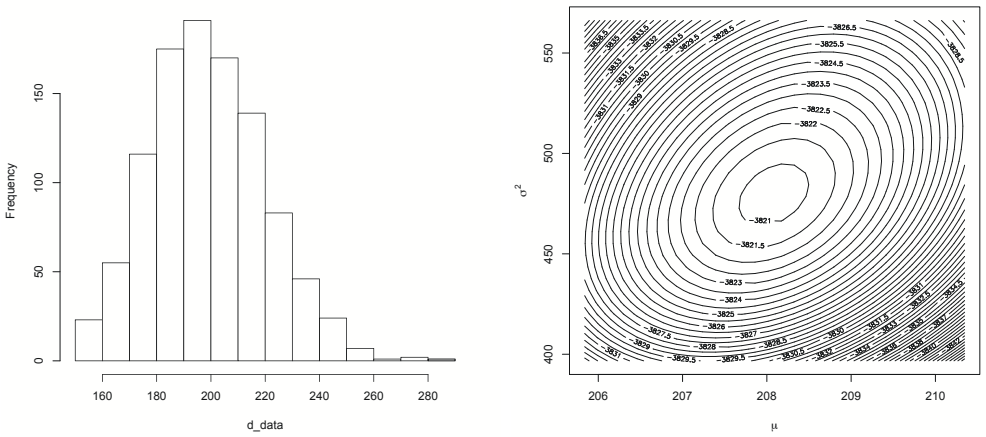
where $\bar{y}$ is the sample mean and S^2 is the total sum of squares of y. The values of hyper-parameters are chosen to be $\alpha = 2$, $\beta = 440.64$, $\gamma = 221.86$, and $\eta^2 = 1$, where the values for β and γ derives from the mean and variance of the dataset respectively. Figure 4.1(b) is a contour plot for the posterior distribution up to the normalization constant. The contour plot shows that a bivariate normal distribution may be reasonable as an approximation to the actual posterior distribution.

The variational solution for the marginal distribution of μ and σ^2 are given in (3.4) and (3.5). VB converges after 14 iterations, taking about 0.05 seconds on a MacBook computer (2 GHz Intel Core 2 Duo processor, 1 GB 667 MHz SDRAM, and Mac OS X 10.4.11).

The distributions for $q(\mu)$ and $q(\sigma^2)$ are given as follows;

Table 4.4: The marginal distributions of VB approximations

Parameter	Distribution (VB)
μ	$N(208.09, 0.32)$
σ^2	$\text{IG}(518.50, 249154.70)$



(a) The histogram of the dataset

(b) The contour for the posterior of μ and σ^2

Figure 4.1: MLB player weights dataset and the posterior contour

Table 4.5: Posterior mean and covariance for the MLB player weights dataset

	Posterior mean: (μ, σ^2)	Posterior covariance: (μ, σ^2)
Gibbs samples	$(208.10, 481.79)$	$\begin{pmatrix} 0.60^2 & 0.34 \times 0.60 \times 22.64 \\ 0.34 \times 0.60 \times 22.64 & 22.64^2 \end{pmatrix}$
VB approx.	$(208.09, 481.46)$	$\begin{pmatrix} 0.56^2 & 0 \\ 0 & 21.18^2 \end{pmatrix}$
Ratios	(1, 1)	(1.13, 1.14)

As a standard comparison, we run a Gibbs sampler, and simulate a sample of size 10^5 from the posterior distribution. Table 4.5 gives a comparison on posterior mean and posterior covariance estimated by the VB approximation and the MCMC sample moments. The third row in Table 4.5 gives the ratios of posterior mean and the ratios of posterior variance. We can see that the means estimated by both methods are almost identical. However the variances approximated by VB are slightly underestimated, (as expected from our discussion above) and it is obvious VB distorts the posterior dependence structure.

We applied the three proposed methods to this problem. All the setting and routines

used here are very similar to the those used in the previous example. The numerical results produced by each method are given in the Appendix. The final results for three methods are given in Table 4.6. We see that all of the methods perform well.

Table 4.6: 3 diagnostics methods for the normal model with MLB Players weights data

	Variance ratios		Correlation coeff.
	μ	σ^2	ρ
Gibbs	1.13	1.14	0.34
Affine	1.09	1.13	0.30
Marginal	1.08	1.20	0.35
Stepwise	1.13	1.14	0.35

For this normal random sample model, MCMC methods, for example, a Gibbs sampler, may be easy to implement, and also can be computationally efficient. We use this model only as an example to demonstrate the performance of the proposed diagnostics methods. The advantages of the VB method will be more obvious when it is applied to models with complex dependence structure.

4.3.3 Finite mixture models

Our last example considers a finite mixture model. Let's revisit the example of a two-component mixtures of normals given in Example 3.8. The VB method converges after 122 iterations. The approximation distributions are given in Table 4.7.

The posterior mean and posterior covariance estimated by the VB approximations and by the MCMC sample moments are given in Table 4.8. The ratios of posterior means and the ratios of posterior variancesare given in the last row in Table 4.8. These ratios indicate that the means estimated by both methods are almost identical. However VB underestimated the actual posterior variances, and again strongly distorts the correlation structure.

Table 4.7: The marginal distributions of VB approximations

Parameter	Distribution (VB)
π	Beta(167.35, 232.67)
$\mu_1 \vert \sigma_1^2$	N(1.13, σ_1^2/168.34)
σ_1^2	IG(85.66, 76.56)
$\mu_2 \vert \sigma_2^2$	N(3.57, σ_2^2/233.66)
σ_2^2	IG(118.33, 50.31)

Table 4.8: Posterior means and covariance for the 2-component mixtures of Normals

	Post. mean: (π, μ_1, μ_2, σ_1^2, σ_1^2)	Post. covariance: (π, μ_1, μ_2, σ_1^2, σ_1^2)
Gibbs samples	(0.418, 1.126, 3.5661, 0.885, 0.430)	variance (0.00208, 0.01902, 0.00349, 0.03696, 0.00431)
		correlation coeff. (0.83, 0.81, 0.94, 0.69, 0.84, 0.73, −0.60, −0.66, −0.76, −0.50)
VB approx.	(0.418, 1.131, 3.571, 0.904, 0.429)	variance (0.00061, 0.00537, 0.00183, 0.00977, 0.00158)
		correlation coeff. (0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00)
Ratios	(1.000, 0.996, 0.999, 0.979, 1.003)	(3.42, 3.54, 1.90, 3.78, 2.73)

The proposed methods are applied to this mixture problem. For the VBAIMH method, we generate 4000 samples to compute acceptance rates. The sample size is much smaller than is used in the Gibbs sampler, to which we use a large sample size to ensure the chain has in fact converged.

In addition, we note that in this example we have targeted the most general form of linear transformation to correct any inadequacy in posterior mean and variance. In fact

a diagnostic test can be designed to be targeted at more special concerns. For example the matrix A in the transformation might be restricted to be of a particular class such as diagonal or banded. This would be particular useful in high dimensional problems where the dimension of the space of A could become problematic.

All numerical results associated with each method are given in the Appendix. The final results are given in Table 4.9.

Table 4.9: 3 diagnostics methods for the two-component mixtures of Normals

	Variance ratios					Correlation coeff.									
	π	μ_1	μ_2	σ_1^2	σ_2^2	ρ_1	ρ_2	ρ_3	ρ_4	ρ_5	ρ_6	ρ_7	ρ_8	ρ_9	ρ_{10}
Gibbs	3.42	3.54	1.90	3.78	2.73	0.83	0.81	0.94	0.69	0.84	0.73	−0.60	−0.66	−0.76	−0.50
Affine	3.64	4.51	3.53	3.35	2.62	0.76	0.68	0.71	0.66	0.73	0.59	−0.61	−0.61	−0.63	−0.50
Marginal	5.56	4.73	2.37	4.15	2.76	0.63	0.59	0.87	0.47	0.82	0.76	−0.46	−0.66	−0.71	−0.54
Stepwise	3.06	4.12	2.70	3.21	2.56	0.71	0.63	0.65	0.64	0.70	0.53	−0.58	−0.58	−0.60	−0.47

4.4 Discussion

The variational method essentially provides posterior marginal approximations, which can be inaccurate in a number of ways. The present chapter aims to provide fast and easy-to-use diagnostics, which mainly target on inadequacy in the covariance structure. From the above numerical studies we can see all three methods can provide both diagnostics showing the quality of the VB approximation and also, in these examples, good estimates on the actual posterior variances and correlations. These methods are easy to use. They are free of any sophisticated tuning techniques or special expertise, and fast compared with the traditional sampling based methods.

This chapter introduces a novel way to use acceptance rates. The idea is that the acceptance rate can act as a key diagnostic to how close the VB distribution is to the true posterior.

As discussed in Section 4.2.2, EAR could be calibrated as a diagnostics tool to measure the inadequacy in marginal approximations by using VB approximations directly as the proposal distributions in VBAIMH. For the situation when posteriors depart from normality, a low acceptance rate still indicate an inaccurate approximation. However, to quantify a particular form of inaccuracy, a single value of EAR may be diluted by the confounding of many factors: inadequate variance, inadequate skewness, inadequate tail behaviour. In the further research, separating these confounding factors would be a necessary step toward measuring a special form of inadequacy.

For high dimensional problems, the three proposed methods can be designed to target more specific situations. For example, the covariance matrix might be sparse; a subset of the parameters might be of immediate concern. As discussed in Section 4.2.2, Section 4.3.3, the three methods can offer different strategies to address the special form of diagnostics. In the affine method, the transformation matrix A might be restricted to be a particular class. In the marginal method, the projections might be set to particular directions. In the stepwise method, the steps might be applied to a subset of the parameters, conditioning on other parameters.

For each individual method, we have some further comments. The affine transformation based method relies on using approximate linear relationship between the VB approximation and the actual posterior as a diagnostics and potentially a correction. In cases where, for example, strong skewness is present in the posterior the correction will of course not be exact, but it will still be a useful diagnostics tool.

As Leonard et al. (1989) point out, the marginal approximation of Tierney et al. (1989), primarily justified by asymptotically $n \to \infty$, might be insufficient for finite n; they also show a number of examples in which the method of Tierney, Kass, and Kadane introduces excessive skewness in the marginal approximations. We also find the inadequacy in the method of Tierney, Kass, and Kadane in our numerical studies. For example: Table 4.10

gives the values of 15 marginal variances along 15 directions in the mixture problem. The first row is the analytical results calculated from the covariance matrix obtained from a Gibbs samples. The second row is based on the posterior marginal approximation. We can see some discrepancy between this two sets of numbers. Even though, the method based on the marginal approximation still works well. Leonard et al. (1989) proposed a refinement on the method of Tierney et al. (1989), that could be considered in future work.

Table 4.10: The marginal variances in the marginal approx. method for mixtures of Normal

	l_1	l_2	l_3	l_4	l_5	l_6	l_7	l_8	l_9	l_{10}	l_{11}	l_{12}	l_{13}	l_{14}	l_{15}
Gibbs	3.42	0.59	1.32	0.80	6.63	0.92	0.67	0.66	0.5	2.51	0.65	0.71	1.98	0.44	2.99
Approx.	3.70	1	1.85	1.31	4.82	1.39	1.05	0.91	0.79	3.38	0.92	1.13	2.48	0.64	2.96

Another potential concern when using the marginal approximation method is that we need to perform a constrained maximization at each sampling step. The maximization can often be completed in straightforward fashion such as the standard Newton's method or search methods. However a optimization at each sampling step may affect the computational efficiency.

The VBAIMH provides a fast means to calculate the variance of the target distribution. When using this approach two particular issues arise. First, as discussed previously when the target variance of σ_t^2 is greater than the proposal variance of σ_p^2, the EAR is monotone decreasing as σ_t^2 increases. Similarly, when $\sigma_t^2 < \sigma_p^2$ the EAR is also monotone decreasing as σ_t^2 decreases. This means that we need to determine if $\sigma_t^2 < \sigma_p^2$ or not, before determining the value of σ_t^2. In practice, we can assume $\sigma_t^2 > \sigma_p^2$ and pick a value of σ_t^2 from the EAR table, and then use this new value as the proposal variance and run the IMH again. If the new acceptance rate is close to one or increases, this means $\sigma_t^2 > \sigma_p^2$ otherwise $\sigma_t^2 < \sigma_p^2$. If $\sigma_t^2 < \sigma_p^2$ is the case, the true value of σ_t^2 is the reciprocal of the value read from the EAR table.

Second, if the approximate normality of posterior distributions does not hold well, for example in the cases where strong skewness is present, the variance read from the EAR table will confound these non-normality effects and will deviate from the true value. As discussed above, when we pick a value of σ_t^2 from the EAR table and use this new value as the proposal variance to run another IMH algorithm, if the new acceptance rate is not close to one, this implies that the normality does not hold well, usually because of skewness. When this happens, we need to adjust the value read from the EAR table. We usually scale the readings as $c\sigma_t^2$. Based on our various numerical studies, a reasonable choice on the scale c is 0.85.

4.5 Appendix

4.5.1 Proof of Theorem 4.1

Proof: The expected acceptance rate is given by:

$$E[\min\left\{\frac{p(\eta^t)q(\theta^{(t)})}{p(\theta^{(t)})q(\eta^t)},1\right\}]=2\int I\left(\frac{p(\eta)q(\theta)}{p(\theta)q(\eta)}>1\right)p(\theta)q(\eta)d\theta d\eta$$
$$\geqslant 2\int I\left(\frac{p(\eta)q(\theta)}{p(\theta)q(\eta)}>1\right)p(\theta)\frac{p(\eta)}{M}d\theta d\eta=\frac{1}{M}$$

where I is the indicator function. Moreover, if $p < Mq$, it is straightforward to show $\mathrm{KL}(p||q) < \log(M)$.

4.5.2 Stepwise method: a 3-dimension example

All the notations used here are defined in Algorithm 5. We consider a vector parameter $\theta = (\theta_1, \theta_2, \theta_3)$, and denote its posterior variance and correlation as σ_1^2, σ_2^2 and σ_3^2, and ρ_1, ρ_2 and ρ_3

Step 1. Define $s_i^2 = \sigma_i^2 / var_{q_i}(\theta_i), i = 1, 2, 3$. Then, by the linear transformation of $Y = Q\theta$, the variances of Y are given by s_1^2, s_2^2, and s_3^2 respectively, with the correlation coefficients of ρ_1, ρ_2 and ρ_3 unchanged. Based on the posterior normality conditions, the conditional variance of $var(Y_1|Y_2, Y_3)$, $var(Y_2|Y_1, Y_3)$, and $var(Y_3|Y_2, Y_1)$ are given respectively

$$var(Y_1|Y_2, Y_3) = \left(1 - \frac{\rho_1^2 + \rho_2^2 - 2\rho_1\rho_2\rho_3}{1 - \rho_3^2}\right) s_1^2 = m_1^2 \tag{4.4}$$

$$var(Y_2|Y_1, Y_3) = \left(1 - \frac{\rho_1^2 + \rho_3^2 - 2\rho_1\rho_2\rho_3}{1 - \rho_2^2}\right) s_2^2 = m_2^2 \tag{4.5}$$

$$var(Y_3|Y_2, Y_1) = \left(1 - \frac{\rho_2^2 + \rho_3^2 - 2\rho_1\rho_2\rho_3}{1 - \rho_1^2}\right) s_3^2 = m_3^2 \tag{4.6}$$

The value of m_1^2, m_2^2, and m_3^2 can be obtained numerically by using the VBAIMH algorithm. After linear transformation of $Z = MY$, the variance of Z_1, Z_2, and Z_3 are given respectively by

$$\begin{aligned}
var(Z_1) &= \frac{1 - \rho_3^2}{1 - (\rho_1^2 + \rho_2^2 + \rho_3^2) - 2\rho_1\rho_2\rho_3}, \\
var(Z_2) &= \frac{1 - \rho_2^2}{1 - (\rho_1^2 + \rho_2^2 + \rho_3^2) - 2\rho_1\rho_2\rho_3}, \\
var(Z_3) &= \frac{1 - \rho_1^2}{1 - (\rho_1^2 + \rho_2^2 + \rho_3^2) - 2\rho_1\rho_2\rho_3},
\end{aligned}$$

where only ρ_1^2, ρ_2^2, and ρ_3^2 are involved.

Step 2. There are three bivariate random vectors in total in Z: $U_{12} = (Z_1, Z_2|Z_3)$, $U_{13} = (Z_1, Z_3|Z_2)$, and $U_{23} = (Z_2, Z_3|Z_1)$. The two random variables in U_{12} have equal variances, similar for U_{13}, and U_{23}. By the eigen-decomposition, the covariance matrix of U_{12} can be expressed as $var(U_{12}) = R^T \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix} R$, where R is the rotation matrix defined in Algorithm 5, and λ_1 and λ_2 are the eigenvalues of $var(U_{12})$. Thus, the covariance matrix of V_1 is diagonal with λ_1 and λ_2 as the entries. The values of λ_1 and λ_2 can be computed numerically, by running the VBAIMH algorithm.

Step 3. Based on the posterior normality conditions, the correlation coefficient r_1 of U_{12}, r_2 of U_{13}, and r_3 of U_{23} are given respectively by

$$r_1 = \frac{(\rho_1 - \rho_2\rho_3)}{\sqrt{(1-\rho_3^2)(1-\rho_1^2)}} \tag{4.7}$$

$$r_2 = \frac{(\rho_2 - \rho_1\rho_3)}{\sqrt{(1-\rho_3^2)(1-\rho_1^2)}} \tag{4.8}$$

$$r_3 = \frac{(\rho_3 - \rho_1\rho_2)}{\sqrt{(1-\rho_1^2)(1-\rho_1^2)}} \tag{4.9}$$

Given the values of λ_1 and λ_2 in Step 2, we can obtain the value of r_1 by computing $r_1 = \left(\frac{\lambda_1^2}{\lambda_2^2} - 1\right) / \left(\frac{\lambda_1^2}{\lambda_2^2} + 1\right)$; similar to compute r_2 and r_3. Thus, we can obtain and solve a system of three polynomial equations given in (4.7), (4.8), and (4.9) to obtain the values of ρ_1,ρ_2, and ρ_3; further, the value of s_1^2, s_2^2, and s_3^2 can be obtained by solving (4.4), (4.5), and (4.6), and then the value of σ_1^2, σ_2^2, and σ_3^2.

