Variational Bayesian Inference for Comparison of VAR(1) models

Adrian James Houghton

Thesis submitted for the degree of

Doctor of Philosophy

School of Mathematics and Statistics

University of Newcastle upon Tyne

Newcastle upon Tyne

United Kingdom

January 2009

Abstract

Suppose that we wish to determine which models in a candidate set are most likely to have given rise to a set of observed data. Then, it is well-established that, from a Bayesian viewpoint, evaluation of the marginal likelihood for each candidate is a crucial step to this end. For the purposes of model comparison, this will enable subsequent computation of both Bayes' factors and posterior model probabilities. Given its evident significance in this area, it is thus regrettable that analytic calculation of the marginal likelihood is often not possible. To tackle this problem, one recent addition to the literature is the variational Bayesian approach.

In this thesis, it is seen that variational Bayes provides efficient, accurate approximations to both the marginal likelihood and the parameter posterior distribution, conditioned on each model. In particular, the theory is applied to ranking sparse, vector autoregressive graphical models of order 1 in both the zero and non-zero mean case. That is, our primary aim is to estimate the unknown sparsity structure of the autoregressive matrix in the process. Moreover, approximate, marginal posterior information about the coefficients of this matrix is also of interest. To enable rapid exploration of higher-dimensional graphical spaces, a Metropolis-Hastings algorithm is presented so that a random walk can be made between neighbouring graphs. The scheme is then tested on both simulated and real datasets of varying dimension.

Acknowledgements

Initially, I would like to express my gratitude to my supervisor, Darren Wilkinson, for his exceptional guidance, expertise and endless patience in the midst of some rather repetitive questioning during my time at Newcastle. Without his influence, this thesis would look somewhat different. Thanks must also go to my advisor, Richard Boys, for providing additional, insightful input.

Without the immense love, support and encouragement of my mum and dad, I would never have got to where I am today. For that, I will always be greatly appreciative. I also extend these words to the rest of my family.

High praise is reserved for all my friends, in particular those within the Department of Statistics, who have always helped me in times of crisis, provided great encouragement and cheered me up with humour of fine quality.

Finally, I would like to acknowledge the financial support of the Engineering and Physical Sciences Research Council.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Suppose that we possess an observed dataset, which has been generated by an incompletely understood underlying process. Then, an important statistical problem is to find a model that explains the inherent trends in the data well. In this case, such a model can subsequently be utilised to make reasonably accurate, future predictions. In real life situations, it is customarily the case that there will be a huge number of complicated factors that will affect the generation of the data. Thus, a standard philosophy to follow is that a model is merely an approximation to the mechanism giving rise to the data.

Assume we now have a collection of possible models in competition, referred to as a candidate set. Then, the *model selection* task is to choose the 'best' model in the set, given the data. That is, we ideally require the model that forms the most suitable representation of the reality. Unfortunately, the procedure is non-trivial. It is valid to ask at this stage what constitutes such a selection. For instance, a sufficiently complex model (with many parameters) will be able to provide a good fit, *i.e.* the underlying trends will be well reflected. Else, the model is said to *underfit* the data. So, the fit in a simple model can be improved by adding extra parameters and will be equivalent to before if these new parameters are set to zero.

However, as Beal (2003) indicates, model fit alone is an unsatisfactory criterion for choosing between models. In any model, by its definition of being an approximation, it will be practically infeasible to capture exactly each factor that has given rise to the data. Hence, we refer to these factors as noise. A sufficiently complex model, with its exceptional flexibility, can be made to produce an exact fit. However, this is not because the trends are being accurately approximated, but instead the noise is being absorbed into the model. That is, an excessive number of parameters will resultantly fit the noise in the data. So, although such a model may be the best fitting in a candidate set given a dataset, it will provide inadequate predictions of future observations, generated by the same truth, as the noise will vary in these new observations. In this case, the model is said to *overfit* the data.

To summarise, by choosing the most complex model in a candidate set, we are not precisely approximating the intangible reality. Instead, there is a necessary trade-off to be made between the fit of a model to a particular dataset and its complexity, in terms of how well it predicts new observations. These issues are at the forefront for any technique used to select a model given observed data. In this chapter, some of these established methods are presented. In particular, we focus primarily on how a Bayesian tackles the model comparison dilemma.

1.1 A Bayesian perspective

Let $\mathcal{M} = \{M_1, \ldots, M_R\}$ be a set of R candidate models, where each model is a probability distribution. Given the observation of data D, we want to compare the credibility of these candidates. To effect this, the fundamentals of the Bayesian approach to model comparison are now examined, illustrated by, *inter alia*, Kass and Raftery (1995) and Chipman et al. (2001). We first require some initial definitions. If $\boldsymbol{\theta}_i = (\boldsymbol{\theta}_{i1}, \ldots, \boldsymbol{\theta}_{id})^T$ is a set of unknown parameters specific to model M_i, then let $p(D \,|\, \boldsymbol{\theta}_i,\, M_i)$ be the probability density function of D given the value of $\boldsymbol{\theta}_i$ (also referred to as the likelihood function for $\boldsymbol{\theta}_i$).

A Bayesian framework dictates the introduction of *priors* on all unknowns. Thus, in this case, let $p(\boldsymbol{\theta}_i \mid M_i)$ be the prior distribution over the parameters of each model. Moreover, we suppose that $p(M_i)$ is the prior probability assigned to each model itself. Upon the observation of data D, we are able to update our prior beliefs about the probability of each model. Thus, by Bayes' Theorem, the *posterior probability* of model M_i is given by

$$p(M_i \mid D) = \frac{p(D \mid M_i)\, p(M_i)}{p(D)}, \tag{1.1}$$

where the probability of the data, a normalising constant, is equivalent to

$$p(D) = \sum_i p(D \mid M_i)\, p(M_i).$$

Moreover, the term $p(D \mid M_i)$ is referred to as the *marginal likelihood* of data D given model M_i, such that

$$p(D \mid M_i) = \int p(D \mid \boldsymbol{\theta}_i,\, M_i)\, p(\boldsymbol{\theta}_i \mid M_i)\, \mathrm{d}\boldsymbol{\theta}_i. \tag{1.2}$$

It is so named since we marginalise, or integrate, over the model parameter space.

We realise that the model posterior, $p(M_i \mid D)$, is a valuable tool to possess when choosing between models. If our task were to pick the most plausible model, we can easily choose that which maximises the value of the posterior probability. So, we can interpret $p(M_i \mid D)$ as the probability that the model M_i is the mechanism that generated the data initially. In other words, this posterior expresses our beliefs, hence quantifies our uncertainty, about each model after the observation of data. Furthermore, we can derive a posterior distribution for the parameters, specific to each model. This is expressed as

$$p(\boldsymbol{\theta}_i \,|\, D,\, M_i) = \frac{p(D \,|\, \boldsymbol{\theta}_i,\, M_i)\, p(\boldsymbol{\theta}_i \,|\, M_i)}{p(D \,|\, M_i)}. \tag{1.3}$$

Upon examination of (1.1) and (1.3), it is noted that computation of the marginal likelihood enables calculation of not only the posterior over models, but also the posterior over parameters. We shall make a further comment on this relationship in due course.

We now turn to the question of specifying a prior over the set of models, namely $p(M_i)$; the same procedure for $p(\boldsymbol{\theta}_i \,|\, M_i)$ is examined in Section 1.1.1. In both cases however, as noted by Chipman et al. (2001), there are two approaches to consider. On the one hand, we could adopt subjective priors, representing our own personal knowledge or beliefs about the unknowns before data is observed. Although a nice proposal, this framework is most idealistic, especially if there are many candidate models in our set, each with high-dimensional parameters $\boldsymbol{\theta}_i$, and we must somehow quantify our information as probability distributions.

Therefore, a pragmatic Bayesian approach is adopted here. In this case, priors are constructed whereby little or even no prior knowledge is available, hence not affecting the construction of the posterior. Such priors are described as being broad, flat, diffuse or vague (Gelman et al., 1995). So, as regards specification of a model prior, a straightforward procedure is to make all models equally plausible, hence representing prior ignorance. Thus, if there are R candidates in our model set $\mathcal{M}$, our prior could be

$$p(M_i) = \frac{1}{R}. \tag{1.4}$$

The above prior follows a (discrete) uniform distribution, whereby each model has been awarded the same prior probability. An interesting point to notice is that, upon using this prior, (1.1) will simplify such that $p(M_i \,|\, D) \propto p(D \,|\, M_i)$ as the model prior cancels.

Hence, on this basis, the model posterior is computed up to a multiplicative constant, and thus the marginal likelihood can be viewed as *the evidence* for each model. By definition, it is the average probability of the data for a given model, with respect to the prior distribution.

As discussed previously in this chapter, it is critical to find a technique to compare models fairly so that more complex models are penalised sufficiently. The marginal likelihood is able to effect this since, by its definition, it naturally integrates out parameters. Thus, it embodies the principle of *Occam's razor*, which states, in general, that a simpler model for the data is preferred over a more complex alternative.

MacKay (1995b) and Beal (2003) discuss this aspect of Bayesian model comparison. Suppose that we have two competing models, M_k and M_l, the former being a simple model and the latter a more complex offering. Consider the space of all datasets of size N. As M_l will possess additional parameters due to its relative, extra complexity, it will be able to model a wider range of datasets than its simpler counterpart, M_k.

For every dataset, the corresponding marginal likelihood for each model can be computed to assess which is the most plausible. Yet, the marginal likelihood over datasets must integrate to 1. Consequently, M_l, although capable of modelling a plethora of datasets, can assign only small marginal likelihoods to each. On the contrary, M_k can award a higher marginal likelihood value to the limited number of datasets that it can model. Thus, if it is possible for a particular dataset D to be modelled by both M_k and M_l, then $p(D \mid M_k) > p(D \mid M_l)$ and the simpler structure is hence favoured. Initially, this phenomenon was displayed diagrammatically in Section 1.3 of MacKay (1995b).

Unfortunately, despite its importance in model comparison, calculation of the marginal likelihood via (1.2) is difficult since the integral could be intractable (*e.g.* if $\boldsymbol{\theta}_i$ is high-dimensional as we integrate over the model parameters). Analytic computation of this quantity is rare. Therefore, a good approximation to this quantity is desirable and techniques that enable this will be presented later in this chapter. At the present time however, we now illustrate its significance in model comparison.

1.1.1 Bayes' Factors

Kass and Raftery (1995) elucidated a simple, but elegant criterion for comparing models in the Bayesian framework, called a *Bayes' factor*. Assume that we have two models, say M_k and M_l, and we want to discover which is the most plausible, given data D. As above, we specify prior probabilities over the models, namely $p(M_k)$ and $p(M_l)$. For subsequent interpretation purposes, we choose to work on the odds scale.

In general, recall that, if the probability of an event occurring is p, then the *odds*, o, in favour of such an event is given by

$$o = \frac{p}{1-p}. \tag{1.5}$$

This is customarily written as $1 : o$. Hence currently, the prior odds in favour of M_k are

$$\frac{p(M_k)}{1-p(M_k)},$$

where, moreover, the denominator is equal to $p(M_l)$. Then, by use of (1.1), it is evident that the ratio of posteriors, or posterior odds of M_k, is equivalent to

$$\frac{p(M_k \mid D)}{p(M_l \mid D)} = \frac{p(D \mid M_k)}{p(D \mid M_l)} \frac{p(M_k)}{p(M_l)}, \tag{1.6}$$

where, of course, $p(M_l \mid D) = 1 - p(M_k \mid D)$. Then, the *Bayes' factor* for model M_k against model M_l is defined as

$$B_{kl} = \frac{p(D \mid M_k)}{p(D \mid M_l)}. \tag{1.7}$$

As hinted at previously, calculation of B_{kl} is dependent on the two marginal likelihoods. Moreover, the Bayes' factor is seen to be the ratio of the posterior odds of M_k to its prior odds, and can be interpreted as the evidence provided by the data in favour of M_k compared to M_l. Thus, for instance, if $B_{kl} = 1$, we are indifferent between the two models. If $B_{kl} > 1$, then model M_k is preferred, otherwise model M_l. A more definitive interpretation is provided by Kass and Raftery. Here, the authors offer a guideline whereby values of $B_{kl} > 100$ indicate decisive evidence in favour of M_k. Conversely, if approximately $1 < B_{kl} < 3$, then the preference for M_k over M_l is small.

There is no doubt that Bayes' factors are straightforward and easy to interpret, by quantifying our preference for one model over another. However, they are certainly not infallible. Initially, we mention two general criticisms. Kass and Raftery adopt the stance that evaluation of the Bayes' factor is made to determine which of the two models is correct. As the authors themselves identity, many would dispute this claim from a couple of viewpoints. Firstly, as discussed above, a model is only an approximation to the truth. Moreover, by examining only two models, we may ultimately choose a poor model, only since it represents the data better than an even worse model. A second criticism is that computation of the Bayes' factor can be arduous due to its reliance on the marginal likelihood, a point already illuminated. This is exacerbated if the size of the candidate set is large.

The final problem that is raised is the most specific and perhaps that which has caused most debate in the literature. It is concerned with prior specification of the parameters. This is elucidated, for instance, by O'Hagan (1995) and we now follow this author's explanation. Of course, to calculate a Bayes' factor, $p(\boldsymbol{\theta}_i \,|\, M_i)$ must be specified (*c.f.* (1.2)). Suppose we want to represent prior ignorance (also referred to as vague prior knowledge) for our parameters. For instance, to express each value of $\boldsymbol{\theta}_i$ as equally likely *a priori*, we could employ a (continuous) uniform distribution (*c.f.* (1.4)). Yet, if the parameter space is infinite, this distribution is no longer defined since it possesses only finite support. Hence, we could use an *improper* uniform prior given as

$$p(\boldsymbol{\theta}_i \,|\, M_i) \propto 1. \tag{1.8}$$

Clearly, no particular value of $\boldsymbol{\theta}_i$ is favoured since the prior mass is spread equally across all values. However, the distribution is now improper since $\int p(\boldsymbol{\theta}_i \,|\, M_i)\mathrm{d}\boldsymbol{\theta}_i$ (representing the total probability mass) diverges and is not equal to one. Thus, any improper prior will contain no normalising constant. A further example of an improper, vague prior is the *Jeffreys' prior*. Such a choice is characterised by its invariance under reparameterisation, *i.e.* the vagueness of a prior on $\boldsymbol{\theta}_i$ is maintained upon transformation of this parameter vector. So, in this case, prior ignorance is represented by the distribution $p(\boldsymbol{\theta}_i \,|\, M_i) \propto |I(\boldsymbol{\theta}_i \,|\, M_i)|^{\frac{1}{2}}$, where $I(\boldsymbol{\theta}_i \,|\, M_i)$ is *Fisher's information matrix.* For more details, see, for instance, Gelman et al. (1995). We note that the use of improper priors to represent prior ignorance is common and often provide an easy Bayesian update from prior to posterior distribution. They are particularly useful when there is difficulty in attempting to quantify one's prior uncertainty in a distribution. Yet, their use is the only way that an improper posterior may be produced.

Now, generalise (1.8) such that

$$p(\boldsymbol{\theta}_i \,|\, M_i) \propto f_i(\boldsymbol{\theta}_i \,|\, M_i), \tag{1.9}$$

where f_i is any known function whose integral does not converge. For instance, this could be the Jeffreys' prior. Thus, it follows that $p(\boldsymbol{\theta}_i \,|\, M_i) = c_i f_i(\boldsymbol{\theta}_i \,|\, M_i)$ for some unspecified constant c_i. Of course, this normalising constant does not exist due to the divergence of the integral. Then, the parameter posterior is

$$p(\boldsymbol{\theta}_i \,|\, D,\, M_i) = \frac{p(D \,|\, \boldsymbol{\theta}_i,\, M_i)\, f_i(\boldsymbol{\theta}_i \,|\, M_i)}{\displaystyle\int p(D \,|\, \boldsymbol{\theta}_i,\, M_i)\, f_i(\boldsymbol{\theta}_i \,|\, M_i)\, \mathrm{d}\boldsymbol{\theta}_i}. \tag{1.10}$$

This distribution is well-defined, assuming the integral for the marginal likelihood is convergent, since the constants c_i have cancelled. Yet, if the models M_k and M_l are given improper priors similar to (1.9), then the corresponding Bayes' factor is, by definition, equivalent to

$$B_{kl} = \frac{c_k}{c_l} \frac{\displaystyle\int p(D \,|\, \boldsymbol{\theta}_k,\, M_k)\, f_k(\boldsymbol{\theta}_k \,|\, M_k)\, \mathrm{d}\boldsymbol{\theta}_k}{\displaystyle\int p(D \,|\, \boldsymbol{\theta}_l,\, M_l)\, f_l(\boldsymbol{\theta}_l \,|\, M_l)\, \mathrm{d}\boldsymbol{\theta}_l}. \tag{1.11}$$

Unfortunately, the constants now do not cancel and so the Bayes' factor contains a ratio of two unknown constants. This dilemma has caused much consternation amongst Bayesian statisticians. If improper priors are to be persisted with to represent ignorance, then solutions have been sought in the literature. Two of the most common are now presented.

Fractional Bayes' factors

To remove the dependence on $\frac{c_k}{c_l}$ from the usual Bayes' factor in the case of improper priors, O'Hagan (1991, 1995) suggested a new variant. Suppose again we wish to compare the models M_k and M_l. Initially, partition the data such that $D = (D_1,\, D_2)$. The portions, D_1 and D_2, are now employed for two separate purposes: D_1, known as a training sample, is used to learn the parameters $\boldsymbol{\theta}_k$ and $\boldsymbol{\theta}_l$, and D_2 to compare M_k and M_l in a Bayes' factor.

Thus, it is simple to form parameter posterior distributions through D_1, $p(\boldsymbol{\theta}_i \,|\, D_1,\, M_i)$ for $i = k,\, l$, using (1.3). Then, the Bayesian paradigm is implemented in a sequential way so that these parameter posteriors become prior distributions in the wake of the new data D_2, hence resulting in a Bayes' factor calculation. So, O'Hagan (1995) defines the *partial* Bayes' factor for model M_k against model M_l using data D_2, conditional on D_1, as

$$B_{kl}(D_2 \,|\, D_1) = \frac{p(D_2 \,|\, D_1,\, M_k)}{p(D_2 \,|\, D_1,\, M_l)} \tag{1.12}$$

$$= \frac{\displaystyle\int p(D_2 \,|\, \boldsymbol{\theta}_k,\, D_1,\, M_k)\, p(\boldsymbol{\theta}_k \,|\, D_1,\, M_k)\, \mathrm{d}\boldsymbol{\theta}_k}{\displaystyle\int p(D_2 \,|\, \boldsymbol{\theta}_l,\, D_1,\, M_l)\, p(\boldsymbol{\theta}_l \,|\, D_1,\, M_l)\, \mathrm{d}\boldsymbol{\theta}_l}. \tag{1.13}$$

Here, the probability density for D_2, namely $p(D_2 \,|\, \boldsymbol{\theta}_i,\, D_1,\, M_i)$ for $i = k,\, l$, is dependent on the parameters and the dataset D_1, itself previously utilised to learn the parameters. Even if the initial priors, $p(\boldsymbol{\theta}_k \,|\, M_k)$ and $p(\boldsymbol{\theta}_l \,|\, M_l)$, are chosen as improper, the sequential updating of posterior to prior implies that the new 'priors', $p(\boldsymbol{\theta}_k \,|\, D_1,\, M_k)$ and $p(\boldsymbol{\theta}_l \,|\, D_1,\, M_l)$, are proper and any unspecified constants have cancelled using (1.10). The partial Bayes' factor is so-called as comparison of the models requires only a portion of the data, hence differing from the full Bayes' factor, and is well-defined.

This partial Bayes' factor can now subsequently be used to construct a full Bayes' factor, incorporating all data D. In this case, the marginal likelihood of D_2 under M_i, conditioned on D_1, is simply

$$\begin{aligned} p(D_2 \,|\, D_1,\, M_i) &= \frac{p(D_1,\, D_2 \,|\, M_i)}{p(D_1 \,|\, M_i)} \\ &= \frac{\displaystyle\int p(D \,|\, \boldsymbol{\theta}_i,\, M_i)\, p(\boldsymbol{\theta}_i \,|\, M_i)\, \mathrm{d}\boldsymbol{\theta}_i}{\displaystyle\int p(D_1 \,|\, \boldsymbol{\theta}_i,\, M_i)\, p(\boldsymbol{\theta}_i \,|\, M_i)\, \mathrm{d}\boldsymbol{\theta}_i}. \end{aligned} \tag{1.14}$$

Then, it is evident that

$$\frac{p(D \,|\, M_k)}{p(D \,|\, M_l)} = \frac{p(D_1 \,|\, M_k)}{p(D_1 \,|\, M_l)} \frac{p(D_2 \,|\, D_1,\, M_k)}{p(D_2 \,|\, D_1,\, M_l)}$$

and so, by definition of Bayes' factors,

$$B_{kl}(D) = B_{kl}(D_1) B_{kl}(D_2 \,|\, D_1). \tag{1.15}$$

By assigning the prior distribution (1.9) for the parameters specific to each model, it follows from (1.11) that the term $\frac{c_k}{c_l}$ is common in the definition of both the Bayes' factors, $B_{kl}(D)$ and $B_{kl}(D_1)$. Hence, this ratio of unspecified constants cancels from both sides of (1.15). Thus, $B_{kl}(D)$ is now well-defined, as intended. Theoretically, the partial Bayes' factor would appear to possess a solid foundation. Yet, in practice, although no-longer dependent on any unspecified constants, it remains reliant on choosing a training sample of size m from a total of N observations, so that the parameters may be learnt (there are $\binom{N}{m}$ ways to do this). To avert the selection of such a dataset D_1, O'Hagan makes an asymptotic approximation to the partial Bayes' factor.

If we define $b = \dfrac{m}{N}$ and then let both m and N become large, then an approximation is obtained such that

$$p(D_1 \,|\, \boldsymbol{\theta}_i,\, M_i) \approx \left[p(D \,|\, \boldsymbol{\theta}_i,\, M_i)\right]^b,$$

where D_1 and D are datasets with m and N observations respectively. Thus, by consideration of (1.14), an alternative marginal likelihood for D under model M_i is given as

$$p_b(D \,|\, M_i) = \frac{\displaystyle\int p(D \,|\, \boldsymbol{\theta}_i,\, M_i)\, p(\boldsymbol{\theta}_i \,|\, M_i)\, \mathrm{d}\boldsymbol{\theta}_i}{\displaystyle\int \left[p(D \,|\, \boldsymbol{\theta}_i,\, M_i)\right]^b p(\boldsymbol{\theta}_i \,|\, M_i)\, \mathrm{d}\boldsymbol{\theta}_i}. \tag{1.16}$$

Hence finally, motivated by (1.12), the *fractional Bayes' factor*, denoted as $B^b_{kl}(D)$, is equivalent to

$$B^b_{kl}(D) = \frac{p_b(D \,|\, M_k)}{p_b(D \,|\, M_l)}. \tag{1.17}$$

It is apparent that, if we choose a prior over the parameters that is improper, any unspecified constants will now cancel in (1.16), and hence the fractional Bayes' factor will be well-defined. Yet, one outstanding issue still remains. Although there is no need to specifically choose a training dataset D_1, we must however specify the *proportion*, b, of D_1. This is the main problem with fractional Bayes' factors and is discussed further in

O'Hagan (1995). On face value, it appears that the method has replaced one problem (the unspecified ratio $\frac{c_k}{c_l}$) with another (how to select a value for b).

Posterior Bayes' Factors

An alternative framework in the context of using Bayes' factors with improper priors is developed by Aitkin (1991). Again, the author is able to construct a methodology, which removes the dependence of any unspecified constants in the comparison of the models M_k and M_l. Firstly, reconsider (1.2), representing the marginal likelihood of the data D, given model M_i. As noted previously, an alternative perspective shows that this equation can also be viewed as the prior mean of the density function.

Consequently, Aitkin suggests that when comparing models, to avert the dilemma caused by arbitrary constants, we can average the density, $p(D \,|\, \boldsymbol{\theta}_i,\, M_i)$, with respect to the parameter posterior distribution, $p(\boldsymbol{\theta}_i \,|\, D,\, M_i)$, instead of the corresponding prior. This seems reasonable since, via (1.10), this posterior is well-defined. Thus, the posterior mean of the likelihood is defined as

$$p_{\text{post}}(D \,|\, M_i) = \int p(D \,|\, \boldsymbol{\theta}_i,\, M_i)\, p(\boldsymbol{\theta}_i \,|\, D,\, M_i)\, \mathrm{d}\boldsymbol{\theta}_i. \tag{1.18}$$

So, a *posterior Bayes' factor* for model M_k against model M_l is then defined as

$$B_{kl}^{\text{post}} = \frac{p_{\text{post}}(D \,|\, M_k)}{p_{\text{post}}(D \,|\, M_l)}. \tag{1.19}$$

Notably, the posterior Bayes' factor is extremely similar in form to the partial Bayes' factor in (1.13), the difference being the latter is dependent on the partition of the data for the purposes of both parameter learning and model comparison. The derivation of both has required a sequential use of Bayes' theorem whereby the parameter posterior has subsequently been applied as a well-defined prior distribution for model comparison. In fact, by substituting in for the parameter posterior, (1.18) can be rewritten as

$$p_{\text{post}}(D \,|\, M_i) = \frac{\displaystyle\int [p(D \,|\, \boldsymbol{\theta}_i,\, M_i)]^2 \; p(\boldsymbol{\theta}_i \,|\, M_i)\, \mathrm{d}\boldsymbol{\theta}_i}{\displaystyle\int p(D \,|\, \boldsymbol{\theta}_i,\, M_i)\, p(\boldsymbol{\theta}_i \,|\, M_i)\, \mathrm{d}\boldsymbol{\theta}_i}. \tag{1.20}$$

When studying (1.16) and (1.20), now notice the similarity between the fractional and posterior Bayes' factors. Thus, akin to before, any outstanding, unspecified constants will cancel from (1.20) and so leave a well-defined Bayes' factor. A consistent criticism of posterior Bayes' factors is the use of the data 'twice' for learning parameters and model comparison, which, as illustrated above, is the significant difference between partial and posterior Bayes' factor methodology. Such a practice lacks any logical foundation. Moreover, it has been shown by Lindley (1991) that the method can be viewed as incoherent via a counter-example.

To summarise this section, the use of Bayes' factors for the purpose of model comparison can become problematic when improper priors are used to illustrate prior ignorance. In response, O'Hagan (1995) and Aitkin (1991) have independently constructed solutions to remove the ratio of unspecified constants, as seen in (1.11). A further technique is developed in Berger and Pericchi (1996), producing a so-called intrinsic Bayes' factor. Using a similar, initial foundation to O'Hagan, the authors reason that, to avoid specifying a training sample D_1, partial Bayes' factors should be computed for *all* training samples and the result then averaged.

A sensible question to ask at this stage is whether it is even necessary to use improper priors to represent prior ignorance. A clear, simple alternative is to specify a *proper* prior distribution (so integrates to 1), which is not concentrated around any one particular value. In other words, we require a prior with a reasonable variance. If both $p(\boldsymbol{\theta}_k \,|\, M_k)$ and $p(\boldsymbol{\theta}_l \,|\, M_l)$ are specified as proper, then calculation of the Bayes' factor is theoretically possible and no dependence on arbitrary constants, seen in (1.11), exists.

Unfortunately, the use of proper, diffuse priors in these circumstances is dangerous since the Bayes' factor may be highly dependent on the arbitrary choice of such a prior variance, and hence inappropriate conclusions may be reached. This is referred to as *Lindley's paradox* and is discussed in more detail in Chapter 3. The fractional and posterior Bayes' factors do not suffer from this paradox in quite the same way as, even if proper priors were specified in each case, both methods have their foundations in using the parameter posterior as a prior for marginal likelihood computation. Thus, specification of a reasonable prior variance will not influence the conclusion of the Bayes' factor in these cases. Yet, each procedure will be influenced by the choice of b and the repetitive use of the data respectively.

The work of O'Hagan and Aitkin is motivated due to the difficulties created with improper priors. Yet, we must question whether much can be gained by the use of the comparison techniques that the authors advocate. In solving one problem, it appears that further issues have been created. Therefore, has much been learnt as regards how to practice Bayesian model comparison? It is evident however that the specification of either a proper or improper prior is a thorny issue when assessing the value of a set of competing models, and a solution is hence required. As mentioned by Aitkin (1991), one possibility would be to carefully apply an informative, proper prior and analyse the sensitivity of results to such a choice. This technique is performed later in this thesis.

1.2 Approximation of the marginal likelihood

The importance of the marginal likelihood in Bayesian model comparison is clear. However, as commented previously, the integral (1.2) is often intractable and so an approximation is necessary. In this section, two of the more popular, analytic techniques for this are considered. Of course, we must stress that such an approximation is also vital in the computation of the normalised parameter posterior distribution, as seen by (1.3).

1.2.1 Laplace's approximation

For the derivation of this method, we follow that given by Beal (2003). Initially, consider the integrand in the definition of the marginal likelihood. By taking logarithms of this expression, we can define

$$h(\boldsymbol{\theta}_i) = \log \left[p(D \,|\, \boldsymbol{\theta}_i,\, M_i)\, p(\boldsymbol{\theta}_i \,|\, M_i) \right]. \tag{1.21}$$

This expression is now expanded using a second-order multivariate Taylor series about its *maximum a posteriori* (MAP) estimate, denoted by $\tilde{\boldsymbol{\theta}}_i$. Clearly, this is the point where the posterior density is maximised, *i.e.* the mode of the posterior distribution. Hence, we achieve

$$\begin{aligned} h(\boldsymbol{\theta}_i) &= h(\tilde{\boldsymbol{\theta}}_i) + (\boldsymbol{\theta}_i - \tilde{\boldsymbol{\theta}}_i)^T\, h'(\tilde{\boldsymbol{\theta}}_i) + \frac{1}{2!}\left(\boldsymbol{\theta}_i - \tilde{\boldsymbol{\theta}}_i\right)^T h''(\tilde{\boldsymbol{\theta}}_i) \left(\boldsymbol{\theta}_i - \tilde{\boldsymbol{\theta}}_i\right) + \ldots \\ &\approx h(\tilde{\boldsymbol{\theta}}_i) + \frac{1}{2}\left(\boldsymbol{\theta}_i - \tilde{\boldsymbol{\theta}}_i\right)^T Hh(\tilde{\boldsymbol{\theta}}_i) \left(\boldsymbol{\theta}_i - \tilde{\boldsymbol{\theta}}_i\right), \end{aligned} \tag{1.22}$$

where $'$ represents differentiation with respect to $\boldsymbol{\theta}_i$. Moreover, $Hh(\tilde{\boldsymbol{\theta}}_i)$ is the Hessian matrix of second partial derivatives for the function h, evaluated at $\tilde{\boldsymbol{\theta}}_i$. Now, notice that $\log p(\boldsymbol{\theta}_i \,|\, D,\, M_i) \propto h(\boldsymbol{\theta}_i)$ and, consequently by (1.3), $[\log p(\boldsymbol{\theta}_i \,|\, D,\, M_i)]' = h'(\boldsymbol{\theta}_i)$. Thus, $h'(\tilde{\boldsymbol{\theta}}_i) = 0$ as $\tilde{\boldsymbol{\theta}}_i$ is a maximum of $h(\boldsymbol{\theta}_i)$, that is, the MAP estimate. Via (1.21) and (1.22), it follows that the log marginal likelihood is given by

$$\begin{aligned} \log p(D \,|\, M_i) &= \log \int \exp\left\{h(\boldsymbol{\theta}_i)\right\}\, \mathrm{d}\boldsymbol{\theta}_i \\ &= \log \left[\exp\left\{h(\tilde{\boldsymbol{\theta}}_i)\right\} \int \exp\left\{\frac{1}{2}\left(\boldsymbol{\theta}_i - \tilde{\boldsymbol{\theta}}_i\right)^T Hh(\tilde{\boldsymbol{\theta}}_i) \left(\boldsymbol{\theta}_i - \tilde{\boldsymbol{\theta}}_i\right)\right\} \mathrm{d}\boldsymbol{\theta}_i \right] \end{aligned}$$

$$\approx h(\tilde{\boldsymbol{\theta}}_i) + \log\left[(2\pi)^{d_i/2}\left(|W^{-1}|\right)^{1/2}\right], \tag{1.23}$$

where d_i is the dimension of $\boldsymbol{\theta}_i$ and $W = -Hh(\tilde{\boldsymbol{\theta}}_i)$. In other words, we have approximated $\exp\{h(\boldsymbol{\theta}_i)\} = p(D \mid \boldsymbol{\theta}_i,\, M_i)\, p(\boldsymbol{\theta}_i \mid M_i)$ via a multivariate normal distribution (see Appendix A) with mean vector $\tilde{\boldsymbol{\theta}}_i$, the MAP estimate, and covariance matrix W^{-1}, and then subsequently integrated. Finally, by substituting (1.21) into (1.23) and taking exponentials, the Laplace approximation is given by

$$p(D \mid M_i)_{\text{Lap}} = p(D \mid \tilde{\boldsymbol{\theta}}_i,\, M_i)\, p(\tilde{\boldsymbol{\theta}}_i \mid M_i)\, (2\pi)^{d_i/2}|W|^{-1/2}. \tag{1.24}$$

This approximation is based on the fact that, for a large dataset, the parameter posterior distribution can be approximately normally distributed (Gelman et al., 1995). Hence, using Laplace seems reasonable if the posterior is unimodal and almost symmetric. Further, it is an enticing option due to the ease of computing the MAP estimate. Yet, on the contrary, we may expect an inaccurate approximation to the marginal likelihood, and hence posterior, if the sample size is small. Moreover, notice that the Hessian matrix is of dimension $d_i \times d_i$. So, such a method may suffer from a computational perspective if $\boldsymbol{\theta}_i$ is high-dimensional. Finally, as Beal (2003) also mentions, this method may not capture the position of the posterior probability mass well since the MAP estimate maximises the posterior density. So, we will obtain a more effective approximation if $p(\boldsymbol{\theta}_i \mid D,\, M_i)$ is tightly peaked about its mode, where all the mass is situated.

1.2.2 Bayesian Information Criterion (BIC)

A further procedure applied to approximate the marginal likelihood is the *Bayesian Information Criterion* (Schwarz, 1978), also termed *Schwarz's Information Criterion (SIC).* This criterion is viewed purely as a means to compare candidate models, and not to construct an approximate, parameter posterior distribution. As we shall see, it contains terms to evaluate both the fit and complexity of any particular model, as discussed in the introduction to this chapter.

The criterion can be derived directly from the Laplace approximation as Ghahramani (2004) demonstrates. Note initially that the Hessian matrix of h, evaluated at the MAP estimate, is equivalent to

$$\begin{aligned} Hh(\tilde{\boldsymbol{\theta}}_i) &= \Big[\log p(D \,|\, \boldsymbol{\theta}_i,\, M_i) + \log p(\boldsymbol{\theta}_i \,|\, M_i) \Big]''_{\boldsymbol{\theta}_i = \tilde{\boldsymbol{\theta}}_i} \\ &= \Big[\sum_{t=1}^{N} \log p(\mathbf{x}_t \,|\, \boldsymbol{\theta}_i,\, M_i) + \log p(\boldsymbol{\theta}_i \,|\, M_i) \Big]''_{\boldsymbol{\theta}_i = \tilde{\boldsymbol{\theta}}_i} \end{aligned} \tag{1.25}$$

where we possess a dataset $D = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$. So, it is evident that the Hessian matrix is dependent on N. Consequently, by taking logarithms of (1.24), and then rejecting all terms that are independent of the sample size N, we obtain

$$\log p(D \,|\, M_i)_{\text{Lap}} = \log p(D \,|\, \tilde{\boldsymbol{\theta}}_i,\, M_i) - \frac{1}{2} \log |W|. \tag{1.26}$$

Here, $\log p(D \,|\, \tilde{\boldsymbol{\theta}}_i,\, M_i)$ will be a sum of N terms. Consequently, it is realised that the Hessian matrix is of order $\mathcal{O}(N)$ since, for each entry of $Hh(\tilde{\boldsymbol{\theta}}_i)$, N summations must again be made. Then, by definition of $\mathcal{O}$ notation, we can specify $W \approx NW_0$ for sufficiently large N, where W_0 is a fixed constant matrix. Thus, it follows immediately that, as W is of dimension $d_i \times d_i$,

$$\frac{1}{2} \log |W| \approx \frac{d_i}{2} \log N + \frac{1}{2} \log |W_0|, \tag{1.27}$$

where $|NW_0| = N^{d_i} |W_0|$ (Harville, 1997). Now, the term $\frac{1}{2} \log |W_0|$ is also fixed with respect to N. By dropping this and substituting (1.27) into (1.26), the BIC, as presented by Schwarz (1978), is defined to be

$$\log p(D \mid M_i)_{\mathrm{BIC}} = \log p(D \mid \tilde{\boldsymbol{\theta}}_i,\, M_i) - \frac{d_i}{2} \log N. \tag{1.28}$$

In (1.28), it is customary that the log-likelihood, $\log p(D \mid \boldsymbol{\theta}_i,\, M_i)$, is evaluated not at the MAP estimate $\tilde{\boldsymbol{\theta}}_i$, but instead at $\hat{\boldsymbol{\theta}}_i$, the *maximum likelihood estimate (MLE)*. This is the value of $\boldsymbol{\theta}_i$ for which the likelihood is maximised.

Due to its reliance on calculation of the MLE, this criterion is easy to handle. Moreover, notice that, although working in a Bayesian context, the BIC is defined such that no specification of the prior $p(\boldsymbol{\theta}_i \mid D,\, M_i)$ is required, assuming that the log-likelihood is evaluated at the MLE. Depending on one's perspective, this may be a positive attribute if it is awkward to elicit one's parameter prior knowledge. However, the converse may be true if an informative prior is required. In addition, as the derivation of the BIC given here is reliant on the Laplace approximation, the criterion may suffer if the sample size is insufficiently large.

We realise that the two terms in the BIC expression each serve a purpose. If we interpret the MLE as the value of the parameters for model M_i that makes the data most plausible, $\log p(D \mid \hat{\boldsymbol{\theta}}_i,\, M_i)$ illustrates how well M_i fits the data, a term that is ideally maximised. On the other hand, $\frac{d_i}{2} \log N$ acts to penalise more complex models, determined by the number of parameters, d_i, that each possesses. So, for a candidate set of models, the optimum model choice is that which has the highest value of (1.28).