4.5.3 Numerical results for the example of Normal random sample

For the affine transformation method, the resulted $\hat{A}$ and $\hat{B}$ is given by,

$$\hat{A} = \begin{pmatrix} 1.049 & 0.000 \\ 11.824 & 1.018 \end{pmatrix}; \hat{B} = \begin{pmatrix} -10.119 \\ -2469.079 \end{pmatrix}.$$

For the marginal approximation method, the directional vectors with the corresponding acceptance rates and EAR table readings are given in Table 4.11.

The numerical results for the stepwise method for each step are given in Table 4.12.

4.5.4 Numerical results for mixture of Normals model

For the affine transformation method, the estimated $\hat{A}$ and $\hat{B}$ is given by,

Table 4.11: The marginal approximation method for MLB Players weights data

direction	Acceptance rate	EAR reading: l_i
$\frac{1}{\sqrt{2}}(1,1)$	0.861	1.55
$\frac{1}{\sqrt{2}}(1,-1)$	0.901	0.74
$\frac{1}{\sqrt{2}}(1,0.5)$	0.966	0.893

Table 4.12: The stepwise method for MLB Players weights data

	Marginal var	Acc. rate	EAR readings	Eigenvalue ratio
Step 1	m_1^2	0.991	1.03	-
	m_2^2	0.966	1.13	-
Step 2	$l_{1,1}^2, l_{1,2}^2$	0.881, 0.889	0.694, 1.44	-
Step 3	r_1	-	-	2.075

$$\hat{A} = \begin{pmatrix} 1.907 & 0.000 & 0.000 & 0.000 & 0.000 \\ 4.822 & 1.372 & 0.000 & 0.000 & 0.000 \\ 2.227 & 0.329 & 1.255 & 0.000 & 0.000 \\ 4.867 & 0.862 & 0.235 & 1.210 & 0.000 \\ -1.593 & -0.190 & -0.342 & -0.011 & 1.177 \end{pmatrix}; \hat{B} = \begin{pmatrix} -0.368 \\ -2.399 \\ -2.199 \\ -4.003 \\ 2.026 \end{pmatrix}.$$

For the marginal approximation method, it requires 15 projections. The directional vectors with the corresponding acceptance rates and EAR table readings are given in Table 4.13.

For the stepwise method. The numerical results for Step 1 and 2 are given in Table 4.14.

Table 4.13: The marginal approximation method for mixtures of Normals

direction	Acceptance rate	the EAR table reading: l_i
$\frac{1}{3}(1,1,1,1,1)$	0.582	4.120
$\frac{1}{3}(1,-1,1,1,1)$	0.901	0.735
$\frac{1}{3}(1,1,-1,1,1)$	0.815	1.800
$\frac{1}{3}(1,1,1,-1,1)$	0.906	1.340
$\frac{1}{3}(1,1,1,1,-1)$	0.477	6.500
$\frac{1}{3}(-1,1,1,1,1)$	0.882	1.420
$\frac{1}{3}(-1,-1,1,1,1)$	0.923	1.280
$\frac{1}{3}(-1,1,-1,1,1)$	0.944	0.840
$\frac{1}{3}(-1,1,1,-1,1)$	0.876	0.676
$\frac{1}{3}(-1,1,1,1,-1)$	0.619	3.600
$\frac{1}{3}(1,-1,-1,1,1)$	0.940	0.830
$\frac{1}{3}(1,-1,1,-1,1)$	0.944	1.200
$\frac{1}{3}(1,-1,1,1,-1)$	0.727	2.380
$\frac{1}{3}(1,1,-1,-1,1)$	0.830	0.581
$\frac{1}{3}(1,1,-1,1,-1)$	0.633	3.400

4.5.5 The EAR table

The EAR table, shown in Table 4.15, is composed as follows: the label for rows contains the first two digits of the target variance; the label for columns contains the decimal of the target variance; the values within the table are expected acceptance rates. For example: if one obtains an acceptance rate of 0.5555, then one would look for the rows to find 4 and the columns to 0.6 which yields the target variance is 4.6.

Table 4.14: The stepwise method for mixtures of Normals

	Marginal variance	Acceptance rate	EAR table readings	Eigenvalue ratio
Step 1	m_1^2	0.933	1.230	-
	m_2^2	0.885	1.428	-
	m_3^2	0.923	1.286	-
	m_4^2	0.817	1.512	-
	m_5^2	0.853	1.425	-
Step 2	$\lambda_{1,1}^2, \lambda_{1,2}^2$	0.917, 0.879	0.776, 1.462	-
	$\lambda_{2,1}^2, \lambda_{2,2}^2$	0.935, 0.919	0.816, 1.282	-
	$\lambda_{3,1}^2, \lambda_{3,2}^2$	0.888, 0.848	0.803, 1.303	-
	$\lambda_{4,1}^2, \lambda_{4,2}^2$	0.885, 0.906	1.221, 0.827	-
	$\lambda_{5,1}^2, \lambda_{5,2}^2$	0.930, 0.911	0.802, 1.331	-
	$\lambda_{6,1}^2, \lambda_{6,2}^2$	0.857, 0.795	0.701, 1.660	-
	$\lambda_{7,1}^2, \lambda_{7,2}^2$	0.885, 0.921	1.174, 0.858	-
	$\lambda_{8,1}^2, \lambda_{8,2}^2$	0.906, 0.880	0.887, 1.090	-
	$\lambda_{9,1}^2, \lambda_{9,2}^2$	0.857, 0.890	1.356, 0.754	-
	$\lambda_{10,1}^2, \lambda_{10,2}^2$	0.978, 0.822	0.984, 0.934	-
Step 3	r_1	-	-	0.307
	r_2	-	-	0.222
	r_3	-	-	0.238
	r_4	-	-	-0.192
	r_5	-	-	0.248
	r_6	-	-	0.406
	r_7	-	-	-0.155
	r_8	-	-	0.103
	r_9	-	-	-0.285
	r_{10}	-	-	-0.026

Table 4.15: EAR table: variance versus expected acceptance rate

	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
1	1.0000	0.9697	0.9422	0.9165	0.8936	0.8720	0.8517	0.8330	0.8157	0.7990
2	0.7833	0.7690	0.7553	0.7423	0.7299	0.7182	0.7067	0.6960	0.6860	0.6761
3	0.6671	0.6577	0.6492	0.6406	0.6325	0.6249	0.6175	0.6104	0.6034	0.5969
4	0.5903	0.5842	0.5782	0.5719	0.5664	0.5609	0.5555	0.5502	0.5450	0.5405
5	0.5354	0.5306	0.5261	0.5221	0.5174	0.5131	0.5089	0.5049	0.5011	0.4971
6	0.4937	0.4897	0.4861	0.4827	0.4791	0.4758	0.4725	0.4693	0.4660	0.4634
7	0.4602	0.4571	0.4542	0.4513	0.4485	0.4460	0.4431	0.4404	0.4376	0.4355
8	0.4325	0.4304	0.4279	0.4251	0.4232	0.4207	0.4184	0.4162	0.4139	0.4118
9	0.4098	0.4077	0.4056	0.4034	0.4013	0.3998	0.3975	0.3956	0.3936	0.3916
10	0.3900	0.3883	0.3864	0.3849	0.3829	0.3809	0.3794	0.3776	0.3759	0.3744
11	0.3728	0.3713	0.3698	0.3682	0.3664	0.3651	0.3638	0.3620	0.3608	0.3592
12	0.3581	0.3563	0.3550	0.3538	0.3520	0.3509	0.3497	0.3486	0.3474	0.3458
13	0.3444	0.3432	0.3419	0.3408	0.3396	0.3384	0.3370	0.3361	0.3346	0.3337
14	0.3325	0.3311	0.3302	0.3292	0.3277	0.3271	0.3257	0.3250	0.3235	0.3226
15	0.3217	0.3209	0.3196	0.3185	0.3177	0.3165	0.3157	0.3148	0.3137	0.3127
16	0.3121	0.3108	0.3100	0.3090	0.3082	0.3074	0.3064	0.3054	0.3047	0.3039
17	0.3030	0.3023	0.3013	0.3005	0.2993	0.2989	0.2979	0.2970	0.2964	0.2954
18	0.2947	0.2940	0.2930	0.2924	0.2915	0.2909	0.2901	0.2894	0.2886	0.2877
19	0.2870	0.2861	0.2854	0.2852	0.2844	0.2834	0.2827	0.2822	0.2814	0.2808

Chapter 5

VB on Dirichlet process priors

This chapter shows how the VB method provides a computational efficient technique in the context of hierarchical modelling using Dirichlet process priors. In particular, it shows how the VB method can handle non-conjugacy in its prior specification, which extends to standard approach to these models. It shows, using the parameter separation parameterization, we are able to provide a full variational solution for the Dirichlet process. The numerical results show that the method is very computationally efficient when compared to MCMC. We also provide a VB approximation to the posterior predictive distribution and compare it with results derived from two Markov chain Monte Carlo (MCMC) methods. For the truncated DP, we propose an empirical method to determine the number of distinct components in a finite dimensional DP.

5.1 Introduction

In Bayesian parametric modelling, the prior distribution is usually constructed by assuming it has a particular parametric form. In many ways, though, it is more appealing that the support of the prior is the class of all distribution functions. In particular, this allows greater

flexibility for modelling and inference. The Dirichlet process, introduced by Ferguson (1973), provides a means of specifying a probability measure $P(dF)$ over the space of all (discrete) probability measures. Following this, the DP has become very popular when applied to Bayesian non-parametric inference. Mixture models are among the important applications of the DP, for example, Escobar (1994) and Escobar and West (1995). In particular the clustering property exhibited by the generalized Polya urn representation (Blackwell and MacQueen, 1973) makes the DP a natural choice for the prior distribution in the mixture model.

Markov chain Monte Carlo (MCMC) methods, in the context of a DP prior, have been extensively studied, for example, see Escobar (1994); Escobar and West (1995); West and Escobar (1993); MacEachern (1994). A common aspect of these methods is that they integrate over the random probability measures and use the generalized Polya urn representation of the DP. The Polya urn samplers are restricted to using conjugate base distributions that allow analytic evaluation of the transition probabilities. When non-conjugate priors are used, these methods require an often difficult numerical integration. MacEachern and Müller (1998); Neal (2000) devised approaches for handling non-conjugacy by using a set of auxiliary parameters.

The truncated stick-breaking representation of the DP has also been considered. For example Ishwaran and Zarepour (2000) shows that with a moderate truncation, the finite dimensional DP should be able to achieve an accurate approximation. Based on this representation, Ishwaran and James (2001) proposed a Gibbs sampler to handle non-conjugacy issue.

In recent years, variational Bayesian inference has been applied to DP based problems, for example see Blei and Jordan (2006). Strictly speaking, they used the mean-field method rather than a full variational solution, where the approximating distributional family is specified, and the optimization is only over the variational parameters. In addition, they

also only consider the case where the conjugate base distribution is an exponential family.

The hierarchical principle is a natural way to model dependence amongst model parameters. This chapter considers the one-way random-effects model which has been discussed in previous chapters. This model is simple, but important, and is widely applicable, being an important special case of the hierarchical linear model. However, as MacEachern (1994) pointed out, restricting the prior to be a normal distribution severely constrains the estimate of normal means, producing estimators that shrink each data value toward the same point. Replacing the normal prior by a Dirichlet process has been considered by MacEachern (1994), and Bush and MacEachern (1996) in an MCMC context.

This chapter considers non-conjugate settings for this model and presents a full variational Bayesian solution, where the optimization is in terms of both the distributional family and the parameters of the approximating distribution. The core ingredient for the proposed solution lies on the parameter separation parameterization, which is introduced in Chapter 2. In our solution, we use a truncated stick-breaking representation of the DP. A natural question is raised by given a dataset how to estimate the truncation level for a finite dimensional DP. We propose an empirical method to determine the number of distinct components in a finite dimensional DP.

The posterior predictive distribution for this model is not available in a closed form. For the VB method, even though we can obtain closed-formed posterior approximations and use them to replace the unknown posterior densities in computing the posterior predictive density, it is still not available in a closed form. In this chapter, we show how to use the similar variational method to approximate this quantity.

The rest of the chapter is organized as follows. Section 2 presents the one-way random-effects model with a Dirichlet process prior, and shows how to use Gibbs samplers to simulate samples from the posterior distributions. Section 3 we provide a full variational solution for the one-way random-effects model with Dirichlet process prior. Section 4 discusses how to

approximate the posterior predictive distributions by the MCMCM methods and by the VB method. Numerical studies are presented in Section 5. Conclusions are given in Section 6.

5.2 The one-way random effects model

In this section, we describe the one-way random-effects model which uses a DP prior in a non-conjugate setting, and then show how we can adapt two MCMC methods introduced by Neal (2000), and Ishwaran and James (2001) to obtain the posterior samples.

As shown in Section 3.2.2, the one-way random effects model can be expressed as follows:

$$\begin{aligned} y_{ij}|\theta_j, \sigma^2 &\sim N(\theta_j, \sigma^2), \\ \theta_j|\mu, \tau^2 &\sim N(\mu, \tau^2), \\ (\sigma^2, \mu, \tau^2) &\sim \pi \quad \text{for } i = 1, \ldots, n_j; j = 1, \ldots, J, \end{aligned} \tag{5.1}$$

where π is a prior distribution. When the normal distribution at the middle stage is replaced by a DP, this gives the following model:

$$\begin{aligned} y_{ij}|\theta_j, \sigma^2 &\sim N(\theta_j, \sigma^2), \\ \theta_j|F &\sim F, \\ F|\alpha, F_0 &\sim \mathrm{DP}(\alpha, F_0), \\ \sigma^2 &\sim \pi \quad \text{for } i = 1, \ldots, n_j; j = 1, \ldots, J, \end{aligned} \tag{5.2}$$

where α is a positive real-valued concentration parameter and F_0 is a base distribution. We consider F_0 a normal distribution with mean μ and variance τ^2, both are further treated as random variables. It is worth noting that in this setting F_0 is not conjugate to the likelihood. The realizations of the DP are discrete with probability one, thus the above

model can be viewed as a countably infinite mixture (Ferguson, 1983). When integrating over F in (5.2), we can obtain a representation, referred as the generalized Polya urn scheme, of the prior distribution of θ_j in terms of successive conditional distributions of the following form (Blackwell and MacQueen, 1973):

$$\theta_j|\theta_1,\cdots,\theta_{j-1} = \begin{cases} \theta_l & \text{with probability } \frac{1}{\alpha+j-1} \text{ for each } l \in \{1,\cdots,j-1\} \\ \sim F_0 & \text{with probability } \frac{\alpha}{\alpha+j-1} \end{cases}$$

This representation gives a clear view for the clustering or mixture effects of the DP prior, and constitutes a fundamental ingredient for the Polya urn form of MCMC samplers.

Alternatively, Sethuraman (1994) provides a constructive definition of the random distribution F in the DP:

$$F = \sum_{j=1}^{T} v_j \delta_{\theta_j},$$

where $w_i \overset{\text{iid}}{\sim} \text{Beta}(1,\alpha)$, and v_j is defined as $v_1 = w_1$, $v_j = w_j \prod_{l=1}^{j-1}(1-w_l)$, and $\theta_j \overset{\text{iid}}{\sim} F_0$, and δ_{θ_j} denotes a discrete measure concentrated at θ_j , and $1 \leq T \leq \infty$. This is often referred to as the "stick-breaking" representation. If $T < \infty$, this is referred to as a truncated DP or finite dimensional DP (Ishwaran and Zarepour, 2000).

The exact computation of posterior quantities using model (5.2) is typically infeasible. However, MCMC provides one means to approximate them. Due to the non-conjugate property of model (5.2), we consider using the methods introduced in Neal (2000) and Ishwaran and James (2001) to obtain posterior samples.

First, we consider the method proposed by Neal (2000) and similar to the "no gaps" algorithm proposed earlier by MacEachern and Müller (1998). Let $\zeta = (\zeta_1,\cdots,\zeta_K)$ denote the set of distinct θ_j, where $j = 1,\cdots,J$ and $K \leq J$. Let $c = (c_1,\cdots,c_J)$ denote a vector of indicators defined by $c_j = k$ if and only if $\theta_j = \zeta_k$. The state of the Markov chain consist

of c, ζ, μ, τ^2 and σ^2. Each sampling scan consists of picking a new value for each c_j from its conditional distribution given y, ζ, and all the c_l for $l \neq j$ (written as c_{-j}), and then picking a new value for each ζ_k from its conditional distribution given y and c, and then picking a new value for μ, τ^2 and σ^2 respectively from their conditional distributions.

Algorithm 6 Polya-urn-type Gibbs sampler

Step 1. For $j = 1, \cdots, J$, generate $c_j^{(t)}$ from the distribution of $c_j | y, \zeta, c_{-j}, \mu, \tau^2, \sigma^2$.

- Let k^- be the number of distinct c_l for $l \neq j$, and let $p = k^- + s$. Label c_l with values in $\{1, \cdots, k^-\}$.
- Draw values independently from $F_0(\mu^{(t-1)}, \tau^{2(t-1)})$ for all the $\zeta_a^{(t)}$ for which $k^- + 1 \leq a \leq p$. If the value of $c_j^{(t-1)}$ is a singleton (only associated with one y_j), then ζ_{k^-} equals to $\zeta_{c_j}^{(t-1)}$, otherwise draw a new value for ζ_{k^-} from $F_0(\mu^{(t-1)}, \tau^{2(t-1)})$.
- Draw a value for $c_j^{(t)}$ from $\{1, \cdots, p\}$ with the following probability

$$P(c_j = a | c_{-j}^{(t-1)}, y, \sigma^{2(t-1)}) \propto \begin{cases} m_{-j,a} f(y_j; \zeta_a^{(t-1)}, \sigma^{2(t-1)}), & \text{for } 1 \leq a \leq k^- \\ \frac{\alpha}{s} f(y_j; \zeta_a^{(t)}, \sigma^{2(t-1)}), & \text{for } k^- < a \leq p \end{cases}$$

  where $m_{-j,a}$ is the number of c_l for $l \neq j$ that are equal to c.
- Discard the ζ_a's that are not now associated with any observation, and relabel ζ_k and corresponding c_j.