1.3 Further criteria

The BIC is a classical technique to evaluate the evidence for a set of models. In fact, other such criteria exist and, in this section, we consider briefly two of the more significant options. We shall see that, as previously, each is dependent on assessing model fit and model complexity.

1.3.1 Akaike's Information Criterion (AIC)

Akaike (1974) realised that we need a way to measure the misfit between a model and a truth to judge whether the former is a decent approximation to the latter on the basis of a dataset. From an alternative perspective, we determine the information lost in making such an approximation whereby a good model will minimise this quantity. To quantify this, the *Kullback-Leibler* (KL) divergence (Kullback and Leibler, 1951) from the truth to the model is employed, defined as

$$\begin{aligned}\mathrm{KL}(f\,|\,p) &= \int f(D)\log\left(\frac{f(D)}{p(D\,|\,\boldsymbol{\theta}_i,\,M_i)}\right)\mathrm{d}D \\ &= \mathrm{E}_{f(D)}\{\log f(D)\} - \mathrm{E}_{f(D)}\{\log p(D\,|\,\boldsymbol{\theta}_i,\,M_i)\},\end{aligned} \tag{1.29}$$

where log is the natural logarithm. In addition, the expectations above are taken with respect to $f(D)$, the true density of D, specified without parameters. Clearly, in (1.29), the term $\mathrm{E}_{f(D)}\{\log f(D)\}$ is a constant across models. Hence, minimising $\mathrm{KL}(f\,|\,p)$ is equivalent to finding the model that maximises $J = \mathrm{E}_{f(D)}\{\log p(D\,|\,\boldsymbol{\theta}_i,\,M_i)\}$, referred to as the *relative KL divergence*.

Unfortunately, calculation of J is not possible *per se* as it is dependent upon knowledge of the truth f. Paradoxically, an understanding of this reality would render the derivation of such a criterion unnecessary. Thus, Akaike introduced a fabricated dataset X, independent of D, but arising from the same distribution. It was then shown that the *expected value* of J with respect to $f(X)$ could be estimated, where $\boldsymbol{\theta}_i$ is replaced by the corresponding MLE $\hat{\boldsymbol{\theta}}_i(X)$, dependent on model M_i and constructed using the dataset X (if it were available). In fact, a biased estimator of $\mathrm{E}_{f(X)}\{J\}$ is given by the maximised log-likelihood function, namely $\log p(D \,|\, \hat{\boldsymbol{\theta}}_i(D),\, M_i)$. Moreover, it was further established that the bias of this estimate is asymptotically (for a large dataset) equivalent to d_i, the dimension of the parameter vector. For additional details on this, see, for instance, Burnham and Anderson (2004) or Stoica and Selén (2004). Upon removing the dependence of the MLE upon D, we see that maximising the unbiased estimator, $\log p(D \,|\, \hat{\boldsymbol{\theta}}_i,\, M_i) - d_i$, for the expectation of J, is equivalent to minimising the following, known as *Akaike's information criterion*:

$$\mathrm{AIC} = -2 \log p(D \,|\, \hat{\boldsymbol{\theta}}_i,\, M_i) + 2d_i. \tag{1.30}$$

The 'best' model is deemed to be that which has the smallest AIC value and is interpreted as the model 'closest' to the actual truth. According to Burnham and Anderson (2004), the multiplication here by -2 is for 'historical reasons'. In fact, the BIC, given by (1.28), is also presented similarly, implying that the resulting expression should now be minimised. Clearly, the AIC and BIC have the same goodness of fit term. Yet, the model complexity term is more stringent in the BIC case (if $N \geq 8$, then $d_i \log N > 2d_i$), hence providing an obvious preference for simpler models. However, this could be detrimental when a simpler model is chosen over a more complex one, even if the former is a poor specification. On the other hand, AIC could be susceptible to overfitting the data by showing an affinity for too complex models. Finally, as it is based on asymptotic maximum likelihood theory, the performance of the AIC in datasets of small size may be questionable.

1.3.2 Deviance Information Criterion (DIC)

The final model comparison criterion that is examined was pioneered by Spiegelhalter et al. (2002). The initial foundation for this technique is provided by the classical *deviance*, which is equivalent to the difference in the log-likelihoods between a model and the unknown truth that generated the data. In fact, the deviance D^* is defined as

$$D^*(\boldsymbol{\theta}_i \,|\, M_i) = -2\log p(D \,|\, \boldsymbol{\theta}_i,\, M_i) + 2\log f(D) \tag{1.31}$$

where, again, $f(D)$ is the true density of the data. However, this term is independent of the model M_i. Correspondingly, it is constant, and hence irrelevant, for the purposes of model comparison. Spiegelhalter et al. examine a Bayesian treatment for the problem at hand and thus focus their attention on the posterior distribution of the deviance.

Thus, the posterior mean of $D^*(\boldsymbol{\theta}_i \,|\, M_i)$ could be utilised as a Bayesian measure of model fit, denoted as

$$\begin{aligned} \bar{D}^* &= \int D^*(\boldsymbol{\theta}_i \,|\, M_i)\, p(\boldsymbol{\theta}_i \,|\, D,\, M_i)\, \mathrm{d}\boldsymbol{\theta}_i \\ &= \mathrm{E}_{\boldsymbol{\theta}_i \,|\, D,\, M_i}\{D^*\}. \end{aligned} \tag{1.32}$$

Due to the definition of the deviance, those models that provide a good fit will possess a small value of $\bar{D}^*$. This will occur when the number of parameters is increased so we now require a measure of model complexity to counterbalance this. So, Spiegelhalter et al. denote such a quantity as p_D, taking the form

$$\begin{aligned} p_D &= \mathrm{E}_{\boldsymbol{\theta}_i \,|\, D,\, M_i}\{D^*\} - D^*\left(\mathrm{E}_{\boldsymbol{\theta}_i \,|\, D,\, M_i}\{\boldsymbol{\theta}_i\} \,|\, M_i\right) \\ &= \bar{D}^* - D^*(\bar{\boldsymbol{\theta}}_i \,|\, M_i). \end{aligned}$$

Thus, p_D is equivalent to the difference between the posterior mean of the deviance and the deviance evaluated at the posterior mean of the parameters. Recalling that $D^*(\boldsymbol{\theta}_i \,|\, M_i) = -2\log p(D \,|\, \boldsymbol{\theta}_i,\, M_i)$, our terms for both measure of fit and the penalty for model complexity can now be summed (akin to the AIC and BIC) to form the *Deviance*

Information Criterion:

$$\begin{aligned} \text{DIC} &= \bar{D}^{*} + p_D \\ &= D^{*}(\bar{\boldsymbol{\theta}}_i \mid M_i) + 2p_D, \end{aligned} \tag{1.33}$$

the latter by rearranging the expression for p_D. In the way of both the AIC and BIC, the high-ranking models are those that minimise the DIC and, hence, an optimal model can be chosen. By writing the AIC in terms of the deviance such that AIC $= D^{*}(\hat{\boldsymbol{\theta}}_i \mid M_i) + 2d_i$, Spiegelhalter et al. show that the DIC is a Bayesian generalisation of the AIC.

In the discussion to this paper, some salient points were raised. For instance, Robert and Titterington (2002) noticed that the authors' had used the data once, to construct a posterior distribution for $\boldsymbol{\theta}_i$, and then a second time, to take the posterior mean of the deviance. This is the same criticism as seen for Aitkin's posterior Bayes' factor whereby the dataset is applied to both learning the parameters and for model comparison. Moreover, Brooks (2002) questioned why it was possible that p_D could in fact be negative, leaving it open to interpretation in such a case.

1.4 Outline of thesis and literature review

In this chapter, a variety of procedures have been analysed so that the evidence for each model in a candidate set can be evaluated. Moreover, the potential hazards associated with each method have also been discussed. In Chapter 2, a relatively recent addition to the Bayesian model comparison literature is introduced, referred to as *variational Bayes*. This method is advantageous since we inherently derive separate approximations to both the posterior distribution and the marginal likelihood, suitable for future inference and ranking models respectively. Its theoretical foundation is reliant upon the Kullback-Leibler divergence, previously seen in this chapter to derive the AIC. To conclude this chapter, its performance in posterior approximation is compared to two other, standard techniques.

For the remainder of the thesis, variational Bayes is applied specifically to comparing sparse vector autoregressive (VAR) models of order 1. In Chapter 3, by modelling using sparsity, a candidate set of *zero mean* VAR(1) graphical models (specifically dynamic Bayesian networks) is established, each of which relates to the autoregressive matrix in the VAR process. We proceed to form a lower bound on the marginal likelihood to compare the evidence for such models. A valid question to inquire at this stage would be how to handle the problem if the candidate set of graphical models is large. This is the focus of Chapter 4 and it is answered by constructing a Metropolis-Hastings type algorithm to search quickly and efficiently for high-scoring models in the graphical space. The ideas of Chapter 3 are then mimicked in Chapter 5 by the study of *non-zero mean* VAR(1) models. Examples involving both simulated and real data are then utilised to elucidate the theory of these two chapters. A summary, illustrating the main points of the thesis, is presented in Chapter 6.

The most comprehensive review of the variational approximation is provided by Beal (2003). In this thesis, by considering any model with both parameters and hidden variables, the author develops a *variational Bayesian EM algorithm*, allowing alternate updating of approximate posteriors for these two sets of unknowns. The algorithm is applied to a variety of statistical models, in particular, hidden Markov models, mixtures of factor analysers and linear dynamical systems, using both simulated and real datasets in each case. The current work extends that of Beal by providing the variational treatment to both zero and non-zero mean VAR models (of course, defined without hidden variables). However, as opposed to determining an optimum model order for a VAR(p) process, we wish to evaluate the evidence for a set of *sparse graphical models* of order 1, given a dataset. This is aided by the use of MCMC methods in high-dimensional spaces.

We realise that it is essential to use an approximation technique such as variational Bayes in the context of VAR(1) model comparison since, even for the model that is saturated, it is not possible to derive the marginal likelihood analytically. This fact is shown explicitly in Chapter 3. Furthermore, this approach is able to enforce naturally the specific sparsity constraints placed upon the approximate posterior distribution of the autoregressive matrix for each candidate. It is also important to note that learning a dynamic Bayesian network from data is a problem that has received much coverage in the statistical literature. To rank candidate structures, Friedman et al. (1998) suggested application of the BIC or the so-called Bayesian Dirichlet equivalence (BDe) score, originally developed in the static case by Heckerman et al. (1995). Alternatively, Husmeier (2003) used a MCMC search algorithm to locate the most plausible models in the space, similar to that which is presented in Chapter 4. However, to enable analytic computation of the marginal likelihood, the author was required to discretise the data, leading to a considerable loss of information.

As opposed to the structural search algorithm considered in this thesis, an alternative method would be to put sparsity priors on the coefficients of the autoregressive matrix. This idea is used by, for instance, Lucas et al. (2006) in the circumstance of regression modelling for microarray data. The aim here is for many entries of this matrix to be estimated close to (or even equal to) zero following a variational Bayesian analysis. Thus, if the value of such elements lies below a specified threshold point, no edge is placed on the graph between the corresponding nodes. Although this approach is more efficient, ascertaining possible influences between nodes in this way can be inaccurate and so is not pursued here.

Chapter 2

Variational Bayes

2.1 Introduction

The focus in the previous chapter was to explore techniques in which the marginal likelihood could be approximated, for the primary purpose of model comparison. Our emphasis now turns to the approximation of the posterior distribution over parameters. In this chapter, the dependence of our distributions on each model M_i will be predominantly removed. So, for completeness, our beliefs about the parameter vector $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_d)^T$ upon observing data D are quantified by the distribution

$$p(\boldsymbol{\theta} \mid D) = \frac{p(D \mid \boldsymbol{\theta})\, p(\boldsymbol{\theta})}{\displaystyle\int p(D \mid \boldsymbol{\theta})\, p(\boldsymbol{\theta})\, \mathrm{d}\boldsymbol{\theta}}. \tag{2.1}$$

Of course, the integral in the denominator of this expression could be intractable. So, a straightforward, direct solution to this would be to apply the analytic Laplace approximation. In this chapter, some alternative approaches are examined. For instance, the use of Markov chain Monte Carlo methods enables samples to be drawn from $p(\boldsymbol{\theta} \mid D)$, with which, *inter alia*, understanding can be garnered about the marginal posterior of each component of $\boldsymbol{\theta}$. Furthermore, an expectation-maximisation type algorithm can allow approximation of the afore-mentioned marginal posteriors via MAP estimation. Finally, special detail is devoted to a relatively, recent technique, known as variational Bayes. Each method will be treated theoretically and then compared via example.

2.2 Markov chain Monte Carlo

The use of Markov chain Monte Carlo (MCMC) methodology to understand a posterior distribution has become highly popular in the Bayesian community. Instead of using an analytic technique such as Laplace to approximate (2.1), a Markov chain is simulated whose samples will be draws from the posterior, upon convergence of the chain. When a chain converges to its stationary distribution, it will possess this distribution for all time henceforth. Thus, to simulate from the posterior, we construct a Markov chain whose stationary distribution is the posterior distribution. The summary statistics and the distribution of these posterior samples will approximate the corresponding characteristics of the true posterior.

Here, we present two of the most fundamental MCMC methods: the Gibbs sampler and the Metropolis-Hastings algorithm. In each case, we consider the general case and hence suppose that $\pi(\boldsymbol{\theta})$ is the density of interest, where we allow the possibility that each $\boldsymbol{\theta}_j$ (the j-th component of $\boldsymbol{\theta}$ for $j = 1, \dots d$) could be multi-dimensional. When simulating from a posterior, we let $\pi(\boldsymbol{\theta}) = p(\boldsymbol{\theta} \mid D)$.

2.2.1 The Gibbs sampler

First documented by Geman and Geman (1984), this method relies on sampling from the full conditional distributions for each component of $\boldsymbol{\theta}$, such that a sample from $\pi(\boldsymbol{\theta})$ may be obtained. The full conditionals are denoted as

$$\pi(\boldsymbol{\theta}_j \,|\, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_{j-1}, \, \boldsymbol{\theta}_{j+1}, \ldots, \boldsymbol{\theta}_d) = \pi(\boldsymbol{\theta}_j \,|\, \boldsymbol{\theta}_{-j}), \tag{2.2}$$

for $j = 1, \ldots, d$, and are assumed to be in closed form so that we may sample from them. The algorithm for the Gibbs sampler is as given below.

Algorithm 1 *1. Initialise the iteration counter to $k = 1$. The chain itself is initialised at a starting value $\boldsymbol{\theta}^{(0)} = (\boldsymbol{\theta}_1^{(0)}, \ldots, \boldsymbol{\theta}_d^{(0)})^T$.*

2. By successive simulation from the full conditionals, a new value $\boldsymbol{\theta}^{(k)}$ is obtained from the previous $\boldsymbol{\theta}^{(k-1)}$:

$$\begin{aligned}
\boldsymbol{\theta}_1^{(k)} &\sim \pi(\boldsymbol{\theta}_1 \,|\, \boldsymbol{\theta}_2^{(k-1)}, \ldots, \boldsymbol{\theta}_d^{(k-1)}) \\
\boldsymbol{\theta}_2^{(k)} &\sim \pi(\boldsymbol{\theta}_2 \,|\, \boldsymbol{\theta}_1^{(k)}, \, \boldsymbol{\theta}_3^{(k-1)}, \ldots, \boldsymbol{\theta}_d^{(k-1)}) \\
&\vdots \\
\boldsymbol{\theta}_d^{(k)} &\sim \pi(\boldsymbol{\theta}_d \,|\, \boldsymbol{\theta}_1^{(k)}, \ldots, \boldsymbol{\theta}_{d-1}^{(k)}).
\end{aligned}$$

3. Change the counter from k to $k + 1$ and return to step 2.

Upon convergence of the Markov chain, the simulated iterates will be draws from $\pi(\boldsymbol{\theta})$. Moreover, at this stage, the values of a particular component will be draws from the corresponding marginal posterior distribution for that component.

2.2.2 Metropolis-Hastings algorithm

A complementary methodology is the Metropolis-Hastings algorithm (Hastings, 1970), which allows simulation from the density of interest when this is known only up to a constant of proportionality. We now introduce an arbitrary proposal distribution, $q(\boldsymbol{\theta}, \, \boldsymbol{\phi})$, the notation of which here specifies the probability of a move from $\boldsymbol{\theta}$ to $\boldsymbol{\phi}$. The reverse move is implied by $q(\boldsymbol{\phi}, \boldsymbol{\theta})$. This distribution should be easy to simulate from. The algorithm is as follows:

Algorithm 2 *1. Initialise both the iteration counter to $k = 1$ and the chain itself to starting value $\boldsymbol{\theta}^{(0)}$.*

2. Generate a proposed value $\boldsymbol{\phi}$ from the distribution $q(\boldsymbol{\theta}^{(k-1)}, \boldsymbol{\phi})$.

3. Compute the acceptance probability $\alpha(\boldsymbol{\theta}^{(k-1)}, \boldsymbol{\phi})$ of the proposed move, where

$$\alpha(\boldsymbol{\theta}, \boldsymbol{\phi}) = \min\left\{1, \frac{\pi(\boldsymbol{\phi})q(\boldsymbol{\phi}, \boldsymbol{\theta})}{\pi(\boldsymbol{\theta})q(\boldsymbol{\theta}, \boldsymbol{\phi})}\right\}. \tag{2.3}$$

4. Put $\boldsymbol{\theta}^{(k)} = \boldsymbol{\phi}$ with probability $\alpha(\boldsymbol{\theta}^{(k-1)}, \boldsymbol{\phi})$, otherwise put $\boldsymbol{\theta}^{(k)} = \boldsymbol{\theta}^{(k-1)}$.

5. Change the counter from k to $k + 1$ and return to step 2.

In essence, at each iteration, a new value is generated from the proposal distribution, which may be accepted (indicating that the chain moves) or rejected (hence the chain stays put). The movement of the chain is dependent on the acceptance probability α. By drawing $u \sim \mathcal{U}(0, 1)$, the proposal $\boldsymbol{\phi}$ is accepted if $u < \alpha(\boldsymbol{\theta}^{(k-1)}, \boldsymbol{\phi})$ at iteration k. Once the chain reaches convergence, all simulated values will be draws from $\pi(\boldsymbol{\theta})$, irrespective of the choice of proposal distribution. The method is of particular use in the Bayesian paradigm, as since $\pi(\cdot)$ is only involved in $\alpha(\boldsymbol{\theta}, \boldsymbol{\phi})$ via a ratio, the proportionality constant $p(D)$, required to compute the posterior distribution and itself computationally problematic, will cancel out. Although not utilised in this chapter, the Metropolis-Hastings algorithm will be a crucial tool at our disposal later in this thesis.

When running an MCMC scheme, the period that elapses prior to convergence of the chain, *i.e.* before the stationary distribution has been reached, is referred to as the *burn-in* period. Therefore, if we want to generate samples from $\pi(\boldsymbol{\theta})$, we discard those values simulated during burn-in. In fact, a good method to establish the length of the burn-in period required is to plot all values using a trace plot, a time series plot displaying the values of a component of $\boldsymbol{\theta}$ against the number of iterations. Moreover, without burn-in, such a plot can be used as a crude test for convergence by revealing how well a chain is said to *mix*. A well mixing chain will move freely about a constant mean level with constant variance, exploring the parameter space. Conversely, a poorly mixing chain will not traverse quickly through the space, indicated on the plot by long, 'flat' regions, as a consequence of numerous, proposed moves being rejected.

The use of MCMC methods in Bayesian statistical inference enables posterior approximation in large multivariate problems or where the posterior itself is of non-standard form. Highly accurate results can be obtained if we draw a large number of samples. However, a huge amount of computational time may be required to achieve this. As Lappalainen and Miskin (2000) indicate, this is in contrast, for instance, to the Laplace approximation, which will produce less accurate results, but in shorter time. Moreover, uncertainty will remain as to whether the chain has reached its stationary distribution. An additional discussion is provided to this latter issue in Chapter 4.

It is briefly worth mentioning that, although this chapter is primarily concerned with approximating the parameter posterior, sampling from this distribution by MCMC methods enables a further approximation to the marginal likelihood. A simple estimate was given by Newton and Raftery (1994) as

$$p(D \mid M_i) \approx \left[\frac{1}{B} \sum_{k=1}^{B} p(D \mid \boldsymbol{\theta}_i^{(k)},\, M_i)^{-1} \right]^{-1},$$

where $\boldsymbol{\theta}_i^{(k)} = (\boldsymbol{\theta}_{i1}^{(k)}, \ldots, \boldsymbol{\theta}_{id}^{(k)})^T$, the parameters specific to M_i, and we obtain B draws from the posterior. For further details, the reader is referred to the afore-mentioned paper.

2.3 Expectation-Maximisation (EM) algorithm

We now turn our attention to a method that will approximate analytically each marginal posterior distribution. As above, let $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_d)^T$ be a parameter vector of length d. Consider the density $p(\boldsymbol{\theta}_1 \,|\, D)$ for a dataset D. Then, Gelman et al. (1995) have illustrated that the EM algorithm (developed by Dempster et al. (1977)) can be applied as an iterative procedure to find a mode (MAP estimate) of this marginal posterior density. This is of particular use in circumstances where knowledge of this marginal is limited, and hence cannot be maximised directly to find such an estimate. We denote the resulting estimate by $\tilde{\boldsymbol{\theta}}_1$. Further suppose that we find corresponding marginal posterior modes, $\tilde{\boldsymbol{\theta}}_2, \ldots, \tilde{\boldsymbol{\theta}}_{j-1}, \tilde{\boldsymbol{\theta}}_{j+1}, \ldots, \tilde{\boldsymbol{\theta}}_d$. Then, by deriving the full conditional posteriors, $p(\boldsymbol{\theta}_j \,|\, \boldsymbol{\theta}_{-j}, D)$ for each $j = 1, \ldots d$, the marginal posterior for $\boldsymbol{\theta}_j$ can be approximated via $p(\boldsymbol{\theta}_j \,|\, \boldsymbol{\theta}_1 = \tilde{\boldsymbol{\theta}}_1, \ldots, \boldsymbol{\theta}_{j-1} = \tilde{\boldsymbol{\theta}}_{j-1}, \boldsymbol{\theta}_{j+1} = \tilde{\boldsymbol{\theta}}_{j+1}, \ldots, \boldsymbol{\theta}_d = \tilde{\boldsymbol{\theta}}_d, D)$. Hence, the algorithm has increased value, as opposed to just providing a parameter point estimate of the marginal posterior where the density is highest.

The algorithm itself is a two-stage iterative process, consisting of an E-step (expectation) and M-step (maximisation). To find a marginal posterior mode for $p(\boldsymbol{\theta}_1 \,|\, D)$, we can follow the procedure below, as presented by Gelman et al..

Algorithm 3 *1. Initialise the iteration counter to $k = 1$. Make an initial MAP estimate of $p(\boldsymbol{\theta}_1 \,|\, D)$, say $\boldsymbol{\theta}_1^{(0)}$.*

2. At iteration k, perform the following two stages:

(a) E-step: Determine the log joint posterior density, $\log p(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_d \,|\, D)$. Then, take its expectation with respect to the conditional posterior distribution of $\boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_d$ given the previous estimate $\boldsymbol{\theta}_1^{(k-1)}$, with density denoted by $p(\boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_d \,|\, \boldsymbol{\theta}_1^{(k-1)}, D)$. In other words, derive the expectation

$$\begin{aligned}&\mathrm{E}_{k-1}\left\{\log p(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_d \,|\, D)\right\}\\ &= \int \cdots \int \log p(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_d \,|\, D)\, p(\boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_d \,|\, \boldsymbol{\theta}_1^{(k-1)},\, D)\, \mathrm{d}\boldsymbol{\theta}_2 \ldots \mathrm{d}\boldsymbol{\theta}_d.\end{aligned}$$

(b) M-step: Determine $\boldsymbol{\theta}_1^{(k)}$*, the new value of* $\boldsymbol{\theta}_1$ *that maximises* $\mathrm{E}_{k-1}\left\{\log p(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_d \,|\, D)\right\}$.

3. Change the counter from k *to* $k+1$ *and return to step 2.*

The algorithm alternates between the E-step and M-step until the estimate has converged after, say, K iterations. At this point, we define $\tilde{\boldsymbol{\theta}}_1 := \boldsymbol{\theta}_1^{(K)}$. As Gelman et al. note, the algorithm works since, at each iteration, $\mathrm{E}_{k-1}\left\{\log p(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_d \,|\, D)\right\}$ is maximised, producing an estimate $\boldsymbol{\theta}_1^{(k)}$ that monotonically increases the log marginal posterior density, *i.e.* $\log p(\boldsymbol{\theta}_1^{(k)} \,|\, D) > \log p(\boldsymbol{\theta}_1^{(k-1)} \,|\, D)$. By running until convergence such that $\boldsymbol{\theta}_1^{(K)} = \boldsymbol{\theta}_1^{(K-1)}$, a mode of the marginal posterior is hence found.

It was indicated earlier that, to approximate a marginal posterior for $\boldsymbol{\theta}_j$, the algorithm must be repeated to find the additional modes $\tilde{\boldsymbol{\theta}}_2, \ldots, \tilde{\boldsymbol{\theta}}_{j-1}$, $\tilde{\boldsymbol{\theta}}_{j+1}, \ldots, \tilde{\boldsymbol{\theta}}_d$, although this task could become somewhat laborious. A final point to make is that, if $p(\boldsymbol{\theta}_1 \,|\, D)$ is multimodal, we may not automatically arrive at the global maximum of $p(\boldsymbol{\theta}_1 \,|\, D)$. To negotiate this problem, the algorithm should be initialised at a variety of points in the parameter space, and then let $\tilde{\boldsymbol{\theta}}_1$ be the mode such that $\log p(\tilde{\boldsymbol{\theta}}_1 \,|\, D)$ is maximal.

2.4 Variational Bayesian methods

We now present a final, alternative technique for approximating a posterior distribution, known as *variational Bayes* (also referred to by MacKay (1995a) as *ensemble learning*). In recent years, the literature has become quite rich in this area - see, for instance, Lappalainen and Miskin (2000), Winn (2003), Beal (2003) or Penny et al. (2006). The outline of the method is as follows. For a set of observed data D and a parameter vector $\boldsymbol{\theta}$, this approach forms a parametric approximation of the true posterior, $p(\boldsymbol{\theta} \,|\, D)$. The approximating distribution is known as a *variational distribution* (or an *ensemble*), and is denoted subsequently by $q(\boldsymbol{\theta} \,|\, D)$. We must ensure that this distribution is as 'close' as possible to the true posterior. To effect this, a dissimilarity measure can be employed to gauge the misfit between the two distributions. As mentioned in Section 1.3.1 to derive the AIC, one standard choice of measure is the Kullback-Leibler divergence.

2.4.1 Kullback-Leibler divergence

Given the true and approximating posterior densities $p(\boldsymbol{\theta} \,|\, D)$ and $q(\boldsymbol{\theta} \,|\, D)$, recall from Chapter 1 that the KL divergence from q to p is defined as

$$\mathrm{KL}(q \,|\, p) = \int q(\boldsymbol{\theta} \,|\, D) \log \frac{q(\boldsymbol{\theta} \,|\, D)}{p(\boldsymbol{\theta} \,|\, D)} \, d\boldsymbol{\theta}. \tag{2.4}$$

It merely measures the extent to which the two densities agree. Of course, as previously discussed, to provide a good approximation, we choose the distribution q such that the KL divergence between q and p is minimised. An important property of the KL divergence is that it is always non-negative, a result known as the Gibbs' inequality (Penny et al., 2006), *i.e.* $\mathrm{KL}(q \,|\, p) \geq 0$ with equality if and only if $q = p$. We also note that, although the KL divergence is gauging the distance between q and p, it is not *per se* a true 'distance' metric since it is not symmetric, *i.e.* $\mathrm{KL}(q \,|\, p) \neq \mathrm{KL}(p \,|\, q)$. Therefore, it is relevant whether we minimise the misfit between q and p or *vice versa*. A further mention of this is made below.

2.4.2 Definition of $\mathcal{L}(q)$

As it currently stands, evaluation of (2.4) is not possible since it requires knowledge of $p(\boldsymbol{\theta} \,|\, D)$, which we have assumed to be intractable. However, the true posterior can be simply rewritten as $p(\boldsymbol{\theta} \,|\, D) = \dfrac{p(\boldsymbol{\theta},\, D)}{p(D)}$. By substituting in, we hence obtain

$$\begin{aligned}
\mathrm{KL}(q \,|\, p) &= \int q(\boldsymbol{\theta} \,|\, D) \log \frac{q(\boldsymbol{\theta} \,|\, D) p(D)}{p(\boldsymbol{\theta},\, D)} \,\mathrm{d}\boldsymbol{\theta} \\
&= \int q(\boldsymbol{\theta} \,|\, D) \log \frac{q(\boldsymbol{\theta} \,|\, D)}{p(\boldsymbol{\theta},\, D)} \,\mathrm{d}\boldsymbol{\theta} + \int q(\boldsymbol{\theta} \,|\, D) \log p(D) \,\mathrm{d}\boldsymbol{\theta} \\
&= \int q(\boldsymbol{\theta} \,|\, D) \log \frac{q(\boldsymbol{\theta} \,|\, D)}{p(\boldsymbol{\theta},\, D)} \,\mathrm{d}\boldsymbol{\theta} + \log p(D). \qquad (2.5)
\end{aligned}$$

Yet, alternatively, if the same substitution is used to simplify the reverse KL-divergence $\mathrm{KL}(p \,|\, q)$, the following is reached:

$$\begin{aligned}
\mathrm{KL}(p \,|\, q) &= \int p(\boldsymbol{\theta} \,|\, D) \log \frac{p(\boldsymbol{\theta},\, D)}{q(\boldsymbol{\theta} \,|\, D) p(D)} \,\mathrm{d}\boldsymbol{\theta} \\
&= \int p(\boldsymbol{\theta} \,|\, D) \log \frac{p(\boldsymbol{\theta},\, D)}{q(\boldsymbol{\theta} \,|\, D)} \,\mathrm{d}\boldsymbol{\theta} - \log p(D).
\end{aligned}$$

Hence, calculating $\mathrm{KL}(p \,|\, q)$ instead of $\mathrm{KL}(q \,|\, p)$ provides no benefit since we would now be required to evaluate the expectation of $\log \dfrac{p(\boldsymbol{\theta},\, D)}{q(\boldsymbol{\theta} \,|\, D)}$ under the true posterior $p(\boldsymbol{\theta} \,|\, D)$, which is only known up to a constant. Consequently, as mentioned previously, there is crucial significance attached to how we measure the misfit between the two distributions.

Therefore, we now return to (2.5). Clearly here, the term $\log p(D)$ is a constant, independent of $q(\boldsymbol{\theta} \,|\, D)$. Thus, to minimise the KL divergence, we need only minimise the first term in (2.5). A quantity, referred to only as $\mathcal{L}(q)$ for the present time, is defined to be the negative of this first term such that

$$\mathcal{L}(q) = \int q(\boldsymbol{\theta} \,|\, D) \log \frac{p(\boldsymbol{\theta},\, D)}{q(\boldsymbol{\theta} \,|\, D)} \,\mathrm{d}\boldsymbol{\theta}. \qquad (2.6)$$

In general, the integral in (2.6) is able to be evaluated, as will be discussed shortly. Hence, by taking the negative, we wish to maximise the value of $\mathcal{L}(q)$, which, correspondingly, minimises the KL divergence between the true and approximating posterior. In performing this, an optimal variational distribution will then have been derived.

2.4.3 Approximating the marginal likelihood

In this chapter hitherto, the variational approach has been examined from the perspective of approximating a posterior distribution. However, such methods are yet further attractive since they can play a significant role in the area of model comparison. Suppose we have an available set of candidate models, $\mathcal{M} = \{M_1, \ldots, M_R\}$. By conditioning our distributions upon the model M_i, (2.5) can now be specified as

$$\mathrm{KL}(q_i \,|\, p) = \log p(D \,|\, M_i) - \mathcal{L}_{M_i}(q_i), \tag{2.7}$$

where each $\mathcal{L}_{M_i}(q_i)$ is specific to every M_i, $q_i = q(\boldsymbol{\theta}_i \,|\, D,\, M_i)$ and $\boldsymbol{\theta}_i = (\boldsymbol{\theta}_{i1}, \ldots, \boldsymbol{\theta}_{id})^T$. Moreover, by the Gibbs' inequality of the KL divergence, we also have that

$$\mathcal{L}_{M_i}(q_i) \leq \log p(D \,|\, M_i). \tag{2.8}$$

Equations (2.7) and (2.8) are of critical importance in the variational Bayesian approach. For each model M_i, (2.8) reveals that $\mathcal{L}_{M_i}(q_i)$ provides a lower bound on the logarithm of the marginal likelihood, with the difference between the two being the KL divergence, as indicated by (2.7). Henceforth, $\mathcal{L}(q)$, whether dependent on a model or not, is referred to as the *lower bound* (or variational score). Moreover, to form (2.8), we now see the rationale behind defining the lower bound to be the *negative* of the first term in (2.5).

The crux of the method is that, by maximising $\mathcal{L}_{M_i}(q_i)$, we hence minimise the KL divergence by (2.7), and so ensure that the variational distribution is a good approximation to the true posterior. Furthermore, this implies correspondingly that the bound (2.8) will be made as tight as possible, and thus $\mathcal{L}_{M_i}(q_i)$ will be a good approximation to the log marginal likelihood over models. The process of enforcing the accuracy of the bound to the true value is referred to as bound optimisation. Now, reconsider (1.1) as in Winn (2003). If we again suppose a uniform prior across models such that $p(M_i) = \frac{1}{R}$, then the afore-mentioned variational theory implies that, approximately, the posterior density for M_i is such that

$$p(M_i \mid D) \propto \exp\left\{\mathcal{L}_{M_i}(q_i)\right\},$$

upon optimising q_i. So, we can utilise the lower bound to compare and rank a set of models. Notice that, if the KL divergence is zero, the lower bound will equal the log marginal likelihood, and the approximate posterior will hence be equivalent to the true posterior. Henceforth in this chapter, the dependence of our distributions on M_i is no longer assumed.

2.4.4 Computation of $\mathcal{L}(q)$

Thus far, we have seen that the variational Bayesian framework is often employed to find an optimal approximation to the true posterior distribution, but moreover, the lower bound $\mathcal{L}(q)$ can be utilised as a variational model selection criterion. However, we have not discussed how to calculate $\mathcal{L}(q)$, defined by (2.6). Thus, to ensure this integral is tractable, the variational approximation is required to be of a simpler form than the true posterior, else nothing has been gained. One way to ensure this is to assume $q(\boldsymbol{\theta} \mid D)$ factorises over parameters such that, if again $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_d)^T$, then

$$q(\boldsymbol{\theta} \mid D) = \prod_{j=1}^{d} q(\boldsymbol{\theta}_j \mid D). \tag{2.9}$$

This implies that the set of parameters, $\{\boldsymbol{\theta}_j\}$, have now been constrained to be independent *approximately a posteriori*. In other words, the approximating distribution now possesses a simpler dependency structure than the true posterior. Hence, approximation (2.9) can be substituted into (2.6). Moreover, we suppose that the joint density of data and parameters can also be split into a product of likelihood and prior terms such that, if $D = (\mathbf{x}_1, \dots, \mathbf{x}_N)$, then

$$p(\boldsymbol{\theta},\, D) = p(D \,|\, \boldsymbol{\theta}) p(\boldsymbol{\theta}) = \prod_{t=1}^{N} p(\mathbf{x}_t \,|\, \boldsymbol{\theta}) \prod_{j=1}^{d} p(\boldsymbol{\theta}_j).$$

The prior distributions are thus forced to be independent. Consequently, by taking logarithms as stated in (2.6), the lower bound $\mathcal{L}(q)$ can be written as

$$\mathcal{L}(q) = \int q(\boldsymbol{\theta} \,|\, D) \left[\sum_{t=1}^{N} \log p(\mathbf{x}_t \,|\, \boldsymbol{\theta}) + \sum_{j=1}^{d} \log \frac{p(\boldsymbol{\theta}_j)}{q(\boldsymbol{\theta}_j \,|\, D)} \right] \mathrm{d}\boldsymbol{\theta}. \tag{2.10}$$

At this stage, we can proceed using two separate procedures, known as the *free form* and *fixed form* variational methods (Lappalainen and Miskin, 2000). Both techniques rely on the independence of the variational distributions, as seen in (2.9). Yet, in the free form approach, no distributional form for the variational posteriors is assumed. Here, by writing $\mathcal{L}(q)$ as a functional of $q(\boldsymbol{\theta}_j \,|\, D)$ for all j, the lower bound is maximised by taking a functional derivative with respect to each variational distribution. Thus, we can derive the required distributional forms. For instance, by expressing (2.10) as a functional of $q(\boldsymbol{\theta}_j \,|\, D)$ and optimising, it is shown by Miskin (2000) and Winn (2003) that, in general,

$$\begin{aligned} q(\boldsymbol{\theta}_j \,|\, D) &\propto \exp\left\{ \mathrm{E}_{q(\boldsymbol{\theta}_{\setminus j} \,|\, D)} \left(\log p(\boldsymbol{\theta},\, D) \right) \right\} \qquad (2.11)\\ &= p(\boldsymbol{\theta}_j) \exp\left\{ \mathrm{E}_{q(\boldsymbol{\theta}_{\setminus j} \,|\, D)} \left(\sum_{t=1}^{N} \log p(\mathbf{x}_t \,|\, \boldsymbol{\theta}) \right) \right\}, \end{aligned}$$

where the notation $\boldsymbol{\theta}_{\setminus j}$ refers to all components of $\boldsymbol{\theta}$ except $\boldsymbol{\theta}_j$. The above formula can be used to simplify calculations in this thesis. Moreover, in this method, equations for the parameters of the variationals (referred to as *variational parameters*) are found simultaneously. Then ultimately, the explicit expression for $\mathcal{L}(q)$ is derived by use of (2.6).

On the other hand, by contemplating the situation from a fixed form perspective, a fixed and specified parametric form is assumed for each variational distribution. Such a selection is made to ensure that the joint variational distribution is similar to the true posterior, albeit an approximation as seen by (2.9). Thus, the lower bound $\mathcal{L}(q)$ is calculated initially by evaluation of the necessary integrals. Then, this bound is maximised with respect to the variational parameter set, hence minimising the KL-divergence and deriving expressions for the parameters.