Step 2. For $k = 1, \cdots, |c|$, generate $\zeta_k^{(t)}$ from the distribution of $\zeta_k | y, \mu, \tau^2, \sigma^2$, which is give by

$$p(\zeta_k | y, \mu^{(t-1)}, \tau^{2(t-1)}, \sigma^{2(t-1)}) \propto \prod_{j: c_j = k} \prod_{i=1}^{n_j} \phi(y_{ij}; \zeta_k, \sigma^{2(t-1)}) \phi(\zeta_k; \mu^{(t-1)}, \tau^{2(t-1)}),$$

where $\phi(.)$ denotes the normal density function.

The key feature to handle the issue of non-conjugacy lies that when c_j is updated, a set of size of s temporary auxiliary parameter variables that represent possible values for ζ_k that are not associated with any other observations is introduced. Since the observations y_j are exchangeable, we can assume that we are updating c_j for the last observation, and that

Algorithm 7 Polya-urn-type Gibbs sampler (continued)

Step 3. Generate $\mu^{(t)}$, $\tau^{2(t)}$, and $\sigma^{2(t)}$ from the corresponding full conditional distribution, that are given as follows:

$$p(\sigma^2|y,\zeta_k^{(t)}) \propto \prod_{k=1}^{|c|} \prod_{j:c_j=k} \prod_{i=1}^{n_j} \phi(y_{ij};\zeta_k^{(t)},\sigma^2)\pi(\sigma^2)$$

$$p(\mu|y,\zeta_k^{(t)},\tau^{2(t)}) \propto \prod_{k=1}^{|c|} \phi(\zeta_k^{(t)};\mu,\tau^{2(t)})\pi(\mu)$$

$$p(\tau^2|y,\zeta_k^{(t)},\mu^{(t)}) \propto \prod_{k=1}^{|c|} \phi(\zeta_k^{(t)};\mu^{(t)},\tau^2)\pi(\tau^2),$$

where $\pi(\tau^2)$, $\pi(\sigma^2)$, and $\pi(\mu)$ are corresponding priors.

the c_l for other observations have values in the set $\{1,\cdots,k^-\}$, where k^- is the number of distinct c_l for $l \neq j$. By using the auxiliary variables, the possible values for a new c_j lies in $\{1,\cdots,k^-,k^-+1,\cdots,k^-+s\}$. Once a new value for c_j has been chosen, all the ζ that are not now associated with any observation will be discarded, and ζ_k and the corresponding c_j are relabeled to have the c_j with values in $\{1,\cdots,|c|\}$, where $|c|$ denotes the number of distinct number in c. This Gibbs updating for model (5.2) is summarized in Algorithm 6.

In addition to handling the issue of non-conjugacy, Neal (2000) suggests that this method can improve the mixing of the chain and shorten the autocorrelation time to reduce the sample size used to estimate the posterior quantities. However, it is clear that since F is integrated over, this Polya-urn like sampler still restricts the inference for the posterior of the random F to be based only on the posterior for ζ_k's, that is, there no explicit inference on F is possible. Ishwaran and James (2001) devised a, so called, blocked Gibbs sampler, which uses the stick-breaking representation, to avoid the limitation imposed by the Polya urn like samplers.

The key to the blocked Gibbs sampler lies that it is infeasible to work on an infinite

numbers of components in the stick-breaking representation, and it has to truncate the DP at a curtain level, denoted as B, and discard the components of $B+1$, $B+2$, $\cdots$. Ishwaran and Zarepour (2000) shows that with a moderate truncation the marginal density under a truncated DP prior is indistinguishable from the one based on the infinite DP prior. By using a stick-breaking representation, the one-way random-effects model given in (5.2) under a truncated DP can be written as follows:

$$
\begin{aligned}
& y_{ij}|c_j, \zeta, \sigma^2 \sim N(\zeta_{c_j}, \sigma^2), \ \text{ for } i = 1, \ldots, n_j; j = 1, \ldots, J, \\
& c_j|v \sim \sum_{b=1}^{B} v_b \delta_b; \ \ v_1 = w_1, v_b = w_b \prod_{l=1}^{b-1} (1 - w_l), \\
& w_b \sim \text{Beta}(1, \alpha), \ \text{for } b = 1, \cdots, B-1, \ \text{and } w_B =, 1 \\
& \zeta_b \sim N(\mu, \tau^2); \ \text{for } b = 1, \ldots, B, \\
& (\sigma^2, \mu, \tau^2) \sim \pi.
\end{aligned} \tag{5.3}
$$

In this model, the state of the Markov chain consist of c, ζ, v, μ, τ^2 and σ^2. The blocked Gibbs sampling for model (5.3) is summarized in Algorithm 8.

5.3 Variational method

As an alternative to MCMC methods, Blei and Jordan (2006) provides a the variational solution on DP in the context of mixture models. There are several constraints in their developments. First, it is the mean-field method rather than a full variational solution, where the approximating distributional family is specified, and the optimization is only over the variational parameters. In addition, they consider the cases of the base distribution conjugate to the likelihood function, and both of the prior distribution and the likelihood are considered in exponential families. Third, their variational formalism are developed only with respect to natural parameters, which are often not the parameters of immediate

Algorithm 8 Blocked Gibbs sampler

Step 1. For $j = 1, \cdots, J$, generate $c_j^{(t)}$ from the distribution of $c_j|y, \zeta, v, \sigma^2$, that is given by:

$$p(c_j|y,\zeta, v, \sigma^2) = \sum_{b=1}^{B} p_{b,j}\delta_b, \quad \text{where } p_{b,j} \propto v_b^{(t-1)} \prod_{i=1}^{n_j} \phi(y_{ij}; \zeta_b^{(t-1)}, \sigma^{2(t-1)})$$

Step 2. For $b = 1, \cdots, B$, generate $\zeta_b^{(t)}$ as follows:

- When $\zeta_b^{(t)}$ is not associated with any y_j, draw a new value from $F_0(\mu^{(t-1)}, \tau^{2(t-1)})$.
- Otherwise, draw a new value from the following conditional distribution:

$$p(\zeta_b|y, \mu^{(t-1)}, \tau^{2(t-1)}, \sigma^{2(t-1)}) \propto \prod_{j:c_j=b} \prod_{i=1}^{n_j} \phi(y_{ij}; \zeta_b, \sigma^{2(t-1)}) \phi(\zeta_b; \mu^{(t-1)}, \tau^{2(t-1)}),$$

Step 3. Generate $v^{(t)}$ from the following conditional distribution:

$$v_1^{(t)} = w_1^{(t)}, v_b^{(t)} = w_b^{(t)} \prod_{l=1}^{b-1} (1 - w_l^{(t)}),$$

$$w_b^{(t)} \sim \text{Beta}(M_b, \alpha + \sum_{l=b+1}^{B} M_l);\ M_b \text{ is the number of } c_j \text{ equals to } b$$

Step 4. Generate $\mu^{(t)}$, $\tau^{2(t)}$, and $\sigma^{2(t)}$ from the corresponding full conditional distribution, that are given as follows:

$$p(\sigma^2|y, \zeta_b^{(t)}) \propto \prod_{b=1}^{B} \prod_{j:c_j=b} \prod_{i=1}^{n_j} \phi(y_{ij}; \zeta_b^{(t)}, \sigma^2) \pi(\sigma^2)$$

$$p(\mu|y, \zeta_b^{(t)}, \tau^{2(t)}) \propto \prod_{b=1}^{B} \phi(\zeta_b^{(t)}; \mu, \tau^{2(t)}) \pi(\mu)$$

$$p(\tau^2|y, \zeta_b^{(t)}, \mu^{(t)}) \propto \prod_{b=1}^{B} \phi(\zeta_b^{(t)}; \mu^{(t)}, \tau^2) \pi(\tau^2),$$

where $\pi(\tau^2)$, $\pi(\sigma^2)$, and $\pi(\mu)$ are corresponding priors.

interests.

In Section 2.5.2 we present a special parameterization for a parametric family, which we call the parameter separation parameterization. An important feature of this representation lies that it provides a factorized form, which is the key to make possible the construction of the analytical form of the variational distributions. This parameterization is not constrained to only be in exponential family. It can include the distributions from non-exponential family. Theorem 2.2 in Section 2.5.2 provides a variational formularization on the parameter separation parameterization.

Theorem 2.2 is ready to be used in developing the variational inference for the one-way random-effects model with the DP prior. Here, we consider the stick-breaking representation given in (5.3). We define c_j in (5.3) as $c_j = (c_{j1}, \cdots, c_{jB})$, where c_{jb} is an indicator variable with probability v_b of equalling to one. This probability is given in (5.3). The joint probability of $y, c, v, \zeta, \sigma^2, \mu, \tau^2$ is given as follows:

$$p(y, c, v, \zeta, \sigma^2, \mu, \tau^2) = \prod_{j=1}^{J}\prod_{b=1}^{B}\left\{v_b\prod_{i=1}^{n_j}\phi(y_{ij};\zeta_b,\sigma^2)\right\}^{c_{jb}}\prod_{b=1}^{B}\phi(\zeta_b;\mu,\tau^2)\prod_{b=1}^{B-1}\text{Beta}(w_b;1,\alpha)\pi(\sigma^2)\pi(\mu)\pi(\tau^2), \quad (5.4)$$

where $\pi(\sigma^2)$,$\pi(\mu)$, and $\pi(\tau^2)$ are the prior distributions. To have these priors providing little influence on the posterior distributions, we assign non-informative uniform priors for μ, $\log(\sigma^2)$, and τ^2. If we were to assign a uniform prior distribution for $\log(\tau^2)$, the posterior distribution would be improper. Thus, we get the prior distribution for μ, $\log(\sigma^2)$, and τ^2 is given by $\pi(\sigma^2, \mu, \tau^2) \propto \frac{1}{\sigma^2}$

We denote $q(c, v, \zeta, \sigma^2, \mu, \tau^2)$ as the VB approximation for the posterior distribution of $p(c, v, \zeta, \sigma^2, \mu, \tau^2|y)$. In contrast to the mean field approximation, we do not require any distributional families to q, except for the independence assumption. We assume q has the

following factorization form:

$$q(c, v, \zeta, \sigma^2, \mu, \tau^2) = \prod_{j=1}^{J} q(c_j) \prod_{b=1}^{B} q(v_b) \prod_{b=1}^{B} q(\zeta_b) q(\sigma^2) q(\mu|\tau^2) q(\tau^2). \tag{5.5}$$

It is worth noting that using a full factorization with $q(\mu, \tau^2) = q(\mu)q(\tau^2)$, results in that the convergence of variational parameters fails in the iterative updates.

It is straightforward to check that the distributions at each stage of model (5.3) all have a parameter separate parameterization, and then Theorem 2.2 is can be used. By plugging (5.4) into (2.17) or (2.19), we can obtain the following results:

$$\begin{aligned}
&q(c_j) = \text{Multinomial}(r_{j1}, \cdots, r_{jB}) \\
&\quad r_{jb} \propto \exp\left\{ -\frac{1}{2}\frac{g}{h}\sum_{i=1}^{n_j}(y_{ij} - a_b)^2 - \frac{1}{2}\frac{g}{h}b_b^2 n_j + \psi(c_b) - \psi(c_b + d_b) + \sum_{l=1}^{b-1}(\psi(c_l) - \psi(c_l + d_l)) \right\} \\
&q(\zeta_b) = N(a_b, b_b^2); \quad a_b = \frac{\frac{g}{h}\sum_{j=1}^{J} r_{jb}(\sum_{i=1}^{n_j} y_{ij}) + \frac{k}{s}e}{\frac{g}{h}\sum_{j=1}^{J} r_{jb} n_j + \frac{k}{s}}, b_b^2 = \frac{1}{\frac{g}{h}\sum_{j=1}^{J} r_{jb} n_j + \frac{k}{s}} \\
&q(v_b) = \text{Beta}(c_b, d_b); \quad c_b = \sum_{j=1}^{J} r_{jb} + 1, d_b = \sum_{l=b+1}^{B}\sum_{j=1}^{J} r_{jb} + \alpha \text{ (for } b < B), d_B = \alpha; \\
&q(\mu|\tau^2) = N(e, \frac{\tau^2}{f^2}); \quad e = \frac{\sum_{b=1}^{B} a_b}{B}, f^2 = B \\
&q(\tau^2) = \text{IG}(k, s); \quad k = \frac{B}{2} - \frac{3}{2}, s = \frac{1}{2}\sum_{b=1}^{B}((a_b - e)^2 + b_b^2) \\
&q(\sigma^2) = \text{IG}(g, h); \quad g = \frac{\sum_{j=1}^{J} n_j}{2}, h = \frac{1}{2}\sum_{j=1}^{J}\sum_{b=1}^{B} r_{jb}\left(\sum_{i=1}^{n_j}(y_{ij} - a_b)^2 + b_b^2\right),
\end{aligned} \tag{5.6}$$

where ψ denotes the digamma function, and IG denotes the gamma distribution. The above approximations are well-recognised distributions, and they are easy to use to make further inference on parameters. The VB algorithm requires an iterative updates on the parameters

of r_{jb}, a_b, b_b^2, c_b, d_b, e, f, g, h, k, and s till they converge.

5.4 The predictive distribution

The posterior predictive distribution provides a distribution for a new data point given the observed data, in which it makes use of the entire posterior distribution. Suppose $y^* = (y_1^*, \cdots, y_{n^*}^*)$ is a new observation, then the posterior predictive distribution of y^* given y is defined as

$$p(y^*|y) = \int p(y^*|\Theta)p(\Theta|y)d\Theta, \tag{5.7}$$

where Θ refers the model parameters. For the one-way random-effects model with a DP prior this quantity is intractable however MCMC methods provide a straightforward approximation. Having a sample of T points from the posterior, we can estimate it by

$$p(y^*|y) = \frac{1}{T}\sum_{t=1}^{T} p(y^*|\Theta^{(t)}), \tag{5.8}$$

where $\Theta^{(t)}$ is the sample drawn from the posterior distribution after the chain reaches its stationary distribution. For Algorithm 6, $p(y^*|\Theta^{(t)})$ is given as follows:

$$p(y^*|\Theta^{(t)}) = \sum_{k=1}^{|c^{*(t)}|} P(c^{*(t)} = k) f(y^*|\zeta_k^{(t)}, \sigma^{2(t)})$$

where again $|c^{*(t)}|$ denotes the number of values which $c^{*(t)}$ takes. For Algorithm 8, it is given as follows:

$$p(y^*|\Theta^{(t)}) = \sum_{b=1}^{B} v_b^{(t)} f(y^*|\zeta_b^{(t)}, \sigma^{2(t)})$$

For the VB method, it is natural to use the VB approximations to replace the unknown posterior distributions in (5.7). Thus, we can have the following approximation for the posterior predictive distribution:

$$\begin{aligned} p(y^*|y) &\approx \int \left(\sum_{b=1}^{B} v_b f(y^*|\zeta_b, \sigma^2) \right) dQ(v, \zeta, \sigma^2) \\ &= \sum_{b=1}^{B} \mathrm{E}_{q(v_b)}[v_b] \int \left(f(y^*|\zeta_b, \sigma^2) \right) dQ(\zeta_b) dQ(\sigma^2) \end{aligned} \tag{5.9}$$

where Q is the VB approximation. Unfortunately, although we have obtained the simple and well-recognised distributions for $Q(\zeta_b)$ and $Q(\sigma^2)$, the integrals in (5.9) are still not available in a closed form. However, we can apply the variational principle again to obtain a lower bounds on this quantity, and propose using this lower bound as an approximation for the posterior predictive distribution.

We denote L_b as $L_b = \int \left(f(y^*|\zeta_b, \sigma^2) \right) dQ(\zeta_b) dQ(\sigma^2)$. If we regard $Q(\zeta_b)$ and $Q(\sigma^2)$ as prior distributions, then L_b can be regarded as a marginal likelihood, that can be approximated by the variational method. We denote $v(\zeta_b)$ and $v(\sigma^2)$ as the variational approximations which result from treating $Q(\zeta_b)$ and $Q(\sigma^2)$ as priors. Again, Theorem 2.2 can be used to obtain the distributional forms for $v(\zeta_b)$ and $v(\sigma^2)$, and gives the following results:

$$\begin{aligned} v(\zeta_b) &= N(A_b, B_b^2); \quad A_b = \frac{\frac{G}{H} \sum_{i=1}^{n^*} y_i^* + \frac{a_b}{b_b^2}}{\frac{G}{H} n^* + \frac{1}{b_b^2}}, B_b^2 = \frac{1}{\frac{G}{H} n^* + \frac{1}{b_b^2}} \\ v(\sigma^2) &= \mathrm{IG}(G, H); \quad G = g + \frac{n^*}{2}, H = h + \frac{1}{2} S^* + \frac{n^*}{2} ((A_b - \bar{y^*})^2 + B_b^2), \end{aligned}$$

where n^* is the number of observations in y^*, and $\bar{y^*}$ is the mean of y^*, and S^* is the total sum of squares of y^*, and a_b, b_b^2, g, and h are given in (5.6).

Once the variational parameters of A_b, B_b^2, G, and H converge, we can obtain a lower bound of the logarithm of L_b, denoted as F_b, which is given as follows:

$$\begin{aligned}
F_b=&\int \frac{q(\zeta_b)}{v(\zeta_b)}dV(\zeta_b)+\int \frac{q(\sigma^2)}{v(\sigma^2)}dV(\sigma^2)+\int \log(f(y^*|\zeta_b,\sigma^2))dV(\zeta_b,\sigma^2)\\
=&\log\left(\frac{1}{b_b}\right)-\log\left(\frac{1}{B_b}\right)-\frac{1}{2b_b^2}((A_b-a_b)^2+B_b^2)\\
&+(G-g)(\log H-\psi(G))+G\left(1-\frac{h}{H}\right)+\log\frac{h^g}{\Gamma(g)}+\log\frac{H^G}{\Gamma(G)}\\
&-\frac{n^*}{2}(\log 2\pi+\log H-\psi(G))-\frac{1}{2}\frac{G}{H}\left(\sum_{i=1}^{n^*}(y_i^*-A_b)^2-n^*B_b^2\right),
\end{aligned}$$

where $\Gamma(.)$ is the gamma function.

Once we obtain the values of each F_b for $b=1,\cdots,B$, we can obtain a lower bound for (5.9)

$$\sum_{b=1}^{B}\mathrm{E}_{q(v_b)}[v_b]L_b\geq\sum_{b=1}^{B}\mathrm{E}_{q(v_b)}[v_b]\exp(F_b)\equiv F.$$

Thus, we propose to use F as an approximation for the posterior predictive distribution of $p(y^*|y)$.

5.5 Numerical studies

We examine the performance of the VB method by comparing it with the two MCMC methods on simulated data. To generate the data, we set μ and τ^2 for the base distribution in (5.2) to be $\mu=0$ and $\tau^2=16$ and σ^2 equal to 0.64. We use the truncated stick-breaking representation to construct the random distribution F. For demonstration purposes, we simply truncate F at level 5, shown in Table 5.1. A data set of 60 groups data are generated

from F , and each group contains 80 data points. We use 50 groups as the observed data and 10 groups as the future data.

Table 5.1: A random distribution F, truncated at level 5

ζ_b	-2.22	-0.54	1.01	4.28	7.10
$P(\zeta_b)$	0.35	0.14	0.13	0.13	0.26

In the VB learning, we assume we have no knowledge about the distribution F, and also mis-specify the truncation level to 10. The algorithm converges after 19 iterations.

Table 5.2: The VB approximations for the random distribution F

$\mathrm{E}[v_1]$	$\mathrm{E}[v_2]$	$\mathrm{E}[v_3]$	$\mathrm{E}[v_4]$	$\mathrm{E}[v_5]$	$\mathrm{E}[v_6]$	$\mathrm{E}[v_7]$	$\mathrm{E}[v_8]$	$\mathrm{E}[v_9]$	$\mathrm{E}[v_{10}]$
$\mathrm{E}[\zeta_1]$	$\mathrm{E}[\zeta_2]$	$\mathrm{E}[\zeta_3]$	$\mathrm{E}[\zeta_4]$	$\mathrm{E}[\zeta_5]$	$\mathrm{E}[\zeta_6]$	$\mathrm{E}[\zeta_7]$	$\mathrm{E}[\zeta_8]$	$\mathrm{E}[\zeta_9]$	$\mathrm{E}[\zeta_{10}]$
0.167	0.16	0.12	0.12	0.01	0.01	0.01	0.13	0.13	0.11
-2.24	-2.24	-0.55	0.97	2.06	2.06	2.06	4.23	7.12	7.12

Table 5.2 gives the expected values for v_b and ζ_b under the VB approximations. We can see a clear pattern. The expected probability weights for the component 5, 6, and 7, are close to zero. This may suggest they can be ruled out from the true model. The component 1 and 2 share the exact same value of -2.24, which is close to the value of component 1 in Table 5.1, and the cumulated expected probability weight of 0.327 is also close to 0.35 in Table 5.1. We can observe a similar situation for component 9 and 10. Thus, by combining same components (with same values) and ruling out the empty components (with very small probability weights), we can conclude that VB picks up 5 components for the random distribution F.

For the Polya-urn type Gibbs sampler (Algorithm 6), we run 2×10^5 iterations. We use the last 20% data, which we believe the chain has reached its stationary distribution. To reduce the serial correlation effect, we pick the every 25^{th} data point. The frequencies of the

distinct number of ζ_b are given in Table 5.3. We see that the posterior probability favors 5, 6, or 7 components, and 6 components has the largest probability.

Table 5.3: Posterior probabilities for the number of ζ

# of ζ	5	6	7	8	9	10
P(# of ζ)	0.270	0.386	0.254	0.068	0.018	0.002

For the blocked Gibbs sampler (Algorithm 8), we run 2.5×10^6 iterations. The last 20% data is used. To reduce the serial correlation effect, we pick the every 25^{th} data point. Even with the order constraints on ζ, the chain still shows the signs of label switching. Thus, a single value of v_b or ζ_b may lose the interpretability.

Finally, we compare the posterior predictive distribution approximated by the three methods. We compute the log predictive likelihoods, shown in Table 5.4, for the 10 groups of future data. For the Gibbs samplers, additional 2,500 samples are collected and used in the computation. We see that the three methods give very similar values. The mean values are given as -95.95, -97.30, -97.32 respectively. A t test, for the log predictive likelihoods computing by Algorithm 8 and by VB, is performed, and it can not reject the hypothesis that the true difference in means is equal to 0 at a p-value equal to 0.9923, and we also can obtain a p-value equal to 0.5049 for Algorithm 8 versus Algorithm 6,

Table 5.4: Log predictive likelihood for 10 groups of future data

Polya-urn	-96.19	-98.43	-89.45	-97.35	-104.31	-95.64	-90.36	-99.84	-92.86	-95.11
Blocked	-97.40	-99.67	-90.59	-98.53	-105.84	-96.76	-91.50	-100.82	-95.53	-96.32
VB	-97.29	-99.88	-90.46	-98.74	-105.90	-96.88	-91.37	-100.62	-95.47	-96.54

5.6 Discussion

The variational Bayes method provides a computational efficient technique to approximate certain posterior quantities in the context of hierarchical modelling using Dirichlet process priors. To avoid the limitation in the existing variational formalism which relies on conjugate exponential families, we consider VB in a new framework. The parameter separation parameterization (Section 2.5.2) gives a factorization which allows flexible dependence structures. Based on this new framework, we provide a full variational solution for the Dirichlet process with non-conjugate base prior. The numerical results show that the VB method is very computationally efficient. Moreover, the comparison with two different MCMC methods shows that VB provides accurate approximations for the posterior predictive distribution. Finally, we propose an empirical method to estimate the truncation level for the truncated DP.

Chapter 6

Variational Bayes for Regime-switching Lognormal Models

This chapter describes how to apply the VB method to the regime-switching log-normal model and how it provides a computationally fast solution to quantify the uncertainty in the model specification and parameter specitication. The results show that the method can recover exactly the model structure, gives the reasonable point estimates, and is very computationally efficient. The potential problems of the method in quantifying the parameter uncertainty are discussed. To remedy these problems, the methods proposed in Chapter 4 are used to compute the true posterior covariance matrix.

6.1 Introduction

Switching between different states or regimes is a common phenomenon in many time series, and regime-switching models, originally proposed by Hamilton (1989), have been used to model these switching processes. Of particular interest to this chapter is the regime-switching lognormal model (RSLN) proposed by Hardy (2001). As demonstrated in Hardy (2002), the

maximum likelihood estimate (MLE) does not give a simple method to deal with parameter uncertainty. The asymptotic normality of maximum likelihood estimators may not apply for sample sizes commonly found in practice. Hence, to understand parameter uncertainty Hardy (2002) considered the RSLN model in a Bayesian framework using the Metropolis-Hastings algorithm. Furthermore, model uncertainty, in particular selecting the correct number of regimes, is a major issue. Hence, model selection criteria have to be used to choose the best model. Hardy (2001) found that a two-regime RSLN model maximized the Bayes information Criterion (BIC) (Schwarz, 1978) for both monthly TSE 300 total return data and S&P 500 total return data, however, according to the Akaike Information Criterion (AIC) (Akaike, 1974), a three regime model was the optimal on S&P data. To account for the model uncertainty associated with the number of regimes, Hartman and Heaton (2011) offered a dynamic estimation of the number of regimes using a Chinese restaurant process.

MCMC methods make possible the computation of all posterior quantities, however there are a number of practical issues associated with their implementation. Detailed discussions can be found in Chapter 1 in particular computational speed is one of the main advantages of VB.

This chapter shows how to apply the VB method to the RSLN model and presents a solution to investigate the model specification problem. In particular it looks at how to find the appropriate number of regimes. While the simplification in the dependence gives computation advantages it also comes at a cost. For example we also found that the posterior variance may be underestimated, and the correlation structure is distorted. We will use the techniques introduced in Chapter 4 to approximate the true posterior covariance matrix.

Moreover, through the numerical results, we can observe that the VB approximations tend to present an approximately symmetric and bell shaped pattern. In this chapter, we aim to explore the asymptotic properties of the VB method.

The chapter is organized as follows. Section 6.2 presents the VB method and its applica-

tion in the RSLN model. Numerical studies on simulated data and real data are provided in Section 6.3, where the VB method is compared with both the criterion-based model selection procedure and the MCMC method. Section 6.4 uses the three method proposed in Chapter 4 to estimate the true posterior covariance matrix. Section 6.5 discusses the asymptotic normality. Conclusions are available in the last section.

6.2 Variational Bayes

6.2.1 Variational Bayes for the RSLN model

A regime-switching lognormal (RSLN) model (Hardy, 2001) with a fixed finite number, K, regimes can be described as a bivariate discrete time process with the observed data sequence $W_{1:T} = \{W_t\}_{t=1}^T$, and the unobserved regime sequence $S_{1:T} = \{S_t\}_{t=1}^T$, where $S_t \in \{1, \cdots, K\}$ and T is the number of observations. The logarithm of the observed data W_t, denoted as $Y_t = \log W_t$, is normally distributed with its mean μ_i and variance σ_i^2 dependent on an unobserved *regime* S_t. The sequence of $S_{1:T}$ is assumed to follow a first order Markov chain (Figure 6.1). In the most basic models the transition probabilities of the hidden regimes and the model for Y_t given S_t are assumed to be invariant over time.

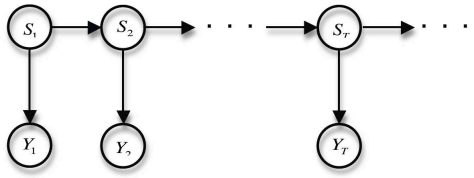


Figure 6.1: A graphical representation of a RSLN model

The parameters in a RSLN model include the initial probabilities of the first regime, $\pi = (\pi_i)_{i=1}^K$, the transition probabilities of the hidden regimes, $A = (a_{ij})$, and the means and

variances $\zeta = \{(\mu_i, \sigma_i^2)\}_{i=1}^K$ for the observed data conditionally on the state. We denote the complete parameter vector by $\tau = (\pi, A, \zeta)$. The joint distribution of $Y_{1:T}$ and $S_{1:T}$ given τ, also referred to as the *complete likelihood function*, is given by

$$P(Y_{1:T}, S_{1:T}|\tau) = p(S_1|\pi) \prod_{t=1}^{T-1} p(S_{t+1}|S_t; A) \prod_{t=1}^{T} p(Y_t|S_t; \zeta). \tag{6.1}$$

By adopting an indicator variable $S_{t,i}$, where $S_{t,i} = 1$, if the process is in state i at time t, and zero otherwise, the joint distribution of (6.1) can be written as

$$P(Y_{1:T}, S_{1:T}|\tau) = \prod_{i=1}^{K} \pi_i^{S_{1,i}} \prod_{t=1}^{T-1} \prod_{i=1}^{K} \prod_{j=1}^{K} a_{ij}^{S_{t,i}, S_{t+1,j}} \prod_{t=1}^{T} \prod_{i=1}^{K} \phi_i(y_t; \mu_i, \sigma_i^2)^{S_{t,i}} \tag{6.2}$$

In the Bayesian framework, we choose the following prior distributions for π, A, ζ,

$$\begin{aligned}
p(\pi) &= \mathrm{Dir}(\pi_1, \cdots, \pi_K; u_1^\pi, \cdots, u_K^\pi),\ u_i^\pi = \frac{C^\pi}{K}, C^\pi > 0, \\
p(A) &= \prod_{i=1}^{K} p(a_i) = \prod_{i=1}^{K} \mathrm{Dir}(a_{i1}, \cdots, a_{iK}; u_1^A, \cdots, u_K^A),\ u_i^A = \frac{C^A}{K}, C^A > 0, \\
p(\zeta) &= \prod_{i}^{K} \mathrm{N}(\mu_i|\sigma_i^2; \gamma, \eta^2)\mathrm{IG}(\sigma_i^2; \alpha, \beta)
\end{aligned} \tag{6.3}$$

where C^π, C^A, γ, η^2, α, and β are hyper-parameters. These priors are chosen due to the conjugacy considerations. To have these priors providing little influence on the posterior distributions, we assign small numbers to C^π and C^A, for example, 0.1; we set γ equal to data mean, β and η^2 equal to data variance, and α equal to 2.

The joint posterior distribution of $\pi, A, \{\mu_i, \sigma_i^2\}_{i=1}^K$, and $S_{1:T}$ can be obtained as,

$$\begin{aligned}
&P(\pi, A, \{\mu_i, \sigma_i^2\}_{i=1}^K, S_{1:T}|y_{1:T}) \propto \\
&p(S_1|\pi) \prod_{t=1}^{T-1} p(S_{t+1}|S_t; A) \prod_{t=1}^{T} p(y_t|S_t; \{\mu_i, \sigma_i^2\}_{i=1}^K) p(\pi) p(A) p(\{\mu_i, \sigma_i^2\}_{i=1}^K).
\end{aligned} \tag{6.4}$$

The following important marginal posterior distributions are then given by

$$
\begin{aligned}
p(\tau, S_{1:T}|Y_{1:T}) &= \frac{1}{Z} P(Y_{1:T}, S_{1:T}|\tau)p(\tau), && (6.5)\\
p(\tau|Y_{1:T}) &= \frac{1}{Z}\sum_{S_{1:T}} p(Y_{1:T}, S_{1:T}|\tau)p(\tau), && (6.6)\\
p(S_{1:T}|Y_{1:T}) &= \int p(S_{1:T}|\tau, Y_{1:T})p(\tau|Y_{1:T})\, d\tau, && (6.7)
\end{aligned}
$$

where Z is the marginal likelihood, $p(\tau)$ is a prior distribution. All posteriors of (6.5), (6.6), (6.7) are intractable in most practical cases, since the normalising factor Z involves sums over exponentially large numbers of paths.

In VB, we consider an approximation which is factorized as:

$$
q(\tau, S_{1:T}) = q(\pi)\prod_{i=1}^{K} q(a_i)\prod_{i}^{K} q(\mu_i|\sigma_i^2)q(\sigma_i^2)q(S_{1:T}),
$$

where a_i is the i^{th} row vector of the transition probability matrix A.

Given that the likelihood can be expressed in (6.2), It is straightforward to check that it follows a parameter separate parameterization. Then Applying Theorem 2.2 yields the following results,

$$
\begin{aligned}
\hat{q}(S_{1:T}) &= \frac{\prod_{i=1}^{K}\pi_i^{*S_{1,i}}\prod_{t=1}^{T-1}\prod_{i=1}^{K}\prod_{j=1}^{K} a_{ij}^{*S_{t,i},S_{t+1,j}}\prod_{t=1}^{T}\prod_{i=1}^{K}\theta^{*S_{t,i}}}{\tilde{Z}}, && (6.8)\\
\hat{q}(\boldsymbol{\pi}) &= \mathrm{Dir}(\pi_1, ..., \pi_K; w_1^{\pi}, ..., w_K^{\pi})\ ; w_i^{\pi} = u_i^{\pi} + w_i^{s}, && (6.9)\\
\hat{q}(A) &= \prod_{i}^{K}\mathrm{Dir}(a_{i1}, ..., a_{iK}; w_{i1}^{A}, ..., w_{iK}^{A})\ ; w_{ij}^{A} = u_j^{A} + v_{ij}^{s}, && (6.10)\\
\hat{q}(\mu_i|\sigma_i^2) &= \mathrm{N}\left(\gamma_i', \eta_i'^2\right), \gamma_i' = \frac{\eta^2\gamma + p_i^s}{\eta^2 + q_i^s}, \eta_i'^2 = \frac{\sigma_i^2}{\eta^2 + q_i^s}, && (6.11)\\
\hat{q}(\sigma_i^2) &= \mathrm{IG}\left(\alpha_i', \beta_i'\right), \alpha_i' = \alpha + \frac{q_i^s}{2}, \beta_i' = \beta + \frac{r_i^s}{2} + \frac{\eta^2}{2}\frac{(q_i^s - \gamma p_i^s)^2}{(q_i^s + \eta^2)^2} && (6.12)
\end{aligned}
$$

Where

$$\pi_i^* = e^{\mathrm{E}_{q(\boldsymbol{\pi})}[\log \pi_i]} = \exp\big(\psi(w_i^\pi) - \psi(\sum_i w_i^\pi)\big), \tag{6.13}$$

$$a_{ij}^* = e^{\mathrm{E}_{q(A)}[\log(a_{ij})]} = \exp\big(\psi(w_{ij}^A) - \psi(\sum_{j=1} w_{ij}^A)\big), \tag{6.14}$$

$$\begin{aligned}\theta_{i,t}^* &= e^{\mathrm{E}_{q(\boldsymbol{\zeta})}[\log \phi_i(y_t)]} \\ &= \exp\bigg(-\frac{1}{2}\log 2\pi - \frac{1}{2}(\log\beta_i' - \psi(\alpha_i')) - \frac{1}{2}\bigg((y_t-\gamma_i')^2\frac{\alpha_i'}{\beta_i'} + \frac{1}{\eta^2+q_i^s}\bigg)\bigg),\end{aligned} \tag{6.15}$$

$$\begin{aligned}w_i^s &= \mathrm{E}_{q(S_{1:T})}[S_{1,i}], \quad v_{ij}^s = \sum_{t=1}^{T-1}\mathrm{E}_{q(S_{1:T})}\left[S_{t,i}S_{t+1,j}\right], \quad u_i^s = \sum_{t=1}^{T-1}\mathrm{E}_{q(S_{1:T})}[s_{t,i}]y_t^2, \\ p_i^s &= \sum_{t=1}^{T-1}\mathrm{E}_{q(S_{1:T})}[s_{t,i}]y_t, \quad q_i^s = \sum_{t=1}^{T-1}\mathrm{E}_{q(S_{1:T})}[s_{t,i}], \quad r_i^s = \sum_{t=1}^{T-1}(\gamma_i'-y_t)^2 E_{q(S_{1:T})}[s_{t,i}]\end{aligned} \tag{6.16}$$

6.2.2 Interpretation of results

First, all approximating distributions above turn out to lie in well-known parametric families. The only unknown quantities are the parameters of these distributions. The evaluation of parameters of $q(\pi)$, $q(A)$, $q(\mu_i|\sigma_i^2)$, and $q(\sigma_i^2)$ requires the knowledge of $q(S_{1:T})$, and also the evaluation of π_i^*, a_{ij}^* and $\theta_{i,t}^*$ requires the knowledge of $q(\pi)$, $q(A)$, $q(\mu_i|\sigma_i^2)$, and $q(\sigma_i^2)$. This structure leads to an iterative updating scheme, described in Algorithm 9.