It is customarily the case that, independent of which method is used, the algebraic expressions for the variational parameters will be dependent on each other. Hence, to find the optimal values of the parameters that maximise $\mathcal{L}(q)$, each equation must be iterated to convergence. Moreover, we realise that the value of $\mathcal{L}(q)$, due to its maximisation, will monotonically increase (or remain unchanged) at each iteration. As this quantity is also bounded above, the algorithm is guaranteed to converge. In fact, as Beal (2003) comments, a local maximum of the lower bound will eventually be reached. In the following example, both methods described here will be elucidated.

2.5 A univariate example

Suppose that we have a set of observed, one-dimensional data $D = (x_1, \ldots, x_N)$ such that we model $x_t \sim \mathcal{N}(m,\, v)$ for all $t = 1, \ldots, N$. Assuming the x_t to be independent, we write $p(D \,|\, m,\, v) = \prod_{t=1}^{N} p(x_t \,|\, m,\, v)$. We wish to infer m and v. Moreover, define prior distributions over m and v so that

$$p(m) = \mathcal{N}(m \,|\, \mu_m,\, \sigma_m) \tag{2.12}$$

$$p(v) = \mathcal{IG}(v \,|\, a,\, b), \tag{2.13}$$

where an inverse gamma prior is defined over v. The choice of priors stems from the fact that these are the typical semi-conjugate choices for a Gaussian distribution with unknown mean and variance. That is the case when, although the usual Bayesian update is non-conjugate using independent priors, the two full conditional posterior distributions for both mean and variance are of standard form and follow the same distributional form as the respective priors. This will become clear when constructing a Gibbs sampler for this example later in the chapter. Allowing our prior beliefs about m and v to separate into independent specifications implies that knowledge of one of the parameters does not inform us about the distribution of the other.

It is evident that, for this model, analytic analysis of the posterior distribution is intractable. This follows since

$$\begin{aligned} p(m,\, v \,|\, D) &\propto \left(\prod_{t=1}^{N} p(x_t \,|\, m,\, v) \right) p(m)\, p(v) \\ &= v^{-\frac{N}{2}} \exp\left\{ -\frac{v^{-1}}{2} \sum_{t=1}^{N} (x_t - m)^2 \right\} \times \exp\left\{ -\frac{\sigma_m^{-1}}{2} (m - \mu_m)^2 \right\} \\ &\quad \times v^{-(a+1)} \exp\left\{ -bv^{-1} \right\} \\ &= v^{-(a+\frac{N}{2}+1)} \exp\left\{ -\frac{v^{-1}}{2} \sum_{t=1}^{N} (x_t - m)^2 - \frac{\sigma_m^{-1}}{2} (m - \mu_m)^2 - bv^{-1} \right\}. \end{aligned} \tag{2.14}$$

Clearly, this density will not factorise for m and v and hence these parameters are not independent *a posteriori*. To combat this problem, three techniques are now employed to learn an approximate, marginal posterior for both m and v, given the data: variational Bayesian methods, Gibbs sampling and the EM algorithm. Initially, both the free form and fixed form variational procedures are applied to elucidate the theory and reveal how the methods can coincide. Throughout this example, the results of Appendix A are of particular importance.

2.5.1 Free form variational method

In this instance, optimal variationals, initially for m and later for v, are found without assuming any distributional form. These are denoted by $q(m \,|\, D)$ and $q(v \,|\, D)$ respectively. In fact, we only allow independence between these distributions such that $q(m,\, v \,|\, D) = q(m \,|\, D)\, q(v \,|\, D)$. This assumption will crucially simplify the following computation. Of course, it is also an approximation to the truth, which, as recognised above, does not factorise.

To commence, we write $\mathcal{L}(q)$ as a functional of both $q(m \,|\, D)$ and $q(v \,|\, D)$. Therefore, in this example, the lower bound is defined as

$$\mathcal{L}(q) = \iint q(m,\, v \,|\, D) \log \left[\frac{p(D,\, m,\, v)}{q(m,\, v \,|\, D)} \right] \mathrm{d}m \, \mathrm{d}v. \tag{2.15}$$

This expression is straightforward to manipulate since the prior and variational distributions are independent. Hence, the lower bound is also equivalent to

$$\begin{aligned} \mathcal{L}(q) = & \iint q(m \,|\, D) q(v \,|\, D) \left\{ \sum_{t=1}^{N} \log p(x_t \,|\, m,\, v) \right\} \mathrm{d}m \, \mathrm{d}v + \int q(m \,|\, D) \log p(m) \, \mathrm{d}m \\ & + \int q(v \,|\, D) \log p(v) \, \mathrm{d}v - \int q(m \,|\, D) \log q(m \,|\, D) \, \mathrm{d}m \\ & - \int q(v \,|\, D) \log q(v \,|\, D) \, \mathrm{d}v. \end{aligned} \tag{2.16}$$

Here, parameters have been integrated out when necessary, a process aided by the factorisation of $q(m,\, v \,|\, D)$. Resultantly, by recombining integrals, the lower bound can now be expressed as a functional of both variational distributions. As a functional of $q(m \,|\, D)$, we obtain

$$\mathcal{L}(q) = \int q(m \,|\, D) \Bigg[\int q(v \,|\, D) \left\{ \sum_{t=1}^{N} \log p(x_t \,|\, m,\, v) \right\} \mathrm{d}v \\ + \log p(m) - \log q(m \,|\, D) \Bigg] \mathrm{d}m \, + \, \text{const.} \quad (2.17)$$

Moreover, expressing in terms of $q(v \,|\, D)$ provides

$$\mathcal{L}(q) = \int q(v \,|\, D) \Bigg[\int q(m \,|\, D) \left\{ \sum_{t=1}^{N} \log p(x_t \,|\, m,\, v) \right\} \mathrm{d}m \\ + \log p(v) - \log q(v \,|\, D) \Bigg] \mathrm{d}v \, + \, \text{const.} \quad (2.18)$$

Each functional contains a constant term that is independent of $q(m \,|\, D)$ and $q(v \,|\, D)$ correspondingly. Recall that, in this method, we take a functional derivative of $\mathcal{L}(q)$ with respect to both variational distributions. Consequently, at that stage, these constants will disappear.

By examining both equations, we can derive the distributional forms for both $q(m \,|\, D)$ and $q(v \,|\, D)$ respectively. The integrals are tackled in order. So, we can substitute in for both $\log p(x_t \,|\, m,\, v)$ and $\log p(m)$ in (2.17) such that

$$\mathcal{L}(q) = \int q(m \,|\, D) \Bigg[\sum_{t=1}^{N} \int q(v \,|\, D) \left\{ -\frac{1}{2} \log 2\pi v - \frac{v^{-1}}{2} (x_t - m)^2 \right\} \mathrm{d}v \\ - \frac{1}{2} \log 2\pi\sigma_m - \frac{\sigma_m^{-1}}{2} (m - \mu_m)^2 - \log q(m \,|\, D) \Bigg] \mathrm{d}m + \text{const.}$$

Again here, there are terms, independent of m. After taking a functional derivative of the lower bound, such terms will clearly play no part in determining the form of $q(m \,|\, D)$.

Henceforth, they are dropped and included in a new constant term. Thus, we acquire

$$\mathcal{L}(q) = \int q(m \,|\, D)\Bigg[-\frac{1}{2}\sum_{t=1}^{N}(x_t - m)^2 \int q(v \,|\, D) v^{-1}\,\mathrm{d}v \\ - \frac{\sigma_m^{-1}}{2}(m - \mu_m)^2 - \log q(m \,|\, D)\Bigg]\,\mathrm{d}m + \text{const.}' \quad (2.19)$$

Of course here, $\int q(v \,|\, D)v^{-1}\,\mathrm{d}v = \mathrm{E}_{q(v \,|\, D)}\left\{v^{-1}\right\}$. This expectation can be defined when the variational distribution for v has been derived. Subsequently, $\mathcal{L}(q)$, written as a functional of $q(m \,|\, D)$, will no longer depend on v as this parameter will have been integrated out. This is important since we assumed independence between the variational distributions. At this time, notice that since $q(m \,|\, D)$ is a density function, it must integrate to 1. That is, we look for the variational distribution that optimises the lower bound, subject to the constraint that it is normalised. To enforce this, Lappalainen and Miskin (2000) use a Lagrange multiplier. Thus, correspondingly, a new functional, $\tilde{\mathcal{L}}(q)$, is formed such that

$$\tilde{\mathcal{L}}(q) = \mathcal{L}(q) + \nu_m\left(\int q(m \,|\, D)\,\mathrm{d}m - 1\right), \quad (2.20)$$

and ν_m is the required Lagrangian. Via (2.19), we differentiate $\tilde{\mathcal{L}}(q)$ with respect to the distribution $q(m \,|\, D)$. Taking the functional derivative and equating to zero results in

$$\frac{\partial \tilde{\mathcal{L}}(q)}{\partial q(m \,|\, D)} = -\frac{\mathrm{E}_{q(v \,|\, D)}\left\{v^{-1}\right\}}{2}\sum_{t=1}^{N}(x_t - m)^2 - \frac{\sigma_m^{-1}}{2}(m - \mu_m)^2 - \log q(m \,|\, D) - 1 + \nu_m = 0.$$

Rearranging in terms of $q(m \,|\, D)$ and dropping constant terms, we arrive at

$$q(m \,|\, D) \propto \exp\Bigg\{ -\frac{1}{2}\Bigg[m^2\left(N\mathrm{E}_{q(v \,|\, D)}\left\{v^{-1}\right\} + \sigma_m^{-1}\right) \\ - 2m\left(\mathrm{E}_{q(v \,|\, D)}\left\{v^{-1}\right\}\sum_{t=1}^{N}x_t + \sigma_m^{-1}\mu_m\right)\Bigg]\Bigg\}$$

$$\propto \exp\left\{-\frac{1}{2}\left(N\mathrm{E}_{q(v\,|\,D)}\left\{v^{-1}\right\}+\sigma_m^{-1}\right)\left(m-\frac{\mathrm{E}_{q(v\,|\,D)}\left\{v^{-1}\right\}\sum_{t=1}^{N}x_t+\sigma_m^{-1}\mu_m}{N\mathrm{E}_{q(v\,|\,D)}\left\{v^{-1}\right\}+\sigma_m^{-1}}\right)^2\right\}$$

via completing the square. It is thus apparent that the variational distribution for m is a Gaussian distribution such that

$$q(m\,|\,D)=\mathcal{N}(m\,|\,{\mu_m}',\,{\sigma_m}'). \tag{2.21}$$

Moreover, the variational parameters, ${\mu_m}'$ and ${\sigma_m}'$, are defined as

$${\mu_m}'=\frac{\mathrm{E}_{q(v\,|\,D)}\left\{v^{-1}\right\}\sum_{t=1}^{N}x_t+\sigma_m^{-1}\mu_m}{N\mathrm{E}_{q(v\,|\,D)}\left\{v^{-1}\right\}+\sigma_m^{-1}} \tag{2.22}$$

$${\sigma_m}'=\frac{1}{N\mathrm{E}_{q(v\,|\,D)}\left\{v^{-1}\right\}+\sigma_m^{-1}}. \tag{2.23}$$

This process is now repeated to find the optimum form for $q(v\,|\,D)$. Hence, reconsider (2.18). On this occasion, by substituting in for $\log p(x_t\,|\,m,\,v)$ and $\log p(v)$, the lower bound is expressed as

$$\begin{aligned}\mathcal{L}(q)=\int q(v\,|\,D)\Bigg[\sum_{t=1}^{N}\int q(m\,|\,D)\left\{-\frac{1}{2}\log 2\pi v-\frac{v^{-1}}{2}(x_t-m)^2\right\}\mathrm{d}m\\ +a\log b-\log\Gamma(a)-(a+1)\log v-bv^{-1}-\log q(v\,|\,D)\Bigg]\,\mathrm{d}v+\text{const.}\end{aligned}$$

As we strive to find the variational distribution for v, those terms that are independent of this parameter can again be dropped as before. Consequently, we obtain

$$\begin{aligned}\mathcal{L}(q)=\int q(v\,|\,D)\Bigg[-\frac{N}{2}\log v-\frac{v^{-1}}{2}\sum_{t=1}^{N}\int q(m\,|\,D)(x_t-m)^2\,\mathrm{d}m\\ -(a+1)\log v-bv^{-1}-\log q(v\,|\,D)\Bigg]\,\mathrm{d}v+\text{const.}'\end{aligned} \tag{2.24}$$

with the new constant term specified. In addition, with $q(m \,|\, D)$ now known, this expression can be simplified yet further. Therefore, by (A.2),

$$\begin{aligned}\sum_{t=1}^{N} \int q(m \,|\, D)(x_t - m)^2 \, \mathrm{d}m &= \sum_{t=1}^{N} \mathrm{E}_{q(m \,|\, D)} \left\{(x_t - m)^2\right\} \\ &= \sum_{t=1}^{N} \left(\mathrm{E}_{q(m \,|\, D)} \left\{x_t - m\right\}\right)^2 + \sum_{t=1}^{N} \mathrm{Var}_{q(m \,|\, D)} \left\{x_t - m\right\} \\ &= \sum_{t=1}^{N} (x_t - {\mu_m}')^2 + N{\sigma_m}'. \end{aligned} \tag{2.25}$$

So, we have an expression for $\mathcal{L}(q)$ that is now independent of m. We seek the variational $q(v \,|\, D)$ that maximises the lower bound, with respect to the density function integrating to 1. Hence again, the Lagrangian ν_v is applied to construct $\tilde{\mathcal{L}}(q)$ such that

$$\tilde{\mathcal{L}}(q) = \mathcal{L}(q) + \nu_v \left(\int q(v \,|\, D) \, \mathrm{d}v - 1 \right).$$

It is now possible to optimise $\tilde{\mathcal{L}}(q)$ with respect to $q(v \,|\, D)$. Thus, by differentiating and setting to zero, we achieve

$$\begin{aligned}\frac{\partial \tilde{\mathcal{L}}(q)}{\partial q(v \,|\, D)} = -\left(a + \frac{N}{2} + 1\right) \log v - v^{-1} \left(b + \frac{1}{2} \sum_{t=1}^{N} (x_t - {\mu_m}')^2 + \frac{N}{2} {\sigma_m}' \right) \\ - \log q(v \,|\, D) - 1 + \nu_v = 0.\end{aligned}$$

A simple rearrangement here provides

$$q(v \,|\, D) \propto v^{-(a + \frac{N}{2} + 1)} \exp \left\{ -v^{-1} \left(b + \frac{1}{2} \sum_{t=1}^{N} (x_t - {\mu_m}')^2 + \frac{N}{2} {\sigma_m}' \right) \right\}.$$

Hence, we have finally that the approximate posterior $q(v \,|\, D)$ is an inverse gamma distribution given by

$$q(v \mid D) = \mathcal{IG}(v \mid a', b'), \tag{2.26}$$

where a' and b' have been found to be equivalent to

$$a' = a + \frac{N}{2} \tag{2.27}$$

$$b' = b + \frac{1}{2}\sum_{t=1}^{N}(x_t - \mu_m{}')^2 + \frac{N}{2}\sigma_m{}'. \tag{2.28}$$

Consequently, we can now compute $\mathrm{E}_{q(v \mid D)}\{v^{-1}\}$, upon which the equations for $\mu_m{}'$ and $\sigma_m{}'$ depend. It is evident that $\mathrm{E}_{q(v \mid D)}\{v^{-1}\} = \frac{a'}{b'}$ via (A.5). To summarise, we have a set of variational parameters, $\{\mu_m{}', \sigma_m{}', a', b'\}$. When inspecting the corresponding algebraic expressions, it follows that these equations, and hence the variational distributions for m and v, are dependent upon each other, as commented upon in Section 2.4.4. Therefore, we solve these equations iteratively. That is, we update the parameter values by continuous use of (2.22), (2.23), (2.27) and (2.28) until convergence. The resulting distributions are then optimal in terms of minimising KL divergence, given the approximation (2.9). In the free form method, this procedure of updating each variational distribution with respect to all others in the approximation will be seen in further examples in this thesis.

2.5.2 Fixed form variational method

The second variational procedure is now applied to learn an approximate posterior for both m and v, given a dataset. Consequently, fixed distributional forms for both variational distributions must be chosen. On this occasion, it is straightforward to make such selections since we can use the distributions derived by the free form approach. Hence, we have

$$
\begin{aligned}
q(m \mid D) &= \mathcal{N}(m \mid {\mu_m}', {\sigma_m}') \\
q(v \mid D) &= \mathcal{IG}(v \mid a', b'),
\end{aligned}
$$

and these distributions are required to be independent as before, thus simplifying computation. We can return to (2.16) and compute $\mathcal{L}(q)$ initially, using the now known variationals. At this point, we realise an overlap between the free and fixed variational approaches. The free form method concludes by deriving $\mathcal{L}(q)$ to obtain an estimate of the log evidence, provided this calculation is deemed necessary. However initially, in this fixed form case, the procedure is carried out in identical fashion since $q(m \mid D)$ and $q(v \mid D)$ have been fixed to follow the same distributions as previously suggested by the free form approach.

The sum of integrals in (2.16) are now tackled in order. Thus firstly, we acquire

$$
\begin{aligned}
&\iint q(m \mid D) q(v \mid D) \left\{ \sum_{t=1}^{N} \log p(x_t \mid m, v) \right\} \mathrm{d}m \, \mathrm{d}v \\
&= \sum_{t=1}^{N} \left[\iint q(m \mid D) q(v \mid D) \left\{ -\frac{1}{2} \log 2\pi v - \frac{v^{-1}}{2} (x_t - m)^2 \right\} \mathrm{d}m \, \mathrm{d}v \right] \\
&= -\frac{N}{2} \log 2\pi - \frac{N}{2} \int q(v \mid D) \log v \, \mathrm{d}v \\
&\quad - \frac{1}{2} \sum_{t=1}^{N} \int q(v \mid D) v^{-1} \, \mathrm{d}v \int q(m \mid D)(x_t - m)^2 \, \mathrm{d}m \\
&= -\frac{N}{2} \log 2\pi - \frac{N}{2} \left[\log b' - \psi(a') \right] - \frac{a'}{2b'} \sum_{t=1}^{N} (x_t - {\mu_m}')^2 - \frac{Na'}{2b'} {\sigma_m}'.
\end{aligned}
$$

In the last line, (2.25), (A.5) and (A.6) have been applied. Further, by the definition of the prior and variational distribution for m,

$$
\begin{aligned}
&\int q(m \,|\, D) \log p(m) \,\mathrm{d}m \\
&= \int q(m \,|\, D) \left\{ -\frac{1}{2} \log 2\pi\sigma_m - \frac{\sigma_m^{-1}}{2} (m - \mu_m)^2 \right\} \mathrm{d}m \\
&= -\frac{1}{2} \log 2\pi\sigma_m - \frac{\sigma_m^{-1}}{2} \left[\left(\mathrm{E}_{q(m \,|\, D)} \left\{ m - \mu_m \right\} \right)^2 + \mathrm{Var}_{q(m \,|\, D)} \left\{ m - \mu_m \right\} \right] \\
&= -\frac{1}{2} \log 2\pi\sigma_m - \frac{\sigma_m^{-1}}{2} \left[\left({\mu_m}' - \mu_m \right)^2 + {\sigma_m}' \right].
\end{aligned}
$$

In a similar fashion for v, we obtain

$$
\begin{aligned}
&\int q(v \,|\, D) \log p(v) \,\mathrm{d}v \\
&= \int q(v \,|\, D) \left\{ a \log b - \log \Gamma(a) - (a+1) \log v - b v^{-1} \right\} \mathrm{d}v \\
&= a \log b - \log \Gamma(a) - (a+1) \left[\log b' - \psi(a') \right] - \frac{b a'}{b'}.
\end{aligned}
$$

Moreover, it is apparent that

$$
\begin{aligned}
&\int q(m \,|\, D) \log q(m \,|\, D) \,\mathrm{d}m \\
&= \int q(m \,|\, D) \left\{ -\frac{1}{2} \log 2\pi{\sigma_m}' - \frac{({\sigma_m}')^{-1}}{2} (m - {\mu_m}')^2 \right\} \mathrm{d}m \\
&= -\frac{1}{2} \log 2\pi{\sigma_m}' - \frac{({\sigma_m}')^{-1}}{2} \left[\left(\mathrm{E}_{q(m \,|\, D)} \left\{ m - {\mu_m}' \right\} \right)^2 + \mathrm{Var}_{q(m \,|\, D)} \left\{ m - {\mu_m}' \right\} \right] \\
&= -\frac{1}{2} \log 2\pi{\sigma_m}' - \frac{1}{2}.
\end{aligned}
$$

Finally, in a comparable way to $\int q(v \,|\, D) \log p(v) \,\mathrm{d}v$, it follows that

$$
\begin{aligned}
&\int q(v \,|\, D) \log q(v \,|\, D) \,\mathrm{d}v \\
&= \int q(v \,|\, D) \left\{ a' \log b' - \log \Gamma(a') - (a'+1) \log v - b' v^{-1} \right\} \mathrm{d}v
\end{aligned}
$$

$$= -\log\Gamma(a') - \log b' + (a'+1)\psi(a') - a'.$$

By collecting all these integrals together, we can substitute into (2.16), and hence obtain an expression for $\mathcal{L}(q)$. This gives the lower bound to be

$$\begin{aligned}\mathcal{L}(q) = &-\frac{N}{2}\log 2\pi - \frac{N}{2}\left[\log b' - \psi(a')\right] - \frac{a'}{2b'}\sum_{t=1}^{N}(x_t - {\mu_m}')^2 - \frac{Na'}{2b'}{\sigma_m}' \\ &-\frac{1}{2}\log 2\pi\sigma_m - \frac{\sigma_m^{-1}}{2}\left[({\mu_m}' - \mu_m)^2 + {\sigma_m}'\right] + a\log b - \log\Gamma(a) \\ &-(a+1)\left[\log b' - \psi(a')\right] - \frac{ba'}{b'} + \frac{1}{2}\log 2\pi{\sigma_m}' + \frac{1}{2} + \log\Gamma(a') \\ &+\log b' - (a'+1)\psi(a') + a'. \end{aligned} \tag{2.29}$$

Consequently, we can find the maximum of the lower bound by setting its gradient to zero. That is, we effect partial differentiation of $\mathcal{L}(q)$ with respect to the variational parameter set $\{{\mu_m}', {\sigma_m}', a', b'\}$. Thus, we examine initially maximisation with respect to ${\mu_m}'$. This then yields

$$\begin{aligned}\frac{\partial\mathcal{L}(q)}{\partial{\mu_m}'} &= \frac{\partial}{\partial{\mu_m}'}\left\{-\frac{\sigma_m^{-1}}{2}\left({\mu_m}' - \mu_m\right)^2 - \frac{a'}{2b'}\sum_{t=1}^{N}(x_t - {\mu_m}')^2\right\} \\ &= -\sigma_m^{-1}\left({\mu_m}' - \mu_m\right) + \frac{a'}{b'}\left(\sum_{t=1}^{N}x_t - N{\mu_m}'\right).\end{aligned}$$

By setting to zero and then rearranging, we find that the update equation for ${\mu_m}'$ is identical to (2.22) as required, recalling the definition of $\mathrm{E}_{q(v\,|\,D)}\{v^{-1}\}$. Differentiating with respect to ${\sigma_m}'$ leads to

$$\begin{aligned}\frac{\partial\mathcal{L}(q)}{\partial{\sigma_m}'} &= \frac{\partial}{\partial{\sigma_m}'}\left\{-\frac{Na'}{2b'}{\sigma_m}' - \frac{\sigma_m^{-1}}{2}{\sigma_m}' + \frac{1}{2}\log 2\pi{\sigma_m}'\right\} \\ &= -\frac{Na'}{2b'} - \frac{\sigma_m^{-1}}{2} + \frac{({\sigma_m}')^{-1}}{2}.\end{aligned}$$

On this occasion, equating to zero and solving for ${\sigma_m}'$ provides (2.23). Effecting the same procedure in terms of $a\,'$ implies

$$\begin{aligned}\frac{\partial\mathcal{L}(q)}{\partial a\,'} &= \frac{\partial}{\partial a\,'}\Bigg\{\frac{N}{2}\psi(a\,') - \frac{a\,'}{2b\,'}\sum_{t=1}^{N}(x_t - {\mu_m}')^2 - \frac{Na\,'}{2b\,'}{\sigma_m}' + (a+1)\psi(a\,') \\ &\qquad - \frac{ba\,'}{b\,'} + \log\Gamma(a\,') - (a\,'+1)\psi(a\,') + a\,'\Bigg\} \\ &= \psi_1(a\,')\left[a + \frac{N}{2} - a\,'\right] - \frac{1}{b\,'}\left[b + \frac{1}{2}\sum_{t=1}^{N}(x_t - {\mu_m}')^2 + \frac{N}{2}{\sigma_m}'\right] + 1, \end{aligned} \tag{2.30}$$

where the *trigamma* function (Johnson et al., 1992), denoted by $\psi_1(z)$ for some $z \in \mathbb{R}$, is defined to be

$$\psi_1(z) = \frac{\mathrm{d}^2}{\mathrm{d}z^2}\log\Gamma(z) = \frac{\mathrm{d}}{\mathrm{d}z}\psi(z).$$

Finally, the partial differentiation of $\mathcal{L}(q)$ with respect to $b\,'$ offers

$$\begin{aligned}\frac{\partial\mathcal{L}(q)}{\partial b\,'} &= \frac{\partial}{\partial b\,'}\Bigg\{ - \frac{N}{2}\log b\,' - \frac{a\,'}{2b\,'}\sum_{t=1}^{N}(x_t - {\mu_m}')^2 - \frac{Na\,'}{2b\,'}{\sigma_m}' \\ &\qquad - (a+1)\log b\,' - \frac{ba\,'}{b\,'} + \log b\,'\Bigg\} \\ &= \frac{a\,'}{(b\,')^2}\left[b + \frac{1}{2}\sum_{t=1}^{N}(x_t - {\mu_m}')^2 + \frac{N}{2}{\sigma_m}'\right] - \frac{1}{b\,'}\left[a + \frac{N}{2}\right]. \end{aligned} \tag{2.31}$$

By inspecting (2.30) and (2.31), it is apparent that these expressions are zeroed by specifications (2.27) and (2.28) for $a\,'$ and $b\,'$. Therefore, when choosing the variational distributional forms in the fixed form method to be those suggested by the free form method, the two variational approaches, as expected, have coincided. However, as Miskin (2000) indicates, one can make incorrect choices for the approximating posteriors in the fixed form algorithm, hence affecting subsequent results.

2.5.3 The Gibbs sampler

Here, we are only required to find the two full conditional posterior distributions for m and v, namely $p(m \,|\, v,\, D)$ and $p(v \,|\, m,\, D)$. From (2.14), it is evident that

$$\begin{aligned} p(m \,|\, v,\, D) &\propto \exp\left\{-\frac{1}{2}\left[v^{-1}\sum_{t=1}^{N}(x_t - m)^2 + \sigma_m^{-1}(m-\mu_m)^2\right]\right\} \\ &\propto \exp\left\{-\frac{1}{2}\left(Nv^{-1}+\sigma_m^{-1}\right)\left(m - \frac{v^{-1}\sum_{t=1}^{N} x_t + \sigma_m^{-1}\mu_m}{Nv^{-1}+\sigma_m^{-1}}\right)^2\right\}, \end{aligned}$$

and hence

$$p(m \,|\, v,\, D) = \mathcal{N}\left(\frac{v^{-1}\sum_{t=1}^{N} x_t + \sigma_m^{-1}\mu_m}{Nv^{-1}+\sigma_m^{-1}},\, \frac{1}{Nv^{-1}+\sigma_m^{-1}}\right). \tag{2.32}$$

Furthermore,

$$p(v \,|\, m,\, D) \propto v^{-(a+\frac{N}{2}+1)} \exp\left\{-v^{-1}\left(b + \frac{1}{2}\sum_{t=1}^{N}(x_t - m)^2\right)\right\},$$

whereby

$$p(v \,|\, m,\, D) = \mathcal{IG}\left(a + \frac{N}{2},\, b + \frac{1}{2}\sum_{t=1}^{N}(x_t - m)^2\right). \tag{2.33}$$

The semi-conjugacy of the problem is now realised, *i.e.* the full conditionals for m and v are of standard form, and follow the same distributions as the corresponding priors. Once initialised anywhere such that the posterior has support, the sampler then produces alternate simulations from the full conditionals and a bivariate Markov chain is hence defined. Upon convergence, the corresponding samples will be draws from the density of interest, $p(m,\, v \,|\, D)$. Moreover, the values for each component are simulations from the corresponding marginal posterior distribution.

2.5.4 EM algorithm

Two separate EM algorithms are now constructed to find the modes $\tilde{m}$ and $\tilde{v}$ of the respective marginal posterior densities, $p(m \,|\, D)$ and $p(v \,|\, D)$. Hence, via the full conditionals (2.32) and (2.33), the unknown marginals can be approximated by $p(m \,|\, v = \tilde{v},\, D)$ and $p(v \,|\, m = \tilde{m},\, D)$.

From (2.14), it is clear that the logarithm of the joint posterior density is

$$\log p(m,\, v \,|\, D) = -\left(a + \frac{N}{2} + 1\right) \log v - \frac{v^{-1}}{2} \sum_{t=1}^{N} (x_t - m)^2 - \frac{\sigma_m^{-1}}{2} (m - \mu_m)^2 - b v^{-1}. \quad (2.34)$$

Suppose that we are currently at iteration k. Initially, we use Algorithm 3 to derive an expression for $m^{(k)}$. So, in the E-step, we take the expectation of (2.34) with respect to $p(v \,|\, m^{(k-1)},\, D)$, where $m^{(k-1)}$ is the marginal posterior mode at the previous iteration. Denoting $\mathrm{E}_{m^{(k-1)}}\{\cdot\}$ to be the expectation with respect to $p(v \,|\, m^{(k-1)},\, D)$, the following is yielded:

$$\begin{aligned}\mathrm{E}_{m^{(k-1)}} \{\log p(m,\, v \,|\, D)\} = -\left(a + \frac{N}{2} + 1\right) \mathrm{E}_{m^{(k-1)}} \{\log v\} - \frac{1}{2} \mathrm{E}_{m^{(k-1)}} \left\{v^{-1}\right\} \sum_{t=1}^{N} (x_t - m)^2 \\ - \frac{\sigma_m^{-1}}{2} (m - \mu_m)^2 - b \,\mathrm{E}_{m^{(k-1)}} \left\{v^{-1}\right\}. \quad (2.35)\end{aligned}$$

Here, evaluation of $\mathrm{E}_{m^{(k-1)}} \{\log v\}$ is not required since, being independent of m, the corresponding term in (2.35) will disappear under differentiation in the M-step. Resultantly, we need only compute $\mathrm{E}_{m^{(k-1)}} \{v^{-1}\}$, which, by (A.5) and our previous derivation of the full conditional for v, is equivalent to

$$\mathrm{E}_{m^{(k-1)}} \left\{v^{-1}\right\} = \frac{a + \frac{N}{2}}{b + \frac{1}{2} \sum_{t=1}^{N} \left[x_t - m^{(k-1)}\right]^2}. \quad (2.36)$$

We can now proceed to the M-step. By differentiating (2.35) with respect to m, we hence achieve

$$\frac{\partial}{\partial m}\mathrm{E}_{m^{(k-1)}}\left\{\log p(m,\, v\,|\, D)\right\} = \mathrm{E}_{m^{(k-1)}}\left\{v^{-1}\right\}\sum_{t=1}^{N}(x_t - m) - \sigma_m^{-1}(m - \mu_m).$$

By equating to zero and solving for m, the current marginal posterior mode estimate for $p(m\,|\,D)$ is

$$m^{(k)} = \frac{\mathrm{E}_{m^{(k-1)}}\left\{v^{-1}\right\}\sum_{t=1}^{N} x_t + \sigma_m^{-1}\mu_m}{N\mathrm{E}_{m^{(k-1)}}\left\{v^{-1}\right\} + \sigma_m^{-1}}, \tag{2.37}$$

substituting in (2.36). The same procedure is performed to determine $v^{(k)}$. As a consequence, we now calculate the expectation of (2.34) with respect to $p(m\,|\,v^{(k-1)},\, D)$ in the E-step. Hence, we obtain

$$\begin{aligned}\mathrm{E}_{v^{(k-1)}}\left\{\log p(m,\, v\,|\, D)\right\} = &-\left(a + \frac{N}{2} + 1\right)\log v - \frac{v^{-1}}{2}\sum_{t=1}^{N}\mathrm{E}_{v^{(k-1)}}\left\{(x_t - m)^2\right\} \\ &- \frac{\sigma_m^{-1}}{2}\mathrm{E}_{v^{(k-1)}}\left\{(m - \mu_m)^2\right\} - bv^{-1}.\end{aligned} \tag{2.38}$$

We realise that, similar to before, computation of $\mathrm{E}_{v^{(k-1)}}\left\{(m - \mu_m)^2\right\}$ is not necessary. Therefore, we have

$$\begin{aligned}&\mathrm{E}_{v^{(k-1)}}\left\{(x_t - m)^2\right\} \\ &= \left(\mathrm{E}_{v^{(k-1)}}\left\{x_t - m\right\}\right)^2 + \mathrm{Var}_{v^{(k-1)}}\left\{m\right\} \\ &= \left(x_t - \frac{\left[v^{(k-1)}\right]^{-1}\sum_{t=1}^{N} x_t + \sigma_m^{-1}\mu_m}{N\left[v^{(k-1)}\right]^{-1} + \sigma_m^{-1}}\right)^2 + \frac{1}{N\left[v^{(k-1)}\right]^{-1} + \sigma_m^{-1}},\end{aligned} \tag{2.39}$$

due to (2.32). In the M-step, maximising (2.38) with respect to v implies

$$\begin{aligned}\frac{\partial}{\partial v}\mathrm{E}_{v^{(k-1)}}\left\{\log p(m,\, v\,|\, D)\right\} = &-\left(a + \frac{N}{2} + 1\right)v^{-1} \\ &+ \frac{v^{-2}}{2}\sum_{t=1}^{N}\mathrm{E}_{v^{(k-1)}}\left\{(x_t - m)^2\right\} + bv^{-2}.\end{aligned}$$

So, this equation is zeroed when

$$v^{(k)} = \frac{b + \frac{1}{2}\sum_{t=1}^{N} \mathrm{E}_{v^{(k-1)}}\left\{(x_t - m)^2\right\}}{a + \frac{N}{2} + 1}, \tag{2.40}$$

substituting in (2.39). By iterating equations (2.37) and (2.40) separately K times until convergence, we will obtain $\tilde{m} = m^{(K)}$ and $\tilde{v} = v^{(K)}$, the modes for $p(m \,|\, D)$ and $p(v \,|\, D)$ respectively. Thus, an approximation to these two marginal posteriors is given by $p(m \,|\, v = \tilde{v},\, D)$ and $p(v \,|\, m = \tilde{m},\, D)$.

It is worth briefly mentioning that the above derivation can also be used to obtain the expressions for the variational parameters seen in Section 2.5.1. By application of (2.11), it is apparent that both (2.22) and (2.23) can be read off from equation (2.35) without any additional work, similarly (2.27) and (2.28) from (2.38).

2.5.5 A numerical example

We illustrate the theory above with a simple, numerical example. Suppose that our dataset consists of $N = 20$ samples, simulated from a univariate Gaussian distribution with mean $m = 2$ and variance $v = 1$. In addition, the priors for m and v were given the following specifications:

$$p(m) = \mathcal{N}(m \,|\, 0,\, 10,000)$$
$$p(v) = \mathcal{IG}(v \,|\, 1,\, 0.001).$$

Thus, both priors are deemed to be diffuse as each has been assigned a huge variance. In fact, as the variance of an $\mathcal{IG}(a,\, b)$ distribution is defined only for $a > 2$, the above distribution for v has infinite variance. So importantly, we do not favour any particular value of the parameters *a priori*. A more thorough discussion of vague, inverse gamma prior specification, in particular, is offered in Chapter 3.

The variational Bayes approach, EM algorithm and Gibbs sampler were then run for this example. In the variational case, equations (2.22) and (2.23), for ${\mu_m}'$ and ${\sigma_m}'$ respectively, are both dependent upon $a\,'$ and $b\,'$. So, an arbitrary, initial choice of $a\,' = b\,' = 1$ was made for the algorithm to commence. Similarly, the EM algorithm and Gibbs sampler were both initialised such that $m^{(0)}$ and $v^{(0)}$ were points simulated from the respective prior distributions. The sampler was run for $10,000$ iterations, the first 1000 of which were discarded as burn-in. Convergence of the variational Bayes algorithm was extremely rapid, taking no more than 4 iterations.

Figure 2.1 shows the plots of the approximate marginal posteriors for the two parameters, illustrating the three methods. To recap, the variational posteriors are the distributions (2.21) and (2.26) with variational parameters whose update equations have been run until convergence. The marginals via the EM algorithm are the full conditionals (2.32) and (2.33), dependent upon the posterior modes $\tilde{v}$ and $\tilde{m}$ respectively. Finally, kernel density estimates are plotted for the draws obtained via the Gibbs sampler. Inspection of plots (a) and (b) in Figure 2.1 clearly illustrate the similarity of the distributions produced by the three approaches. Moreover, each marginal is centred at values very close to the true values of the parameters. This is impressive since a dataset of only small size was used to infer m and v. Hence in this case, the variational Bayes method appears to produce results, considered equally as good as two other rival approximations.

Further evidence for the worth of the variational approach is offered in Figure 2.2. Here, contour lines are plotted for both the joint variational distribution, $q(m,\, v \,|\, D)$, and the true posterior (2.14), known up to a multiplicative constant. The figure clearly shows that the contours for these distributions are centred in almost the equivalent position and, moreover, are similar in shape. However, due to the independence assumption that $q(m,\, v \,|\, D) = q(m \,|\, D) q(v \,|\, D)$, the variational approximation is not quite able to fully capture correlations between m and v, seen in the truth. Yet, it does correctly show that

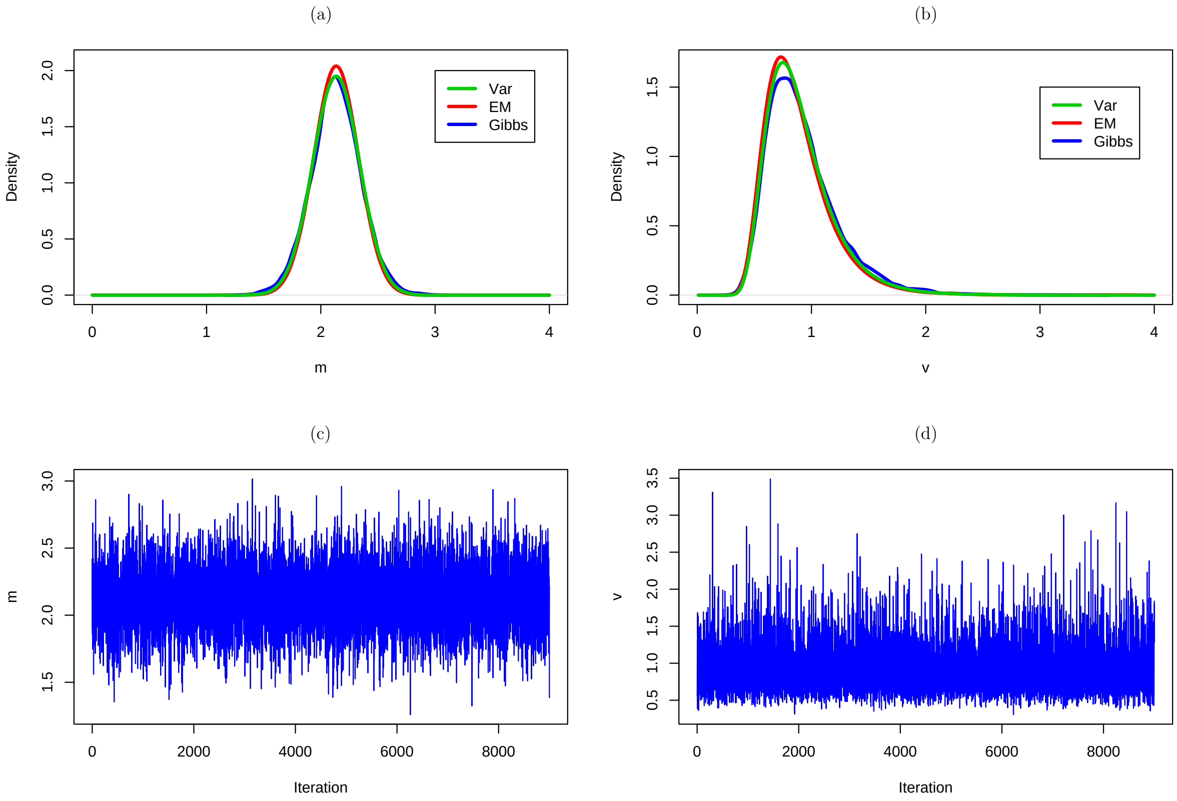


Figure 2.1: (a) and (b): Approximate marginal posterior distributions for m and v respectively, using variational Bayes, the EM algorithm and the Gibbs sampler; (c) and (d): Corresponding trace plots for m and v produced by the Gibbs sampler

the posterior density is not symmetric about the mode value of v.

2.6 A multivariate example

In the previous section, *inter alia*, a variational Bayesian approach was used to infer approximate distributions for the unknown mean and variance of a univariate, Gaussian sample. In fact, the above numerical example has shown the method to produce fast and accurate results. These ideas are now extended to the corresponding multivariate case. This will motivate subsequent chapters in this thesis, whereby variational Bayes is applied to vector autoregressive models of order 1.

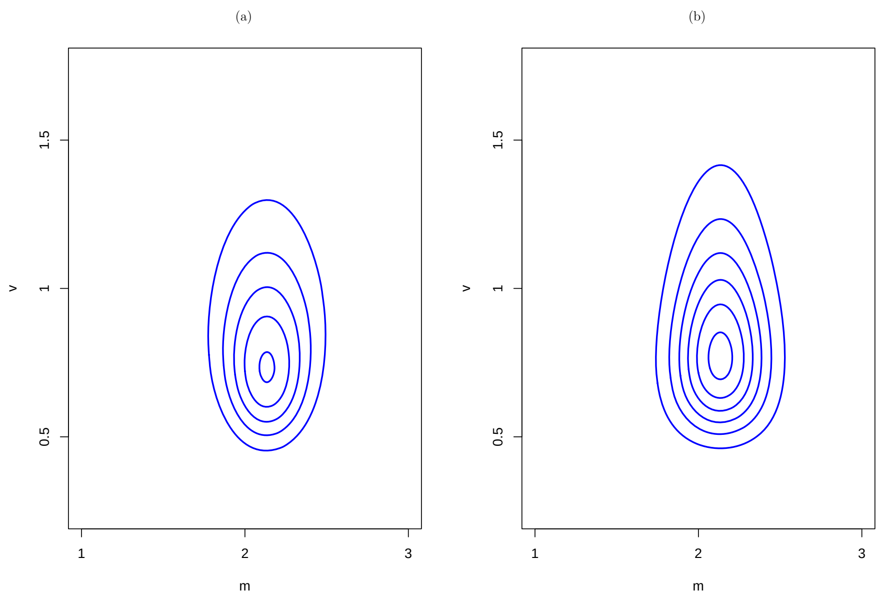


Figure 2.2: Contour plots for: (a) Variational posterior, (b) True posterior

Consequently, we now possess a dataset $D = (\mathbf{x}_1, \ldots, \mathbf{x}_N)$, where each $\mathbf{x}_t$ is an independent, d-dimensional random vector such that $\mathbf{x}_t \sim \mathcal{N}(\mathbf{m}, V)$ for all $t = 1, \ldots, N$. Here, $\mathbf{m}$ is a mean vector and V a $d \times d$ covariance matrix. The specification of priors for $\mathbf{m}$ and V is now

$$p(\mathbf{m}) = \mathcal{N}(\mathbf{m} \,|\, \boldsymbol{\mu}_m, \, \Sigma_m) \tag{2.41}$$

$$p(V) = \mathcal{IW}(V \,|\, A, \, r), \tag{2.42}$$

where the prior for V follows an *inverse Wishart* distribution with parameters A, a $d \times d$ matrix, and a scalar r. Further details of this distribution are provided in Appendix A. These independent prior distributions merely generalise the univariate, semi-conjugate choices, seen in Section 2.5, to higher dimensions.

To emphasise the need to approximate the joint posterior in this case, it follows that

$$p(\mathbf{m}, \, V \,|\, D) \propto \left(\prod_{t=1}^{N} p(\mathbf{x}_t \,|\, \mathbf{m}, \, V) \right) p(\mathbf{m}) \, p(V)$$

$$= |V|^{-(r+d+N+1)/2} \exp\left\{ -\frac{1}{2}\left[\sum_{t=1}^{N} (\mathbf{x}_t - \mathbf{m})^T V^{-1} (\mathbf{x}_t - \mathbf{m}) \right.\right.$$
$$\left.\left. + (\mathbf{m} - \boldsymbol{\mu}_m)^T \Sigma_m^{-1} (\mathbf{m} - \boldsymbol{\mu}_m) + \mathrm{Tr}\left[V^{-1}A\right] \right]\right\}.$$

Of course, this is akin to (2.14) and shows that the above density will not factorise, hence no distributional form can be found for the marginal posteriors of $\mathbf{m}$ and V. Thus, we can use variational Bayesian techniques to infer respective variational distributions, $q(\mathbf{m} \,|\, D)$ and $q(V \,|\, D)$. Again, we form an approximation such that the joint variational posterior factorises into the corresponding variational marginals.

We proceed by mimicking the free form method of Section 2.5.1, hence assuming no variational distributional form. Consequently, the lower bound is now given by

$$\mathcal{L}(q) = \iint q(\mathbf{m},\, V \,|\, D) \log \frac{p(D,\, \mathbf{m},\, V)}{q(\mathbf{m},\, V \,|\, D)} \,\mathrm{d}\mathbf{m}\,\mathrm{d}V. \tag{2.43}$$

Writing $\mathcal{L}(q)$ as a functional of both $q(\mathbf{m} \,|\, D)$ and $q(V \,|\, D)$ is elementary via studying the analogous univariate expressions, (2.17) and (2.18). We again notice the importance here of independence between each prior and variational distribution. Hence, as a functional of $q(\mathbf{m} \,|\, D)$, the following is acquired:

$$\mathcal{L}(q) = \int q(\mathbf{m} \,|\, D) \left[\int q(V \,|\, D) \left\{ \sum_{t=1}^{N} \log p(\mathbf{x}_t \,|\, \mathbf{m},\, V) \right\} \mathrm{d}V \right.$$
$$\left. + \log p(\mathbf{m}) - \log q(\mathbf{m} \,|\, D) \right] \mathrm{d}\mathbf{m} + \text{const.} \tag{2.44}$$

The corresponding expression in terms of $q(V \,|\, D)$ is

$$\mathcal{L}(q) = \int q(V \,|\, D) \Bigg[\int q(\mathbf{m} \,|\, D) \left\{ \sum_{t=1}^{N} \log p(\mathbf{x}_t \,|\, \mathbf{m},\, V) \right\} \mathrm{d}\mathbf{m} \\ + \log p(V) - \log q(V \,|\, D) \Bigg] \mathrm{d}V \;+\; \text{const.} \quad (2.45)$$

The distributional form for $q(\mathbf{m} \,|\, D)$ is derived initially. By substituting in the appropriate terms, (2.44) can be rewritten as

$$\mathcal{L}(q) = \int q(\mathbf{m} \,|\, D) \Bigg[\sum_{t=1}^{N} \int q(V \,|\, D) \Bigg\{ -\frac{d}{2} \log 2\pi - \frac{1}{2} \log |V| \\ - \frac{1}{2} (\mathbf{x}_t - \mathbf{m})^T V^{-1} (\mathbf{x}_t - \mathbf{m}) \Bigg\} \mathrm{d}V - \frac{d}{2} \log 2\pi - \frac{1}{2} \log |\Sigma_m| \\ - \frac{1}{2} (\mathbf{m} - \boldsymbol{\mu}_m)^T \Sigma_m^{-1} (\mathbf{m} - \boldsymbol{\mu}_m) - \log q(\mathbf{m} \,|\, D) \Bigg] \mathrm{d}\mathbf{m} + \text{const.}$$

By dropping all terms independent of $\mathbf{m}$, we then obtain

$$\mathcal{L}(q) = \int q(\mathbf{m} \,|\, D) \Bigg[-\frac{1}{2} \sum_{t=1}^{N} (\mathbf{x}_t - \mathbf{m})^T \, \mathrm{E}_{q(V \,|\, D)} \left\{ V^{-1} \right\} (\mathbf{x}_t - \mathbf{m}) \, \mathrm{d}V \\ - \frac{1}{2} (\mathbf{m} - \boldsymbol{\mu}_m)^T \Sigma_m^{-1} (\mathbf{m} - \boldsymbol{\mu}_m) - \log q(\mathbf{m} \,|\, D) \Bigg] \mathrm{d}\mathbf{m} + \text{const.}'$$

The term $\mathrm{E}_{q(V \,|\, D)} \left\{ V^{-1} \right\}$ can be computed upon determining the variational posterior for V. The functional $\tilde{\mathcal{L}}(q)$ is now formed using the Lagrangian $\nu_{\mathbf{m}}$ in the way akin to (2.20), hence ensuring that $q(\mathbf{m} \,|\, D)$ is normalised. We now seek the optimal $q(\mathbf{m} \,|\, D)$ that maximises $\tilde{\mathcal{L}}(q)$. Hence, by differentiating, we obtain

$$\frac{\partial \tilde{\mathcal{L}}(q)}{\partial q(\mathbf{m} \,|\, D)} = -\frac{1}{2} \sum_{t=1}^{N} (\mathbf{x}_t - \mathbf{m})^T \, \mathrm{E}_{q(V \,|\, D)} \left\{ V^{-1} \right\} (\mathbf{x}_t - \mathbf{m}) \\ - \frac{1}{2} (\mathbf{m} - \boldsymbol{\mu}_m)^T \Sigma_m^{-1} (\mathbf{m} - \boldsymbol{\mu}_m) - \log q(\mathbf{m} \,|\, D) - 1 + \nu_{\mathbf{m}} = 0.$$

Rearranging this expression then implies

$$\begin{aligned}
q(\mathbf{m}\,|\,D) \propto \exp\Bigg\{ &-\frac{1}{2}\Bigg[\mathbf{m}^T\left(N\mathrm{E}_{q(V\,|\,D)}\left\{V^{-1}\right\}+\Sigma_m^{-1}\right)\mathbf{m} \\
&-\mathbf{m}^T\left(\mathrm{E}_{q(V\,|\,D)}\left\{V^{-1}\right\}\sum_{t=1}^{N}\mathbf{x}_t+\Sigma_m^{-1}\boldsymbol{\mu}_m\right)-\left(\sum_{t=1}^{N}\mathbf{x}_t^T\mathrm{E}_{q(V\,|\,D)}\left\{V^{-1}\right\}+\boldsymbol{\mu}_m^T\Sigma_m^{-1}\right)\mathbf{m}\Bigg]\Bigg\} \\
\propto \exp\Bigg\{ &-\frac{1}{2}\left[\mathbf{m}-\left(N\mathrm{E}_{q(V\,|\,D)}\left\{V^{-1}\right\}+\Sigma_m^{-1}\right)^{-1}\left(\mathrm{E}_{q(V\,|\,D)}\left\{V^{-1}\right\}\sum_{t=1}^{N}\mathbf{x}_t+\Sigma_m^{-1}\boldsymbol{\mu}_m\right)\right]^T \\
&\times\left[N\mathrm{E}_{q(V\,|\,D)}\left\{V^{-1}\right\}+\Sigma_m^{-1}\right] \\
&\times\left[\mathbf{m}-\left(N\mathrm{E}_{q(V\,|\,D)}\left\{V^{-1}\right\}+\Sigma_m^{-1}\right)^{-1}\left(\mathrm{E}_{q(V\,|\,D)}\left\{V^{-1}\right\}\sum_{t=1}^{N}\mathbf{x}_t+\Sigma_m^{-1}\boldsymbol{\mu}_m\right)\right]\Bigg\}.
\end{aligned}$$

Hence, it follows that the variational posterior for $\mathbf{m}$ is a multivariate Gaussian distribution such that

$$q(\mathbf{m}\,|\,D)=\mathcal{N}(\mathbf{m}\,|\,\boldsymbol{\mu}_m{}',\,\Sigma_m{}'), \tag{2.46}$$

with update equations for the variational parameters specified as

$$\boldsymbol{\mu}_m{}'=\left(N\mathrm{E}_{q(V\,|\,D)}\left\{V^{-1}\right\}+\Sigma_m^{-1}\right)^{-1}\left(\mathrm{E}_{q(V\,|\,D)}\left\{V^{-1}\right\}\sum_{t=1}^{N}\mathbf{x}_t+\Sigma_m^{-1}\boldsymbol{\mu}_m\right) \tag{2.47}$$

$$\Sigma_m{}'=\left(N\mathrm{E}_{q(V\,|\,D)}\left\{V^{-1}\right\}+\Sigma_m^{-1}\right)^{-1}. \tag{2.48}$$

The identical course is now taken for $q(V\,|\,D)$. By substituting in for the prior on V and likelihood, (2.45) is now given by

$$\begin{aligned}
\mathcal{L}(q)=\int q(V\,|\,D)\Bigg[\sum_{t=1}^{N}\int q(\mathbf{m}\,|\,D)\Bigg\{&-\frac{d}{2}\log 2\pi-\frac{1}{2}\log|V| \\
&-\frac{1}{2}(\mathbf{x}_t-\mathbf{m})^TV^{-1}(\mathbf{x}_t-\mathbf{m})\Bigg\}\,\mathrm{d}\mathbf{m}-\log k+\frac{r}{2}\log|A| \\
&-\frac{r+d+1}{2}\log|V|-\frac{1}{2}\mathrm{Tr}\left[V^{-1}A\right]-\log q(V\,|\,D)\Bigg]\,\mathrm{d}V+\text{const.}
\end{aligned}$$

where k is defined by (A.10). Dropping all terms independent of V provides

$$\begin{aligned}\mathcal{L}(q) = \int q(V \,|\, D)\Bigg[&-\frac{N}{2}\log|V| - \frac{1}{2}\sum_{t=1}^{N}\mathrm{E}_{q(\mathbf{m}\,|\,D)}\left\{(\mathbf{x}_t-\mathbf{m})^T V^{-1}(\mathbf{x}_t-\mathbf{m})\right\} \\ &-\frac{r+d+1}{2}\log|V| - \frac{1}{2}\mathrm{Tr}\left[V^{-1}A\right] - \log q(V \,|\, D)\Bigg]\,\mathrm{d}V + \text{const.}' \end{aligned} \quad (2.49)$$

By knowledge of $q(\mathbf{m} \,|\, D)$, we can also compute

$$\begin{aligned}\mathrm{E}_{q(\mathbf{m}\,|\,D)}\left\{(\mathbf{x}_t-\mathbf{m})^T V^{-1}(\mathbf{x}_t-\mathbf{m})\right\} &= \mathrm{E}_{q(\mathbf{m}\,|\,D)}\left\{\mathbf{x}_t-\mathbf{m}\right\}^T V^{-1}\mathrm{E}_{q(\mathbf{m}\,|\,D)}\left\{\mathbf{x}_t-\mathbf{m}\right\} \\ &\quad + \mathrm{Tr}\left[V^{-1}\mathrm{Var}_{q(\mathbf{m}\,|\,D)}\left\{\mathbf{x}_t-\mathbf{m}\right\}\right] \\ &= (\mathbf{x}_t-\boldsymbol{\mu}_m{}')^T V^{-1}(\mathbf{x}_t-\boldsymbol{\mu}_m{}') + \mathrm{Tr}\left[V^{-1}\Sigma_m{}'\right]. \end{aligned} \quad (2.50)$$

Here, we have utilised the identity to find the expectation of a quadratic form, *i.e.*

$$\mathrm{E}\left\{\mathbf{w}^T P \mathbf{w}\right\} = \mathrm{E}\left\{\mathbf{w}\right\}^T P\,\mathrm{E}\left\{\mathbf{w}\right\} + \mathrm{Tr}\left[P\,\mathrm{Var}\left\{\mathbf{w}\right\}\right], \quad (2.51)$$

where $\mathbf{w}$ is a random vector and P is a compatible, fixed matrix (Rice, 1995).

By forming $\tilde{\mathcal{L}}(q)$ with the Lagrangian ν_V, we differentiate with respect to $q(V \,|\, D)$:

$$\begin{aligned}\frac{\partial\tilde{\mathcal{L}}(q)}{\partial q(V \,|\, D)} = &-\frac{N}{2}\log|V| - \frac{1}{2}\sum_{t=1}^{N}(\mathbf{x}_t-\boldsymbol{\mu}_m{}')^T V^{-1}(\mathbf{x}_t-\boldsymbol{\mu}_m{}') - \frac{N}{2}\mathrm{Tr}\left[V^{-1}\Sigma_m{}'\right] \\ &-\frac{r+d+1}{2}\log|V| - \frac{1}{2}\mathrm{Tr}\left[V^{-1}A\right] - \log q(V \,|\, D) - 1 + \nu_V = 0.\end{aligned}$$

Notice that $(\mathbf{x}_t-\boldsymbol{\mu}_m{}')^T V^{-1}(\mathbf{x}_t-\boldsymbol{\mu}_m{}') = \mathrm{Tr}\left[V^{-1}(\mathbf{x}_t-\boldsymbol{\mu}_m{}')(\mathbf{x}_t-\boldsymbol{\mu}_m{}')^T\right]$ since

$$\mathrm{Tr}\left[PQR\right] = \mathrm{Tr}\left[RPQ\right] = \mathrm{Tr}\left[QRP\right]$$

for compatible matrices P, Q, R (Harville, 1997).

Thus, defining $S\,' = \frac{1}{N}\sum_{t=1}^{N}(\mathbf{x}_t - \boldsymbol{\mu}_m{}')(\mathbf{x}_t - \boldsymbol{\mu}_m{}')^T$, it hence follows that

$$q(V \mid D) \propto |V|^{-(r+N+d+1)/2} \exp\left\{-\frac{1}{2}\mathrm{Tr}\left[V^{-1}\left(A + N\Sigma_m{}' + NS\,'\right)\right]\right\}.$$

Therefore, the variational for V is distributed as

$$q(V \mid D) = \mathcal{IW}(V \mid A\,',\, r\,'), \tag{2.52}$$

with $A\,'$ and $r\,'$ expressed as

$$A\,' = A + N\Sigma_m{}' + NS\,' \tag{2.53}$$

$$r\,' = r + N. \tag{2.54}$$

Hence, we can now express $\mathrm{E}_{q(V \mid D)}\left\{V^{-1}\right\} = r\,'(A\,')^{-1}$ via (A.11) and this result is substituted into both (2.47) and (2.48). Finally, update equations (2.47), (2.48), (2.53) and (2.54) are iterated until converged values of $\boldsymbol{\mu}_m{}'$, $\Sigma_m{}'$, $A\,'$ and $r\,'$ are found, hence defining the variational distributions for $\mathbf{m}$ and V.

In the task of deriving variational posteriors for the unknown mean and variance of a Gaussian sample, it is evident that the univariate case in Section 2.5 has been naturally extended in this section to higher dimensions. Clearly, the choice of multivariate normal and inverse Wishart priors for $\mathbf{m}$ and V respectively simply generalises the semi-conjugate specifications seen previously. This is further true in terms of the variational distributions ultimately derived in both circumstances. In the following chapter, similar variational multivariate theory will be required to score sparse vector autoregressive models.

2.7 Summary

The purpose of this chapter was to introduce the theory behind variational Bayes, founded upon minimising the KL divergence between the approximating, variational distribution and the true posterior. Hence, as Kullback-Leibler is a global measure, an analytic, global approximation to this distribution will be provided that is optimal over the whole parameter space. This is in contrast to Laplace's approximation that only makes a local (Gaussian) approximation to the posterior at the MAP estimate. Furthermore, we noted that $\mathcal{L}(q)$, a bound on the log marginal likelihood for each model, can be utilised as a variational model comparison criterion. This feature is exploited in the remainder of this thesis.

The main problem with the variational Bayesian approach is that, for computational reasons, the true posterior is assumed to factorise. This implies that we cannot determine any *a posteriori* dependencies between parameters. However, in contrast, we have seen that variational Bayes is a fast and computationally efficient procedure. In addition, the example of Section 2.5.5 has moreover revealed its accuracy, relative to two competing alternatives: the Gibbs sampler and the EM algorithm.

The option of using either a free form or fixed form variational method would also appear attractive. As has been mentioned, care must be employed when choosing a distributional form for the variationals in the latter case to ensure a reliable approximation. Yet, by not making such an assumption, the free form procedure will always ensure the best, possible variationals are selected. On the other hand, the fixed form method is often straightforward to apply for more complicated models where the free form approach is intractable, *i.e.* in situations where integration over the model parameters cannot be performed. Although that is not the case in this thesis, both of these procedures will have an important role to play in Chapters 3 and 5.

Chapter 3

Model comparison of VAR(1) models

3.1 Introduction

By definition, a time series is a set of data values that are measured at equally spaced, successive time points. For instance, a simple example would be to monitor the average price of houses in a particular region each month. By modelling such data accurately, we hope to be able to predict future events in the series. A popular way to effect this would be to use an autoregressive (AR) model, defined so that there exists a linear dependence on previous data values. Moreover, if the data is of dimension d, then the time series is now multivariate (*i.e.* there are d time series), and can be modelled via a *vector autoregressive (VAR)* process, possessing either a zero or non-zero mean.

The Bayesian treatment of VAR models has traditionally focussed on learning the optimum model order and the model parameters, given a set of time series data. Such analysis is analytically intractable. To tackle this, the variational Bayesian algorithm has been often applied as an approximation. For instance, in the context of choosing the model order p, its use has been seen when modelling via a zero mean, univariate autoregressive model, with noise given by both a Gaussian distribution (Penny and Roberts, 2000) and a mixture of Gaussians (Roberts and Penny, 2002) and, moreover, a zero mean VAR(p) model (Penny and Roberts, 2002). In addition, the method has allowed approximation of the parameter posterior distributions found in dynamic linear models, leading to identification and subsequent graphical display of potential interactions between genes (Beal et al., 2005).

In contrast, in the rest of this thesis, an analogous treatment is provided to the particular situation of VAR models with fixed model order 1, but *sparse* matrix of VAR coefficients, denoted as A. Hence, it can be shown that the VAR process can be represented graphically. Moreover, the use of sparsity as a modelling tool implies that we are able to develop a model comparison problem. This is due to the construction of a candidate set of potential 'A-graphs', corresponding to sparse 'A-matrices'. Our task therefore is, given data modelled using a VAR(1) process, to find the models that appear the most likely from the set. That is, we want to estimate the unknown sparsity structure of A, the autoregressive matrix. In this chapter, the zero mean case is considered, in Chapter 5, non-zero mean models. Of course, we already know that the variational framework is particularly attractive for this purpose as we can use $\mathcal{L}_{M_i}(q_i)$, a tractable lower bound on the logarithm of the marginal likelihood, inherent within the algorithm, to rank candidate models.

3.2 VAR(1) graphical models

We commence by studying the family of VAR(p) models, in particular the VAR(1) process, and showing how to model using sparsity. The *zero mean* VAR(p) process of dimension d is expressed as

$$\mathbf{y}_t = \sum_{i=1}^{p} \mathbf{y}_{t-i} A(i) + \mathbf{e}_t. \tag{3.1}$$

So, as Penny and Roberts (2002) indicate, the new, t-th value of the multivariate time series, $\mathbf{y}_t$, is explained via a linear combination of the p previous data values of the series. Here, $\mathbf{y}_t = (y_{t1}, y_{t2}, \ldots, y_{td})$, a $(1 \times d)$ vector, each $A(i)$ is a $d \times d$ matrix of coefficients and $\mathbf{e}_t = (e_{t1}, e_{t2}, \ldots, e_{td})$ is a $(1 \times d)$ noise-vector, distributed as

$$\mathbf{e}_t \sim \mathcal{N}(\mathbf{0}, \Gamma).$$

Moreover, with $\mathbf{e}_t$ independent to $\mathbf{e}_u$ for $t \neq u$ (*i.e.* $\mathrm{Cov}(\mathbf{e}_t, \mathbf{e}_u) = 0$) and with zero mean, such vectors are defined to be Gaussian white noise. We now focus in particular on the following zero mean VAR(1) model:

$$\mathbf{y}_t = \mathbf{y}_{t-1}A + \mathbf{e}_t, \tag{3.2}$$

where $\mathbf{e}_t$ is distributed as above. The covariance matrix is now defined as $\Gamma = \sigma^2 \mathrm{I}_d$ for unknown parameter σ^2 and $d \times d$ identity matrix I_d. This specification implies that all off-diagonal covariances are zero, *i.e.* $\mathrm{Cov}(e_{ij}, e_{kl}) \neq 0$ if and only if $(i, j) = (k, l)$ for $i, k = 1, \ldots N$ and $j, l = 1, \ldots d$, assuming N samples are collected. Moreover, $\mathrm{Var}(e_{ij}) = \sigma^2$. We define Γ in this way since, in a Bayesian analysis, it allows a simple, univariate prior specification on σ^2 as opposed to needing a more complicated specification on a matrix, such as an inverse Wishart prior (*c.f.* Appendix A).

For a detailed analysis of VAR(p) models, refer to Lütkepohl (2005). We now draw special attention to the matrix A of VAR(1) coefficients. For the purposes of what follows, it is assumed that A is a *sparse* matrix. Hence, by definition, it will consist of many, in particular off-diagonal, elements constrained to be zero, with only a few, unspecified non-zero entries. A matrix of this ilk allows us to take advantage of the substantial number of zeroes that it possesses. For instance, it can be related to a graphical structure. We realise that, in a different context, the pattern of zeroes in the concentration matrix of an arbitrary multivariate Gaussian distribution provides the conditional independence structure, which subsequently characterises an existent graphical (Gaussian) model. This is detailed in Appendix B.

Similarly, the sparse matrix A in a VAR(1) process, containing a clear zero structure, can be represented graphically also. To do this, we must model (3.2) using a dynamic graphical structure (Ghahramani, 1997; Friedman, Murphy, and Russell, 1998; Mihajlovic and Petkovic, 2001). Of course, the same procedure can be applied in the non-zero mean case. We realise that a Bayesian network (a graphical model with directed edges) is used to describe the conditional dependencies between a fixed set of random variables in a static situation (see Appendix B).

Conversely, a dynamic Bayesian network is a special case of the afore-mentioned static graphical model, specifically orientated towards modelling time series. Each time point, at which the values of a set of random variables are observed, is often referred to as a time slice. Within a network, directed edges connect nodes from one slice to the next, denoting the dependencies of the corresponding variables. Such edges are sometimes called inter-edges. A convention is adopted whereby inter-edges point in the direction of time, hence illustrating that one variable can cause another, only if the latter is in the future.

Moreover, dynamic Bayesian networks can also contain edges within each slice, known as intra-edges. In this case, the conditional dependencies between variables in a single time slice are represented by a static Bayesian network. In other words, a dynamic Bayesian network can be viewed as merely a collection of static Bayesian networks, linked by inter-edges. Each dynamic network would contain not only the identical graphical structure for every time slice, but, moreover, the identical dependencies between slices. Thus, notice that the term 'dynamic' does not refer to the network changing over time slices, but instead to the dynamic process being modelled.

VAR models can be represented as continuous-state dynamic Bayesian networks since each node is a continuous random variable. In what follows, we consider no intra-edges in the graphical model. However, such edges, given as undirected connections, can be used to specify the zero structure in the corresponding concentration matrix of the noise vector $\mathbf{e}_t$ (Eichler, 2001). In our case, the VAR model has order equal to 1. Thus, consider a dynamic Bayesian network between times $t-1$ and t, where each component of $\mathbf{y}_{t-1} = (y_{t-1,1},\, y_{t-1,2},\, \ldots, y_{t-1,d})$ and $\mathbf{y}_t$ is a node. We use inter-edges to connect nodes in these two successive time slices together and this pattern is repeated over all slices. The network is correspondingly said to have order 1. By aggregating the nodes y_{ti} for each $i = 1, \ldots, d$ across all time points t (in particular, from $t-1$ to t) into a single node, say y_i, in the time series graph, we can hence form a causality graph (Dahlhaus and Eichler, 2003). Thus, each node represents one component of the whole time series. If a component is dependent upon its own past, we allow this to be expressed by a directed self-loop.

It has previously been documented that in such a time series graph with $p = 1$, for $a, b = 1, \ldots, d$, an edge exists between nodes $y_{t-1,a}$ and y_{tb} if and only if the element a_{ba} of the autoregressive matrix is non-zero (Eichler, 2001; Murphy, 2002; Dahlhaus and Eichler, 2003). We note that such a result generalises to a VAR(p) model. Hence, there is a clear link between the causality graph and A in this circumstance as the former is defined through the sparsity structure of the latter. Due to this relationship, the causality graph is resultantly referred to as an *A-graph* throughout the remainder of this thesis. The subsequent example elucidates the situation.

3.2.1 Example

Suppose $d = 2$. Consider Figure 3.1, showing a time series graph for a VAR(1) process. As mentioned erstwhile for such a model, inter-edges are used to define a structure between successive time slices, here shown repeated. By letting $\mathbf{y}_i = (y_{i1},\, y_{i2})$ where $i = t-2,\, t-1,\, t$, the nodes on the graph are clearly specified by representing these random variables.

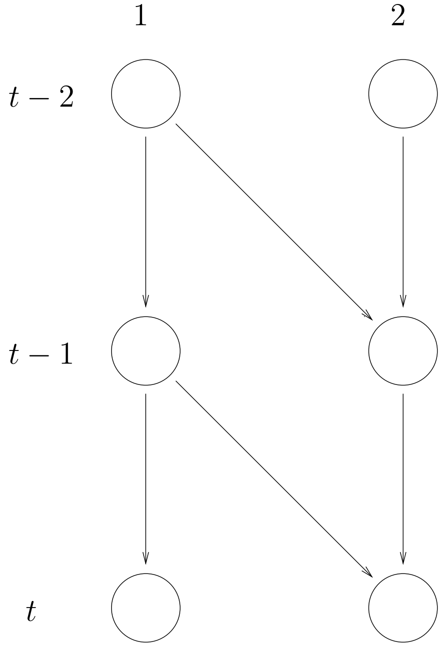


Figure 3.1: Time series graph for a VAR(1) model

By concentrating only on one pair of successive slices, we can hence produce a causality graph across all time slices for this process, with nodes y_1 and y_2, as given below.

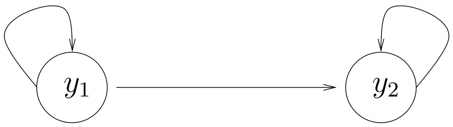


Figure 3.2: Causality graph for the VAR(1) process

Notice the use of self-loops for both nodes here. Using the above result of correspondence between the dynamic Bayesian network, hence causality graph, and sparse A-matrix, we have the specification in this circumstance that $A = \begin{pmatrix} * & 0 \\ * & * \end{pmatrix}$, whereby $*$ represents a free, non-zero element. Thus, the graph of Figure 3.2 is termed an A-graph.

3.3 Scoring zero mean VAR(1) graphical models

It is clear that the variational Bayesian method is of high relevance in terms of scoring models. We now apply this to the particular case of zero mean VAR(1) models. Hence, using the theory of Section 3.2, we construct a candidate set of graphical models, say $\mathcal{M} = \{M_1, M_2, \ldots\}$. Each graphical model M_i relates to an A-graph, G_i, which, in turn, corresponds to a sparse A-matrix, denoted by $A^{(i)}$. Recall that there exists an edge between two nodes of a given G_i if and only if the correct corresponding element of $A^{(i)}$ is non-zero. We can quantify the evidence for each prospective graphical model with the corresponding marginal likelihood, denoted as $p(\{\mathbf{y}_t\} \mid M_i)$. However, as previously discussed, we can approximate this quantity using a variational Bayesian framework and, in particular, the tractable lower bound $\mathcal{L}_{M_i}(q_i)$. Henceforth, we assume not only dependence of the lower bound, but also conditioning in our distributions upon the graphical model M_i, although not stated explicitly.

The subsequent set-up follows that of Penny and Roberts (2002). Assume there exists $t = 1, \ldots, N$ independent samples of the time series. Therefore, to take account of and store these samples, we rewrite (3.2) using matrix notation, and hence form a multivariate linear model. Firstly, define $\mathbf{x}_t = [\mathbf{y}_{t-1}]$ for all $t = 1, \ldots, N$. Then, we form matrices Y, X and E, all of which have dimension $N \times d$, such that the t-th row of each matrix is respectively given by $\mathbf{y}_t$, $\mathbf{x}_t$ and $\mathbf{e}_t$. Consequently, using the definitions of these vectors, we obtain a matrix equation such that

$$\begin{pmatrix} y_{11} & \ldots & y_{1d} \\ \vdots & \ddots & \vdots \\ y_{t1} & \ldots & y_{td} \\ \vdots & \ddots & \vdots \\ y_{N1} & \ldots & y_{Nd} \end{pmatrix} = \begin{pmatrix} x_{11} & \ldots & x_{1d} \\ \vdots & \ddots & \vdots \\ x_{t1} & \ldots & x_{td} \\ \vdots & \ddots & \vdots \\ x_{N1} & \ldots & x_{Nd} \end{pmatrix} \begin{pmatrix} a_{11} & \ldots & a_{1d} \\ \vdots & \ddots & \vdots \\ a_{d1} & \ldots & a_{dd} \end{pmatrix} + \begin{pmatrix} e_{11} & \ldots & e_{1d} \\ \vdots & \ddots & \vdots \\ e_{t1} & \ldots & e_{td} \\ \vdots & \ddots & \vdots \\ e_{N1} & \ldots & e_{Nd} \end{pmatrix}.$$

So, we can succinctly denote this as

$$Y = XA + E. \tag{3.3}$$

At the end-points, we take $\mathbf{x}_1 = (0, 0, \ldots, 0)$ and $\mathbf{x}_N = \mathbf{y}_{N-1}$. $\mathbf{x}_N$ is specified directly through the definition of $\mathbf{x}_t$. Correspondingly, we set $\mathbf{x}_1 = \mathbf{y}_0$ to equal the mean of the stationary distribution. Of course, (3.2) is such that $\mathrm{E}(\mathbf{y}_t) = \mathbf{0}$ for all t, *i.e.* all $\mathbf{y}_t$ possess this mean, regardless of t.

Next, by using the afore-mentioned matrix notation, we can now resultantly compute the probability of the data, using ideas from Lütkepohl (2005) and Box and Tiao (1992). Assume a given data set $D = \{X, Y\}$. By using the vec operator, we now rewrite (3.3) according to

$$\begin{aligned} \mathrm{vec}(Y) &= \mathrm{vec}(XA + E) \\ &= \mathrm{vec}(XA) + \mathrm{vec}(E) \\ &= (\mathrm{I}_d \otimes X)\mathrm{vec}(A) + \mathrm{vec}(E) \\ \Longrightarrow \mathbf{y} &= (\mathrm{I}_d \otimes X)\mathbf{a} + \mathbf{e}, \end{aligned} \tag{3.4}$$

where $\mathbf{y}$, $\mathbf{e}$ are both $dN \times 1$ vectors and $\mathbf{a}$ is a $d^2 \times 1$ vector. That is, for example, $\mathbf{y}$ is formed by stacking the columns of Y one under the other, similarly for $\mathbf{e}$ and $\mathbf{a}$.