The main computational effort in Algorithm 9 is computing $\mathrm{E}_{q(S_{1:T})}[S_{t,i}]$ and $\mathrm{E}_{q(S_{1:T})}\left[S_{t,i}S_{t+1,j}\right]$ which have no simple tractable forms. We note that the distributional form of $q(S_{1:T})$ given in (6.8) has a very similar structure as the conditional distribution of $p(S_{1:T}|Y_{1:T},\tau)$ for which the forward-backward algorithm (Baum et al., 1970) is commonly used to compute $\mathrm{E}_{p(S_{1:T}|Y_{1:T},\tau)}[S_{t,i}|Y_{1:T},\tau]$ and $\mathrm{E}_{p(S_{1:T}|Y_{1:T},\tau)}\left[S_{t,i}S_{t+1,j}|Y_{1:T},\tau\right]$. Therefore, we also use the forward-backward algorithm to compute $\mathrm{E}_{q(S_{1:T})}[S_{t,i}]$ and $\mathrm{E}_{q(S_{1:T})}\left[S_{t,i}S_{t+1,j}\right]$.

The conditional distribution of $q(\mu_i|\sigma_i^2)$ is $\mathrm{N}\left(\mu_i|\sigma_i^2;\gamma_i',\frac{\sigma_i^2}{\kappa_i}\right)$, then the marginal distribution of μ_i is the location-scale t distribution, denoted as $t_{2\alpha_i'}\left(\mu_i;\gamma_i',\frac{\kappa_i}{\beta_i'/\alpha_i'}\right)$, where the density

Algorithm 9 Variational Bayes algorithm for the RSLN model

Initialize $w_i^{s(0)}$, $p_i^{s(0)}$, $q_i^{s(0)}$, $v_{ij}^{s\ (0)}$, and $r_i^{s(0)}$ at step 0

while $w_i^{\pi(t-1)}$, $w_{ij}^{A(t-1)}$, $\gamma_i'^{(t-1)}$, $\alpha_i'^{(t-1)}$, $\beta_i'^{(t-1)}$, $\pi_i^{*(t-1)}$, $a_{ij}^{*\ (t-1)}$, and $\theta_{i,t}^{*\ (t-1)}$ do not converge **do**

1. Compute $w_i^{\pi(t)}$, $w_{ij}^{A(t)}$, $\gamma_i'^{(t)}$, $\kappa_i{}^{(t)}$, $\alpha_i'^{(t)}$, and $\beta_i'^{(t)}$at step t by

$$w_i^{\pi(t)} = \frac{C_\pi^K}{K} + w_i^{s(t-1)}, \quad w_{ij}^{A(t)} = \frac{C_\pi^A}{K} + v_{ij}^{s\ (t-1)}, \quad \gamma_i'^{(t)} = \frac{\eta^2\gamma + p_i^{s(t-1)}}{\eta^2 + q_i^{s(t-1)}},$$

$$\kappa_i{}^{(t)} = \eta^2 + q_i^{s(t-1)}, \quad \alpha_i'^{(t)} = \alpha + \frac{q_i^{s(t-1)}}{2}, \quad \beta_i'^{(t)} = \beta + \frac{r_i^{s(t-1)}}{2} + \frac{\eta^2}{2}(\gamma_i'^{(t)} - \gamma)^2$$

2. Compute $\pi_i^{*(t)}$, $\theta_{i,t}^{*\ (t)}$ and $a_{ij}^{*\ (t)}$ at step t by

$$\pi_i^{*(t)} = \exp\left\{\psi(w_i^{\pi(t)}) - \psi(\sum_i w_i^{\pi(t)})\right\}, \quad a_{ij}^{*\ (t)} = \exp\left\{\psi(w_{ij}^{A(t)}) - \psi(\sum_{j=1} w_{ij}^{A(t)})\right\}$$

$$\theta_{i,t}^{*\ (t)} = \exp\left\{-\frac{1}{2}\log 2\pi - \frac{1}{2}(\log\beta_i'^{(t)} - \psi(\alpha_i'^{(t)})) - \frac{1}{2}\left((y_t - \gamma_i'^{(t)})^2\frac{\alpha_i'^{(t)}}{\beta_i'^{(t)}} + \frac{1}{\kappa_i{}^{(t)}}\right)\right\}$$

3. Compute $w_i^{s(t)}$, $v_{ij}^{s\ (t)}$, $p_i^{s(t)}$, $q_i^{s(t)}$, and $r_i^{s(t)}$ at step t by

$$w_i^{s(t)} = \mathrm{E}_{q^{(t)}(S_{1:T})}[S_{1,i}],\ v_{ij}^{s\ (t)} = \sum_{t=1}^{T-1}\mathrm{E}_{q^{(t)}(S_{1:T})}\left[S_{t,i}S_{t+1,j}\right],\ p_i^{s(t)} = \sum_{t=1}^{T-1}\mathrm{E}_{q^{(t)}(S_{1:T})}[s_{t,i}]y_t,$$

$$q_i^{s(t)} = \sum_{t=1}^{T-1}\mathrm{E}_{q^{(t)}(S_{1:T})}[s_{t,i}],\ r_i^{s(t)} = \sum_{t=1}^{T-1}(\gamma_i'^{(t)} - y_t)^2 E_{q^{(t)}(S_{1:T})}[s_{t,i}]$$

$t \Leftarrow t+1$

end while

function of $t_\nu(x;\mu,\lambda)$ is defined as $p(x|\nu,\mu,\lambda) = \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})}\left(\frac{\lambda}{\pi\nu}\right)^{\frac{1}{2}}\left[1+\frac{\lambda(x-\mu)^2}{\nu}\right]^{-\frac{\nu+1}{2}}$, for $x,\mu \in (-\infty,+\infty)$, and $\nu,\lambda > 0$.

To start the Algorithm 9, we need initial values for w_i^s, p_i^s, q_i^s, v_{ij}^s, and r_i^s at step 0. In general, we choose symmetric values for most of the parameters, and choose random values for others. For example, one setting we use is assigning $w_i^s = \frac{1}{K}$, $q_i^s = \frac{T}{K}$, $p_i^s = \frac{T}{K}\gamma$ where γ is the hyper-parameter (see (6.3)), and assign random numbers for v_{ij}^s, and r_i^s.

6.3 Numerical studies

6.3.1 Simulated data

In this section we applied the VB solutions to four sets of simulated data, which are used in Hartman and Heaton (2011). Through these simulated studies, we will test the performance of VB on the detecting the number of regimes and compare it with those of the BIC and the sticky HDP-HMM, described in Hartman and Heaton (2011).

To estimate the number of regimes, we construct a matrix, called the *relative magnitude matrix* (RMM), defined as $A' = \left(\hat{a}'_{ij}\right)$ where $\hat{a}'_{ij} = \frac{w_{ij}^A}{w_0^A}$, $w_0^A = \sum_{i=1}^K \sum_{j=1}^K w_{ij}^A$, and w_{ij}^A is the parameter of $q(A)$ given in (6.10). Our model selection procedure is to fit a VB with a large number of regimes and examine the rows and columns in the RMM. If the values of the entries in the i^{th} row and the i^{th} column of A' are all equal to $\frac{C^A/K}{T-1+C^A\times K}$, then we will declare the regime i nonexistent. This method is validated by the following observations. It can shown that the parameter of v_{ij}^s in w_{ij}^A is equal to the number of times the process leaves regime i and enters regime j. Therefore, for the i^{th} regime, the values of zero for all of v_{ji}^s and v_{ij}^s with $j = 1,\cdots,K$ indicate that there is no transition process entering or leaving regime i.

Table 6.1 specifies the parameters for the four cases, and we generate 671 observations for

each case (equal to the number of months from 1956 January to 2011 September). The parameters used in Case 1 are identical to the maximum likelihood estimates for TSX monthly return data from 1956 to 1999 (Hardy, 2001). Case 2 only has one regime present. Case 3 is similar to case 1, but the two regimes have the same mean. Case 4 adds a third regime. For each case, we use MLE to fit a 1-regime, 2-regime, 3-regime, and 4-regime RSLN model and report the corresponding BIC and log-likelihood scores. We then mis-specify the number of regimes and run a 4-regime VB algorithm.

Table 6.1: Parameters of simulated data

Case	Regime 1 (μ_i, σ_i)	Regime 2 (μ_i, σ_i)	Regime 3 (μ_i, σ_i)	Transition probability
1	(0.012, 0.035)	(-0.016, 0.078)	-	$\begin{pmatrix} 0.963 & 0.037 \\ 0.210 & 0.790 \end{pmatrix}$
2	(0.014, 0.050)	-	-	-
3	(0.000, 0.035)	(0.000, 0.078)	-	$\begin{pmatrix} 0.963 & 0.037 \\ 0.210 & 0.790 \end{pmatrix}$
4	(0.012, 0.035)	(-0.016, 0.078)	(0.04, 0.01)	$\begin{pmatrix} 0.953 & 0.037 & 0.01 \\ 0.210 & 0.780 & 0.01 \\ 0.80 & 0.190 & 0.01 \end{pmatrix}$

Table 6.2 shows the number of iterations which VB takes to converge in each case. On average, VB converges after a hundred iterations. The results of the BIC with the log-likelihood (in parentheses), the relative magnitude matrices, and the posterior probabilities for the models with the different number of regimes estimated by the sticky HDP-HMM (cited from Hartman and Heaton, 2011) are given in Table 6.3. In Case 1, the BIC favors the two-regime model. The posterior probability estimated by the sticky HDP-HMM for the one-regime model is the largest, but there is still a large probability for the two regime model. The relative magnitude matrix clearly shows there are only two regimes whose $\hat{a}'_{ij}$ are not negligible. This implies VB removes excess transition and emission processes, and

discovers the exact number of hidden regimes. In Case 2 and Case 3, both VB and the BIC can select the correct number of regimes, and the posterior probability for the one-regime model estimated by the sticky HDP-HMM is still the largest. In Case 4, VB does not detect the third regime. The transition probability to this regime is only 0.01, and the means and standard deviations of regime 1 makes the rare data from regime 3 easily merged within the data from regime 1. From Table 6.3, it is clear that for all the cases the log-likelihood always increases as the number of regimes increase.

Table 6.2: Computational efficiency of VB

	Case 1	Case 2	Case 3	Case 4
Iterations to converge	62	182	132	94

6.3.2 Real data

In this section, we apply the VB solution to the TSX monthly total return index in the period from January 1956 to December 1999 (528 observations in total and studied in Hardy, 2001, 2002).

A four-regime VB is implemented first. VB converges after 100 iterations about 34.284 seconds (on a MacBook, 2 GHz Intel Core 2 Duo processor, 1 GB 667 MHz SDRAM, and Mac OS X 10.4.11). The relative magnitude matrix, given in Table 6.4, clearly shows that VB identifies two regimes. This matches both the BIC and AIC-based results (Hardy, 2001). Based on this results, we then fit a two-regime VB which converges after 83 iterations in about 14.241 seconds. Table 6.5 gives the marginal distributions for all the parameters.

Tabel 6.6 gives the maximum likelihood estimates (cited from Hardy, 2001) and mean parameters computed by the MCMC method (cited from Hardy, 2002) and mean parameters computed by VB. It clearly shows that the point estimates by VB are similar to those by

Table 6.3: The estimated number of regimes by VB, BIC, and the sticky HDP-HMM

Case	No. of reg.	MLE BIC(log likehood)	HDP-HMM post. prob.	VB relative magnitude matrix
1	1	1108.875(1115.384)	0.647	$\begin{pmatrix} 0.14357 & 0.00004 & 0.00004 & 0.03153 \\ 0.00004 & 0.00004 & 0.00004 & 0.00004 \\ 0.00004 & 0.00004 & 0.00004 & 0.00004 \\ 0.03018 & 0.00004 & 0.00004 & 0.79428 \end{pmatrix}$
	2	1158.227(1174.499)	0.214	
	3	1156.370(1182.405)	0.088	
	4	1153.150(1188.948)	< 0.052	
2	1	1045.448(1051.957)	0.864	$\begin{pmatrix} 0.99944 & 0.00004 & 0.00004 & 0.00004 \\ 0.00004 & 0.00004 & 0.00004 & 0.00004 \\ 0.00004 & 0.00004 & 0.00004 & 0.00004 \\ 0.00004 & 0.00004 & 0.00004 & 0.00004 \end{pmatrix}$
	2	1038.360(1054.632)	0.109	
	3	1030.733(1056.768)	0.020	
	4	1026.882(1062.680)	< 0.006	
3	1	1110.903(1117.411)	0.629	$\begin{pmatrix} 0.11322 & 0.00004 & 0.00004 & 0.02647 \\ 0.00004 & 0.00004 & 0.00004 & 0.00004 \\ 0.00004 & 0.00004 & 0.00004 & 0.00004 \\ 0.02659 & 0.00004 & 0.00004 & 0.83327 \end{pmatrix}$
	2	1139.214(1155.486)	0.221	
	3	1131.904(1157.719)	0.098	
	4	1121.921(1157.940)	< 0.052	
4	1	1044.819(1051.328)	0.641	$\begin{pmatrix} 0.22643 & 0.00004 & 0.00004 & 0.05518 \\ 0.00004 & 0.00004 & 0.00004 & 0.00004 \\ 0.00004 & 0.00004 & 0.00004 & 0.00004 \\ 0.05377 & 0.00004 & 0.00004 & 0.66417 \end{pmatrix}$
	2	1092.610(1108.881)	0.203	
	3	1087.435(1113.470)	0.094	
	4	1080.240(1116.038)	< 0.06	

Table 6.4: Estimations of the number of regimes for TSX data

	Jan. 1956 - Dec. 1999
R. M. M.	$\begin{pmatrix} 0.11496 & 0.00005 & 0.00005 & 0.02803 \\ 0.00005 & 0.00005 & 0.00005 & 0.00005 \\ 0.00005 & 0.00005 & 0.00005 & 0.00005 \\ 0.02853 & 0.00005 & 0.00005 & 0.82791 \end{pmatrix}$

Table 6.5: The marginal distributions of the parameters estimated by VB

Parameter	Distribution	Mean	*s.d.*	Transition Probability
μ_1	$t_{454.61}(0.0123, 370778.19)$	0.0123	0.00165	-
σ_1^2	IG(227.30, 0.28)	0.00122(0.0349)	0.00008	-
μ_2	$t_{80.39}(-0.0161, 12987.55)$	-0.0161	0.00889	-
σ_2^2	IG(40.20, 0.24)	0.00603(0.0777)	0.00098	-
$p_{1,2}$	Beta(15.21, 434.78)	0.0338	0.00851	$\begin{pmatrix} 0.9662 & 0.0338 \\ 0.1969 & 0.8031 \end{pmatrix}$
$p_{2,1}$	Beta(15.00, 61.21)	0.1969	0.04525	

MLE and MCMC. The numbers in parenthesis in Table 6.6 are the standard deviations computed by the three methods respectively. It is worth to note that Hardy (2001, 2002) reports the standard deviations for σ_1 and σ_2, rather than σ_1^2 and σ_2^2. Our derivations are based on σ_1^2 and σ_2^2. For a easy comparison, we transform the standard deviations for σ_1 and σ_2 to the standard deviations for σ_1^2 and σ_2^2 by delta methods. The numbers in parenthesis in Table 6.6 in the rows of σ_1 and σ_2 are the standard deviations for σ_1^2 and σ_2^2.

6.4 Diagnostics

The posterior variance and covariance estimated by the VB approximations and by the MCMC sample moments (cited from Hardy, 2002) are given in Table 6.7, and the ratios of

Table 6.6: The parameter estimates and their standard deviations given by VB, MLE and MCMC

	VB	MLE	MCMC
μ_1	0.0123(0.00165)	0.0123(0.002)	0.0122(0.002)
σ_1	0.0349(0.00008)	0.0347(0.00007)	0.0351(0.00014)
$p_{1,2}$	0.0338(0.00851)	0.0371(0.012)	0.0334(0.012)
μ_2	-0.0161(0.00889)	-0.0157(0.010)	-0.0164(0.010)
σ_2	0.0777(0.00098)	0.0778(0.00140)	0.0804(0.00145)
$p_{2,1}$	0.1969(0.04525)	0.2101(0.086)	0.2058(0.065)

the two sets of estimates are given in the last row in 6.7. These ratios indicate that VB underestimated the actual posterior variances, and again strongly distorts the correlation structure.

Table 6.7: Posterior variance and covariance estimated by VB and MCMC

	Post. cov.: (μ_1, σ_1,$p_{1,2}$, μ_2, σ_2,$p_{2,1}$)
MCMC samples	s.d.:
	0.002, 0.00014, 0.012, 0.010, 0.00145, 0.065
	correlation coeff.:
	-0.16, 0.17, -0.34,-0.10, -0.11, 0.08, -0.17, 0.22,-0.25,-0.15, 0.06,-0.04,0.34,-0.14,0.12
VB approx.	s.d.:
	0.0017, 0.00008, 0.0085, 0.0089, 0.0010, 0.045
	correlation coeff.:
	0, 0, 0, 0, 0, 0, 0, 0, 0, 0,0, 0, 0, 0, 0
Ratios	1.49, 2.95, 2.16, 1.43, 1.74, 2.13

The methods proposed in Chapter 4 can be used to reconstruct the true posterior covariance matrix. For the optimal affine transformations, a random sample of size 600 is generated from (6.10), (6.11) and (6.12). There are 27 parameters in total in the linear transformation. Maximizing the posterior probability over these parameters can be done with the Newton's method or standard search methods. For the marginal and stepwise methods, we use 4000

samples to compute the acceptance rates in VBAIMH, which significant improve the computational efficiency, compared with other MCMC methods, for example, the random walk Metropolis-Hastings.