Here, we have used a core property of the vec operator: $\mathrm{vec}(P+Q) = \mathrm{vec}(P)+\mathrm{vec}(Q)$, for compatible matrices P and Q (Petersen and Pedersen, 2007). In addition, we define $\otimes$ to be a Kronecker product (Henderson and Searle, 1981). Furthermore, suppose specifically that P and Q are matrices of dimensions $m \times p$ and $p \times r$ respectively. Then, notice the result from Henderson and Searle (1979) that

$$\mathrm{vec}(PQ) = (\mathrm{I}_r \otimes P)\mathrm{vec}(Q) = (Q^T \otimes P)\mathrm{vec}(\mathrm{I}_p) = (Q^T \otimes \mathrm{I}_m)\mathrm{vec}(P). \tag{3.5}$$

Recall that $\mathbf{e}_t \sim \mathcal{N}(\mathbf{0},\, \sigma^2 \mathrm{I}_d)$. We now determine the mean vector and covariance matrix of $\mathbf{e}$. Clearly, $\mathrm{E}(\mathbf{e}) = \mathbf{0}$. Moreover, to derive the covariance, we define $\mathbf{e}_{(s)} = (e_{1s},\, e_{2s},\, \dots, e_{Ns})^T$, the error vector of the component $s = 1, \dots, d$ for each of the N data samples, *i.e.* the s-th column of E. Thus,

$$\mathrm{Var}(\mathbf{e}) = \mathrm{Var}\left(\begin{array}{c} \mathbf{e}_{(1)} \\ \hline \vdots \\ \hline \mathbf{e}_{(d)} \end{array}\right) = \begin{pmatrix} \mathrm{Var}\left(\mathbf{e}_{(1)}\right) & \dots & \mathrm{Cov}\left(\mathbf{e}_{(1)},\, \mathbf{e}_{(d)}\right) \\ \vdots & \ddots & \vdots \\ \mathrm{Cov}\left(\mathbf{e}_{(d)},\, \mathbf{e}_{(1)}\right) & \dots & \mathrm{Var}\left(\mathbf{e}_{(d)}\right) \end{pmatrix}.$$

Assume henceforth that $i,\, k = 1, \dots N$ and $j,\, l,\, r,\, s = 1, \dots d$. We know that $\mathrm{Var}(e_{ij}) = \sigma^2$, for all $i,\, j$. Thus, for each s, it follows that $\mathrm{Var}(\mathbf{e}_{(s)}) = \sigma^2 \mathrm{I}_N$, an $N \times N$ matrix. Moreover, from before, $\mathrm{Cov}(e_{ij},\, e_{kl}) = 0$ if and only if $(i,\, j) \neq (k,\, l)$. Therefore, $\mathrm{Cov}\left(\mathbf{e}_{(r)},\, \mathbf{e}_{(s)}\right) = 0$ for all $r \neq s$.

As E possesses d columns, consequently $\mathrm{Var}(\mathbf{e}) = \mathrm{I}_d \otimes \sigma^2 \mathrm{I}_N$ by definition of the Kronecker product. Therefore, $\mathbf{e} \sim \mathcal{N}(\mathbf{0},\, \mathrm{I}_d \otimes \sigma^2 \mathrm{I}_N)$. In other words, the probability density function for $\mathbf{e}$ is denoted by

$$p(\mathbf{e} \,|\, \sigma^2) = (2\pi)^{-\frac{dN}{2}} \left|\mathrm{I}_d \otimes \sigma^2 \mathrm{I}_N\right|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}\mathbf{e}^T (\mathrm{I}_d \otimes \sigma^2 \mathrm{I}_N)^{-1} \mathbf{e}\right\}. \tag{3.6}$$

Ultimately, to find the likelihood of the data, we rearrange (3.4) in terms of $\mathbf{e}$ and substitute into the exponent of the above. By concentrating solely on this exponent for the time being, this provides

$$\begin{aligned} & \exp\left\{-\frac{1}{2}\mathbf{e}^T (\mathrm{I}_d \otimes \sigma^{-2} \mathrm{I}_N) \mathbf{e}\right\} \\ = & \exp\left\{-\frac{1}{2}\left[\mathbf{y} - (\mathrm{I}_d \otimes X)\mathbf{a}\right]^T (\mathrm{I}_d \otimes \sigma^{-2} \mathrm{I}_N) \left[\mathbf{y} - (\mathrm{I}_d \otimes X)\mathbf{a}\right]\right\} \\ = & \exp\left\{-\frac{1}{2}\left[\mathrm{vec}(Y) - \mathrm{vec}(XA)\right]^T (\mathrm{I}_d \otimes \sigma^{-2} \mathrm{I}_N) \left[\mathrm{vec}(Y) - \mathrm{vec}(XA)\right]\right\} \end{aligned}$$

$$= \exp\left\{-\frac{1}{2}\left[\mathrm{vec}(Y - XA)\right]^T (\mathrm{I}_d \otimes \sigma^{-2}\mathrm{I}_N) \left[\mathrm{vec}(Y - XA)\right]\right\}$$
$$= \exp\left\{-\frac{1}{2}\mathrm{Tr}\left[(Y - XA)^T\sigma^{-2}\mathrm{I}_N(Y - XA)\mathrm{I}_d\right]\right\}$$
$$= \exp\left\{-\frac{1}{2}\mathrm{Tr}\left[\sigma^{-2}g(A)\right]\right\}, \tag{3.7}$$

where, to ease notation, we let $g(A) = (Y - XA)^T(Y - XA)$, a $d \times d$ matrix. In addition, notice, for compatible matrices P, Q and R, the use of the identity $\mathrm{Tr}(P^TQPR) = [\mathrm{vec}(P)]^T(R^T \otimes Q)\mathrm{vec}(P)$ (Henderson and Searle, 1979). Furthermore, when surveying (3.6), we realise that

$$\left|\mathrm{I}_d \otimes \sigma^2\mathrm{I}_N\right|^{-\frac{1}{2}} = \left|\sigma^2(\mathrm{I}_d \otimes \mathrm{I}_N)\right|^{-\frac{1}{2}} = \left[(\sigma^2)^{dN}|\mathrm{I}_d|^N|\mathrm{I}_N|^d\right]^{-\frac{1}{2}} = (\sigma^2)^{-\frac{dN}{2}}.$$

Here, we use the identity that if P is $m \times m$ and Q is $r \times r$, then $|P \otimes Q| = |P|^r\,|Q|^m$ (Muirhead, 1982). Finally then, the probability of the data is given by the expression

$$p(D\,|A,\,\sigma^2) = (2\pi\sigma^2)^{-\frac{dN}{2}} \exp\left\{-\frac{1}{2}\mathrm{Tr}\left[\sigma^{-2}g(A)\right]\right\}. \tag{3.8}$$

3.3.1 Priors

To perform in a Bayesian framework, we specify prior distributions over the parameter set $\boldsymbol{\theta} = \{A,\, \sigma^2\}$. In fact here, a prior is assigned over $\mathbf{a} = \mathrm{vec}(A)$ where $\mathbf{a}$ is a d^2-vector. Note that the use of the vec operator to convert a matrix to a vector is a simple way to define any distribution over a matrix. Thus, we have

$$p(\mathbf{a}) = \mathcal{N}(\mathbf{a}\,|\,\mathbf{0},\, C^*) \tag{3.9}$$
$$p(\sigma^2) = \mathcal{IG}(\sigma^2\,|\,\alpha,\, \beta). \tag{3.10}$$

As was noted in Chapter 2, the above, independent priors seem reasonable as they are the semi-conjugate specifications for a normally distributed random sample with both mean and variance unknown.

At this stage, attention is drawn to C^*. We wish to evaluate $\mathcal{L}(q)$ for different choices of sparsity of A, corresponding to different graphical structures. Furthermore, we must carry such a sparsity choice through the whole problem, implied by the A-matrices. As a result, a prior distribution is chosen on $\operatorname{vec} A$ that imposes the sparsity structure, and which appropriately distinguishes different priors. So, we construct a matrix $C = (c_{ij})$ such that, for each choice of $A = (a_{ij})$ and $\forall i,\ j$,

$$c_{ij} = \begin{cases} c & \text{if } a_{ij} \neq 0 \\ 0 & \text{if } a_{ij} = 0 \end{cases}, \tag{3.11}$$

for some fixed constant c. Accordingly, define $C^* = \operatorname{diag}\{\operatorname{vec}(C)\}$, a natural choice for the covariance matrix of size $d^2 \times d^2$. Hence, the sparsity structure is maintained by constraining C to be of the same form as A, and thus the prior distribution will vary, dependent upon the sparsity structure for each A-matrix. So, we have effectively specified a prior only on the non-zero components of $\mathbf{a}$. Whenever an element of sparse matrix A is equal to zero (*i.e.* not present in the problem), the corresponding variance element of C^* is thus constrained to zero. Moreover, the constant c represents the prior variance of those elements of $\mathbf{a}$ that are present. Each non-zero element is given the equivalent prior variance since we have no extra prior knowledge about the value of one these elements over another.

This construction, for constraining certain prior variance elements to zero, is simple and elegant to apply. In addition, realise that C^* is usually not of full rank, *i.e.* its columns/rows do not form a linearly independent set. This is unless A is a dense matrix, *i.e.* strictly no zero elements, in which case the diagonal elements of C^* are all non-zero. When C^*

is rank deficient, this complicates subsequent analysis in terms of matrix inversion and computing the logarithms of determinants. We shall make further mention of this, and of maintaining sparsity structure, later.

3.3.2 Free form method

Recall that, in the variational approach, we intend to approximate each true posterior by a variational distribution. In this case, two approximate posteriors are considered, namely $q(\mathbf{a}\,|\,D)$ and $q(\sigma^2\,|\,D)$. We continue initially by using the free form method to find the variational posteriors for σ^2 and then $\mathbf{a}$. However, we shall also make use of the fixed form method, as will be seen in due course.

So, by a free form perspective, recollect from Chapter 2 that one and only one assumption is made, which aids the subsequent calculation: that these variational distributions are independent, *i.e.* $q(\mathbf{a},\,\sigma^2\,|\,D) = q(\mathbf{a}\,|\,D)\,q(\sigma^2\,|\,D)$. However, the true posterior $p(\mathbf{a},\,\sigma^2\,|\,D)$ does not factorise in this way, and hence this assumption is an approximation. This is clear as follows:

$$
\begin{aligned}
p(\mathbf{a},\,\sigma^2\,|\,D) &\propto p(D\,|\,A,\,\sigma^2)\,p(\mathbf{a})\,p(\sigma^2)\\
&= (\sigma^2)^{-\frac{dN}{2}}\exp\left\{-\frac{1}{2}\mathrm{Tr}\left[\sigma^{-2}g(A)\right]\right\}\times\exp\left\{-\frac{1}{2}\mathbf{a}^T{C^*}^{-1}\mathbf{a}\right\}\\
&\quad\times(\sigma^2)^{-(\alpha+1)}\exp\left\{-\beta(\sigma^2)^{-1}\right\}\\
&= (\sigma^2)^{-(\alpha+\frac{dN}{2}+1)}\exp\left\{-\frac{(\sigma^2)^{-1}}{2}\mathrm{Tr}\left[g(A)\right]-\frac{1}{2}\mathbf{a}^T{C^*}^{-1}\mathbf{a}-\beta(\sigma^2)^{-1}\right\}.
\end{aligned}
$$

Notice that $g(\cdot)$ is defined only in terms of matrix A, not vector $\mathbf{a}$, hence also the likelihood, (3.8). However, $\mathrm{Tr}\left[g(A)\right]$ can be written in terms of $\mathbf{a}$, as will be seen later. Nevertheless, the term $\frac{(\sigma^2)^{-1}}{2}\mathrm{Tr}\left[g(A)\right]$ implies that this posterior density will not factorise, and hence $\mathbf{a}$ and σ^2 are not *a posteriori* independent.

Consequently, the lower bound in this context is given as

$$\mathcal{L}(q) = \iint q(\mathbf{a}, \sigma^2 \,|\, D) \, \log \left[\frac{p(D \,|\, A, \, \sigma^2) \, p(\mathbf{a}, \, \sigma^2)}{q(\mathbf{a}, \sigma^2 \,|\, D)} \right] \mathrm{d}\mathbf{a} \, \mathrm{d}\sigma^2. \tag{3.12}$$

Due to the independence of both prior and approximating posterior distributions, this can now be rewritten as a sum of integrals:

$$\begin{aligned} \mathcal{L}(q) = \iint & q(\mathbf{a} \,|\, D) q(\sigma^2 \,|\, D) \log p(D \,|\, A, \, \sigma^2) \, \mathrm{d}\mathbf{a} \, \mathrm{d}\sigma^2 + \int q(\mathbf{a} \,|\, D) \log p(\mathbf{a}) \, \mathrm{d}\mathbf{a} \\ & + \int q(\sigma^2 \,|\, D) \, \log p(\sigma^2) \, \mathrm{d}\sigma^2 - \int q(\mathbf{a} \,|\, D) \log q(\mathbf{a} \,|\, D) \, \mathrm{d}\mathbf{a} \\ & - \int q(\sigma^2 \,|\, D) \log q(\sigma^2 \,|\, D) \, \mathrm{d}\sigma^2, \end{aligned} \tag{3.13}$$

integrating out parameters where appropriate. Thus, by writing $\mathcal{L}(q)$ as a functional of $q(\mathbf{a} \,|\, D)$, we derive

$$\mathcal{L}(q) = \int q(\mathbf{a} \,|\, D) \left[\int q(\sigma^2 \,|\, D) \log p(D \,|\, A, \, \sigma^2) \, \mathrm{d}\sigma^2 + \log p(\mathbf{a}) - \log q(\mathbf{a} \,|\, D) \right] \mathrm{d}\mathbf{a} \, + \, \text{const.} \tag{3.14}$$

As a functional of $q(\sigma^2 \,|\, D)$, the lower bound is:

$$\mathcal{L}(q) = \int q(\sigma^2 \,|\, D) \left[\int q(\mathbf{a} \,|\, D) \log p(D \,|\, A, \, \sigma^2) \, \mathrm{d}\mathbf{a} + \log p(\sigma^2) - \log q(\sigma^2 \,|\, D) \right] \mathrm{d}\sigma^2 + \text{const.} \tag{3.15}$$

We inspect both of these equations in turn. Firstly, derive the variational distribution, $q(\sigma^2 \,|\, D)$. By substituting in for both $\log p(D \,|\, A, \, \sigma^2)$ and $\log p(\sigma^2)$ in (3.15), we acquire

$$\begin{aligned} \mathcal{L}(q) = \int q(\sigma^2 \,|\, D) \Bigg[& \int q(\mathbf{a} \,|\, D) \left\{ -\frac{dN}{2} \log 2\pi\sigma^2 - \frac{1}{2} \mathrm{Tr} \left[(\sigma^2)^{-1} g(A) \right] \right\} \mathrm{d}\mathbf{a} \\ & + \alpha \log \beta - \log \Gamma(\alpha) - (\alpha + 1) \log \sigma^2 - \beta (\sigma^2)^{-1} - \log q(\sigma^2 \,|\, D) \Bigg] \mathrm{d}\sigma^2 + \text{const.} \end{aligned}$$

By dropping terms independent of σ^2, a new constant term is formed, and resultantly, we obtain

$$\begin{aligned}\mathcal{L}(q) = \int q(\sigma^2 \,|\, D) \bigg[& -\frac{dN}{2} \log \sigma^2 - \frac{(\sigma^2)^{-1}}{2} \mathrm{E}_{q(\mathbf{a}\,|\,D)} \left\{ \mathrm{Tr}\ [g(A)] \right\} \\ & - (\alpha + 1)\ \log \sigma^2 - \beta (\sigma^2)^{-1} - \log q(\sigma^2 \,|\, D) \bigg] \, \mathrm{d}\sigma^2 + \text{const.}' \end{aligned} \tag{3.16}$$

Our attention is now focussed on the term $\mathrm{E}_{q(\mathbf{a}\,|\,D)} \{\mathrm{Tr}\ [g(A)]\}$. As was commented upon previously, Tr $[g(A)]$ can be denoted in terms of $\mathbf{a}$ and, since here we take the expectation with respect to the variational distribution of vec(A), this would be beneficial. Therefore,

$$\begin{aligned}\mathrm{Tr}\ [g(A)] &= \mathrm{Tr}\ \left[(Y - XA)^T (Y - XA) \right] \\ &= \left[\mathrm{vec}(Y - XA) \right]^T \left[\mathrm{vec}(Y - XA) \right] \\ &= \left[\mathrm{vec}(Y) - \mathrm{vec}(XA) \right]^T \left[\mathrm{vec}(Y) - \mathrm{vec}(XA) \right] \\ &= \left[\mathbf{y} - (\mathrm{I}_d \otimes X)\, \mathbf{a} \right]^T \left[\mathbf{y} - (\mathrm{I}_d \otimes X)\, \mathbf{a} \right] \qquad (3.17) \\ &=: h(\mathbf{a}), \end{aligned}$$

where $\mathbf{y}$, $\mathbf{a}$ are as given previously, and the function h is defined to ease notation. Moreover, we have used the following identity:

$$\mathrm{Tr}(P^T Q) = [\mathrm{vec}(P)]^T \mathrm{vec}(Q) \tag{3.18}$$

for compatible matrices P, Q (Henderson and Searle, 1979). We now take the expectation of (3.17). Multiplying out the brackets and use of (2.51) consequently provides

$$\begin{aligned}\mathrm{E}_{q(\mathbf{a}\,|\,D)} \{\mathrm{Tr}\ [g(A)]\} &= \mathrm{E}_{q(\mathbf{a}\,|\,D)} \left\{ \mathbf{y}^T\mathbf{y} - \mathbf{a}^T (\mathrm{I}_d \otimes X^T)\mathbf{y} - \mathbf{y}^T (\mathrm{I}_d \otimes X)\mathbf{a} + \mathbf{a}^T (\mathrm{I}_d \otimes X^T X)\mathbf{a} \right\} \\ &= \mathbf{y}^T\mathbf{y} - \mathrm{E}_{q(\mathbf{a}\,|\,D)} \left\{ \mathbf{a}^T \right\} (\mathrm{I}_d \otimes X^T)\mathbf{y} - \mathbf{y}^T (\mathrm{I}_d \otimes X) \mathrm{E}_{q(\mathbf{a}\,|\,D)} \{\mathbf{a}\} \\ &\quad + \mathrm{E}_{q(\mathbf{a}\,|\,D)} \left\{ \mathbf{a}^T (\mathrm{I}_d \otimes X^T X)\mathbf{a} \right\} \end{aligned}$$

$$
\begin{aligned}
&= \mathbf{y}^T\mathbf{y} - \boldsymbol{\rho}^T(\mathrm{I}_d \otimes X^T)\mathbf{y} - \mathbf{y}^T(\mathrm{I}_d \otimes X)\boldsymbol{\rho} + \boldsymbol{\rho}^T(\mathrm{I}_d \otimes X^T X)\boldsymbol{\rho} \\
&\qquad + \mathrm{Tr}\left[(\mathrm{I}_d \otimes X^T X)\tau\right] \\
&= h(\boldsymbol{\rho}) + \mathrm{Tr}\left[(\mathrm{I}_d \otimes X^T X)\tau\right], \qquad (3.19)
\end{aligned}
$$

where we define $\boldsymbol{\rho} = \mathrm{E}_{q(\mathbf{a}\,|\,D)}\{\mathbf{a}\}$ and $\tau = \mathrm{Var}_{q(\mathbf{a}\,|\,D)}\{\mathbf{a}\}$. When deriving the variational posterior for $\mathbf{a}$, algebraic forms for $\boldsymbol{\rho}$ and τ will be found. Notice that

$$(P \otimes Q)(R \otimes S) = PR \otimes QS \qquad (3.20)$$

for matrices P compatible with R, Q with S (Harville, 1997).

We can hence substitute (3.19) back into (3.16) to give an expression for $\mathcal{L}(q)$, a functional of $q(\sigma^2\,|\,D)$, which no longer depends upon $\mathbf{a} = \mathrm{vec}(A)$. Now, in accordance with the technique of Chapter 2, a Lagrangian ν_{σ^2} can be used to ensure that the distribution $q(\sigma^2\,|\,D)$ is normalised. Hence, the new functional $\tilde{\mathcal{L}}(q)$ is formed, given by

$$\tilde{\mathcal{L}}(q) = \mathcal{L}(q) + \nu_{\sigma^2}\left(\int q(\sigma^2\,|\,D)\,\mathrm{d}\sigma^2 - 1\right). \qquad (3.21)$$

Thus, we determine the maximum of $\tilde{\mathcal{L}}(q)$ by computing the functional derivative with respect to $q(\sigma^2\,|\,D)$ and setting to zero. This gives

$$
\begin{aligned}
\frac{\partial\tilde{\mathcal{L}}(q)}{\partial q(\sigma^2\,|\,D)} = -\frac{dN}{2}\log\sigma^2 - \frac{(\sigma^2)^{-1}}{2}\left(h(\boldsymbol{\rho}) + \mathrm{Tr}\left[(\mathrm{I}_d \otimes X^T X)\tau\right]\right) \\
- (\alpha+1)\,\log\sigma^2 - \beta(\sigma^2)^{-1} - \log q(\sigma^2\,|\,D) - 1 + \nu_{\sigma^2} = 0.
\end{aligned}
$$

By dropping constant terms and manipulating, we obtain

$$q(\sigma^2\,|\,D) \propto (\sigma^2)^{-(\alpha+\frac{dN}{2}+1)}\,\exp\left\{-(\sigma^2)^{-1}\left(\beta + \frac{1}{2}\left(h(\boldsymbol{\rho}) + \mathrm{Tr}\left[(\mathrm{I}_d \otimes X^T X)\tau\right]\right)\right)\right\}.$$

Hence, it is immediately apparent that

$$q(\sigma^2 \,|\, D) = \mathcal{IG}(\sigma^2 \,|\, \gamma,\, \delta), \tag{3.22}$$

with variational parameters expressed as

$$\gamma = \alpha + \frac{dN}{2} \tag{3.23}$$
$$\delta = \beta + \frac{1}{2}\left(h(\boldsymbol{\rho}) + \mathrm{Tr}\left[(\mathrm{I}_d \otimes X^T X)\tau\right]\right), \tag{3.24}$$

where the function $h(\cdot)$ is defined in (3.17) and $\tau = \mathrm{Var}_{q(\mathbf{a}\,|\,D)}\left\{\mathbf{a}\right\}$. Notice that it is also possible for (3.23) and (3.24) to be derived from the equations (2.53) and (2.54) in Chapter 2.

Now, return to (3.14) and follow the identical procedure to find $q(\mathbf{a}\,|\,D)$. Substituting in for $\log p(D\,|\,A,\,\sigma^2)$ and $\log p(\mathbf{a})$ implies that

$$\begin{aligned}\mathcal{L}(q) = \int q(\mathbf{a}\,|\,D)\Bigg[\int q(\sigma^2\,|\,D)\left\{-\frac{dN}{2}\log 2\pi\sigma^2 - \frac{1}{2}\mathrm{Tr}\left[(\sigma^2)^{-1}g(A)\right]\right\}\,\mathrm{d}\sigma^2 \\ -\frac{d^2}{2}\log 2\pi - \frac{1}{2}\log|C^*| - \frac{1}{2}\mathbf{a}^T {C^*}^{-1}\mathbf{a} - \log q(\mathbf{a}\,|\,D)\Bigg]\,\mathrm{d}\mathbf{a} + \mathrm{const.}\end{aligned}$$

This time, those terms that are independent of $\mathbf{a}$ will disappear under a functional derivative with respect to $q(\mathbf{a}\,|\,D)$. Therefore, we arrive at

$$\begin{aligned}\mathcal{L}(q) = \int q(\mathbf{a}\,|\,D)\Bigg[-\frac{1}{2}\mathrm{Tr}\left[g(A)\right]\mathrm{E}_{q(\sigma^2\,|\,D)}\left\{(\sigma^2)^{-1}\right\} \\ -\frac{1}{2}\mathbf{a}^T {C^*}^{-1}\mathbf{a} - \log q(\mathbf{a}\,|\,D)\Bigg]\,\mathrm{d}\mathbf{a} + \mathrm{const.}'\end{aligned} \tag{3.25}$$

Furthermore, we already know that $\mathrm{E}_{q(\sigma^2\,|\,D)}\left\{(\sigma^2)^{-1}\right\} = \frac{\gamma}{\delta}$. By forming $\tilde{\mathcal{L}}(q)$ with the Lagrange multiplier $\nu_{\mathbf{a}}$ to enforce normality, we maximise this functional, now with respect to $q(\mathbf{a}\,|\,D)$. So, we acquire

$$\frac{\partial \tilde{\mathcal{L}}(q)}{\partial q(\mathbf{a}\,|\,D)} = -\frac{\gamma}{2\delta}\mathrm{Tr}\left[g(A)\right] - \frac{1}{2}\mathbf{a}^T {C^*}^{-1}\mathbf{a} - \log q(\mathbf{a}\,|\,D) - 1 + \nu_{\mathbf{a}} = 0.$$

In a similar fashion to before, this can be rewritten as

$$q(\mathbf{a}\,|\,D) \propto \exp\left\{-\frac{\gamma}{2\delta}\mathrm{Tr}\left[g(A)\right] - \frac{1}{2}\mathbf{a}^T {C^*}^{-1}\mathbf{a}\right\}. \tag{3.26}$$

At this stage, we can express Tr $[g(A)]$ more usefully in terms of $\mathbf{a}$ by using (3.17), as this is the parameter for which we require the variational distribution. This hence removes the dependence of (3.26) on A. In other words, we have

$$\begin{aligned}
q(\mathbf{a}\,|\,D) \propto \exp\Bigg\{ &-\frac{1}{2}\Bigg[\frac{\gamma}{\delta}\mathbf{y}^T\mathbf{y} - \frac{\gamma}{\delta}\mathbf{a}^T(\mathrm{I}_d \otimes X^T)\,\mathbf{y} - \frac{\gamma}{\delta}\mathbf{y}^T(\mathrm{I}_d \otimes X)\mathbf{a} \\
&+ \frac{\gamma}{\delta}\mathbf{a}^T(\mathrm{I}_d \otimes X^T X)\mathbf{a} - \mathbf{a}^T {C^*}^{-1}\mathbf{a}\Bigg]\Bigg\} \\
\propto \exp\Bigg\{ &-\frac{1}{2}\Bigg[\left(\mathbf{a} - \frac{\gamma}{\delta}\left[\frac{\gamma}{\delta}(\mathrm{I}_d \otimes X^T X) + {C^*}^{-1}\right]^{-1}(\mathrm{I}_d \otimes X^T)\mathbf{y}\right)^T \\
&\times \left(\frac{\gamma}{\delta}(\mathrm{I}_d \otimes X^T X) + {C^*}^{-1}\right) \\
&\times \left(\mathbf{a} - \frac{\gamma}{\delta}\left[\frac{\gamma}{\delta}(\mathrm{I}_d \otimes X^T X) + {C^*}^{-1}\right]^{-1}(\mathrm{I}_d \otimes X^T)\mathbf{y}\right)\Bigg]\Bigg\},
\end{aligned}$$

via completing the square. Consequently, the variational posterior of $\mathbf{a}$ is distributed such that

$$q(\mathbf{a}\,|\,D) = \mathcal{N}(\mathbf{a}\,|\,\boldsymbol{\rho},\,\tau), \tag{3.27}$$

and algebraic equations for $\boldsymbol{\rho}$ and τ have been derived such that

$$\boldsymbol{\rho} = \frac{\gamma}{\delta}\left[\frac{\gamma}{\delta}\left(\mathrm{I}_d \otimes X^T X\right) + {C^*}^{-1}\right]^{-1}(\mathrm{I}_d \otimes X^T)\mathbf{y} \tag{3.28}$$

$$\tau = \left[\frac{\gamma}{\delta}\left(\mathrm{I}_d \otimes X^T X\right) + {C^*}^{-1}\right]^{-1}. \tag{3.29}$$

In a way akin to the variational distribution on σ^2, we realise that (3.28) and (3.29) are applications of equations (2.47) and (2.48). At this time, the expression for δ is now fully defined. Optimal solutions for $\{\gamma,\ \delta,\ \boldsymbol{\rho},\ \tau\}$ can be found by iteratively updating these parameter values until convergence, using equations (3.23), (3.24), (3.28) and (3.29).

However, here we realise a problem. The above variational posterior for $\mathbf{a}$ is fine when the matrix A is dense. Yet, we also need to examine the circumstance when A is sparse. Therefore, we must maximise the functional $\tilde{L}(q)$ with respect to some of the elements of $\mathbf{a}$ being zero, dependent on the sparsity structure of each A-matrix. Recall that, when specifying the prior $p(\mathbf{a})$, a matrix C was created to have the identical zero structure of a given A. Consequently, some of the prior variance elements of C^* were necessarily constrained to zero.

Now similarly, for these same entries of A, we are to enforce the corresponding variational posterior mean and variance elements of $\boldsymbol{\rho}$ and τ respectively to be zero. Thus, for each graphical structure, $\boldsymbol{\rho}$ will be of the same form in terms of its dimension and sparsity structure as $\mathbf{a}$, likewise τ with the diagonal matrix C^*. To apply this constraint, a clean and direct solution is now offered.

3.3.3 Fixed form method

In the previous section, we have derived the variational distribution for $\mathbf{a}$, and hence the variational parameters, $\boldsymbol{\rho}$ and τ, in the dense case. However, dealing with a prescribed sparsity structure using a free form approach is difficult. Fortunately, to handle this problem, it turns out to be relatively straightforward to adopt the fixed form variational procedure.

Thus, now suppose that we assume fixed parametric forms for the variational distributions of both $\mathbf{a}$ and σ^2. To effect this, as suggested in Chapter 2, we can simply use the parametric families suggested by the free form method, *i.e.*

$$\begin{aligned} q(\mathbf{a} \,|\, D) &= \mathcal{N}(\mathbf{a} \,|\, \boldsymbol{\rho},\, \tau) \\ q(\sigma^2 \,|\, D) &= \mathcal{IG}(\sigma^2 \,|\, \gamma,\, \delta). \end{aligned}$$

Then, as mentioned previously, the lower bound is derived initially using the known variational distributions. Ordinarily thereafter, we would optimise $\mathcal{L}(q)$ with respect to the variational parameter set. However, update equations for γ and δ, namely (3.23) and (3.24), have erstwhile been computed via the free form approach, and $q(\sigma^2 \,|\, D)$ does not depend upon the sparsity of A. Thus, we need only to consider maximising with respect to $\boldsymbol{\rho}$ and τ. At this point, the sparsity constraint can be enforced.

Consider once again (3.13). With assumed knowledge of the variational posteriors, these integrals can now be computed in turn. Therefore firstly, by (3.8), we obtain

$$\begin{aligned} &\iint q(\mathbf{a} \,|\, D) q(\sigma^2 \,|\, D) \, \log p(D \,|\, A,\, \sigma^2) \, \mathrm{d}\mathbf{a} \, \mathrm{d}\sigma^2 \\ &= \iint q(\mathbf{a} \,|\, D) q(\sigma^2 \,|\, D) \, \left[-\frac{dN}{2} \log 2\pi\sigma^2 - \frac{1}{2} \mathrm{Tr} \left[(\sigma^2)^{-1} g(A) \right] \right] \, \mathrm{d}\mathbf{a} \, \mathrm{d}\sigma^2 \\ &= -\frac{dN}{2} \log 2\pi - \frac{dN}{2} \int q(\sigma^2 \,|\, D) \, \log \sigma^2 \, \mathrm{d}\sigma^2 \\ &\quad - \frac{1}{2} \int q(\mathbf{a} \,|\, D) \mathrm{Tr} \left[g(A) \right] \, \mathrm{d}\mathbf{a} \int q(\sigma^2 \,|\, D) (\sigma^2)^{-1} \, \mathrm{d}\sigma^2 \\ &= -\frac{dN}{2} \log 2\pi - \frac{dN}{2} \left[\log \delta - \psi(\gamma) \right] - \frac{\gamma}{2\delta} \left(h(\boldsymbol{\rho}) + \mathrm{Tr} \left[(\mathrm{I}_d \otimes X^T X) \tau \right] \right). \end{aligned}$$

Furthermore, it is realised that the final line requires the use of (3.19) and (A.6). Moreover, by (2.51),

$$\begin{aligned} &\int q(\mathbf{a} \,|\, D) \, \log p(\mathbf{a}) \, \mathrm{d}\mathbf{a} \\ &= \int q(\mathbf{a} \,|\, D) \left[-\frac{d^2}{2} \log 2\pi \, - \, \frac{1}{2} \log |C^*| \, - \, \frac{1}{2} \mathbf{a}^T {C^*}^{-1} \mathbf{a} \right] \, \mathrm{d}\mathbf{a} \end{aligned}$$

$$\begin{aligned}
&= -\frac{d^2}{2}\log 2\pi - \frac{1}{2}\log|C^*| - \frac{1}{2}\mathrm{E}_{q(\mathbf{a}\,|\,D)}\left\{\mathbf{a}\right\}^T {C^*}^{-1}\mathrm{E}_{q(\mathbf{a}\,|\,D)}\left\{\mathbf{a}\right\} \\
&\quad - \frac{1}{2}\mathrm{Tr}\left[{C^*}^{-1}\mathrm{Var}_{q(\mathbf{a}\,|\,D)}\left\{\mathbf{a}\right\}\right] \\
&= -\frac{d^2}{2}\log 2\pi - \frac{1}{2}\log|C^*| - \frac{1}{2}\boldsymbol{\rho}^T {C^*}^{-1}\boldsymbol{\rho} - \frac{1}{2}\mathrm{Tr}\left[{C^*}^{-1}\tau\right].
\end{aligned}$$

Similarly, using previous results,

$$\begin{aligned}
&\int q(\sigma^2\,|\,D)\,\log p(\sigma^2)\,\mathrm{d}\sigma^2 \\
&= \int q(\sigma^2\,|\,D)\,\left[\alpha\log\beta - \log\Gamma(\alpha) - (\alpha+1)\log\sigma^2 - \beta(\sigma^2)^{-1}\right]\,\mathrm{d}\sigma^2 \\
&= \alpha\log\beta\ -\ \log\Gamma(\alpha) - (\alpha+1)[\log\delta - \psi(\gamma)] - \frac{\beta\gamma}{\delta}.
\end{aligned}$$

In addition, by knowledge of the variational for **a**, we have that

$$\begin{aligned}
&\int q(\mathbf{a}\,|\,D)\,\log q(\mathbf{a}\,|\,D)\,\mathrm{d}\mathbf{a} \\
&= \int q(\mathbf{a}\,|\,D)\,\left[-\frac{d^2}{2}\log 2\pi - \frac{1}{2}\log|\tau| - \frac{1}{2}(\mathbf{a}-\boldsymbol{\rho})^T\tau^{-1}(\mathbf{a}-\boldsymbol{\rho})\right]\,\mathrm{d}\mathbf{a} \\
&= -\frac{d^2}{2}\log 2\pi - \frac{1}{2}\log|\tau| - \frac{1}{2}\,\mathrm{E}_{q(\mathbf{a}\,|\,D)}\left\{[\mathbf{a}-\boldsymbol{\rho}]^T\right\}\tau^{-1}\mathrm{E}_{q(\mathbf{a}\,|\,D)}\left\{\mathbf{a}-\boldsymbol{\rho}\right\} \\
&\quad - \frac{1}{2}\mathrm{Tr}\left[\tau^{-1}\mathrm{Var}_{q(\mathbf{a}\,|\,D)}\left\{\mathbf{a}-\boldsymbol{\rho}\right\}\right] \\
&= -\frac{d^2}{2}\log 2\pi - \frac{1}{2}\log|\tau| - \frac{d^2}{2}.
\end{aligned}$$

Finally, again by (A.6), it is clear that

$$\begin{aligned}
&\int q(\sigma^2\,|\,D)\,\log q(\sigma^2\,|\,D)\,\mathrm{d}\sigma^2 \\
&= \int q(\sigma^2\,|\,D)\,\left[\gamma\log\delta\ \ - \log\Gamma(\gamma) - (\gamma+1)\log\sigma^2 - \delta(\sigma^2)^{-1}\right]\,\mathrm{d}\sigma^2 \\
&= -\log\Gamma(\gamma) - \log\delta + (\gamma+1)\psi(\gamma) - \gamma.
\end{aligned}$$

Therefore, by simplifying all integral computations from (3.13), the lower bound is found to be

$$\begin{aligned}\mathcal{L}(q) = &-\frac{dN}{2}\log 2\pi - \frac{dN}{2}\log\delta + \frac{dN}{2}\psi(\gamma) - \frac{\gamma}{2\delta}\left(h(\boldsymbol{\rho}) + \mathrm{Tr}\left[(\mathrm{I}_d \otimes X^T X)\tau\right]\right) \\ &- \frac{1}{2}\log|C^*| - \frac{1}{2}\boldsymbol{\rho}^T {C^*}^{-1}\boldsymbol{\rho} - \frac{1}{2}\mathrm{Tr}\left[{C^*}^{-1}\tau\right] + \alpha\log\beta - \log\Gamma(\alpha) - \alpha\log\delta \\ &+ \alpha\psi(\gamma) - \frac{\beta\gamma}{\delta} + \frac{1}{2}\log|\tau| + \frac{d^2}{2} + \log\Gamma(\gamma) - \gamma\psi(\gamma) + \gamma. \end{aligned} \tag{3.30}$$

Resultantly, we have calculated $\mathcal{L}(q)$, the quantity required to approximate the log marginal likelihood for each graphical model M_i, using the fixed form method.

Having derived the lower bound, to optimise, the partial differentiation of $\mathcal{L}(q)$ with respect to $\boldsymbol{\rho}$ and τ is now examined. Of course, if $\mathcal{L}(q)$ is maximised with respect to γ and δ, then, by setting to zero and manipulating, we acquire the same iterative equations for these variational parameters, namely (3.23) and (3.24), as in the free form method. However, recall that the sparsity structure only needs to be enforced for the variational distribution $q(\mathbf{a}\,|\,D)$. Thus, first by maximising with respect to $\boldsymbol{\rho}$, it is established that

$$\begin{aligned}\frac{\partial\mathcal{L}(q)}{\partial\boldsymbol{\rho}} &= \frac{\partial}{\partial\boldsymbol{\rho}}\left\{-\frac{\gamma}{2\delta}h(\boldsymbol{\rho}) - \frac{1}{2}\boldsymbol{\rho}^T {C^*}^{-1}\boldsymbol{\rho}\right\} \\ &= \frac{\partial}{\partial\boldsymbol{\rho}}\Bigg\{-\frac{\gamma}{2\delta}\left[\mathbf{y}^T\mathbf{y} - \boldsymbol{\rho}^T(\mathrm{I}_d\otimes X^T)\mathbf{y} - \mathbf{y}^T(\mathrm{I}_d\otimes X)\boldsymbol{\rho} + \boldsymbol{\rho}^T(\mathrm{I}_d\otimes X^T X)\boldsymbol{\rho}\right] \\ &\qquad - \frac{1}{2}\boldsymbol{\rho}^T {C^*}^{-1}\boldsymbol{\rho}\Bigg\} \\ &= \frac{\partial}{\partial\boldsymbol{\rho}}\left\{\boldsymbol{\rho}^T\left[-\frac{\gamma}{2\delta}(\mathrm{I}_d\otimes X^T X) - \frac{1}{2}{C^*}^{-1}\right]\boldsymbol{\rho} + \frac{\gamma}{\delta}\mathbf{y}^T(\mathrm{I}_d\otimes X)\boldsymbol{\rho}\right\},\end{aligned}$$

where $\boldsymbol{\rho}^T(\mathrm{I}_d\otimes X^T)\mathbf{y} = \left[\mathbf{y}^T(\mathrm{I}_d\otimes X)\boldsymbol{\rho}\right]^T = \mathbf{y}^T(\mathrm{I}_d\otimes X)\boldsymbol{\rho}$, as we transpose a scalar quantity.

Maximising this expression *per se* is fairly straightforward, requiring some standard matrix calculus. However, this is a logical point at which we can introduce the constraint of some of the elements of $\boldsymbol{\rho}$ being zero, dependent on the sparsity of each A-matrix, as explained earlier. Thus, we observe the benefit of using the fixed form method in this context. Accordingly, the problem reduces to a *quadratic programming* (QP) problem. Classically, this features the minimisation of a quadratic function, subject to a set of linear constraints. For more details on this topic, see Fletcher (2000). In this case, the constraint is able to be handled more easily and, in fact, the problem is specified as:

$$\max_{\boldsymbol{\rho}} \boldsymbol{\rho}^T H \boldsymbol{\rho} + \mathbf{c}^T \boldsymbol{\rho} \tag{3.31}$$

subject to:

some elements of $\boldsymbol{\rho}$ constrained to be zero,

where

$$\begin{aligned} H &= -\frac{\gamma}{2\delta}\,(\mathrm{I}_d \otimes X^T X) - \frac{1}{2} {C^*}^{-1} \\ \mathbf{c}^T &= \frac{\gamma}{\delta}\,\mathbf{y}^T(\mathrm{I}_d \otimes X). \end{aligned}$$

Suppose that a given A-matrix contains η free, non-zero elements with a prescribed sparsity structure. This structure is moreover inherent within $\boldsymbol{\rho}$ and thus, to solve the QP problem, we subsequently maximise with respect to the non-zero elements of $\boldsymbol{\rho}$. To effect this, we initially must permute rows and columns of (3.31) so that the first η elements of $\boldsymbol{\rho}$ are now these non-zeroes. The corresponding η-vector is defined to be $\boldsymbol{\rho}_1$. Henceforth, we practise in terms of block matrices.

Thus, after permuting rows and columns, define $\boldsymbol{\rho}_{\text{perm}} = \begin{bmatrix} \boldsymbol{\rho}_1 \\ \mathbf{0} \end{bmatrix}$. Moreover, let $H_{\text{perm}} = \begin{bmatrix} \mathbf{H}_{11} & \mathbf{H}_{12} \\ \mathbf{H}_{21} & \mathbf{H}_{22} \end{bmatrix}$, whereby $\mathbf{H}_{11}$ is of dimension $\eta \times \eta$, $\mathbf{H}_{12}$ is $\eta \times (d^2 - \eta)$, $\mathbf{H}_{21}$ $(d^2 - \eta) \times \eta$ and $\mathbf{H}_{22}$ is a $(d^2 - \eta) \times (d^2 - \eta)$ matrix. In other words, $\mathbf{H}_{11}$ is the submatrix obtained by deleting the i-th row and column of H, corresponding to a zero element in the i-th position of $\boldsymbol{\rho}$ for all i. Likewise, $\mathbf{c}_{\text{perm}}^T = \begin{bmatrix} \mathbf{c}_1 \\ \mathbf{c}_2 \end{bmatrix}^T$, where $\mathbf{c}_1$, a vector of dimension η, is the analogous subvector of $\mathbf{c}$, and $\mathbf{c}_2$ is of size $d^2 - \eta$.

Hence, by substituting into (3.31), the problem reduces to

$$\begin{aligned} &\max_{\boldsymbol{\rho}_1} \begin{bmatrix} \boldsymbol{\rho}_1 \\ \mathbf{0} \end{bmatrix}^T \begin{bmatrix} \mathbf{H}_{11} & \mathbf{H}_{12} \\ \mathbf{H}_{21} & \mathbf{H}_{22} \end{bmatrix} \begin{bmatrix} \boldsymbol{\rho}_1 \\ \mathbf{0} \end{bmatrix} + \begin{bmatrix} \mathbf{c}_1 \\ \mathbf{c}_2 \end{bmatrix}^T \begin{bmatrix} \boldsymbol{\rho}_1 \\ \mathbf{0} \end{bmatrix} \\ = &\max_{\boldsymbol{\rho}_1} \boldsymbol{\rho}_1^T \mathbf{H}_{11} \boldsymbol{\rho}_1 + \mathbf{c}_1^T \boldsymbol{\rho}_1. \end{aligned}$$

Optimising this expression with respect to the non-zero vector, $\boldsymbol{\rho}_1$, is now elementary since

$$\begin{aligned} &\frac{\partial}{\partial \boldsymbol{\rho}_1} \left\{ \boldsymbol{\rho}_1^T \mathbf{H}_{11} \boldsymbol{\rho}_1 + \mathbf{c}_1^T \boldsymbol{\rho}_1 \right\} \\ &= (\mathbf{H}_{11} + {\mathbf{H}_{11}}^T) \boldsymbol{\rho}_1 + \mathbf{c}_1 \\ &= 2\mathbf{H}_{11} \boldsymbol{\rho}_1 + \mathbf{c}_1 \\ &= \left[-\frac{\gamma}{\delta} \left(\mathrm{I}_d \otimes X^T X \right) - {C^*}^{-1} \right]_{11} \boldsymbol{\rho}_1 + \left[\frac{\gamma}{\delta} \left(\mathrm{I}_d \otimes X^T \right) \mathbf{y} \right]_1, \end{aligned}$$

using the definitions of H and $\mathbf{c}^T$, in addition to the continuing subscript notation. We apply standard results for the matrix calculus required here (Petersen and Pedersen, 2007). Observe that as H is symmetric, then by only removing rows and corresponding columns, $\mathbf{H}_{11}$ is also a symmetric matrix. Ultimately, setting to zero and solving for $\boldsymbol{\rho}_1$ provides

$$\boldsymbol{\rho}_1 = \frac{\gamma}{\delta} \left(\left[\frac{\gamma}{\delta} \left(\mathrm{I}_d \otimes X^T X \right) + {C^*}^{-1} \right]_{11} \right)^{-1} \left[\left(\mathrm{I}_d \otimes X^T \right) \mathbf{y} \right]_1. \tag{3.32}$$

Finally, it is recognised that once $\boldsymbol{\rho}_1$ has been found, then $\boldsymbol{\rho}$ is reformed by re-introducing the sparsity structure, in accordance with the specific A-matrix. We realise that this reconstruction is needed as, for instance, (3.24) is strictly reliant upon the full vector $\boldsymbol{\rho}$.

In summary, when A is dense, we need only utilise (3.28) to find $\boldsymbol{\rho}$. However, when A is sparse, $\boldsymbol{\rho}_1$ must be computed initially using (3.32), before re-constructing $\boldsymbol{\rho}$. This is easily performed by choosing the correct block of H (*i.e.* $\mathbf{H}_{11}$) and $\mathbf{c}$ (*i.e.* $\mathbf{c}_1$) every time. Thus, notice the obvious analogy between the dense and the sparse case from (3.28) and (3.32).

So, we have constrained, according to the specific sparsity structure of each A-matrix, some elements of the variational posterior mean vector, $\boldsymbol{\rho}$, to be zero. The same operation can now be performed to constrain to zero the corresponding elements of τ. Customarily, when differentiating (3.30) with respect to τ, we would need to find

$$\frac{\partial \mathcal{L}(q)}{\partial \tau} = \frac{\partial}{\partial \tau} \left\{ -\frac{\gamma}{2\delta} \mathrm{Tr}\left[(\mathrm{I}_d \otimes X^T X)\tau \right] - \frac{1}{2} \mathrm{Tr}\left[C^{*^{-1}} \tau \right] + \frac{1}{2} \log |\tau| \right\}.$$

Clearly, in this case, we cannot enforce the sparsity constraint by using simple quadratic programming as was seen with $\boldsymbol{\rho}$. Consequently, an alternative approach is to attempt the problem again in component form, and find an expression for the elements of τ that correspond to those elements of C^*, which have non-zero, prior variance. Obviously, by the sparsity structure, if an element of C^* is constrained to have zero prior variance, then the corresponding element of τ will have zero variational posterior variance.

The prior over σ^2 is maintained to be of the same form as before since σ^2 is unaffected by sparsity. Yet, as C^* is a diagonal matrix, we can use a well-known property of the multivariate Gaussian distribution to denote the prior over $\mathbf{a}$ as a product of independent, univariate Gaussians, *i.e.*

$$p(\mathbf{a}) = \prod_{(p,q) \in I} \mathcal{N}(a_{pq} \,|\, 0,\, C^*_{(p,q)}), \tag{3.33}$$

where I is the set of those elements of $\mathbf{a}$ (corresponding to A) for which $a_{pq} \neq 0$. We then can use a *fixed form* method to proceed. Previously, we assumed that the joint variational posterior, $q(\mathbf{a},\, \sigma^2 \,|\, D)$ could be factorised, an approximation to the true posterior. Now, a further approximation is made at the component level to $p(\mathbf{a},\, \sigma^2 \,|\, D)$, namely that, moreover, the variational distribution for $\mathbf{a}$ can be factorised into a product of univariate Gaussian distributions, *i.e.* we let

$$q(\mathbf{a} \,|\, D) = \prod_{(p,q)\in I} \mathcal{N}(a_{pq} \,|\, \rho_{(p,q)},\, \tau_{(p,q)}), \tag{3.34}$$

where $\rho_{(p,q)}$ is the variational posterior mean that corresponds to element a_{pq} of $\mathbf{a}$, similarly $\tau_{(p,q)}$. We now continue in parallel with the fixed form method in the multivariate case by evaluating (3.13) at the component level.

Initially, the likelihood (3.8) is rewritten in component form. Realising, for an $m \times r$ matrix P, that by Harville (1997),

$$\mathrm{Tr}\left[P^T P\right] = \sum_{j=1}^{m}\sum_{k=1}^{r} p_{jk}^2, \tag{3.35}$$

consequently we acquire, by definition of $g(A)$ and matrix multiplication,

$$\begin{aligned} p(D \,|\, \{a_{ij}\},\, \sigma^2) &= (2\pi\sigma^2)^{-\frac{dN}{2}} \exp\left\{-\frac{(\sigma^2)^{-1}}{2}\mathrm{Tr}\left[(Y - XA)^T(Y - XA)\right]\right\} \\ &= (2\pi\sigma^2)^{-\frac{dN}{2}} \exp\left\{-\frac{(\sigma^2)^{-1}}{2}\sum_{j=1}^{N}\sum_{k=1}^{d}\left([Y - XA]_{jk}\right)^2\right\} \\ &= (2\pi\sigma^2)^{-\frac{dN}{2}} \exp\left\{-\frac{(\sigma^2)^{-1}}{2}\sum_{j=1}^{N}\sum_{k=1}^{d}\left(y_{jk} - \sum_{i=1}^{d} x_{ji}a_{ik}\right)^2\right\}. \end{aligned}$$

Now, examine the first integral of (3.13):

$$\iint q(\mathbf{a} \,|\, D)q(\sigma^2 \,|\, D)\, \log p(D \,|\, \{a_{ij}\},\, \sigma^2)\, \mathrm{d}\mathbf{a}\, \mathrm{d}\sigma^2$$

$$= \iint q(\mathbf{a} \,|\, D) q(\sigma^2 \,|\, D) \left[-\frac{dN}{2} \log 2\pi\sigma^2 - \frac{(\sigma^2)^{-1}}{2} \sum_{j=1}^{N} \sum_{k=1}^{d} \left(y_{jk} - \sum_{i=1}^{d} x_{ji} a_{ik} \right)^2 \right] \mathrm{d}\mathbf{a} \, \mathrm{d}\sigma^2$$

$$= -\frac{dN}{2} \log 2\pi - \frac{dN}{2} \left[\log \delta - \psi(\gamma) \right] - \frac{\gamma}{2\delta} \sum_{j=1}^{N} \sum_{k=1}^{d} \int q(\mathbf{a} \,|\, D) \left(y_{jk} - \sum_{i=1}^{d} x_{ji} a_{ik} \right)^2 \mathrm{d}\mathbf{a}, \tag{3.36}$$

using (A.5) and (A.6). The final integral in (3.36) is equivalent to

$$\mathrm{E}_{q(\mathbf{a} \,|\, D)} \left\{ \left(y_{jk} - \sum_{i=1}^{d} x_{ji} a_{ik} \right)^2 \right\}$$

$$= y_{jk}^2 - 2 y_{jk} \mathrm{E}_{q(\mathbf{a} \,|\, D)} \left\{ \sum_{i=1}^{d} x_{ji} a_{ik} \right\} + \mathrm{E}_{q(\mathbf{a} \,|\, D)} \left\{ \left[\sum_{i=1}^{d} x_{ji} a_{ik} \right]^2 \right\}$$

$$= y_{jk}^2 - 2 y_{jk} \sum_{i=1}^{d} x_{ji} \mathrm{E}_{q(\mathbf{a} \,|\, D)} \{ a_{ik} \} + \sum_{i=1}^{d} \mathrm{Var}_{q(\mathbf{a} \,|\, D)} \{ x_{ji} a_{ik} \} + \left[\sum_{i=1}^{d} x_{ji} \mathrm{E}_{q(\mathbf{a} \,|\, D)} \{ a_{ik} \} \right]^2$$

$$= y_{jk}^2 - 2 y_{jk} \sum_{i=1}^{d} x_{ji} \rho_{(i,k)} + \sum_{i=1}^{d} x_{ji}^2 \tau_{(i,k)} + \left[\sum_{i=1}^{d} x_{ji} \rho_{(i,k)} \right]^2 ,$$

noting that, as τ is diagonal, the covariance between elements of $\mathbf{a}$ with respect to $q(\mathbf{a} \,|\, D)$ is always zero. Furthermore, we obtain, by (3.33), that

$$\int q(\mathbf{a} \,|\, D) \log p(\mathbf{a}) \, \mathrm{d}\mathbf{a}$$

$$= \sum_{(p,q) \in I} \int q(\mathbf{a} \,|\, D) \left[-\frac{1}{2} \log 2\pi - \frac{1}{2} \log C^*_{(p,q)} - \frac{1}{2} \frac{a_{pq}^2}{C^*_{(p,q)}} \right] \mathrm{d}\mathbf{a}$$

$$= - \sum_{(p,q) \in I} \left[\frac{1}{2} \log 2\pi + \frac{1}{2} \log C^*_{(p,q)} + \frac{1}{2 C^*_{(p,q)}} \left(\mathrm{Var}_{q(\mathbf{a} \,|\, D)} \{ a_{pq} \} + \left[\mathrm{E}_{q(\mathbf{a} \,|\, D)} \{ a_{pq} \} \right]^2 \right) \right]$$

$$= - \sum_{(p,q) \in I} \left[\frac{1}{2} \log 2\pi + \frac{1}{2} \log C^*_{(p,q)} + \frac{1}{2} \frac{\tau_{(p,q)}}{C^*_{(p,q)}} + \frac{1}{2} \frac{\rho^2_{(p,q)}}{C^*_{(p,q)}} \right].$$

Moreover, by (3.34),

$$\begin{aligned}
&\int q(\mathbf{a} \,|\, D) \log q(\mathbf{a} \,|\, D) \,\mathrm{d}\mathbf{a} \\
&= \sum_{(p,q)\in I} \int q(\mathbf{a} \,|\, D) \left[-\frac{1}{2}\log 2\pi - \frac{1}{2}\log \tau_{(p,q)} - \frac{1}{2}\frac{(a_{pq}-\rho_{(p,q)})^2}{\tau_{(p,q)}} \right] \mathrm{d}\mathbf{a} \\
&= -\sum_{(p,q)\in I} \left[\frac{1}{2}\log 2\pi + \frac{1}{2}\log \tau_{(p,q)} \right. \\
&\qquad \left. + \frac{1}{2\tau_{(p,q)}} \left(\left[\mathrm{E}_{q(\mathbf{a} \,|\, D)}\{a_{pq}-\rho_{(p,q)}\} \right]^2 + \mathrm{Var}_{q(\mathbf{a} \,|\, D)}\{a_{pq}-\rho_{(p,q)}\} \right) \right] \\
&= -\sum_{(p,q)\in I} \left[\frac{1}{2}\log 2\pi + \frac{1}{2}\log \tau_{(p,q)} + \frac{1}{2} \right].
\end{aligned}$$

The two other integrals of (3.13) are independent of $\mathbf{a}$, and so are calculated as previous. Therefore, in component form, the lower bound is equivalent to

$$\begin{aligned}
\mathcal{L}(q) = &-\frac{dN}{2}\log 2\pi - \frac{dN}{2}\left[\log\delta - \psi(\gamma)\right] - \frac{\gamma}{2\delta}\sum_{j=1}^{N}\sum_{k=1}^{d} y_{jk}^2 + \frac{\gamma}{\delta}\sum_{i=1}^{d}\sum_{j=1}^{N}\sum_{k=1}^{d} y_{jk}x_{ji}\rho_{(i,k)} \\
&- \frac{\gamma}{2\delta}\sum_{i=1}^{d}\sum_{j=1}^{N}\sum_{k=1}^{d} x_{ji}^2\tau_{(i,k)} - \frac{\gamma}{2\delta}\sum_{j=1}^{N}\sum_{k=1}^{d}\left[\sum_{i=1}^{d} x_{ji}\rho_{(i,k)}\right]^2 - \frac{1}{2}\sum_{(p,q)\in I}\log C^*_{(p,q)} \\
&- \frac{1}{2}\sum_{(p,q)\in I}\frac{\tau_{(p,q)}}{C^*_{(p,q)}} - \frac{1}{2}\sum_{(p,q)\in I}\frac{\rho^2_{(p,q)}}{C^*_{(p,q)}} + \alpha\log\beta - \log\Gamma(\alpha) - \alpha\log\delta + \alpha\psi(\gamma) \\
&- \frac{\beta\gamma}{\delta} + \frac{1}{2}\sum_{(p,q)\in I}\log\tau_{(p,q)} + \sum_{(p,q)\in I}\frac{1}{2} + \log\Gamma(\gamma) - \gamma\psi(\gamma) + \gamma. \qquad (3.37)
\end{aligned}$$

Finally, we optimise $\mathcal{L}(q)$ by differentiating with respect to the component $\tau_{(p,q)}$ to obtain

$$\begin{aligned}
\frac{\partial\mathcal{L}(q)}{\partial\tau_{(p,q)}} &= \frac{\partial}{\partial\tau_{(p,q)}}\left\{ -\frac{\gamma}{2\delta}\sum_{i=1}^{d}\sum_{j=1}^{N}\sum_{k=1}^{d} x_{ji}^2\tau_{(i,k)} - \frac{1}{2}\sum_{(p,q)\in I}\frac{\tau_{(p,q)}}{C^*_{(p,q)}} + \frac{1}{2}\sum_{(p,q)\in I}\log\tau_{(p,q)} \right\} \\
&= -\frac{\gamma}{2\delta}\sum_{j=1}^{N} x_{jp}^2 - \frac{1}{2C^*_{(p,q)}} + \frac{1}{2\tau_{(p,q)}}.
\end{aligned}$$

By equating to zero and solving for $\tau_{(p,q)}$, we have that each non-zero diagonal element of τ is given by

$$\tau_{(p,q)} = \left(\frac{1}{C^*_{(p,q)}} + \frac{\gamma}{\delta} \sum_{j=1}^{N} x_{jp}^2 \right)^{-1} . \tag{3.38}$$

Consequently, we can express the diagonal elements of τ such that

$$\tau_{(p,q)} = \begin{cases} \left(\dfrac{1}{C^*_{(p,q)}} + \dfrac{\gamma}{\delta} \displaystyle\sum_{j=1}^{N} x_{jp}^2 \right)^{-1} & \text{if } a_{pq} \neq 0 \\ 0 & \text{if } a_{pq} = 0 \end{cases} .$$

Clearly, by differentiating (3.37) with respect to γ and δ, we reach the same update equations as before, but in component form. A genuine question to ask at this stage would be why not use this method to find an expression for any $\rho_{(p,q)}$, corresponding to a non-zero element a_{pq}, rather than the quadratic programming method, as examined earlier. In this case, maximising with respect to $\rho_{(p,q)}$, we acquire

$$\begin{aligned} \frac{\partial \mathcal{L}(q)}{\partial \rho_{(p,q)}} &= \frac{\partial}{\partial \rho_{(p,q)}} \Bigg\{ \frac{\gamma}{\delta} \sum_{i=1}^{d} \sum_{j=1}^{N} \sum_{k=1}^{d} y_{jk} x_{ji} \rho_{(i,k)} - \frac{\gamma}{2\delta} \sum_{j=1}^{N} \sum_{k=1}^{d} \left[\sum_{i=1}^{d} x_{ji} \rho_{(i,k)} \right]^2 \\ &\qquad - \frac{1}{2} \sum_{(p,q) \in I} \frac{\rho^2_{(p,q)}}{C^*_{(p,q)}} \Bigg\} \\ &= \frac{\gamma}{\delta} \sum_{j=1}^{N} y_{jq} x_{jp} - \frac{\gamma}{2\delta} \sum_{j=1}^{N} \sum_{t=1}^{d} x_{jp} x_{jt} \rho_{(t,q)} - \frac{\rho_{(p,q)}}{C^*_{(p,q)}}, \end{aligned}$$

where we note that $\sum_{j=1}^{N} \sum_{k=1}^{d} \left[\sum_{i=1}^{d} x_{ji} \rho_{(i,k)} \right]^2 = \sum_{j=1}^{N} \sum_{k=1}^{d} \sum_{t=1}^{d} \sum_{i=1}^{d} x_{ji} x_{jt} \rho_{(i,k)} \rho_{(t,k)}$. However, by equating to zero and solving for $\rho_{(p,q)}$, we realise that the expression is not independent of $\sum_{t=1}^{d} \rho_{(t,q)}$. Hence, the quadratic programming method to constrain $\boldsymbol{\rho}$ according to sparsity is preferred.

To summarise, we have derived the distributional form for both variational posteriors, namely $q(\mathbf{a} \,|\, D)$ and $q(\sigma^2 \,|\, D)$, together with update equations for the set of variational parameters, initially γ and δ, then $\boldsymbol{\rho}$ (also $\boldsymbol{\rho}_1$) and τ (using $\tau_{(p,q)}$). These update equations are run until convergence, hence finding the parameter values for our variational distributions. At each iteration for every graph, we can evaluate the lower bound (3.30) until convergence, thereupon giving a good approximation of the log marginal likelihood. This provides the evidence needed to rank the graphical structures available from the candidate set of models.

3.4 Other issues

3.4.1 Problems with computation

We comment briefly upon the computational issues of the matrices C^* and τ. Notice that in (3.30), the expression for $\mathcal{L}(q)$, we must compute $\log|C^*|$, $\log|\tau|$ and ${C^*}^{-1}$. Yet, these afore-mentioned matrices by construction have all off-diagonal entries equal to zero and, unless A is dense, contain some zero elements on the leading diagonal. Thus, when A is sparse, the determinant of C^* and τ will typically be zero and, hence, the logarithm of the determinant is undefined. Moreover by definition, both matrices will be consequently singular, implying a problem in calculating the inverse of C^*.

However, these dilemmas can be overcome. To understand this, suppose that $\mathbf{X}$ is a random vector that follows a $\mathcal{N}_r(\mathbf{m},\, V)$ distribution, where the subscript r is used to emphasise the dimension of $\mathbf{X}$. If V is singular (with rank $k < r$), then the standard multivariate normal density function does not exist on $\mathbb{R}^r$. However, it does exist on a k-dimensional subspace of $\mathbb{R}^r$ where the distribution has support. In addition, the density of $\mathbf{X}$ on this subspace is defined by Rao and Mitra (1972) as

$$p(\mathbf{x}\,|\,\mathbf{m},\, V) = \frac{(2\pi)^{-k/2}}{(\lambda_1 \cdots \lambda_k)^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\mathbf{m})^T V^{-}(\mathbf{x}-\mathbf{m})\right\}, \qquad (3.39)$$

where $\lambda_1, \ldots, \lambda_k$ are the non-zero eigenvalues of V, and V^- is any generalised inverse of V (see Appendix C). Thus, we refer to (3.39) as the density of a singular $\mathcal{N}_r(\mathbf{m}, V)$ distribution of rank k.

Of course, there is a clear correspondence between (3.39) and its non-singular counterpart, (A.7). That is, the singular density may be computed on the subspace by alternative calculation of the determinant and inverse of V in the density of full rank. We use this relationship to justify the following analysis. If a given A-matrix has η non-zero elements as before, then by construction, C^*, and consequently τ, will have rank $\eta < d^2$ (recall that the rank of any diagonal matrix is equivalent to the number of non-zero diagonal elements that it possesses). So, analogous to (3.39) where V was again a singular, positive semidefinite matrix, the determinant of C^* (and hence τ) can also be calculated as the product of its non-zero eigenvalues (Neudecker, 1995), namely its η, non-zero diagonal entries. If $c^*_{11}, \ldots, c^*_{\eta\eta}$ are these elements, then resultantly, by taking logarithms, we easily have

$$\log |C^*| = \sum_{i=1}^{\eta} \log c^*_{ii}. \tag{3.40}$$

The case is similar for τ.

Additionally, if it were non-singular, the inverse of C^* would trivially be the matrix with the diagonal elements of C^* replaced by their reciprocals. Yet, when A is sparse, a generalised inverse can be utilised for this procedure, as seen in (3.39). Moreover, a generalised inverse of a diagonal matrix is formed by reciprocating only the *non-zero*, diagonal entries (Harville, 1997). In fact, this is the Moore-Penrose inverse of the matrix, denoted in this case by ${C^*}^+$ — additional details are again provided in Appendix C. Consequently, in (3.30), the lower bound expression is altered so that ${C^*}^{-1}$ is replaced by ${C^*}^+$.

Finally, the lower bound contains terms that remain constant across different models.

Thus, we can rewrite (3.30) as

$$\begin{aligned}\mathcal{L}(q) \propto -\frac{dN}{2}\log\delta - \frac{\gamma}{2\delta}\left(h(\boldsymbol{\rho}) + \mathrm{Tr}\left[(\mathrm{I}_d \otimes X^T X)\tau\right]\right) - \frac{1}{2}\log|C^*| \\ - \frac{1}{2}\boldsymbol{\rho}^T {C^*}^+ \boldsymbol{\rho} - \frac{1}{2}\mathrm{Tr}\left[{C^*}^+ \tau\right] - \alpha\log\delta - \frac{\beta\gamma}{\delta} + \frac{1}{2}\log|\tau|,\end{aligned} \tag{3.41}$$

and hence use the above expression to provide evidence for the competing, candidate models.

3.4.2 Specification of priors

We discuss specification of the parameter values for both prior distributions included in the model. Such issues were succinctly touched upon in Section 2.5.5. In the case of $p(\sigma^2)$, it is customary to want the prior to have little influence over the resulting approximate posterior distribution. Thus, we aim to use a vague prior. A popular choice is to apply the relatively flat, proper prior $\mathcal{IG}(\epsilon,\, \epsilon)$ for low values of ϵ and, in particular, when $\epsilon = 0.001$ (Spiegelhalter et al., 1995). By Appendix A, any $\mathcal{IG}(a,\, b)$ distribution is defined only when $a,\, b > 0$, and so this specification is thus considered to be a 'just' proper prior. As mentioned previously, the $\mathcal{IG}(0.001,\, 0.001)$ distribution is described as vague since it has very large (in this case, infinite) variance.

However, it has been seen, in the context of hierarchical models, that the resultant posterior distribution can be highly sensitive to the choice of ϵ, when the variance parameter in question is estimated to be small *a posteriori* (Gelman, 2006). In this case, the author showed that, in particular for one dataset, the prior was perversely not at all vague. We realise that this prior is highly peaked for small σ^2, and so may show a preference for lower values of σ^2 in the Bayesian update. Perhaps a more attractive choice would be to use a $\mathcal{IG}(1,\, \epsilon)$ density for $\epsilon \to 0$. When $\epsilon = 0.001$ say, this prior has a maximum around $\sigma^2 = 0$ as before. Yet, this peak is now extremely sharp. As such, the density reaches negligible values quicker than before, implying that the prior is then flatter.

Customarily, for the choice of a Gaussian prior, we again represent prior ignorance by choosing the distribution to have large variance. However, when choosing between models, this policy may be susceptible to *Lindley's paradox*. Introduced initially in Chapter 1, we now explain the paradox, following that of Shafer (1982). Suppose a random quantity X follows a Gaussian distribution with unknown mean μ and known variance ω^2. On the basis of observing a dataset D of size n with sample mean $\bar{x}$, we want to evaluate the evidence for two models, which may have given rise to the data. These are:

$$M_k : X \sim \mathcal{N}(\mu_0, \omega^2)$$
$$M_l : X \sim \mathcal{N}(\mu, \omega^2).$$

According to M_l, we place a diffuse prior over μ, centred at μ_0, *e.g.* $\mu \sim \mathcal{N}(\mu_0, d^2)$ for large d. We then can compute the Bayes' factor for model M_k against M_l, as illustrated in Chapter 1:

$$B_{kl} = \frac{p(D \,|\, M_k)}{p(D \,|\, M_l)}.$$

Yet, as Shafer illustrates, by allowing the prior variance d^2 to become sufficiently large, the Bayes' factor, in turn, will become significantly greater than 1, and hence, M_k will be favoured always ahead of M_l. Here, we realise that model M_l can be written as $X \sim \mathcal{N}(\mu_0, d^2 + \omega^2)$. Thus, with increasing d, M_l becomes more complex and B_{kl} works in favour of the simpler M_k. The paradox arises since a standard, hypothesis test, in particular $Z = \dfrac{\bar{x} - \mu_0}{\omega/\sqrt{n}}$, may show strong evidence against model M_k, whilst simultaneously, the Bayesian assessment can display exactly the reverse conclusion. In other words, even if the sample mean of the data is significantly different from μ_0, then, by making the prior as flat as necessary, this will override the evidence, provided by the data, and hence suggest that $\mu = \mu_0$. This conflicts with the interpretation of the Bayes' factor as being the odds of the two models, implied solely by the data. For further discussion, see, for instance, Robert (1993), Berger and Sellke (1987) and Aitkin (1991).

The converse of the paradox is the scenario when the prior variance is allowed to approach zero. Then, the prior, now highly informative for μ, will be tightly-peaked around μ_0, implying that those values of μ, in the vicinity of μ_0, will be given very high prior probability. If the sample mean $\bar{x}$ is reasonably close to the value of μ_0, then these same values will also possess high likelihood. Define $L(\mu \mid D, M_l) = p(D \mid \mu, M_l)$ as the likelihood function, specific to M_l. Hence, the marginal likelihood for this model, namely $p(D \mid M_l) = \displaystyle\int L(\mu \mid D, M_l)\, p(\mu \mid M_l)\, \mathrm{d}\mu$, denoting the average of the likelihood with respect to the prior, will increase, hence reducing the Bayes' factor to below 1, and thus favouring M_l.

In the above example of Lindley's paradox, we can consider M_k to be the simpler model and M_l the more complex model. In effect, we choose a model either with mean equal to μ_0 or with unknown mean μ. This situation corresponds to the current, zero mean VAR(1) case if we take $\mu_0 = 0$. Then, we choose each element of the $d \times d$ matrix A to be either a zero or a free entry, whereby the latter follows a prior distribution of the form given by (3.33). That is, in each case, we again examine a simpler or more complex model respectively. When specifying the prior, it seems reasonable to take the mean as zero, hence the prior distribution is centred around the simpler model for each component.

By carefully specifying a value not too small for the prior variance on each non-zero element of $\mathbf{a}$, we will be able to penalise more complex A-matrices in the candidate set, *i.e.* those with more unspecified, non-zero entries. Although such models with more parameters will be better at fitting the data, by penalising, it will enable us to choose the model with optimum structure. On the other hand, the use of a diffuse prior on $\mathbf{a}$ may lead to favouring an A-matrix of simpler structure. Hence, there is a justification to use a deliberately chosen informative prior so that the problem is not susceptible to Lindley's paradox. Moreover, the specification of c from (3.11) must be a compromise between always favouring a simpler A-matrix (large c) and a more complex matrix (small c). A full prior sensitivity analysis, examining these issues, is provided in the next chapter.

3.5 Toy example

To illustrate the above procedure, we consider a simple toy example, based upon an arbitrary, zero mean VAR(1) model. In previous work, we have specified distributions over $\mathbf{a} = \operatorname{vec} A$ and found that $\mathrm{E}_{q(\mathbf{a}\,|\,D)}(\mathbf{a}) = \boldsymbol{\rho}$. In the following example, we unstack the d^2-vector $\boldsymbol{\rho}$ to form a $d \times d$ matrix called $\hat{A}$. This procedure is illustrated further in Section 5.2.2.

An easy case is supposed whereby $d = 2$. A dataset of size $N = 250$ was simulated from the zero mean VAR(1) model (3.2) with specifications such that $A = \begin{pmatrix} 0 & 0.7 \\ 0.3 & 0 \end{pmatrix}$ and $\sigma^2 = 0.1$. We choose A with care to ensure that all its eigenvalues have modulus less than 1. In this case, the VAR process is said to be *stable* and, hence, the dataset does not explode for increasing N (Lütkepohl, 2005). As mentioned previously, we again let $\mathbf{x}_1 = (0,\, 0)$ and $\mathbf{x}_{250} = \mathbf{y}_{249}$. This choice of A thus defines the A-graph below.

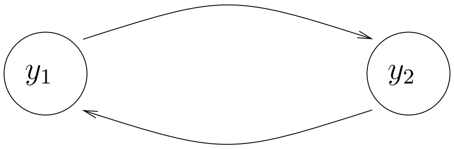


Figure 3.3: A-graph for the true zero mean VAR(1) model

When $d = 2$, there are $2^4 = 16$ directed A-graphs on two vertices. However, the null model, represented by a completely sparse graph, is ignored. In such a case, C^* and τ would both be zero matrices, implying that taking the logarithm of their determinants, as required in the computation of $\mathcal{L}_{M_i}(q_i)$, would be undefined (*c.f.* (3.40)). As a consequence, we construct a candidate set of 15 graphs, referred to in terms of their corresponding A-matrices. Thus, a zero element in a given matrix implies no edge between the corresponding vertices on the graph, as explained earlier. So, given the data, we aim to select the optimum model from the set. The prior distributions over $\mathbf{a} = \text{vec}(A)$ and σ^2 were specified to be

$$\begin{aligned} p(\mathbf{a}) &= \mathcal{N}(\mathbf{a} \,|\, \mathbf{0},\, C^*) \;\text{ where }\; c_{ij} \in \{0,\, 0.5\} \\ p(\sigma^2) &= \mathcal{IG}(\sigma^2 \,|\, 1,\, 0.001). \end{aligned}$$

Clearly, the prior on σ^2 remains the same throughout whereas, for that on $\mathbf{a}$, the covariance matrix C^* changes according to the sparsity structure of each A-matrix.

Recall that we wish to find parameter values for variational distributions given by

$$\begin{aligned} q(\mathbf{a} \,|\, D) &= \mathcal{N}(\mathbf{a} \,|\, \boldsymbol{\rho},\, \tau) \\ q(\sigma^2 \,|\, D) &= \mathcal{IG}(\sigma^2 \,|\, \gamma,\, \delta). \end{aligned}$$

Consequently, for each candidate model, the lower bound, $\mathcal{L}_{M_i}(q_i)$, was evaluated at each iteration, and update equations for the variational parameters were run until convergence. For this medium-sized dataset, this took merely 4 iterations in each case, identical to what was seen in the example of Section 2.5.5. As the update equations for $\boldsymbol{\rho}_1$ and $\tau_{(p,q)}$ (hence defining $\boldsymbol{\rho}$ and τ) depend upon γ and δ, the algorithm was initialised arbitrarily with $\gamma = \delta = 1$.

The results of the example are shown in Table 3.1. Initially, we report back the log marginal likelihood estimation by $\mathcal{L}_{M_i}(q_i)$. As we would hoped, the model which relates to the true choice of A, in terms of sparsity structure, was chosen. The more complex models with A-matrices that contained at least the two, true free elements were also well-favoured. Recall that the prior parameters for $\mathbf{a}$ were chosen to avoid Lindley's paradox. The specification made appears to be a valid one since neither the simpler, nor more complex models are favoured, ahead of the truth.

Now, we examine the estimates of the true A and σ^2, namely $\hat{A}$ and $\mathrm{E}_{q(\sigma^2 \,|\, D)}\{\sigma^2\}$ respectively. Notice that all the non-zero elements of each $\hat{A}$ are very similar to each other, but not mathematically exact. This stems from the variational algorithm and, in particular, the definition of $\boldsymbol{\rho}_1$. For each candidate model, H_{11} and $\mathbf{c}_1$ will vary in dimension and value of elements, dependent on the sparsity structure, hence creating such a difference. Moreover, C^* will also change from model to model. With a reasonably-sized dataset chosen, each $\hat{A}$-matrix is similar to the original choice of A. This is since, with $N = 250$, the likelihood term dominates the prior distribution. That is, with a prior chosen so that no favouritism exists for either simpler or more complex models, most of the information about $\mathbf{a}$ was passed through to the variational distribution $q(\mathbf{a} \,|\, D)$, in this approximate Bayesian update, via the data.

The corresponding estimates for σ^2, found using (A.4), are all reasonably accurate to the true value. However, there is less concordance between these values than the $\hat{A}$-matrices. It is apparent that any candidate that possessed a sparsity structure similar to the truth produced better estimates of σ^2 than the remaining models. We are already aware that γ is constant across models (*c.f.* (3.23)). Thus, if the wrong model is chosen, the value of δ has become more inaccurate when compared with that for the true model, hence $\mathrm{E}_{q(\sigma^2 \,|\, D)}\{\sigma^2\}$ also. Suppose we attempt to fit a model with the wrong sparsity pattern to the given set of data. Then, the noise variance, which determines the extent to which the data fluctuates about the mean of the process (in this case, zero), must be readjusted to cope with this model misspecification. That is, the model error, created by attempting to model the data incorrectly, is 'pushed' exclusively into the estimate of σ^2.