The final results are given in Table 6.8, where the upper part of the table gives the ratios of the variances estimated by MCMC, the affine transformation, the marginal method and the stepwise method versus those estimated by VB; the lower part of the table gives correlation coefficients estimated by MCMC, the affine transformation, the marginal method and the stepwise method. The fact that all variance ratios are greater than one indicates that VB underestimates the actual posterior variances.

In this example, we work through a complete cycle of the stepwise method to obtain the exact estimations on variances and correlation coefficients, which provide a quantitative correction on VB approximations. In fact, each step of the stepwise method can provide a qualitative diagnostics. Table 6.10 gives all numerical results in the stepwise method. The values of m_i in Step 1 are all greater than 1, which indicates that the VB variances are smaller than the true ones, since a conditional variance always penalizes the marginal variance. In Step 2, the values of the pair of $\lambda_{k,1}$ and $\lambda_{k,2}$ provide possible information about the sign of the correlation.

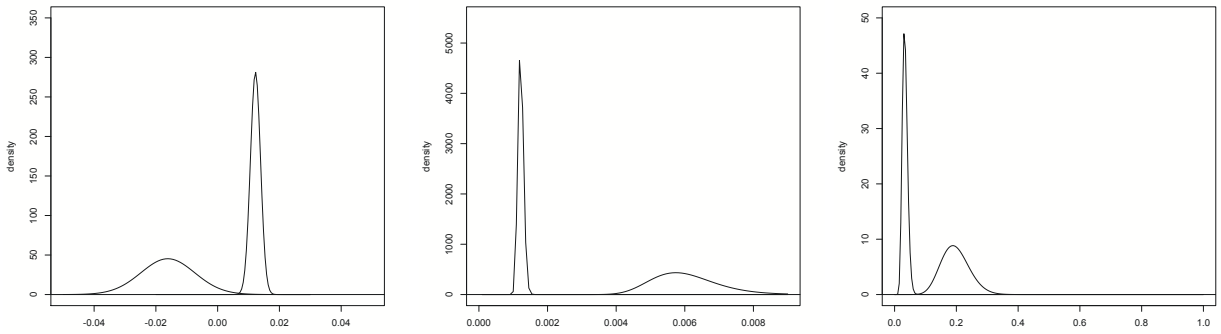
6.5 Normality

The variational approximations for μ_i, σ_i^2, and p_{ij} (or a_i in the general notation given in (6.10)) are given in Table 6.5, where we can see that the approximation for μ_i results in a t distribution, for σ_i^2 it is an inverse Gamma, and for p_{ij} it is a Beta. However, the density functions plotted in Figure 6.2 show that all the figures present an approximately symmetric and bell shaped pattern.

In this section, we aim to explain this phenomenon through exploring the asymptotic

Table 6.8: 3 diagnostics methods

	MCMC	Affine	Marginal	Stepwise
	Variance ratios (over VB variances)			
μ_1	1.49	1.44	1.97	1.22
σ_1^2	2.95	1.60	1.76	1.63
$p_{1,2}$	2.16	2.81	1.41	2.21
μ_2	1.43	1.20	1.82	1.34
σ_2^2	1.74	1.57	2.00	1.49
$p_{2,1}$	2.13	2.34	2.15	1.97
	Correlation coefficients			
ρ_{12}	-0.1630	-0.1217	-0.1175	-0.1266
ρ_{13}	0.1681	0.2228	0.1220	0.1367
ρ_{23}	-0.3438	-0.2970	-0.3831	-0.3388
ρ_{14}	-0.1043	-0.1294	-0.1874	-0.1275
ρ_{24}	-0.1094	-0.0903	-0.0649	-0.0865
ρ_{34}	0.0796	0.0221	0.0856	0.0507
ρ_{15}	-0.1678	-0.1856	-0.1061	-0.1328
ρ_{25}	0.2235	0.1793	0.1008	0.1390
ρ_{35}	-0.2517	-0.1604	-0.2890	-0.2160
ρ_{45}	-0.1476	-0.0747	-0.0116	-0.0231
ρ_{16}	0.0552	0.0528	0.0942	0.0640
ρ_{26}	-0.0374	-0.0690	0.0461	-0.0772
ρ_{36}	0.3385	0.3985	0.5947	0.3518
ρ_{46}	-0.1433	-0.1154	-0.1664	-0.0989
ρ_{56}	0.1238	0.1291	0.1434	0.1023



(a) Left: μ_2, Right: μ_1 (b) Left: σ_1^2, Right: σ_2^2 (c) Left: $p_{1,2}$, Right: $p_{2,1}$

Figure 6.2: The VB marginal distributions of the parameters

behavior of the VB approximations. Theorem 6.1 provided in Chen (1985) will be our main tool.

Theorem 6.1 (Chen, 1985) *Let* $\{f_T(\theta), T = 1...\}$ *be a sequence of probability density function defined on the parameter space* Ω*. It is assumed that, for each* T*, there exists a strictly local maximum,* m_T*, of* f_T*, and let* $\Sigma_T = (-L_T''(m_T))^{-1}$*, where* $L_T = \log(f_T)$*. Then, the following three conditions of (C1), (C2), and (C3) are the necessary and sufficient conditions for the standardized* Z_T*, defined as*

$$Z_T = \Sigma_T^{-1/2}(\theta - m_T),$$

to converge in distribution to the standard normal distribution:
(C1) $\sigma_T^2 \to 0$*, as* $n \to \infty$*, where* σ_T^2 *is the largest eigenvalue of* Σ_T
(C2) For any $\epsilon > 0$*, there exists an* N *and* $\delta > 0$*, such that, for any* $n > N$ *and* $\theta \in \Omega$ *and* $|\theta - m_T| < \delta$*,* L_T'' *exists and satisfies* $I - A(\epsilon) \leq L_T''(\theta)(L_T''(m_T))^{-1} \leq I + A(\epsilon)$*, where* I *is the identity matrix and* $A(\epsilon)$ *is a a positive semidefinite symmetric matrix whose largest eigenvalue tends to zero as* $\epsilon \to 0$*.*
(C3) For any $\delta > 0$*, there exists an* S*, and two positive numbers* c *and* p*, for any* $T > S$

and $\theta \in \Omega$ *and* $|\theta - m_T| < \delta$, $L_T(\theta) - L_T(m_T) < -c((\theta - m_T)^T \Sigma_T^{-1}(\theta - m_T))^p$.

In the applications to variational approximations of the above general results, $f_T(\theta) \equiv q(\theta)$, where $q(\theta)$ is the VB density of θ, dependent implicitly on the relevant data y_T. As the original author points out, the asymptotic results given in Theorem 6.1 are non-probabilistic in nature; the data y_T are not regarded as random due to sampling plans, but simply form a well-defined deterministic sequence, and hence the other quantities related to the data y_T.

We would like to make it clear that the investigation in this section is explorational to explain the symmetric and bell shapes in Figure 6.2, rather than rigorous proofs, as the reasons stated above. Based on certain assumptions, we can show that $q(\sigma_i^2)$, $q(a_i)$, and $q(\mu_i)$ in RSLN models satisfy the three conditions given in Theorem 6.1. The following several identities will be used in the derivations. Given the definitions in (6.16), q_i^s can be interpreted as the number of times the process in regime i. Given the model (the number of total regimes) is correct, q_i^s is proportional to the total number of observations T. Therefore, we consider $q_i^s = O(T)$. By the similar argument, we consider $v_{ij}^s = O(T)$ as well.

First, we consider $q(\sigma_i^2)$. Since $q(\sigma_i^2) = \text{IG}(\alpha_i', \beta_i')$, it can be shown that $m_T = \frac{\beta_i'}{\alpha_i'+1}$ and $(L_T''(m_T))^{-1} = \frac{\beta_i'^2}{(\alpha_i'+1)^3}$. It is obvious that $\alpha_i' = O(T)$. We define $c = \max_t(\gamma_i' - y_t)^2$, and then we can obtain $\beta_i' \leq O(1) + O(1) + \frac{c}{2}q_i^s = O(T)$, and $\frac{\beta_i'^2}{(\alpha_i'+1)^3} \to 0$, as $T \to \infty$. For condition 2, we have $L_T''(\theta)(L_T''(m_T))^{-1} = 1$. In fact, the probability density function of the inverse gamma distribution is strictly concave in terms of σ_i^2, then $L_T(\sigma_i^2) - L_T(m_T) = L_T'(\xi)(\sigma_i^2 - m_T) < -\frac{L_T'(\xi)}{2}|\sigma_i^2 - m_T| = -c((\frac{\sigma_i^2 - m_T}{\Sigma_T})^2)^{1/2}$, $\xi \in (\sigma_i^2, m_T)$ and $c = \frac{L_T'(\xi)\Sigma_T}{2}$

Second, we consider $q(a_i)$. Since $q(a_i) = \text{Dir}(a_i; w_{i1}^A, ..., w_{iK}^A)$, it can be shown that $m_T = \left(\frac{w_{i1}^A - 1}{\sum_{j=1}^K w_{ij}^A - K}, \cdots, \frac{w_{iK}^A - 1}{\sum_{j=1}^K w_{ij}^A - K}\right)$ and Σ_T is a diagonal matrix with the jj^{th} entry of $\frac{w_{ij}^A - 1}{(\sum_{j=1}^K w_{ij}^A - K)^2}$. Since $w_{ij}^A = O(T)$ for all $j = 1, \cdots, K$, it is obvious that the largest eigenvalue of Σ_t goes to zero as $T \to \infty$. For condition 2, we have $L_T''(\theta)(L_T''(m_T))^{-1} = I_K$. In fact, $(\log(q(a_i)))''$ is a diagonal matrix with the jj^{th} entry of $-\frac{w_{ij}^A - 1}{a_{ij}^2}$. When T is large,

$(\log(q(a_i)))''$ is negative definite. Thus, $q(a_i)$ is a strictly concave function. By the similar argument above, condition 3 is satisfied.

Last, as discussed above, $q(\mu_j)$ is $t_{2\alpha_i'}\left(\mu_i; \gamma_i', \frac{\kappa_i}{\beta_i'/\alpha_i'}\right)$. The degree of the t distribution is $2\alpha_j' = O(T)$. As T becomes large, it approaches a normal distribution.

6.6 Conclusions

The variational Bayes method provides a new framework to approximate the posterior quantities. We applied this method to the regime-switching log-normal model and provide solutions to account for both model uncertainty and parameter uncertainty. The numerical results show that our method can recover exactly the number of regimes and gives the reasonable point estimates. The VB method is also demonstrated to be very computationally efficient.

The application on the TSX monthly total return index data in the period from January 1956 through December 1999 confirms the similar results in the literature in finding the number of regimes. Finally, we address the potential problem of the method in estimating the posterior covariance structure, and use the three methods proposed in Chapter 4 to compute the covariance matrix.

6.7 Appendix

For the affine transformation method, the estimated $\hat{A}$ and $\hat{B}$ is given by,

$$\hat{A} = \begin{pmatrix} 1.200 & & & & & \\ -0.008 & 1.257 & & & & \\ 1.864 & -45.563 & 1.570 & & & \\ -0.722 & -11.953 & 0.026 & 1.076 & & \\ -0.156 & 2.657 & -0.014 & -0.013 & 1.208 & \\ 2.195 & -52.512 & 3.292 & -1.028 & 11.013 & 1.361 \end{pmatrix}; \hat{B} = \begin{pmatrix} -0.002 \\ 0.000 \\ 0.017 \\ 0.024 \\ -0.002 \\ -0.220 \end{pmatrix}.$$

For the marginal approximation method, it requires 21 projections. The directional vectors with the corresponding acceptance rates and EAR table readings are given in Table 6.9.

For the stepwise method. The numerical results for Step 1 and 2 are given in Table 6.10.

Table 6.9: The marginal approximation method for the regime-switching lognormal model

direction	Acceptance rate	EAR reading: l_i
$\frac{1}{3}(1,1,1,1,1,1)$	0.852	1.6
$\frac{1}{3}(1,-1,1,1,1,1)$	0.811	1.84
$\frac{1}{3}(1,1,-1,1,1,1)$	0.843	1.66
$\frac{1}{3}(1,1,1,-1,1,1)$	0.867	1.50
$\frac{1}{3}(1,1,1,1,-1,1)$	0.755	2.16
$\frac{1}{3}(1,1,1,1,1,-1)$	0.845	0.62
$\frac{1}{3}(-1,1,1,1,1,1)$	0.873	1.48
$\frac{1}{3}(-1,-1,1,1,1,1)$	0.897	1.36
$\frac{1}{3}(-1,1,-1,1,1,1)$	0.877	1.48
$\frac{1}{3}(-1,1,1,-1,1,1)$	0.884	1.44
$\frac{1}{3}(-1,1,1,1,-1,1)$	0.873	1.48
$\frac{1}{3}(-1,1,1,1,1,-1)$	0.849	0.64
$\frac{1}{3}(1,-1,-1,1,1,1)$	0.817	0.56
$\frac{1}{3}(1,-1,1,-1,1,1)$	0.841	1.60
$\frac{1}{3}(1,-1,1,1,-1,1)$	0.819	1.76
$\frac{1}{3}(1,-1,1,1,1,-1)$	0.890	0.73
$\frac{1}{3}(1,1,-1,-1,1,1)$	0.802	1.52
$\frac{1}{3}(1,1,-1,1,-1,1)$	0.865	0.65
$\frac{1}{3}(1,1,-1,1,1,-1)$	0.854	1.62
$\frac{1}{3}(1,1,1,-1,-1,1)$	0.768	2.08
$\frac{1}{3}(1,1,1,-1,1,-1)$	0.865	0.65

Table 6.10: The stepwise method for the regime-switching lognormal model

	Variance	Acceptance rate	EAR readings	Ratio
Step 1	m_1^2	0.951	1.15	-
	m_2^2	0.862	1.42	-
	m_3^2	0.817	1.61	-
	m_4^2	0.928	1.28	-
	m_5^2	0.807	1.35	-
	m_6^2	0.831	1.63	-
Step 2	$\lambda_{1,1}^2, \lambda_{1,2}^2$	0.929, 0.936	1.29, 0.91	-
	$\lambda_{2,1}^2, \lambda_{2,2}^2$	0.949, 0.931	0.97, 1.30	-
	$\lambda_{3,1}^2, \lambda_{3,2}^2$	0.950, 0.960	1.18, 0.89	-
	$\lambda_{4,1}^2, \lambda_{4,2}^2$	0.844, 0.875	1.61, 0.90	-
	$\lambda_{5,1}^2, \lambda_{5,2}^2$	0.966, 0.962	0.98, 1.14	-
	$\lambda_{6,1}^2, \lambda_{6,2}^2$	0.871, 0.873	1.51, 0.77	-
	$\lambda_{7,1}^2, \lambda_{7,2}^2$	0.921, 0.943	1.27, 0.91	-
	$\lambda_{8,1}^2, \lambda_{8,2}^2$	0.921, 0.774	0.88, 2.09	-
	$\lambda_{9,1}^2, \lambda_{9,2}^2$	0.952, 0.907	0.99, 1.04	-
	$\lambda_{10,1}^2, \lambda_{10,2}^2$	0.926, 0.930	0.94, 1.29	-
	$\lambda_{10,1}^2, \lambda_{11,2}^2$	0.858, 0.819	1.53, 0.77	-
	$\lambda_{11,1}^2, \lambda_{12,2}^2$	0.917, 0.783	0.76, 2.00	-
	$\lambda_{13,1}^2, \lambda_{13,2}^2$	0.866, 0.884	0.96, 0.96	-
	$\lambda_{14,1}^2, \lambda_{14,2}^2$	0.930, 0.961	1.19, 0.94	-
	$\lambda_{15,1}^2, \lambda_{15,2}^2$	0.889, 0.810	0.82, 1.22	-
Step 3	r_1	-	-	-0.127
	r_2	-	-	0.137
	r_3	-	-	-0.338
	r_4	-	-	-0.128
	r_5	-	-	-0.087
	r_6	-	-	0.057
	r_7	-	-	-0.133
	r_8	-	-	0.139
	r_9	-	-	-0.216
	r_{10}	-	-	-0.023
	r_{11}	-	-	0.064
	r_{12}	-	-	-0.077
	r_{13}	-	-	0.352
	r_{14}	-	-	-0.099
	r_{15}	-	-	0.102

Chapter 7

Future Work

In this chapter, we will summarize what has been achieved in the thesis. We also outline a number of future research problems.

7.1 Summary of the thesis

This thesis aims to give insight into the general applicability and usefulness of the VB method and also study its applications to various real-world problems. The work is split between theoretical and applied work, and mainly includes the following aspects:

1. *Its general applicability and properties*
2. *Diagnostics for VB approximations*
3. *Variational applications*

In Chapter 2, we discuss the working principle of variational methods and introduce the important issue of variational learning. Two techniques, which are used to obtain the

VB approximations, are described. To overcome the limitations in the current variational development, we proposed a factorization scheme with a more general dependency structure than is traditional in VB, and propose a special parameterization for a parametric family called the parameter separation parameterization. We provide the variational formulation based on these new frameworks.

In Chapter 3, we mainly address the question: how good is the VB approximation in any given example. In particular, we study the mean structure approximations by the VB method, and show how it is possible using VB approximations to approach model selection tasks such as determining the dimensionality of a model, or variable selection. We also studied the local convergence properties of the method, the effects of choosing improper priors, and the effects of different choices of factorization schemes.

In Chapter 4, we mainly address the question: how well the approximations represent the actual posterior distributions, in particular how good is the approximation to the covariance structure, and how can it be improved. We propose three methods to diagnose VB approximations. These methods are quick and easy to use, and require no sophisticated tuning expertise. In numerical studies, we work through a series of statistical problems or models to examine these methods.

In Chapter 5, we consider VB in Bayesian non-parametric problems. We shows how the VB method provides a computationally efficient technique in the context of hierarchical modelling using a Dirichlet process priors, in particular, handling non-conjugacy in its prior specification. It shows, by using the parameter separation parameterization, that we are able to provide a full variational solution for the Dirichlet process. We also provide a VB approximation to the posterior predictive distribution and compare it with results derived from two Markov chain Monte Carlo methods. For the truncated DP, we propose an empirical method to determine the number of distinct components in a finite dimensional DP.