Specification	Posterior means		$\mathcal{L}_{M_i}(q_i)$
A-matrix	$\hat{A}$-matrix	$\mathrm{E}_{q(\sigma^2 \mid D)}\{\sigma^2\}$	
$\begin{pmatrix} \ast & 0 \\ 0 & 0 \end{pmatrix}$	$\begin{pmatrix} -0.066 & 0 \\ 0 & 0 \end{pmatrix}$	0.135	-1136.791
$\begin{pmatrix} 0 & 0 \\ 0 & \ast \end{pmatrix}$	$\begin{pmatrix} 0 & 0 \\ 0 & -0.041 \end{pmatrix}$	0.135	-1137.165
$\begin{pmatrix} 0 & \ast \\ 0 & 0 \end{pmatrix}$	$\begin{pmatrix} 0 & 0.674 \\ 0 & 0 \end{pmatrix}$	0.109	-1084.561
$\begin{pmatrix} 0 & 0 \\ \ast & 0 \end{pmatrix}$	$\begin{pmatrix} 0 & 0 \\ 0.341 & 0 \end{pmatrix}$	0.126	-1119.862
$\begin{pmatrix} \ast & \ast \\ 0 & 0 \end{pmatrix}$	$\begin{pmatrix} -0.066 & 0.674 \\ 0 & 0 \end{pmatrix}$	0.109	-1086.926
$\begin{pmatrix} \ast & 0 \\ \ast & 0 \end{pmatrix}$	$\begin{pmatrix} -0.074 & 0 \\ 0.342 & 0 \end{pmatrix}$	0.126	-1122.113
$\begin{pmatrix} 0 & \ast \\ 0 & \ast \end{pmatrix}$	$\begin{pmatrix} 0 & 0.676 \\ 0 & -0.052 \end{pmatrix}$	0.110	-1087.169
$\begin{pmatrix} 0 & 0 \\ \ast & \ast \end{pmatrix}$	$\begin{pmatrix} 0 & 0 \\ 0.341 & -0.041 \end{pmatrix}$	0.126	-1122.622
$\begin{pmatrix} \ast & 0 \\ 0 & \ast \end{pmatrix}$	$\begin{pmatrix} -0.066 & 0 \\ 0 & -0.041 \end{pmatrix}$	0.135	-1139.534
$\begin{pmatrix} 0 & \ast \\ \ast & 0 \end{pmatrix}$	$\begin{pmatrix} 0 & 0.675 \\ 0.341 & 0 \end{pmatrix}$	0.100	-1065.786
$\begin{pmatrix} \ast & \ast \\ 0 & \ast \end{pmatrix}$	$\begin{pmatrix} -0.066 & 0.676 \\ 0 & -0.052 \end{pmatrix}$	0.109	-1089.534
$\begin{pmatrix} \ast & 0 \\ \ast & \ast \end{pmatrix}$	$\begin{pmatrix} -0.074 & 0 \\ 0.342 & -0.041 \end{pmatrix}$	0.126	-1124.873
$\begin{pmatrix} \ast & \ast \\ \ast & 0 \end{pmatrix}$	$\begin{pmatrix} -0.074 & 0.675 \\ 0.343 & 0 \end{pmatrix}$	0.100	-1067.992
$\begin{pmatrix} 0 & \ast \\ \ast & \ast \end{pmatrix}$	$\begin{pmatrix} 0 & 0.676 \\ 0.341 & -0.052 \end{pmatrix}$	0.100	-1068.393
$\begin{pmatrix} \ast & \ast \\ \ast & \ast \end{pmatrix}$	$\begin{pmatrix} -0.074 & 0.676 \\ 0.343 & -0.052 \end{pmatrix}$	0.100	-1070.599

Table 3.1: Lower bounds and posterior means for each zero mean VAR(1) model

Chapter 4

Searching the graphical space

4.1 Motivation

In the previous chapter, we constructed a candidate set of individual VAR(1) models, each dependent on the sparsity structure of an A-matrix and represented by an A-graph. Given a set of observed data, these sparse models were able to be scored by evaluating, in each case, $\mathcal{L}_{M_i}(q_i)$, a lower bound approximation to the logarithm of the marginal likelihood, derived using the variational Bayesian method. Thus, we were able to select the most plausible models from the set. At present, this approach is applicable, but, conversely, rather limited.

This is because we can only compute the lower bound, or variational score, for a set of predetermined graphs individually. As a consequence, it becomes a computational impossibility to consider all the candidate models within the graphical space in this way, when the dimension of the VAR(1) models, d, increases. In fact, explicitly, the number of possible candidate models, each represented by a graph on d nodes, is $2^{d^2} - 1$, excluding the null model. Obviously, for large d, the task of computing a lower bound for each candidate individually is somewhat arduous! It would hence be more beneficial if the whole process was fully automated, and thus we possessed an efficient way to traverse through the graphical space quickly to find high scoring graphs.

In this chapter, we consider such an automated system. In particular, two such methods are developed. Initially, a customary hill-climbing algorithm is contemplated. In this circumstance, we are able to manoeuvre through the graphical space by comparing the values of two lower bounds, and accepting the A-graph that effects the higher $\mathcal{L}(q)$, with probability 1. The alternative is to make a random walk across the space so that moves to neighbouring graphs are reliant upon Markov chain Monte Carlo (MCMC) techniques and, in particular, the Metropolis-Hastings algorithm (previously documented in Chapter 2). Thus, we accept a move to a new graph on the basis of an acceptance probability, α. However, due to exclusion of the null graph from our candidate set, care must be shown when specifying α. We shall consider each approach in turn.

4.2 Hill-climbing

As the above introduction suggests, the hill-climbing algorithm (Russell and Norvig, 2003) is the simpler of the two approaches. In general, this a straightforward search method used in a large state space that, at each iteration, will move to a neighbour of a given current state, whenever the new state is of increased value. In this case, the algorithm is said to 'climb' in an uphill direction until it reaches a local maximum, *i.e.* a point where no neighbour has higher value. The algorithm is known as *greedy* as it always chooses the best available state at each iteration without thinking any further ahead.

Consider the algorithm from the perspective of scoring sparse VAR(1) models. Thus, at some iteration, say that we have accepted a model from the graphical space, represented by an A-graph, together with an associated lower bound value. Then, at the next iteration, we propose a new model, by randomly choosing a graph within the neighbourhood of the current, accepted graph, *i.e.* by the addition or deletion of a single edge. We then evaluate the lower bound for the proposal, and compare the value to that of the accepted graph. If the variational score between the two models improves (*i.e.* increases), then the proposed graph, with its corresponding lower bound, is accepted categorically, otherwise we reject and return to the graph at the previous iteration. We continue until all neighbouring graphs have lower scores, at which point the lower bound of the accepted graph is a local maximum. This procedure can be represented by a formal hill-climbing algorithm as given below.

Algorithm 4 *1. Initialise the iteration counter to $k = 1$. For an initial graphical model M_0, relating to a directed A-graph on d vertices, G_0 (itself corresponding to matrix $A^{(0)}$), run update equations for the variational parameters until convergence, and hence evaluate the converged lower bound, $\mathcal{L}_{M_0}(q_0)$.*

2. At iteration k, propose a modified graphical model, M_ϕ, to the current model, M_{k-1}, such that we have exactly one of the following:

(a) a new edge is randomly added to the current graph, G_{k-1}.

(b) a randomly selected edge is deleted from this graph.

That is, randomly and independently, simulate two integers from the sequence $1, \ldots, d$, namely i, j. Examine the corresponding entry of the matrix $A^{(k-1)}$. If $a_{ji}^{(k-1)} = 0$ (the value of a_{ji} at iteration $k-1$), add the corresponding directed edge from i to j to the existing G_{k-1}, and let $a_{ji}^{(\phi)} = *$ (an unspecified, non-zero). Otherwise, if non-zero, delete this edge and let $a_{ji}^{(\phi)} = 0$.

3. Evaluate the variational score, $\mathcal{L}_{M_\phi}(q_\phi)$, for the proposed model M_ϕ.

4. Set $\mathcal{L}_{M_k}(q_k) = \mathcal{L}_{M_\phi}(q_\phi)$, hence $M_k = M_\phi$, i.e. accept the new variational score and model M_ϕ, if $\mathcal{L}_{M_\phi}(q_\phi) > \mathcal{L}_{M_{k-1}}(q_{k-1})$. Otherwise, set $\mathcal{L}_{M_k}(q_k) = \mathcal{L}_{M_{k-1}}(q_{k-1})$ and thus $M_k = M_{k-1}$.

5. *Change the counter from k to $k+1$ and return to step 2.*

So, this algorithm attempts to locate a locally optimum graph by searching throughout the graphical space, accepting and rejecting moves as appropriate. This procedure is illustrated by a simple example.

4.2.1 Example

Suppose $d = 10$. A dataset of size $N = 250$ was simulated from the VAR(1) model (3.2) with specifications $A = \text{diag}(0.6)$, a 10×10 diagonal matrix of coefficients, and $\sigma^2 = 0.1$. This implies that the true choice of A can be represented by an A-graph such that each of the 10 nodes has a directed self-loop, and no other edges exist. We realise now that the dimension of the graphical space is $2^{10^2} - 1 \approx 1.268 \times 10^{30}$.

The prior distributions over $\mathbf{a}$ and σ^2 were again given by

$$\begin{aligned} p(\mathbf{a}) &= \mathcal{N}(\mathbf{a}\,|\,\mathbf{0},\, C^*) \;\text{ where }\; c_{ij} \in \{0,\, 0.5\} \\ p(\sigma^2) &= \mathcal{IG}(\sigma^2\,|\,1,\, 0.001). \end{aligned}$$

Algorithm 4 was then implemented for this example by choosing three distinct, initial models, denoted as $M_{0,1}$, $M_{0,2}$, $M_{0,3}$, each represented by a corresponding A-graph, namely $G_{0,1}$, $G_{0,2}$, $G_{0,3}$. Then, $G_{0,1}$ was specified to be a graph with only one edge, a self-loop on node y_1, whereas $G_{0,2}$ was given as the complete graph. Finally, $G_{0,3}$ had directed edges in both directions between nodes y_i and y_{i+1} for all $i = 1, \ldots, 9$. This latter graph corresponds to an A-matrix with non-zero elements down the first sub-diagonal and first super-diagonal (the diagonals immediately below and above the main diagonal respectively) and zeroes elsewhere.

Then, in each case, the algorithm was run for $10,000$ iterations. For the three starting graphs, $G_{0,1}$, $G_{0,2}$ and $G_{0,3}$, a local maximum of $\mathcal{L}_{M_i}(q_i)$ was located after 1041, 540 and 1041 iterations, requiring 39, 98 and 58 accepted moves respectively to reach this value. The local maximum found by each model was, in fact, the variational score associated with the graph from which the data was simulated. Given the dimension of the graphical space, we cannot be certain that we have reached a *global* maximum as there may be other graphs that are erroneously preferred to the truth. However, since the same optimum has been reached from three different starting points, there is a good chance that this maximum cannot be improved upon, and is indeed global. The convergence patterns of the three models is shown in Figure 4.1 below.

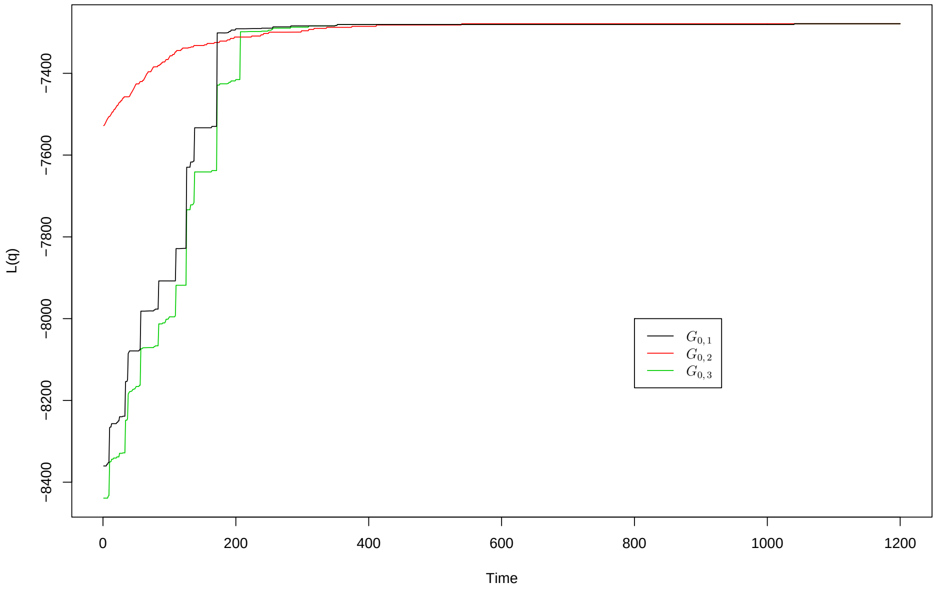


Figure 4.1: Convergence of hill-climbing algorithm for different, initial graphs

The plot shows that the climb taken by $G_{0,2}$ to reach the maximum was much smoother and quicker than that of $G_{0,1}$ and $G_{0,3}$, whose paths were actually quite akin to each other. Clearly, the complete graph was already well-favoured by the data. For $G_{0,2}$, it is no surprise however that the number of accepted moves needed to reach convergence was greater than for both $G_{0,1}$ and $G_{0,3}$ as it was the most distinct initial graph from the truth. We notice that both $G_{0,1}$ and $G_{0,3}$, due to the similarity of their routes, became stuck at the same local maximum, after around 400 iterations, before reaching the final lower bound value. This is indicated by the flatness of their convergence at this stage.

The hill-climbing algorithm is an efficient search tool that can be used in large state spaces and, as the above example illustrates, can perform well in reasonable time. It is clear that as the state space decreases in size, the number of iterations required to reach a local maximum will also lessen. However, hill-climbing does possess some intrinsic problems. For instance, we can never be certain of finding a global maximum. We can easily get stuck at a local maximum, where all neighbouring states are of lesser value. However, this point could be significantly worse than the global maximum, and even other local maxima in the space. Moreover, the algorithm can reach a flat part of the state space known as a plateau. In this case, all neighbours will be of the same quality, no uphill moves can be made, and hence the algorithm is again trapped. So, the success of the algorithm is dependent upon the shape of the state space. To tackle these issues, several different forms of hill-climbing have been constructed. For details, see, for example, Russell and Norvig (2003).

4.3 Random walks

As we have seen in the previous section, application of the hill-climbing algorithm consists of comparing the variational scores for a current and proposed model at each iteration. However, unfortunately, no information is provided about the model posterior distribution. Of course, this is a key concept because it encapsulates our post-data beliefs about each model. So, an alternative method, which will allow exploration of this distribution, would be to make a random walk across the graphical space. As mentioned in Section 4.1, the acceptance of a proposed move to a new graph is determined by the Metropolis-Hastings algorithm. It is already evident that the greediness of hill-climbing is, in fact, its downfall, *i.e.* we can never see beyond exactly any one move. So, a further advantage of random walks is now that we may be willing to accept ‘downhill’ moves which, although result in a decrease of variational score, may lead to a greater increase at a subsequent iteration, and hence escape local maxima. Other such authors to search the graphical space in this way include Giudici and Green (1999) and Jones et al. (2005).

A random walk on a single graph involves randomly selecting, and hence moving to, an adjacent node to the current node on the graph, with equal probability for each of these neighbours, hence forming a sequence of selected nodes. That is, the next node is chosen from the (discrete) uniform distribution. We can illustrate this by considering Figure 4.2. Here, we let each undirected edge represent a two-way directed edge between a pair of nodes. Suppose we are at node 1. Then, as this node possesses two neighbours (nodes 2 and 4), the probability of moving to either neighbour is $\frac{1}{2}$. However, if we are at node 2, the corresponding probability would be $\frac{1}{3}$ and so on.

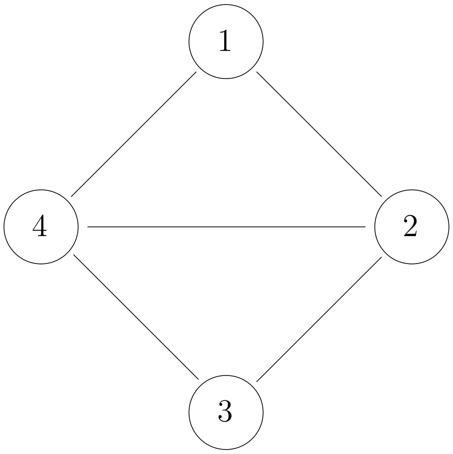


Figure 4.2: A simple graph on which to make a random walk

In general, notice that the probability of moving to a new node at time $t+1$ is dependent only upon the node at which we reside at time t, and none of the previous history of the random walk. Thus, the sequence of visited nodes forms a Markov chain with transition matrix probabilities, p_{ij}, indicating a move from node i to node j, such that

$$p_{ij} = \begin{cases} \frac{1}{d_i} & \text{if } (i,\, j) \in E \\ 0 & \text{otherwise} \end{cases},$$

as noted by Häggström (2002). Here, d_i is the number of neighbours of node i and E is the set of edges. For a comprehensive review of random walks, see, for instance, Lovász (1993).

Here however, we examine not movement between adjacent nodes on one graph, but instead between adjacent graphs in the space. So, in terms of models, for a current model, say M_j, represented by an A-graph, G_j, a proposed, neighbouring model, M_l, is considered, whose graph G_l differs from that corresponding to M_j by the addition or deletion of a uniformly selected edge. We allow acceptance of the proposed graphical model by using the Metropolis-Hastings algorithm. For further details, the reader is referred back to Chapter 2.

By considering a move from a current to a proposed model, an acceptance probability (*c.f.* (2.3)) can be specified such that

$$\alpha(M_j,\, M_l) = \min\left\{1,\, \frac{p(M_l \mid D)\, q(M_l,\, M_j)}{p(M_j \mid D)\, q(M_j,\, M_l)}\right\}. \tag{4.1}$$

Our distribution of interest is now given as the posterior over models, and $q(\cdot\,,\,\cdot)$ is again the proposal distribution, entering $\alpha(\cdot\,,\,\cdot)$ via a ratio, namely

$$\frac{q(M_l,\, M_j)}{q(M_j,\, M_l)}.$$

The denominator specifies the probability of the move from current model M_j to proposed model M_l whereas, on the numerator, that of the reverse move. However, as was seen in

Chapter 1, we have no knowledge of the model posterior as, by Bayes' Theorem, this is dependent upon the marginal likelihood, $p(D \mid M_i)$, for each model M_i. Despite this, we have already used the variational algorithm to bound, and hence approximate, $p(D \mid M_i)$ whereby $\mathcal{L}_{M_i}(q_i) \leq \log p(D \mid M_i)$. Thus, the ratio of model posteriors in the acceptance probability (4.1) can be rewritten as

$$\frac{p(M_l \mid D)}{p(M_j \mid D)} = \frac{p(D \mid M_l)\, p(M_l)}{p(D \mid M_j)\, p(M_j)} \approx \frac{\exp\{\mathcal{L}_{M_l}(q_l)\}\, p(M_l)}{\exp\{\mathcal{L}_{M_j}(q_j)\}\, p(M_j)}. \tag{4.2}$$

As was mentioned previously in Chapter 2, the constant of proportionality, $p(D)$, is eliminated in the acceptance probability by the ratio of posteriors. An important point to raise here is that we can only be sure of sampling from an approximation to the true model posterior. Of course, by minimising the KL divergence between the variational and true posterior, we suppose that the lower bound is close to the log marginal likelihood (and hence the samples are of sufficient quality). However, this is only an assumption and does not illustrate formally the accuracy of the bound.

To combat this problem, Miskin (2000) and Beal (2003) discuss the use of importance sampling from the varaiational approximation to estimate $\log p(D \mid M_i)$. That is, the logarithm of integral (1.2) is approximated by taking importance samples from the variational distribution $q(\boldsymbol{\theta}_i \mid D,\, M_i)$. This idea seems sensible since the variational should be representative of the true posterior by free form optimisation. However, Beal (2003) indicates several drawbacks with this approach, most notably that importance sampling performs poorly in high dimensions and can even fail in one dimension. In any case, we would expect the difference between the log marginal likelihood and lower bound to be similar for any model M_i. As a consequence, any inaccuracy in the approximation will cancel from the ratio of bounds in (4.2), and so such a comparison simulation is not employed here.

In returning to (4.2), a form for the model priors must also be specified. However, further discussion is required before making such a choice. We reconsider (3.9) and (3.10). Although not stated explicitly, these prior distributions are conditioned on each model M_i in the candidate set, similarly the variational distributions (3.22) and (3.27). Yet moreover, it is recalled that, in the previous chapter, sparsity modelling was used to induce many zeroes in the parameter matrix A. That is, across Metropolis-Hastings iterations, we expect models to be chosen with few existent edges, implying that the number of causations between nodes will be sparse. By modelling in this manner, a prior over the coefficients a_{ij} of the autoregressive matrix A, where $i,\, j = 1, \ldots, d$, across models is induced automatically, which dictates that each a_{ij} may either be zero or non-zero. Such a specification is termed a *sparsity prior* (also termed by Lucas et al. (2006) as a ‘point-mass mixture’ prior), taking the form

$$p(a_{ij}) = p\delta_0(a_{ij}) + (1-p)\mathcal{N}(a_{ij} \,|\, 0,\, c). \tag{4.3}$$

Here, c is defined from (3.11), and $p = \mathrm{P}(a_{ij} = 0) = 1 - \mathrm{P}(a_{ij} \neq 0)$ is the ‘in-out’, prior probability that a coefficient is zero. Moreover, $\delta_0(\cdot)$ is the *Dirac delta function*, which, for any $z \in \mathbb{R}$, possesses the properties

$$\int_{-\infty}^{+\infty} \delta_0(z)\,\mathrm{d}z = 1$$
$$\delta_0(z) = \begin{cases} 0 & \text{if } z \neq 0 \\ \infty & \text{if } z = 0 \end{cases}.$$

That is, $\delta_0(z)$ has a peak of infinite height at $z = 0$, and vanishes elsewhere on the real line such that it integrates to unity. Thus, (4.3) dictates that a_{ij} is a point mass at zero *a priori* with probability p, whereas a Gaussian distribution is followed with probability $1-p$. So, this prior, not conditional on each model, mixes a probability mass at $a_{ij} = 0$ with a distribution over non-zero values of a_{ij}. By updating (4.3), approximate marginal posterior information can be provided about each a_{ij}, and this is discussed in the next section.

Presently however, we examine how this sparsity prior affects the specification of the model prior distributions. Since p is the common, prior probability that no edge exists between any pair of nodes, a form for $p(M_i)$ is determined, assuming that all edges are *a priori* independent. If a model M_i has η non-zero elements in its associated A-matrix (*i.e.* there are η edges on the graph), then, as A contains d^2 elements, a prior distribution can be specified such that

$$p(M_i) = p^{d^2-\eta}(1-p)^{\eta}. \tag{4.4}$$

For example, suppose that a model is represented by the matrix $A = \begin{pmatrix} 0 & 0 \\ * & 0 \end{pmatrix}$. Then, the prior for this model is equivalent to $\mathrm{P}(a_{11},\, a_{12},\, a_{22} = 0) \times \mathrm{P}(a_{21} \neq 0) = p^3(1-p)$.

From (4.4), we notice that all models will be equally likely *a priori* if $p = 0.5$. In such a case, no preference is given to any particular model. However, if we choose $p > 0.5$, this is no-longer true since more complex models will be penalised, and sparse models favoured. Correspondingly, in the sparsity prior (4.3), this would then imply that all nodes have lower prior probability of association with another node. For the subsequent examples in this chapter, varying choices of p will be made to gauge the effect produced on the models accepted across iterations.

Ultimately, we return to (4.2), and specify the ratio of model priors in this expression by inspecting two distinct circumstances. Suppose that there are η edges on the graph G_j, corresponding to the current model M_j. Then firstly, assume that the proposed model M_l is defined such that a uniformly chosen edge is *added* to the graph. Thus, the model prior ratio provides

$$\frac{p(M_l)}{p(M_j)} = \frac{p^{d^2-\eta-1}(1-p)^{\eta+1}}{p^{d^2-\eta}(1-p)^{\eta}} = \frac{1-p}{p}.$$

So, in this case, the acceptance probability is given by

$$\alpha(M_j,\, M_l) = \min\left\{1,\, \left[\frac{1-p}{p} \times \frac{\exp\{\mathcal{L}_{M_l}(q_l)\}\, q(M_l,\, M_j)}{\exp\{\mathcal{L}_{M_j}(q_j)\}\, q(M_j,\, M_l)}\right]\right\}. \tag{4.5}$$

Secondly, let M_l now be specified whereby an edge is randomly *deleted* from the current graph. Hence, on this occasion, the afore-mentioned ratio yields

$$\frac{p(M_l)}{p(M_j)} = \frac{p^{d^2-\eta+1}(1-p)^{\eta-1}}{p^{d^2-\eta}(1-p)^{\eta}} = \frac{p}{1-p}.$$

Of course, the proposed model is subsequently accepted on the basis of

$$\alpha(M_j,\, M_l) = \min\left\{1,\, \left[\frac{p}{1-p} \times \frac{\exp\{\mathcal{L}_{M_l}(q_l)\}\, q(M_l,\, M_j)}{\exp\{\mathcal{L}_{M_j}(q_j)\}\, q(M_j,\, M_l)}\right]\right\}. \tag{4.6}$$

Notice that (4.5) and (4.6) are equivalent, only when $p = 0.5$, as mentioned previously.

We also need to define a form for the proposal distribution, $q(\cdot\,,\,\cdot)$. In general, care must be shown when making such a choice. If the proposal results in candidates being regularly rejected, the chain will move infrequently, and so mixing will be poor. Similarly, if too many candidates are accepted, achieved by proposing only small moves, exploration of the graphical space will again take a long time. A good proposal will avert both these extremes.

If we examine the random walk on a single graph again, such proposal probabilities are easily specified. For instance, reconsider Figure 4.2. If we were proposing the move from node 1 to node 2, then using the same notation for proposal densities as above, the ratio of proposals in the Metropolis-Hastings acceptance probability would be

$$\frac{q(2,\, 1)}{q(1,\, 2)} = \frac{\frac{1}{3}}{\frac{1}{2}} = \frac{2}{3},$$

as node 1 and node 2 have two and three neighbours respectively. The case of random walks amongst graphs is analogous to this, where we now use a (discrete) uniform proposal distribution to choose a new graph from the set of neighbouring graphs. For a graph on d nodes, there are typically d^2 neighbours. This is clear since each graph corresponds to a $d \times d$ sparse matrix A, and hence there are d^2 possible ways to change exactly one of the elements in the matrix to form a new graph. However, we must proceed cautiously here since we ignore the null graph, represented by the zero matrix, as pointed out in Section 3.5. Thus, for graphs with just 1 existent edge, there are only $d^2 - 1$ neighbouring graphs.

Consequently, we look at three scenarios for the ratio of proposal densities, each to be considered in turn. Recall that we modify the existent model in two ways: either adding or deleting an edge from the current graph. A 'delete move' from a model M_j, corresponding to a graph G_j with two edges, to a specified model M_l, whose graph G_l possesses a single edge, has ratio of proposals such that

$$\frac{q(M_l,\, M_j)}{q(M_j,\, M_l)} = \frac{\frac{1}{d^2-1}}{\frac{1}{d^2}} = \frac{d^2}{d^2-1}. \tag{4.7}$$

Here, the probability of a move from model M_j to M_l, given by the denominator, follows since G_j has d^2 neighbouring graphs. Moreover, the probability of the reverse move, provided by the numerator, is clear as G_l has only $d^2 - 1$ neighbours. In contrast, an 'add move' from M_j, whose associated graph has one edge, to a given M_l, represented graphically with two edges, has ratio of proposals given by

$$\frac{q(M_l,\, M_j)}{q(M_j,\, M_l)} = \frac{\frac{1}{d^2}}{\frac{1}{d^2-1}} = \frac{d^2-1}{d^2}. \tag{4.8}$$

Any other move has the ratio

$$\frac{q(M_l,\, M_j)}{q(M_j,\, M_l)} = \frac{\frac{1}{d^2}}{\frac{1}{d^2}} = 1, \tag{4.9}$$

as now we can traverse to any of the d^2 neighbouring graphs. A minor issue here is how to proceed if the final edge is selected for deletion, and this is explained in Algorithm 5 below.

Finally, we realise that, in practice, the exponential of the variational lower bound is customarily very small in size. Hence, it makes computationally better sense to work on the log scale. In total, four separate versions are obtained for the log acceptance probability $\log \alpha(\cdot, \cdot)$. If an edge is added to the current graph, the ratio of proposal densities, (4.8) and (4.9), are substituted directly into (4.5), representing respectively the cases when the current graph has one edge, and otherwise. In an analogous way, (4.7) and (4.9) are inserted accordingly into (4.6), when an edge is deleted. In all cases, logarithms are taken of both parts of the acceptance probability. So, the Metropolis-Hastings algorithm can be provided thus. In what follows, we suppose that each current graph has λ edges, whereas every proposed graph has ζ edges, *i.e.* λ and ζ represent the number of non-zeroes in the corresponding A-matrix.

Algorithm 5 *1. Initialise the iteration counter to* $k = 1$*. For an initial graphical model* M_0*, relating to a directed A-graph on* d *vertices,* G_0 *(itself corresponding to matrix* $A^{(0)}$*), run update equations for the variational parameters until convergence, and hence evaluate the converged lower bound,* $\mathcal{L}_{M_0}(q_0)$*.*

2. *At iteration* k*, propose a modified graphical model,* M_ϕ*, to the current model,* M_{k-1}*, such that we have exactly one of the following:*

   (a) a new edge is randomly added to the current graph, G_{k-1}*.*

   (b) a randomly selected edge is deleted from this graph.

   That is, randomly and independently, simulate two integers from the sequence $1, \ldots, d$*, namely* i*,* j*. Examine the corresponding entry of the matrix* $A^{(k-1)}$*. If* $a_{ji}^{(k-1)} = 0$*, add the corresponding directed edge from* i *to* j *to the existing* G_{k-1}*, and let* $a_{ji}^{(\phi)} = *$

(an unspecified, non-zero). Otherwise, if non-zero, delete this edge and let $a_{ji}^{(\phi)} = 0$*. If the last edge is chosen for deletion, additional pairs of integers are simulated until an edge is found that can be added. During this time, the algorithm remains at iteration* k*.*

3. *Evaluate the variational score,* $\mathcal{L}_{M_\phi}(q_\phi)$*, for the proposed model* M_ϕ*.*

4. *Calculate the log acceptance probability* $\log \alpha(M_{k-1},\, M_\phi)$ *of the proposed move, where:*

   (a) *an edge is added to the graph* G_{k-1}*.*

   - *if* $\lambda = 1$*,*

   $$\begin{aligned}\log \alpha(M_{k-1},\, M_\phi) = \min\{0,\, &\log(1-p) - \log p + \mathcal{L}_{M_\phi}(q_\phi) - \mathcal{L}_{M_{k-1}}(q_{k-1}) \\ &+ \log(d^2-1) - \log d^2\}.\end{aligned}$$

   - *otherwise,*

   $$\log \alpha(M_{k-1},\, M_\phi) = \min\{0,\, \log(1-p) - \log p + \mathcal{L}_{M_\phi}(q_\phi) - \mathcal{L}_{M_{k-1}}(q_{k-1})\}.$$

   (b) *an edge is deleted from the graph* G_{k-1}*.*

   - *if* $\zeta = 1$*,*

   $$\begin{aligned}\log \alpha(M_{k-1},\, M_\phi) = \min\{0,\, &\log p - \log(1-p) + \mathcal{L}_{M_\phi}(q_\phi) - \mathcal{L}_{M_{k-1}}(q_{k-1}) \\ &+ \log d^2 - \log(d^2-1)\}.\end{aligned}$$

   - *otherwise,*

   $$\log \alpha(M_{k-1},\, M_\phi) = \min\{0,\, \log p - \log(1-p) + \mathcal{L}_{M_\phi}(q_\phi) - \mathcal{L}_{M_{k-1}}(q_{k-1})\}.$$

5. *Put* $\mathcal{L}_{M_k}(q_k) = \mathcal{L}_{M_\phi}(q_\phi)$*, hence* $M_k = M_\phi$*, i.e. accept the new variational score and graphical model* M_ϕ*, with log probability* $\log \alpha(M_{k-1},\, M_\phi)$*. Otherwise, put* $\mathcal{L}_{M_k}(q_k) =$

$\mathcal{L}_{M_{k-1}}(q_{k-1})$ and thus $M_k = M_{k-1}$.

6. *Change the counter from k to $k+1$ and return to step 2.*

To clarify, at each iteration, a new model is simulated from the proposal distribution, represented in the algorithm by a corresponding variational score. The score (and hence model) can be either accepted or rejected upon comparison to the lower bound of the current model, determined by the acceptance probability. Notice that, from a computational perspective, a proposed move is accepted if $\log u < \log \alpha(M_{k-1}, M_\phi)$ where $u \sim \mathcal{U}(0, 1)$. It is also realised that, from the definitions of the log acceptance probability, $\log\left(\frac{1-p}{p}\right) = \log\left(\frac{p}{1-p}\right) = 0$ when $p = 0.5$. Thus, the acceptance of models in the scheme will be independent of p in this circumstance.

We hence construct a Markov chain of accepted variational scores, whose values, upon convergence and exponentiating, will be draws from the distribution proportional to $\exp\{\mathcal{L}_{M_i}(q_i)\}\, p(M_i)$, an approximation to the model posterior. Here, $\mathcal{L}_{M_i}(q_i)$ is the distribution of lower bounds across all models, where $i = 1, \ldots, R$. By comparing both Algorithms 4 and 5, it is clear that this version of the Metropolis-Hastings algorithm is merely an extension of the simpler hill-climbing algorithm of earlier. Presently however, a proposed move is dependent upon an acceptance probability, which, as discussed previously, enables the graphical space to be explored with greater effect than would be seen with hill-climbing.

4.3.1 Implementation and analysis

In due course, we will use Algorithm 5 to make a random walk across the graphical space in several examples. Before this, we examine the MCMC theory required to produce and analyse the subsequent results. Initially, recall from Chapter 2 that trace plots can be used to assess not only the duration of the burn-in period, but also the mixing properties of a Markov chain as a way to analyse possible convergence. Note that, in this case prior to convergence, both the accepted lower bound values and associated candidate models are eliminated.

A further plot to utilise when examining for convergence of a chain is that of the autocorrelation function (ACF). We realise that the values generated by using the Metropolis-Hastings sampler, upon convergence, are not independent since, by definition of a Markov chain, each simulated value is dependent on the previous value. For instance here, the current model, which may have been accepted at many iterations previous, is used to generate a proposed model, by the addition or deletion of a uniformly selected edge from the representative graph. Hence, this dependence implies that there will be correlation between the corresponding variational scores for these models.

To quantify this correlation, the ACF at lag h measures the correlation between the whole chain of lower bound values and the same chain, time-shifted by h iterations. Suppose that, after burn-in and upon convergence, the chain is of length n. Then, for example, at lag 10, we study the correlation between the lower bounds of the chain at the iteration sets $\{1, 2, \ldots, n-10\}$ and $\{11, 12, \ldots, n\}$. If $\mathcal{L}_{M_k}(q_k)$ is the lower bound at the k-th iteration, then the lag h ACF is estimated by

$$\hat{r}_h = \frac{\sum_{k=1}^{n-h}(\mathcal{L}_{M_k}(q_k) - \bar{\mathcal{L}}(q))(\mathcal{L}_{M_{k+h}}(q_{k+h}) - \bar{\mathcal{L}}(q))}{\sum_{k=1}^{n}(\mathcal{L}_{M_k}(q_k) - \bar{\mathcal{L}}(q))^2}, \tag{4.10}$$

where $\bar{\mathcal{L}}(q) = \frac{1}{n}\sum_{k=1}^{n}\mathcal{L}_{M_k}(q_k)$.

A high value of the ACF indicates poor mixing as indicated by no rapid movements on the trace plot, and hence a lack of convergence. On the contrary, lower autocorrelations correspond to little dependence between chain values. Therefore, new values of the chain will not remain in the same area of the graphical space as those before, leading to good coverage of the space, and hence a well mixing chain. Thus, the values are seen to be ‘independent’ when there is approximately zero autocorrelation at each lag. The number of independent values represented is called the effective sample size of the chain. A standard way to reduce autocorrelation is by only retaining every t-th value of the chain after burn-in, a method referred to as *thinning.* It is important that t is chosen not to be too large, since, although this would further reduce autocorrelation, we require a chain of sufficient length to conduct analysis.

In our discussion of testing for the convergence of a Markov chain, we have hitherto inspected graphical methods, which are reliable, but lack formality. To this end, several such convergence diagnostics have been developed, two of which are now examined and a third illustrated in Section 4.3.5. Such diagnostics can be located within the `R` package called `CODA` (Plummer et al., 2006).

The Raftery-Lewis test (Raftery and Lewis, 1992) is formulated around estimating a quantile Q of the distribution of interest to within an accuracy of $\pm r$ with probability s. Recall, in general, that the value x_p of a distribution F whereby $F(x_p) = p$, for $0 < p < 1$, is the p-th quantile of this distribution. Having specified Q, r, s, the test breaks the Markov chain into a new sequence such that we obtain a '1' if $\mathcal{L}_{M_k}(q_k) \leq \mathcal{L}_Q(q)$ (the Q-th quantile of the sample distribution of lower bounds) and a '0' otherwise, for all k. This binary sequence generates a two-state Markov chain. Transition probabilities can be estimated from the sample by counting the number of times that state a moves to state b where a, $b = 0$, 1, and normalising so that the row sums in the transition matrix equal one.

Thus, the test subsequently estimates the length of the burn-in period, M, the thinning interval t and the number of additional iterations, N, required to achieve the level of specified accuracy. Moreover, also determined is N_{min}, the number of iterations required had the chain been fully independent. From this, the convergence diagnostic, I, known as the *dependence factor* is derived such that $I = \frac{N}{N_{\text{min}}}$. This measures the increase in the number of iterations needed to achieve convergence as a result of the correlation within the chain. As a rule of thumb, if $I > 5$, then the chain suffers from strong autocorrelation, indicating convergence problems.