In Chapter 6, as a concrete application study, we work through the VB method on regime-switching lognormal models and present solutions to quantify both the uncertainty in the parameters and model specification. Through a series numerical comparison studies with likelihood based methods and MCMC methods on the simulated and real data sets, we show that the VB method can recover exactly the model structure, gives the reasonable point estimates, and is very fast. We also use the methods of Chapter 4 to improved the variance estimation.

In the next a few sections, we present some research problems which are the natural extensions to the development of what have been achieved so far.

7.2 Asymptotic normality

In this thesis, we have studied various statistical models, for example, the Normal random sample with unknown mean and variance, one-way random-effect models, mixtures of normal models, regime-switching lognormal models, and hierarchical regression models. The numerical results showed that we often found that these approximations tend to be normally distributed, at least to a good order of approximation. In particular, in Chapter 6 Section 6.5, we examine the special case, the RSLN model, and the VB approximations satisfy the normality conditions of Chen (1985).

In the classical Bayesian theory, it is well established that under regularity conditions, the posterior distribution of $I_n^{1/2}(\theta - \hat{\theta}_n)$ converges to $N(0, I)$ with probability one under the true model for the data, where $\hat{\theta}_n$ is the maximum likelihood estimate and I_n is the Fisher information matrix evaluated at $\hat{\theta}_n$. This result can be found in Le Cam (1952, 1986); Walker (1969); Chen (1985). A more recent result about the asymptotic property of the variational application can be found in Hall et al. (2011). They studied the asymptotic distributional behavior of Gaussian variational approximate estimators of the parameters in

a single predictor Poisson mixed model. We believe that under certain regularity conditions, this asymptotic result will hold for general VB approximations.

This asymptotic result has practical meanings and will provide theoretical grounds for our development of the VB method. For example, in our proposed VBAIMH method, the choice of a normal as the proposal distribution is motivated by this asymptotic normality observation. Moreover, this result can help us to have a better understanding about the function F. It can be shown under normality conditions the function F turns out to be the BIC.

In this section, we will give one more demonstration to explore the asymptotic behavior of the VB approximations, and we will look at a two-component mixtures of Normal model. Recall the VB approximations for this model are given by

$\hat{q}(z) = \prod_{i=1}^{n} \text{multinomial}(z_i; 1, q_{i1}, q_{i2})$,

where $q_{ij} = \frac{\exp\left(\psi(a_j')-\psi(\sum_{j=1}^{K} a_j')-\frac{1}{2}(\log f_j'-\psi(e_j'))-\frac{1}{2}(c_j'-x_i')^2\frac{e_j'}{f_j'}-\frac{1}{2}\frac{1}{d_j^2+\sum_{i=1}^{n} q_{ij}}\right)}{\sum_{j=1}^{K}\exp\left(\psi(a_j')-\psi(\sum_{j=1}^{K} a_j')-\frac{1}{2}(\log f_j'-\psi(e_j'))-\frac{1}{2}(c_j'-x_i')^2\frac{e_j'}{f_j'}-\frac{1}{2}\frac{1}{d_j^2+\sum_{i=1}^{n} q_{ij}}\right)}$,

$\hat{q}(\pi) = \text{Beta}(a_1', a_2')\ ; a_j' = a_j + \sum_{i=1}^{n} q_{ij}$,

$\hat{q}(\mu_j|\sigma_j^2) = \text{N}\left(c_j', d_j'^2\right), c_j' = \frac{d_j^2 c_j + \sum_{i=1}^{n} x_i q_{ij}}{d_j^2+\sum_{i=1}^{n} q_{ij}}, d_j'^2 = \frac{\sigma_j^2}{d_j^2+\sum_{i=1}^{n} q_{ij}}$,

$\hat{q}(\sigma_j^2) = \text{IG}\left(e_j', f_j'\right), e_j' = e_j + \frac{\sum_{i=1}^{n} q_{ij}}{2}, f_j' = f_j + \frac{d_j^2}{2}(c_j'-c_j)^2 + \frac{1}{2}\sum_{i=1}^{n} q_{ij}(c_j'-x_i)^2$.

We still check the normality conditions described in Theorem 6.1 for $q(\pi)$, $q(\sigma_j^2)$, and $q(\mu_j)$. First, we consider $q(\sigma_j^2)$. Since $q(\sigma_j^2) = \text{IG}\left(e_j', f_j'\right)$, it can be shown that $m_n = \frac{f_j'}{e_j'+1}$ and $(L_n''(m_n))^{-1} = \frac{f_j'^2}{(e_j'+1)^3}$. As the similar argument in Section 6.5, we consider $\sum_{i=1}^{n} q_{ij} = O(n)$, then we have $e_j' = O(n)$. We define $c = \max_i (c_j' - x_i)^2$, then $f_j' \leq O(1) + O(1) + \frac{c}{2}\sum_{i=1}^{n} q_{ij} = O(n)$. Then $\frac{f_j'^2}{(e_j'+1)^3} \to 0$, as $n \to 0$. For condition 2, we have $L_n''(\theta)(L_n''(m_n))^{-1} = 1$. In fact, the probability density function of the inverse gamma distribution is strictly concave, then $L_n(\sigma_j^2) - L_n(m_n) = L_n'(\xi)(\sigma_j^2 - m_n) < -\frac{L_n'(\xi)}{2}|\sigma_j^2 - m_n| = -c\left(\left(\frac{\sigma_j^2 - m_n}{\Sigma_n}\right)^2\right)^{1/2}$, $\xi \in (\sigma_j^2, m_n)$ and $c = \frac{L_n'(\xi)\Sigma_n}{2}$.

Second, we consider $q(\pi)$. Since $q(\pi) = \text{Beta}(a_1^{'}, a_2^{'})$, it can be shown that $m_n = \frac{a_1^{'}-1}{a_1^{'}+a_2^{'}-2}$ and $(L_n^{''}(m_n))^{-1} = \frac{(a_1^{'}-1)(a_2^{'}-1)}{(a_1^{'}+a_2^{'}-2)^3}$. Since $a_1^{'} = O(n)$ and $a_2^{'} = O(n)$, it is obvious that $(L_n^{''}(m_n))^{-1} \to 0$. For condition 2, we have $L_n^{''}(\theta)(L_n^{''}(m_n))^{-1} = 1$. In fact, $q^{''}(\pi) = -(\frac{a_1^{'}-1}{\pi^2} + \frac{a_2^{'}-1}{(1-\pi)^2})$. When n is large, $q^{''}(\pi) < 0$ strictly. Thus, $q(\pi)$ is a strictly concave function. By the similar argument above, condition 3 is satisfied.

Last, we consider $q(\mu)$. The conditional distribution of $q(\mu_j|\sigma_j^2)$ is given by $\text{N}\left(c_j^{'}, d_j^{'2}\right)$, where $c_j^{'} = \frac{d_j^2 c_j + \sum_{i=1}^n x_i q_{ij}}{d_j^2 + \sum_{i=1}^n q_{ij}}, d_j^{'2} = \frac{\sigma_j^2}{d_j^2 + \sum_{i=1}^n q_{ij}}$. The marginal distribution of $q(\mu)$ is a location-scale t distribution, given by $t_{2e_j^{'}}\left(c_j^{'}, \frac{d_j^2 + \sum_{i=1}^n q_{ij}}{f_j^{'}/e_j^{'}}\right)$. The degree of the t distribution is $2e_j^{'} = O(n)$, then as n becomes large, it approaches a normal distribution.

7.3 Consistence of variational estimators

In this thesis, we have seen that the mean structure approximations using the VB method are impressive. A natural question raised here is why VB works very well? In the classical Bayesian theory, it is well-known, for example, see Bickel and Yahav (1969); Ibragimov and Khas' minskii (1973), that under regularity conditions, the Bayes estimator, under mean square error (MSE) risk is asymptotically unbiased and it converges in distribution to the normal distribution: $\sqrt{n}(\delta_n - \theta_0) \to N\left(0, \frac{1}{I(\theta_0)}\right)$, where $I(\theta_0)$ is the fisher information of θ_0. It follows that the Bayes estimator under MSE is asymptotically efficient. It is also well-known that the maximum likelihood estimator (MLE) is asymptotically normal and efficient. Moreover, White (1982) shows that when the true distribution is unknown, the MLE is still a consistent estimator for the parameters, which minimize the Kullback-Leibler from the true distribution to the hypothetic distribution.

Through the various applications in this thesis, we observe that there exist a close relationship between the maximum likelihood and Bayes estimators under variational approx-

imations. We can show it in the following simple example. Recall that the variational approximation for μ, in the normal random sample with unknown mean and variance, is given as

$$q(\mu) \quad = \quad N(\gamma_{vb}, \eta^2_{vb}), \quad \gamma_{vb} = \frac{\frac{\gamma}{\eta^2} + n\bar{y}E_{\hat{q}(\sigma^2)}[\frac{1}{\sigma^2}]}{\frac{1}{\eta^2} + nE_{\hat{q}(\sigma^2)}[\frac{1}{\sigma^2}]}, \quad \eta^2_{vb} = \frac{1}{\frac{1}{\eta^2} + nE_{\hat{q}(\sigma^2)}[\frac{1}{\sigma^2}]},$$

where $\bar{y}$ is the sample mean. The Bayes estimator under MSE is

$$E_{q(\mu)}[\mu] \quad = \quad \gamma_{vb} = \frac{\frac{\gamma}{\eta^2} + n\bar{y}E_{q(\sigma^2)}[\frac{1}{\sigma^2}]}{\frac{1}{\eta^2} + nE_{q(\sigma^2)}[\frac{1}{\sigma^2}]},$$

where γ and η^2 are hyper-parameters. The MLE is $\bar{y}$. Therefore we can get

$$E_{q(\mu)}[\mu] \quad = \quad \frac{\frac{\gamma}{\eta^2}}{\frac{1}{\eta^2} + nE_{q(\sigma^2)}[\frac{1}{\sigma^2}]} + \frac{nE_{q(\sigma^2)}[\frac{1}{\sigma^2}]}{\frac{1}{\eta^2} + nE_{q(\sigma^2)}[\frac{1}{\sigma^2}]}\bar{y},$$

The last equation implies that, for $n \to \infty$, the Bayes estimator is close to the MLE. Motivated by these observations, we believe that the Bayes estimators under variational approximations is also asymptotically consistent.

7.4 Overfitted models

In this thesis we showed how it is possible using VB approximations to approach model selection tasks such as determining the dimensionality, or selections of variables. We apply different methods for the different models. In the finite mixture of normals model, the mixing proportions are used to decide the number of components. In the truncated Dirichlet process, the probability weights are used to decide truncation level. In regime-switching models, a relative magnitude matrix, that was constructed based on the posterior distribution of the transition probability matrix, is used to remove excess transition processes. In the linear

regression model, the regression setup was embedded in a hierarchical framework where latent variables are used to identify subset choices.

A common feature of these methods lies in that we use a larger model, whose dimension or numbers of components are greater than the true model, referred as to a overfitted model. From the numerical results on the above applications, we have seen that the VB approximations tend to empty the extra components, in the sense that either mixing proportions or probability weights or transition probability tend to zero.

Rousseau and Mengersen (2011) proved that for a mixture model, the posterior distribution of the overfitted model will concentrate on a subset of parameters so that the extra components have weight zero. It would be very interesting if this theoretical result can be extended to a broader class of models, for example, the models we have already studied above. If this result can further be verified for the VB method, we will obtain an important theoretical framework for our numerical studies.

7.5 Diagnostics for VB

The variational approximations may be inadequate from a number of perspectives. In Chapter 4 we list some of the most important: 1) By definition the posterior covariance structure is distorted; 2) VB might only converge to a local optimal, and hence it might focus on a single mode of a multimodal solution; 3) They may be inaccurate in higher order posterior moments. Chapter 4 particularly targets on the diagnostics of the covariance structure. In the future research, we are interested in designing other diagnostic tests to detect the other types of inadequacy, for examples, the possible multimodalily, and the errors in higher order moments.

Chapter 4 also discusses how it is possible to apply the proposed methods in the high-dimensional problems. The marginal method has special advantages to target more specific

errors by designing the project on to particular directions.

In the affine transformation method, currently, we have targeted the most general form of linear transformation to correct any inadequacy in posterior mean and variance, by specifying the transformation matrix A to be a lower triangular. The dimension of the space of A could be problematic in high dimensional problems. One possible solution could be to restrict the matrix A to be of a particular class, such as diagonal or more sparse.

In fact, we have applied a diagonal transformation matrix to the example of the 3-dimension Multivariate normal distributions, given in Section 4.3.1. Our reasoning is that by specifying a diagonal transformation matrix we may be able to obtain the posterior marginal variance estimates, while ignoring the estimation of the correlation structures. However, the numerical results show that the recovered marginal variances are only close to the actual conditional variances. This result is interesting. In the future work, we will investigate how to design the transformation matrix to target more specific forms of the diagnostics.

Chapter 4 introduces a novel way to use acceptance rates to estimate posterior variances. As discussed in Chapter 4, we are also interested in further calibrating EAR as a diagnostics tool to measure the inadequacy in marginal approximations, by using VB approximations directly as the proposal distributions in VBAIMH. A key problem lies in when posterior normality is not satisfied well, a single value of EAR may be diluted by the confounding of many errors: wrong variance, wrong skewness, wrong tail behaviour. In the further research, we will investigate how to separate these confounding factors to have EAR targeted on a special form of inadequacy.

The three methods proposed in Chapter 4 mainly considers the most commonly used situation, where VB is fully factorized. If in the cases VB can offer partial correlation structure, our methods should be able to use these additional information. The affine transformation method can adapt the additional information directly. For the marginal and stepwise methods, our current development only uses the knowledge from mean and variance estimates.

Our future work will be interested in how to design the marginal and stepwise methods to adapt additional information.

7.6 Open questions

The numerical results in Chapter 5 show that for Dirichlet processes the posterior predictive likelihoods approximated by VB are very close to those computed by MCMC methods. There are many questions associated with the numerical observations. For example, whether this result is true in general for other models? Under what conditions will VB provide good approximations on posterior predictive distributions? Whether we can have a general formalization on the variational approximation of posterior predictive likelihoods?

Theorem 2.1 provides a general framework for the factorization scheme which contains more general dependency structures. This factorization scheme has been used for many applications in the thesis. We will look for more cases of non-trivial posterior dependence in the models which we can build into the variational model. We will also be interested in other simplification techniques, other than the simplification on model dependency structure, where we can apply the variational principle.

Another interesting question is how we would decide to use VB or MCMC in real-world problems? Currently, we can list a number of advantages of VB over MCMC methods. First, it is the computational advantage. As we have seen throughout the dissertation, VB is not only fast, but also free from many implementation issues. For example, the convergence is straightforward in VB, but it is a major concern for MCMC methods. Given initial values, iterative update in VB is straightforward, but MCMC methods might require many tuning expertise. Second, VB can provide good mean structure approximation. However, VB has several limitations with respect to MCMC methods. As we have already seen in Chapter 4, VB may distort the posterior dependence structure; VB may only converge to a local

optimal in a multimodal solution; VB may be inaccurate in higher order posterior moments. Essentially, VB is a trade-off of learning accuracy to gain the computation efficiency. For 'real-world' problems we need to carefully balance the advantages brought by reduced model complexity against the disadvantages.

Bibliography

Akaike, H. (1974). A new look at the statistical model identification. *Automatic Control, IEEE Transactions 19*, pp. 716–723.

Ans, B., J. Hérault, and C. Jutten (1985). Adaptive neural architectures: detection of primitives. *Proc. of Cognitiva'85*, pp. 593–597.

Attias, H. (1999). Independent factor analysis. *Neural computation 11*(4), pp. 803–851.

Attias, H. (2000). A variational Bayesian framework for graphical models. *Advances in Neural Information Processing Systems*, pp. 209–215.

Barber, D. and C. M. Bishop (1998). Ensemble learning in Bayesian neural networks. *Neural Networks and Machine Learning*, pp. 215–237.

Baum, L. E., T. Petrie, G. Soules, and N. Weiss (1970). A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *The Annals of Mathematical Statistics 41*(1), pp. 164–171.

Beal, M. J. (2003). *Variational algorithms for approximate Bayesian inference*. Ph. D. thesis, University of London.

Beal, M. J. and Z. Ghahramani (2001). Propagation algorithms for variational Bayesian learning. *Neural Information Processing Systems*, pp. 507–513.

Beal, M. J. and Z. Ghahramani (2003). The variational Bayesian EM algorithm for incomplete data: with application to scoring graphical model structures. *Bayesian Statistics*, pp. 453–463.

Beal, M. J. and Z. Ghahramani (2006). Variational Bayesian learning of directed graphical models with hidden variables. *Bayesian Analysis*, pp. 793–832.

Berger, J. O. (1985). *Statistical Decision Theory and Bayesian Analysis*. Springer-Verlag.

Bernardo, J. M. (1976). Algorithm as 103: Psi (digamma) function. *Journal of the Royal Statistical Society: Series C (Applied Statistics) 25*(3), pp. 315–317.

Bernardo, J. M. and A. F. Smith (1994). *Bayesian Theory*. New York: Wiley.

Bickel, P. J. and J. A. Yahav (1969). Some contributions to the asymptotic theory of Bayes solutions. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete 11*(4), pp. 257–276.

Bishop, C. M. (1999). Variational principal components. In *Proceedings Ninth International Conference on Artificial Neural Networks, ICANN'99*, pp. 509–514.

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. 2006 Springer Science+Business Media.

Blackwell, D. and J. B. MacQueen (1973). Ferguson distributions via Pólya urn schemes. *The Annals of Statistics 1*(2), pp. 353–355.

Blei, D. M. and M. I. Jordan (2006). Variational inference for Dirichlet process mixtures. *Bayesian Analysis 1*(1), pp. 121–143.

Box, G. E., W. G. Hunter, and J. S. Hunter (1978). *Statistics for experimenters: an introduction to design, data analysis, and model building*. John Wiley & Sons.

Box, G. E. P. and G. C. Tiao (1973). *Bayesian Inference in Statistical Analysis*. Wiley Classics.

Brooks, S. (1998). Markov chain Monte Carlo method and its application. *Journal of the Royal Statistical Society: series D (the Statistician) 47*(1), pp. 69–100.

Brown, P. J., M. Vannucci, and T. Fearn (1998). Multivariate Bayesian variable selection and prediction. *Journal of the Royal Statistical Society: Series B (Methodological) 60*(3), pp. 627–641.

Bush, C. A. and S. N. MacEachern (1996). A semiparametric Bayesian model for randomised block designs. *Biometrika 83*(2), pp. 275–285.

Chen, C.-F. (1985). On asymptotic normality of limiting density functions with Bayesian implications. *Journal of the Royal Statistical Society: Series B (Methodological)*, pp. 540–546.

Chen, M.-H. (1994). Importance-weighted marginal Bayesian posterior density estimation. *Journal of the American Statistical Association 89*(27), pp. 818–824.

Chen, M.-H., Q.-M. Shao, and J. G. Ibrahim (2000). *Monte Carlo Methods in Bayesian Computation*. Springer.

Chib, S. (1995). Marginal likelihood from the Gibbs output. *Journal of the American Statistical Association 90*(432), pp. 1313–1321.

Chib, S. and I. Jeliazkov (2001). Marginal likelihood from the Metropolis-Hastings output. *Journal of the American Statistical Association 96*(453), pp. 270–281.

Chipman, H. (1996). Bayesian variable selection with related predictors. *Canadian Journal of Statistics 24*, pp. 17–36.

Corduneanu, A. and C. M. Bishop (2001). Variational Bayesian model selection for mixture distributions. In *Artificial intelligence and Statistics*, Volume 2001, pp. 27–34.

Critchley, F., P. Marriott, and M. Salmon (1994). Preferred point geometry and the local differential geometry of the Kullback-Leibler divergence. *The Annals of Statistics 22*(3), pp. 1587–1602.

Dempster, A. P., N. M. Laird, and D. B. Rubin (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, pp. 1–38.

DiCiccio, T. J., R. E. Kass, A. Raftery, and L. Wasserman (1997). Computing Bayes factors by combining simulation and asymptotic approximations. *Journal of the American Statistical Association 92*(439), pp. 903–915.

Diestel, R. (2010). *Graph Theory* (Fourth ed.). Springer-Verlag, Heidelberg.

Draper, N. and H. Smith (1981). *Applied regression analysis. Series in Probability and Mathematical Statistics*. Wiley.

Escobar, M. D. (1994). Estimating normal means with a Dirichlet process prior. *Journal of the American Statistical Association 89*(425), pp. 268–277.

Escobar, M. D. and M. West (1995). Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association 90*(430), pp. 577–588.

Faes, C., J. Ormerod, and M. Wand (2011). Variational Bayesian inference for parametric and nonparametric regression with missing data. *Journal of the American Statistical Association 106*(495), pp. 959–971.

Ferguson, T. S. (1973). A Bayesian analysis of some nonparametric problems. *The Annals of Statistics*, pp. 209–230.

Ferguson, T. S. (1983). Bayesian density estimation by mixtures of normal distributions. *Recent advances in statistics 24*, pp. 287–302.

Feynman, R. P. (1972). *Statistical Mechanics*. W.A. Benjamin Inc.

Gelfand, A. E. and D. K. Dey (1994). Bayesian model choice: asymptotics and exact calculations. *Journal of the Royal Statistical Society: Series B (Methodological)*, pp. 501–514.

Gelfand, A. E. and A. F. M. Smith (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association 85*(410), pp. 398–409.

Gelfand, I. M. and S. V. Fomin (1964). *Calculus of variations*. Prentice-Hall.

Gelman, A., J. B. Carlin, H. S. Stern, and D. B. Rubin. (1995). *Bayesian Data Analysis*. Chapman & HALL/CRC.

Geman, S. and D. Geman (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *Biometrika 6*(3), pp. 721–741.

George, E. I. and R. E. McCulloch (1993). Variable selection via Gibbs sampling. *Journal of the American Statistical Association 88*(423), pp. 881–889.

George, E. I. and R. E. McCulloch (1997). Approaches for Bayesian variable selection. *Statistica Sinica 7*(2), pp. 339–373.

Geweke, J. (1989). Bayesian inference in econometric models using Monte Carlo integration. *Econometrica 60*(6), pp. 1317–1339.

Ghahramani, Z. and M. J. Beal (2000a). Graphical models and variational methods. *Advanced Mean Field Method—Theory and Practice*.

Ghahramani, Z. and M. J. Beal (2000b). A variational inference for Bayesian mixtures of factor analysers. *Advances in Neural Information Processing Systems 12*, pp. 449–455.

Ghahramani, Z. and M. I. Jordan (1997). Factorial hidden Markov models. *Machine learning 29*(2-3), pp. 245–273.

Green, P. J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika 82*, pp. 711–732.

Haff, L. R. (1991). The variational form of certain Bayes estimators. *The Annals of Statistics 19*(3), pp. 1163–1190.

Hall, P., J. T. Ormerod, and M. P. Wand (2011). Theory of Gaussian variational approximation for a poisson mixed model. *Statistica Sinica 21*, pp. 369–389.

Hamilton, J. D. (1989). A new approach to the economic analysis of nonstationary time series and the business cycle. *Econometrica 57*(2), pp. 357–384.

Hammersley, J. and P. Clifford (1971). Markov fields on finite graphs and lattices. 1971. *Unpublished manuscript*.

Hardy, M. R. (2001). A regime-switching model of long-term stock returns. *North American Actuarial Journal 5*(2), pp. 41–53.

Hardy, M. R. (2002). Bayesian risk management for equity-linked insurance. *Scandinavian Actuarial Journal 2002*(3), pp. 185–211.

Hartigan, J. (1964). Invariant prior distributions. *The Annals of Mathematical Statistics 35*(2), pp. 836–845.

Hartman, B. M. and M. J. Heaton (2011). Accounting for regime and parameter uncertainty in regime-switching models. *Insurance: Mathematics and Economics 49*, pp. 429–437.

Hastings, K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika 57*(1), pp. 97–109.

Hinton, G. E. and D. Van Camp (1993). Keeping the neural networks simple by minimizing the description length of the weights. In *Proceedings of the sixth annual conference on Computational learning theory*, pp. 5–13. ACM.

Holden, L., R. Hauge, and M. Holden (2009). Adaptive independent Metropolis Hastings. *The Annals of Applied Probability 19*(1), pp. 395–413.

Hosino, T., K. Watanabe, and S. Watanabe (2005). Stochastic complexity of variational Bayesian hidden Markov models. In *Proceedings 2005 IEEE International Joint Conference on Neural Networks*, Volume 2, pp. 1114–1119. IEEE.

Ibragimov, I. and R. Khas' minskii (1973). Asymptotic behavior of some statistical estimators ii. limit theorems for the a posteriori density and Bayes' estimators. *Theory of Probability & Its Applications 18*(1), pp. 76–91.

Ishwaran, H. and L. F. James (2001). Gibbs sampling methods for stick-breaking priors. *Journal of the American Statistical Association 96*(453), pp. 161–173.

Ishwaran, H. and M. Zarepour (2000). Markov chain Monte Carlo in approximate Dirichlet and beta two-parameter process hierarchical models. *Biometrika 87*(2), pp. 371–390.

Jaakkola, T. S. (1997). *Variational methods for inference and estimation in graphical models*. Ph. D. thesis, MIT.

Jaakkola, T. S. and M. I. Jordan (1999). Variational methods and the QMR-DT database. *Journal of Artificial Intelligence Research 10*, pp. 291–322.

Jeffreys, H. (1961). *Theory of Probability*. Oxford University Press.

Jensen, J. (1906). Sur les fonctions convexes et les inégalités entre les valeurs moyennes. *Acta Mathematica 30*(1), pp. 175–193.

Johnson, R. A. and D. Wichern (2002). Clustering distance methods, and ordination. *Applied Multivariate Statixtical Analysis*, pp. 668–747.

Jordan, M. I., Z. Ghahramani, T. S. Jaakkola, and L. K. Saul (1999). An introduction to variational methods for graphical models. *Machine Learning 37*, pp. 183–233.

Karush, W. (1939). Minima of functions of several variables with inequalities as side constraints. Master's thesis, Dept. of Mathematics, Univ. of Chicago.

Kass, R. E. and A. E. Raftery (1995). Bayes factors. *Journal of the American Statistical Association 90*.

Kim, J.-O. and C. W. Mueller (1978). *Factor Analysis: Statistical Methods and Practical Issues*. Sage Publications.

Kuhn, H. W. and A. W. Tucker (1951). Nonlinear programming. In *Proceedings of the second Berkeley symposium on mathematical statistics and probability*, Volume 5. California.

Lappalainen, H. (1999). Ensemble learning for independent component analysis. In *Proc. Int. Workshop on Independent Component Analysis and Signal Separation (ICA'99)*, pp. 7–12.

Lawrence, N. D. and C. M. Bishop (2000). Variational Bayesian independent component analysis. *Univ of Cambridge Tech Report*.

Le Cam, L. M. (1952). *On some asymptotic properties of maximum likelihood estimates and related Bayes' estimates*. Ph. D. thesis, University of California, Berkeley.

Le Cam, L. M. (1986). *Asymptotic methods in statistical theory*. Springer-Verlag New York, Inc.

Leonard, T., J. S. J. Hsu, and K.-W. Tsui (1989). Bayesian marginal inference. *Journal of the American Statistical Association 84*(408), pp. 1051–1058.

Li, S. (2009). *Markov random field modeling in image analysis*. Springer.

Liu, J. S. (1996). Metropolized independent sampling with comparisons to rejection sampling and importance sampling. *Statistics and Computing 6*, pp. 113–119.

Liu, J. S. (2008). *Monte Carlo strategies in scientific computing*. Berlin: Springer-Verlag.

MacEachern, S. N. (1994). Estimating normal means with a conjugate style Dirichlet process prior. *Communications in Statistics-Simulation and Computation 23*(3), pp. 727–741.

MacEachern, S. N. and P. Müller (1998). Estimating mixture of Dirichlet process models. *Journal of Computational and Graphical Statistics 7*(2), pp. 223–238.

MacKay, D. (1995). Developments in probabilistic modelling with neural networks - ensemble learning. *Neural Networks: Artificial Intelligence and Industrial Applications*, pp. 14–15.

MacKay, D. J. (1997). Ensemble learning for hidden Markov models. Technical report, Technical report, Cavendish Laboratory, University of Cambridge.

McGrory, C. A. and D. Titterington (2007). Variational approximations in Bayesian model selection for finite mixture distributions. *Computational Statistics and Data Analysis 8*, pp. 5352–5367.

McGrory, C. A., D. Titterington, R. Reeves, and A. N. Pettitt (2009). Variational Bayes for estimating the parameters of a hidden Potts model. *Statistics and Computing 19*(3), pp. 329–340.

Metropolis, N., A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller (1953). Equations of state calculations by fast computing machines. *Journal of Chemical Physics 21*(6), pp. 1087–1092.

Miskin, J. W. (2000). Ensemble learning for independent component analysis. In *Advances in Independent Component Analysis*. Citeseer.

Mitchell, T. J. and J. J. Beauchamp (1988). Bayesian variable selection in linear regression. *Journal of the American Statistical Association 83*(404), pp. 1023–1032.

Mohammad-Djafari, A. and H. Ayasso (2009). Variational bayes and mean field approximations for Markov field unsupervised estimation. In *Machine Learning for Signal Processing, 2009. MLSP 2009*, pp. 1–6.

Moulines, E., J.-F. Cardoso, and E. Gassiat (1997). Maximum likelihood for blind separation and deconvolution of noisy signals using mixture models. In *Proceedings of the ICASSP'97*, Volume 5, Munich, Germany, pp. 3617–3620.

Neal, R. M. (2000). Markov chain sampling methods for Dirichlet process mixture models. *Journal of Computational and Graphical Statistics 9*(2), pp. 249–265.

Neal, R. M. (2003). Slice sampling. *The Annals of Statistics 31*(3).

Neal, R. M. and G. Hinton (1998). A view of the EM algorithm that justifies incremental, sparse, and other variants. *In Learning in Graphical Model*.

Ntzoufras, I., J. J. Forster, and P. Dellaportas (2000). Stochastic search variable selection for log-linear models. *Journal of Statistical Computation and Simulation 68*(1), pp. 23–37.

Onge, J. M. S., P. M. Krueger, and R. G. Rogers (2008). Historical trends in height, weight, and body mass: Data from U.S. major league baseball players, 1869-1983. *Economics & Human Biology 6*(3), pp. 482–488.

Ormerod, J. T. and M. P. Wand (2012). Gaussian variational approximate inference for generalized linear mixed models. *Journal of Computational and Graphical Statistics 21*(1), pp. 2–17.

Parisi, G. (1988). *Statistical Field Theory*. Advanced Book Classics.

Penny,, R. C. W. and S. Roberts (2000). An ensemble learning approach to independent component analysis. *IEEE Workshop on Neural Networks for Signal Processing*.

Peskun, P. H. (1973). Optimum Monte Carlo sampling using Markov chains. *Biometrika 60*(3), pp. 607–612.

Propp, J. G. and D. B. Wilson (1996). Exact sampling with coupled Markov chains and applications to statistical mechanics. *Random Structure Algorithms 9*, pp. 223–252.

Raftery, A. E., M. A. Newton, J. M. Satagopan, and P. N. Krivitsky (1994). Estimating the integrated likelihood via posterior simulation using the harmonic mean identity (with discussion). *Bayesian Statistics 8 (edited by J.M. Bernardo et al.)*, pp. 1–45.

Robert, C. P. and G. Casella (1999). *Monte Carlo statistical methods*, Volume 2. Springer New York.

Robert, C. P. and G. Casella (2010). *Monte Carlo Statistical Methods*. Springer.

Roberts, G. O., A. Gelman, and W. R. Gilks (1997). Weak convergence and optimal scaling of random walk Metropolis algorithms. *The Annals of Applied Probability 7*(1), pp. 110–120.

Roberts, G. O. and J. S. Rosenthal (1998). Markov-chain Monte Carlo: Some practical implications of theoretical results. *The Canadian Journal of Statistics 26*(1), pp. 5–20.

Roberts, G. O. and J. S. Rosenthal (2004). General state space Markov chains and mcmc algorithms. *Probability Surveys 1*, pp. 20–71.

Roberts, G. O. and R. L. Tweedie (1999). Bounds on regeneration times and convergence rates for Markov chains. *Stochastic Processes and their Applications 80*, pp. 211–229.

Rockafellar, T. (1997). *Convex analysis*, Volume 28. Princeton university press.

Rosenthal, J. S. (1995). Convergence rates for Markov chains. *SIAM Review 37*(3), pp. 387–405.

Rousseau, J. and K. Mengersen (2011). Asymptotic behaviour of the posterior distribution in overfitted mixture models. *Journal of the Royal Statistical Society: Series B (Methodological) 73*(5), pp. 689–710.

Rue, H., S. Martino, and N. Chopin (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested laplace approximations. *Journal of the Royal Statistical Society: Series B (Methodological) 71*, pp. 319–392.

Saul, L. K., T. Jaakkola, and M. I. Jordan (1996). Mean field theory for sigmoid belief networks. *Journal of Artificial Intelligence Research 4*, pp. 61–76.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics 6*(2), pp. 461–464.

Sethuraman, J. (1994). A constructive definition of Dirichlet priors. *Statistica Sinica 4*, pp. 639–650.

Shun, Z. and P. McCullagh (1995). Laplace approximation of high dimensional integrals. *Journal of the Royal Statistical Society: Series B (Methodological) 57*(4).

Spiegelhalter, D. J., N. G. Best, B. P. Carlin, and A. Van Der Linde (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Methodological) 64*(4), pp. 583–639.

Stephens, M. (2000). Bayesian analysis of mixture models with an unknown number of components - an alternative to reversible jump methods. *The Annals of Statistics 28*(1), pp. 40–74.

Storkey, A. J. (2000). Dynamic trees: A structured variational method giving efficient propagation rules. In *Uncertainty in Artificial Intelligence*, pp. 566–573.

Tierney, L. (1994). Markov chains for exploring posterior distributions. *Annals of Statistics 22*, pp. 1701–1762.

Tierney, L., R. E. Kass, and J. B. Kadane (1989). Approximate marginal densities of nonlinear functions. *Biometrika 76*(3), pp. 425–433.

Wainwright, M. J. and M. I. Jordan (2008). Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning 1*(1-2), pp. 1–305.

Walker, A. (1969). On the asymptotic behaviour of posterior distributions. *Journal of the Royal Statistical Society: Series B (Methodological)*, pp. 80–88.

Wang, B. and D. Titterington (2004). Convergence and asymptotic normality of variational Bayesian approximations for exponential family models with missing values. In *Proceedings of the 20th conference on Uncertainty in artificial intelligence*, pp. 577–584. AUAI Press.

Wang, B. and M. Titterington (2006). Convergence properties of a general algorithm for calculating variational Bayesian estimates for a normal mixture model. *Bayesian Analysis 1*(3), pp. 625 – 650.

Watanabe, K. and S. Watanabe (2005). Stochastic complexity for mixture of exponential families in variational Bayes. *Proceedings of the 16th International Conference on Algorithmic Learning Theory (ALT2005)*, pp. 107–121.

Watanabe, K. and S. Watanabe (2007). Stochastic complexities of general mixture models in variational Bayesian leaning. *Neural Networks 20*(2), pp. 210–219.

Watanabe, S. (2009). *Algebraic Geometry and Statistical Learning Theory*. Cambridge University Press New York, NY, USA.

West, M. and M. D. Escobar (1993). *Hierarchical priors and mixture models, with application in regression and density estimation*. Institute of Statistics and Decision Sciences, Duke University.

White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica: Journal of the Econometric Society*, pp. 1–25.

Williams, C. K. and N. J. Adams (1999). DTs: dynamic trees. *Advances in neural information processing systems*, pp. 634–640.

Winn, J. (2003). *Variational Message Passing and its Applications*. Ph. D. thesis, Department of Physics, University of Cambridge.

Zhao, J.-h. and P. L. Yu (2009). A note on variational Bayesian factor analysis. *Neural Networks 22*(7), pp. 988–997.