An alternative diagnostic for convergence of a Markov chain is given by Heidelberger and Welch (1983). Initially, we test the null hypothesis that the values of the chain come from a stationary distribution. If the null is accepted, then no burn-in is needed; if rejected, the first 10% of the chain is removed and we repeat the test. If the test fails again, we remove the next 10% from the sequence, and continue on until either the null hypothesis is accepted, whereby the burn-in is considered to be the discarded part of the chain, or less than 50% of the chain is left. In the latter case, the 'stationarity test' is deemed to have failed, and hence the requirement for a longer MCMC run.

If the stationarity test is passed, then we conduct a half-width test on the remaining part of the chain by constructing a confidence interval for the mean of the distribution of interest. Subsequently, we find the ratio between half the width of the interval and the sample mean. If the ratio is less than a specified value, ϵ, then this test is also passed. A failure implies that a larger sample is required from which the mean can be estimated with the necessary accuracy.

Thus far, the MCMC output has been pivotal to our analysis. However, we have additional interest in the graphical structures of the models accepted across iterations, information not provided by the variational scores on their own. To proceed, we simply measure the cumulative effect of such models. Thus, we create a $d \times d$ 'counting' matrix, $\hat{\Pi}$, initialised as the zero matrix, which records each edge between any pair of nodes for the accepted graph at every iteration of the thinned chain with burn-in discarded. For example, at iteration k, suppose that there exists an edge between nodes i and j, and this same edge had already been counted a times in the previous $k-1$ iterations. Then, we say that $\hat{\Pi}_{ij}^{(k)} = a + 1$. This process is repeated for other edges between nodes before progressing to iteration $k+1$, and so on.

At the end of the MCMC run, the matrix can be inspected to establish which edges have been accepted most often. It would be beneficial if this task were able to be performed visually. Fortunately, we can utilise the standard `R` function, `image`. This function creates a grid, representing each element of the matrix $\hat{\Pi}$, and each rectangle on the grid is assigned a colour. In the examples forthcoming, a spectrum of colours is applied, ranging from red to white. The lighter the colour, the more often that edge has been accepted between two particular nodes. In other words, if any rectangle is red, that edge has occurred infrequently.

When simulating data ourselves, we can introduce a true *adjacency matrix*, Π, defined such that $\Pi_{ij} = 1$ if $a_{ij} \neq 0$, otherwise zero. Having normalised $\hat{\Pi}$ by the length of the chain n, we wish to employ a formal technique so that we may compare the empirical proportions that any edge exists on the graph to the (0, 1)–matrix of true probabilities. That is, to measure the discrepancy between the truth and the normalised estimate, we can compute the residual sum of squares, denoted by S. Hence, in this instance, we have

$$S = \sum_{i=1}^{d} \sum_{j=1}^{d} \left(\frac{1}{n} \hat{\Pi}_{ij} - \Pi_{ij} \right)^2 . \tag{4.11}$$

As mentioned in Section 4.3, the coefficients a_{ij} of the sparse matrix A across models are a further quantity of interest. Formerly, given a dataset, our principle objective has been to estimate the unknown sparsity structure of A, and discover which nodes on the A-graph have an influence over others. That is, we aspired to determine the pattern of zeroes in the truth. However, the focus now switches to learning the likely values of each a_{ij} on the basis of an MCMC sampler. Thus, upon specifying the sparsity prior (4.3) over these coefficients, we would like to revise our beliefs by inferring both $\mathrm{P}(a_{ij} = 0 \mid D)$ and $p(a_{ij} \mid a_{ij} \neq 0,\, D)$. The former is the posterior probability over models that a particular $a_{ij} = 0$, whereas the latter is the marginal posterior density of a_{ij}, given that it is non-zero in value.

For each coefficient, it is possible to *estimate* $\mathrm{P}(a_{ij} = 0 \mid D)$ by counting the number of times that $a_{ij} = 0$ for the models accepted across Metropolis-Hastings iterations of the thinned, converged chain, and dividing by the length of the run n. It is critical to realise that this estimate is dependent, not only upon the simulation of a new model via the proposal distribution, but also the use of the variational algorithm to determine which model is accepted at each iteration. At this stage, we recollect that the two lower bounds for the proposed and current model are entered into the acceptance probability. Hence, our approximation is denoted by $\mathrm{P}_{\mathrm{var}}(a_{ij} = 0 \,|\, D)$.

We now proceed to infer the marginal $p(a_{ij} \,|\, a_{ij} \neq 0,\, D)$. It is clear that the variational posterior for a_{ij}, which corresponds to the model, M_k, accepted at iteration k of the sampler, is conditioned upon it. So, to estimate this true posterior density of a_{ij} without conditioning, we average these variational densities across iterations whenever $a_{ij} \neq 0$. This technique is known as *Bayesian model averaging* — for a brief overview, see Kass and Raftery (1995). Now, define $n_{a_{ij} \neq 0}$ to be the length of the chain when $a_{ij} \neq 0$. Thus, using (3.34) and previous notation, we calculate

$$\begin{aligned} p_{\mathrm{var}}(a_{ij} \,|\, a_{ij} \neq 0,\, D) &= \frac{1}{n_{a_{ij} \neq 0}} \sum_{a_{ij}^{(k)} \neq 0} q(a_{ij} \,|\, D,\, M_k) \\ &= \frac{1}{n_{a_{ij} \neq 0}} \sum_{a_{ij}^{(k)} \neq 0} \mathcal{N}\left(a_{ij} \,|\, \rho_{(i,j)}^{(k)},\, \tau_{(i,j)}^{(k)}\right). \end{aligned} \tag{4.12}$$

Here, recall that $a_{ij}^{(k)}$ is the value of a_{ij} at iteration k, similarly $\rho_{(i,j)}^{(k)}$, $\tau_{(i,j)}^{(k)}$. Computationally, we can evaluate the densities at the same set of points, and then average the values obtained. Of course, the error associated with this estimate, achieved by simulation, is again a consequence of the use of the variational approximation.

We now require a way to summarise the above, approximate marginal posterior information for a set of coefficients of the matrix A. This can be performed graphically, as seen in Scott and Berger (2006). For each a_{ij}, $\mathrm{P}_{\mathrm{var}}(a_{ij} = 0 \,|\, D)$ is denoted by the height of a black, vertical bar with a circle atop, placed at zero and corresponding to the probability scale on the right-hand side of every graph. Moreover, the density $p_{\text{var}}(a_{ij} \,|\, a_{ij} \neq 0,\, D)$ is also plotted, measured by the scale on the opposite side, and indicating the value of a_{ij}, given that it is non-zero.

4.3.2 Examples

Algorithm 5 was coded in C, and then applied to three simulated data-sets from the VAR(1) model (3.2), each of size $N = 250$ with dimension $d = 10$ and $\sigma^2 = 0.1$. Only the specifications of A and $p = \mathrm{P}(a_{ij} = 0)$ were considered for alteration in each example. Here, the true A-graphs were chosen to be highly symmetric. Of course, the simulation could be extended to test the algorithm on randomly generated, less structured graphs. The prior distributions were again chosen to be

$$\begin{aligned} p(\mathbf{a}) &= \mathcal{N}(\mathbf{a} \,|\, \mathbf{0},\, C^*) \;\text{ where }\; c_{ij} \in \{0,\, 0.5\} \\ p(\sigma^2) &= \mathcal{IG}(\sigma^2 \,|\, 1,\, 0.001). \end{aligned}$$

The Metropolis-Hastings scheme was run for $10,000,000$ iterations in each case, and the output transferred to R for subsequent analysis. It was initialised from graph G_0, containing one self-loop on the node y_1. Of course, this corresponds to the matrix $A^{(0)}$ where $a_{11}^{(0)} = *$, otherwise zero. Using trace plots, the burn-in period was taken to be the first $100,000$ iterations. The remainder was thinned by maintaining every 1000-th iteration, leaving a total of 9900 iterations for each example on which to conduct analysis. A histogram of 30 bins was employed throughout.

Moreover, the Raftery-Lewis test was initialised with $Q = 0.025$, $r = 0.005$, $s = 0.95$, *i.e.* estimate the 2.5% quantile of the cumulative distribution function to within an accuracy of ±0.005 with probability 0.95. These are the default specifications for the function, as quoted in Raftery and Lewis (1992). On the other hand, the Heidelberger-Welch diagnostic was specified to find a 95% confidence interval for the mean, and the half-width ratio to be less than $\epsilon = 0.1$. Moreover, the stationarity test was passed if the p-value calculated was greater than 0.05.

Example 1

Initially, A and p were specified such that $A = \text{diag}(0.8)$ and $p = 0.5$, *i.e.* all models were favoured equally *a priori*. Figure 4.3 shows the trace plot, ACF plot and histogram of the lower bound values, and the `image` plot of $\hat{\Pi}$.

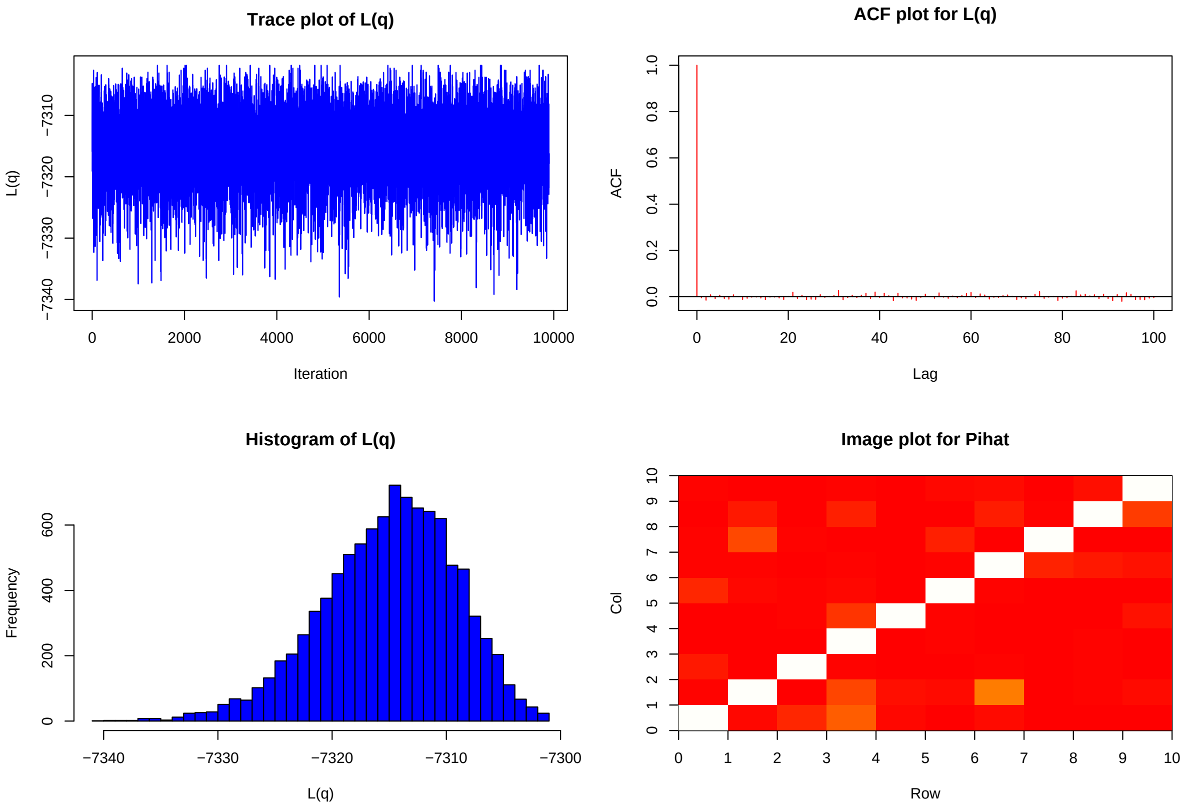


Figure 4.3: Plots for the analysis of the MCMC output in Example 1

The trace plot shows rapid movement throughout the graphical space, and hence that it is mixing well. Moreover, the ACF plot drops immediately to approximately zero autocorrelation, thus indicating the independence of the values of the chain. This is illustrated further by utilising the `effectiveSize` function in the `CODA` library. In this case, the effective sample size, *i.e.* the equivalent independent sample size, is calculated to be 9900, which is the length of the entire chain. Therefore, we can view the output as an independent chain. Given the length of the MCMC run, the stringent thinning has evidently worked. It is stressed that the correlation at lag 0 is always equal to 1, as this is the chain cross-correlated with itself without any time shift. This is clear from (4.10) by setting $h = 0$.

The above analysis provides good evidence of convergence of the chain to the stationary distribution. For the purposes of formality, we now apply the two afore-mentioned convergence diagnostics to the set of lower bounds. So, using `CODA`, the output from the Raftery-Lewis test was as follows.

```
Iterations = 1:9900
Thinning interval = 1
Number of chains = 1
Sample size per chain = 9900

Quantile (q) = 0.025
Accuracy (r) = +/- 0.005
Probability (s) = 0.95

    Burn-in  Total Lower bound  Dependence
    (M)      (N)   (Nmin)       factor (I)
Lq   2       3780  3746         1.01
```

The results suggest that only the first 2 iterations should be taken as additional burn-in, and a further 3780 iterations are necessary to attain the desired level of accuracy. To this end, resultantly, the 9900 iterations actually applied are more than sufficient. Finally, the value of the dependence factor is close to 1, indicating the independence inherent within the chain and evidence for convergence.

The Heidelberger-Welch diagnostic produced

```
     Stationarity start     p-value
     test         iteration
Lq   passed       1         0.119

     Halfwidth Mean  Halfwidth
     test
Lq   passed    -7315 0.117
```

On the output for the stationarity test, we see that the test was passed without the need to discard any of the chain, hence the start iteration is given as 1. Thus, the null hypothesis that the chain has converged is accepted with a p-value greater than the threshold value of 0.05. Moreover, the half-width test above yields the sample mean of the lower bounds, and the size of half of the constructed confidence interval. With both tests passed, once again, the chain of variational scores seems to have converged. It is noted that diagnostics were also considered for several components of $\boldsymbol{\rho}$ and τ, recorded at each iteration and corresponding to the accepted model. In each case, these were consistent with the convergence results for $\mathcal{L}_{M_k}(q_k)$, and so are not shown here. We suggest that studying $\mathcal{L}_{M_k}(q_k)$ is deemed sufficient to test for convergence.

In addition, we realise that the `image` plot of $\hat{\Pi}$ in Figure 4.3 is as expected, with self-loops regularly recognised for all nodes (white rectangles along the main diagonal). With the true adjacency matrix specified as $\Pi = \text{diag}(1)$, this accuracy is amplified by calculating the residual sum of squares to be $S = 0.684$. So, when choosing $p = 0.5$, the true sparsity structure has been identified to a highly acceptable level. Finally, graphical summaries for both $\text{P}_{\text{var}}(a_{ij} = 0 \,|\, D)$ and $p_{\text{var}}(a_{ij} \,|\, a_{ij} \neq 0,\, D)$ are presented in Figure 4.4 for several coefficients of A.

As the true A had non-zero entries only along the diagonal, we would expect the value of $\text{P}_{\text{var}}(a_{ij} = 0 \,|\, D)$, the approximate posterior probability of a point mass at zero, to be high for those off-diagonal coefficients. This is certainly apparent from the above plots. Moreover, each diagonal coefficient possesses the corresponding probability to be approximately zero. It is also noticeable that these same entries possess a density $p_{\text{var}}(a_{ij} \,|\, a_{ij} \neq 0,\, D)$

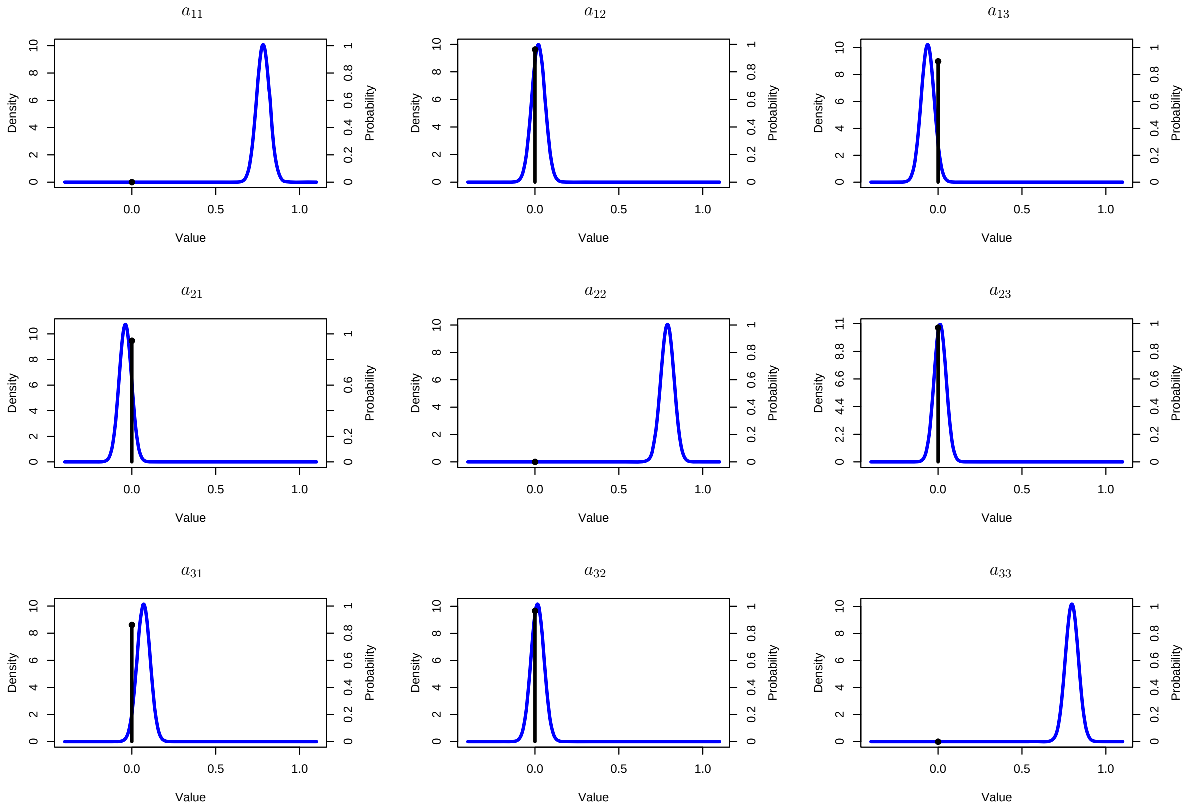


Figure 4.4: Plots showing estimated, marginal posterior distributions for a_{ij}, i, $j = 1, 2, 3$, in Example 1

with mode approximately equal to 0.8. Similarly, the related density for the off-diagonal components is peaked around zero. So, in each case, the truth is being well represented. Thus, the dataset of size $N = 250$ has overridden the prior on $\mathbf{a}$ at each iteration in the variational Bayesian update, hence providing accurate, akin estimates of a_{ij}, when not constrained to zero (*c.f.* Table 3.1 in Section 3.5). At the same time, the variance from prior to variational distribution has decreased, and so there is greater certainty about these values. Therefore, as this will be the case for each component, these marginal density plots possess a similar shape.

Example 2

On this occasion, we again chose $p = 0.5$, but now A as the tridiagonal matrix such that $A = \text{tridiag}(0.2,\ 0.4,\ 0.2)$, using the notation of Saad (2003). That is, A has 0.4s down the main diagonal and 0.2s down the first sub-diagonal and first super-diagonal, with zeroes elsewhere. The graphical analysis of the MCMC output for this data-set is displayed in Figure 4.5.

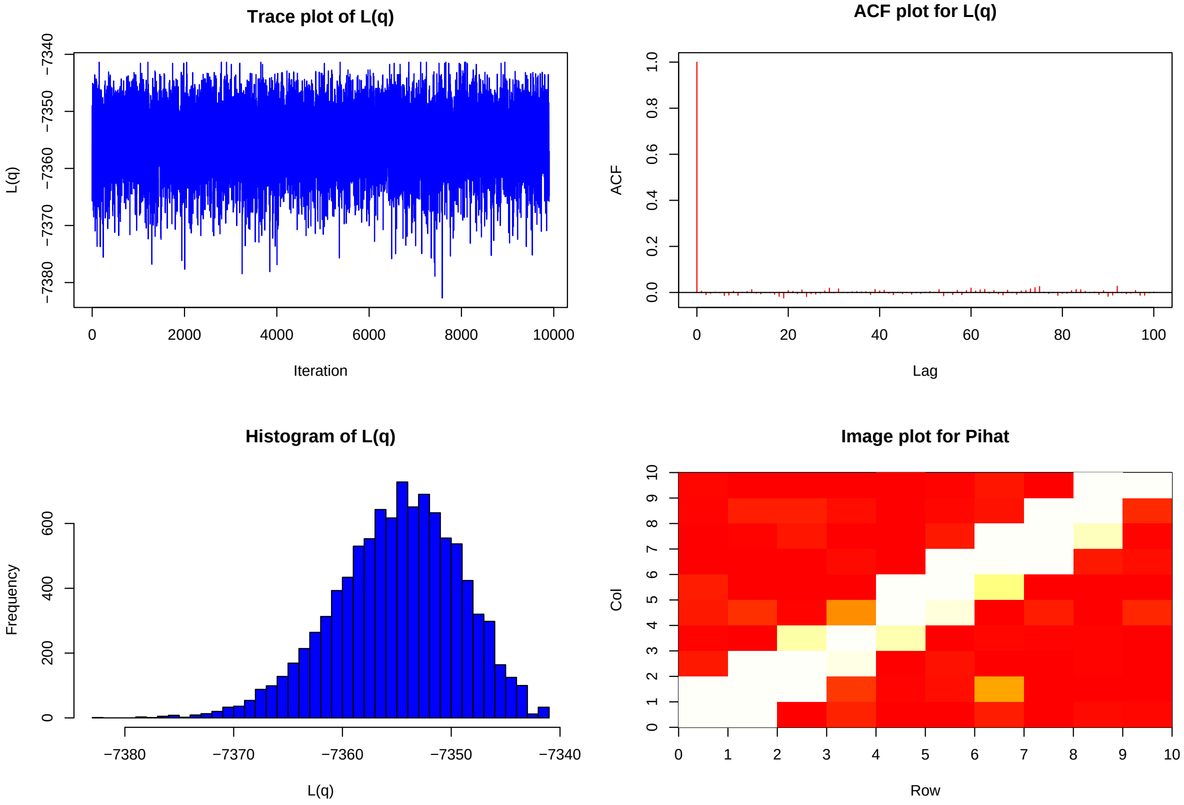


Figure 4.5: Plots for the analysis of the MCMC output in Example 2

The trace and ACF plots reveal quick mixing and an independent chain respectively, implying convergence. This is emphasised by `effectiveSize` being calculated as 9900, similar to above. For completeness, we apply the two convergence diagnostics for the lower bounds. The Raftery-Lewis output shows

```
Iterations = 1:9900
Thinning interval = 1
Number of chains = 1
Sample size per chain = 9900

Quantile (q) = 0.025
Accuracy (r) = +/- 0.005
Probability (s) = 0.95

    Burn-in  Total Lower bound  Dependence
    (M)      (N)   (Nmin)       factor (I)
Lq   2       3812  3746         1.02
```

whereas the Heidelberger-Welch diagnostic provides

```
     Stationarity start     p-value
     test         iteration
Lq   passed       1         0.270

     Halfwidth Mean  Halfwidth
     test
Lq   passed    -7355 0.116
```

Both tests supply good evidence that the stationary distribution of the chain has been acquired. As before, a similar conclusion is reached when applying the diagnostics to components of $\boldsymbol{\rho}$ and τ. With the true value of Π taken as $\Pi = \text{tridiag}(1, 1, 1)$, it was found that $S = 1.746$. Thus, $\hat{\Pi}$ remains highly accurate, as is displayed by the `image` plot whereby the non-zero elements of A have been identified, despite the A-graph that simulated the data containing more edges than previous.

Upon examination of Figure 4.6, we see that the output is encouraging. For instance, both $\text{P}_{\text{var}}(a_{13} = 0 \,|\, D)$ and $\text{P}_{\text{var}}(a_{31} = 0 \,|\, D)$ are very high in value. The same probability for all other coefficients is negligible, whereas the plots $p_{\text{var}}(a_{ij} \,|\, a_{ij} \neq 0,\, D)$ for these elements are peaked, close to the true value in each case. When compared with Example 1, the algorithm here had to determine more non-zero signals in the truth at each iteration, indicating a reason as to why the density estimates were slightly less accurate than before.

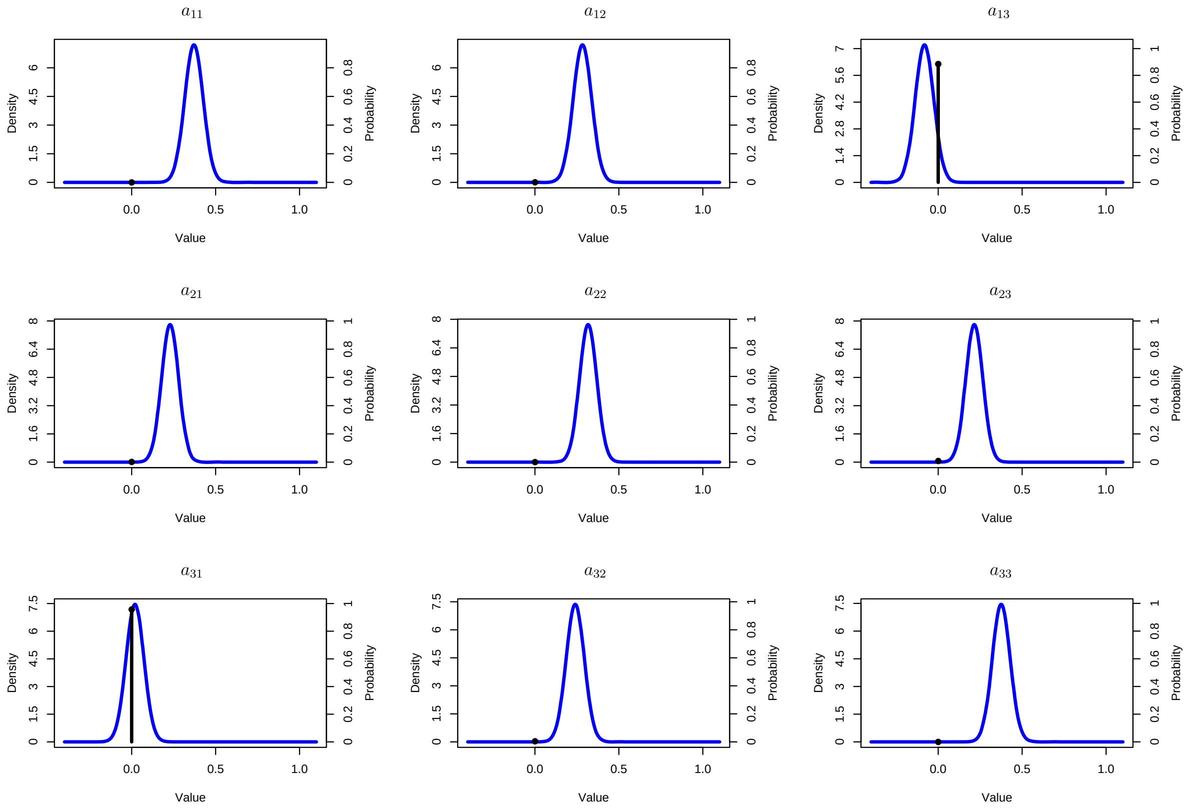


Figure 4.6: Plots showing estimated, marginal posterior distributions for a_{ij}, i, $j = 1, 2, 3$, in Example 2

Example 3

For the final example, A was the sparse matrix given as $A = \text{tridiag}(0.4, 0, 0.4)$. However, we let $p = 0.9$, a larger specification than used before and one which should favour the acceptance of sparse models across iterations. The results are shown graphically in Figure 4.7.

As with the two previous examples, the `effectiveSize` was computed in `CODA` as 9900 and, together with the trace and ACF plots, reveals probable convergence. This is further emphasised by the Raftery-Lewis test on the variational scores, the results of which are given below.

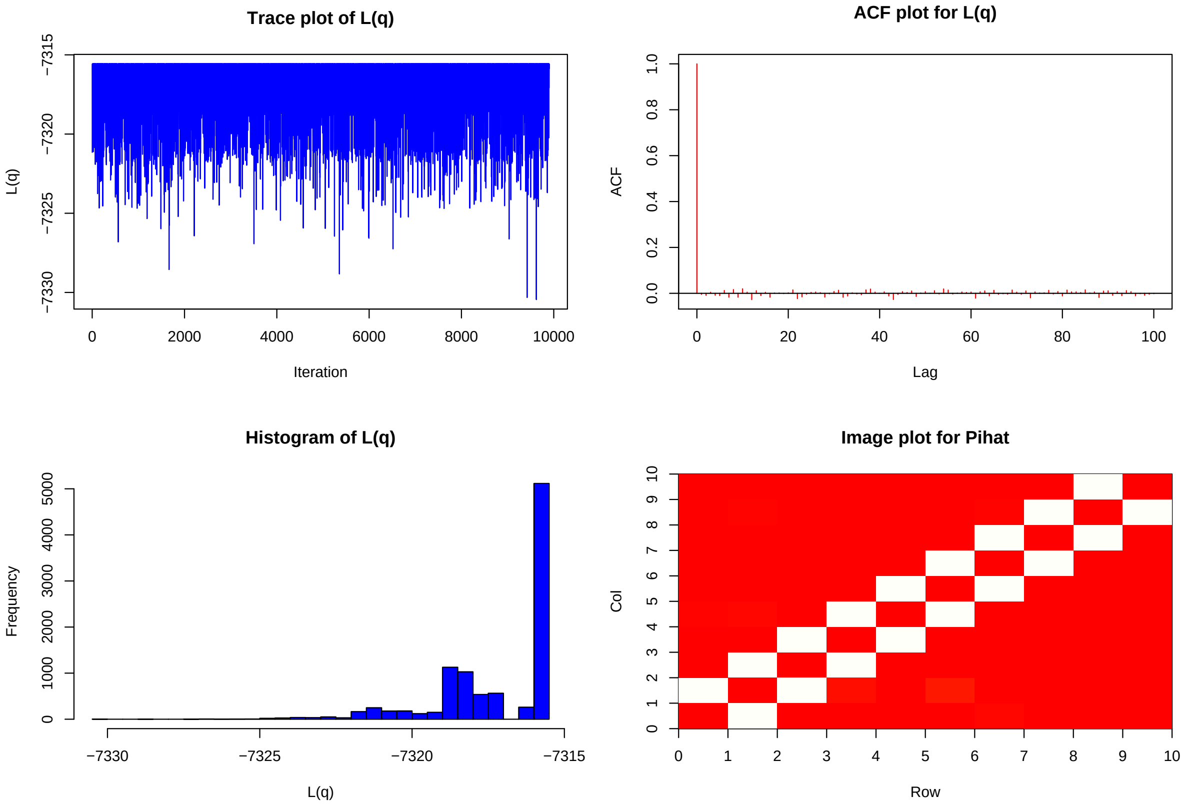


Figure 4.7: Plots for the analysis of the MCMC output in Example 3

```
Iterations = 1:9900
Thinning interval = 1
Number of chains = 1
Sample size per chain = 9900

Quantile (q) = 0.025
Accuracy (r) = +/- 0.005
Probability (s) = 0.95

    Burn-in  Total Lower bound  Dependence
    (M)      (N)   (Nmin)       factor (I)
Lq   2       3843  3746         1.03
```

Moreover, the Heidelberger-Welch diagnostic exhibits additional confirmation:

```
     Stationarity start     p-value
     test         iteration
Lq   passed       1         0.303
```

```
          Halfwidth Mean   Halfwidth
          test
     Lq   passed    -7317 0.0393
```

The histogram in Figure 4.7 is most intriguing. Here, very few moves are being accepted (illustrated by an acceptance rate of just 1.4%), and any movement that is made is to high-ranking models in the neighbourhood of the truth. By specifying $p = 0.9$, the plot of $\hat{\Pi}$ shows only the true edges being selected consistently, and no 'wrong' links identified. This contrasts slightly to the previous examples. After normalising, the proximity of $\hat{\Pi}$ to the true adjacency matrix is evident since now $S = 0.014$.

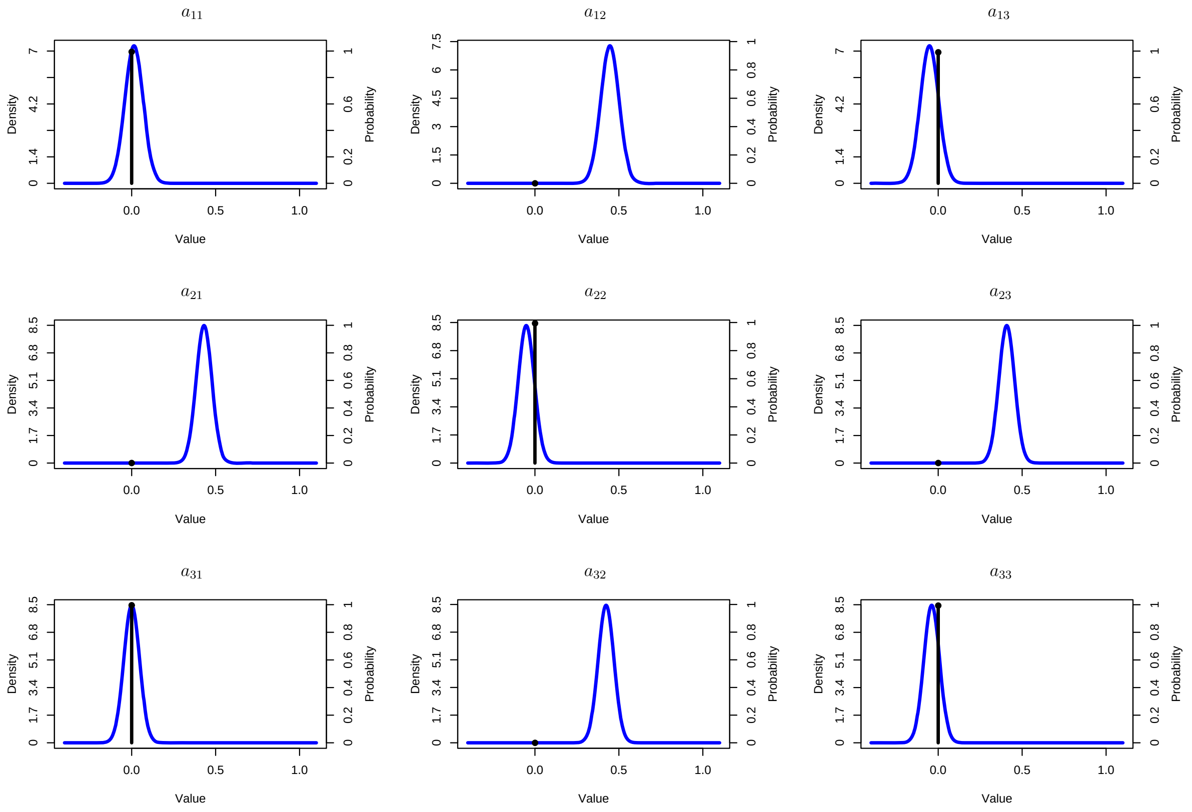


Figure 4.8: Plots showing estimated, marginal posterior distributions for a_{ij}, i, $j = 1, 2, 3$, in Example 3

Ultimately, the trends in the variational marginal posteriors for a set of coefficients of A, given above in Figure 4.8, mimic what has been seen in the previous examples. In other words, those entries in the truth, specified as zero, are again predicted to have a point mass at zero with very high probability. Moreover, for all true non-zero elements, $p_{\text{var}}(a_{ij} \,|\, a_{ij} \neq 0,\, D)$ is centred around the original specification in each case.

4.3.3 Prior sensitivity

The above examples show initially that the Markov chain is reaching its stationary distribution. We have tested this by using a combination of trace and ACF plots, the `effectiveSize` function in `CODA` and two different convergence diagnostics. Moreover, we have emphasised that the true sparsity structure is being recognised across iterations, as seen by the `image` plot and statistic S, although this is affected by the prior specification of p. Finally, approximate marginal posteriors for sets of coefficients a_{ij} have been seen to be accurate, compared to the true specification of A.

Henceforth, a prior sensitivity analysis is conducted. Initially, we alter the prior parameters for both $\mathbf{a}$ and σ^2, whilst using a fixed data-set. This is important, in particular for the informative prior on $\mathbf{a}$, to gauge if such a choice averts Lindley's paradox (see Section 3.4.2). Subsequently, sensitivity to the choice of p is also analysed. On this occasion, the dimension is increased such that $d = 20$. A data-set was simulated with respect to $N = 250$, $A = \text{tridiag}(0.2,\, 0.4,\, 0.2)$ and $\sigma^2 = 0.1$. Six different specifications of non-zero $c_{ij} = c$ and five varying choices of α, β were examined (*cf.* (3.9) and (3.10)). The inverse gamma specifications were amended between $\alpha = 1$, $\beta = 0.001$; $\alpha = 1$, $\beta = 0.01$; $\alpha = 1$, $\beta = 0.1$; $\alpha = 1$, $\beta = 1$ and $\alpha = 10$, $\beta = 10$. Currently, p was fixed at the true proportion of zeroes in the A-matrix, namely $p = 0.855$. Hence, the results obtained will not be affected by this choice.

Algorithm 5 was initialised as in the previous section, and run for 10,000,000 iterations, separately for each of the 30 combinations of c and α, β. Again, in each case, the first 100,000 were treated as burn-in and the remainder thinned by 1000. `Image` plots were then produced for all possibilities. It was found that, for every choice of c, the varying specifications of α, β made often no difference to each `image` plot. So, at each Metropolis-Hastings iteration, a sufficient quantity of data was available to estimate the noise parameter, σ^2, extremely well, regardless of the prior specification. Therefore, with the noise in the data identified, we will be able to establish the correct signals in the truth throughout the scheme, leading to `image` plots that are similar to the truth, and to each other.

Consequently, Figure 4.9 contains the plots for altering c where $\alpha = 1$, $\beta = 0.001$ are now fixed, a flat prior, as mentioned previously. Moreover, Table 4.1 displays the values of S, computed for each specification of c. Here, the impact of Lindley's paradox can be ascertained. To produce frame (a), the most diffuse prior distribution was used, *i.e.* the variance for each non-zero component, a_{ij}, was large. In addition, by recalling that every a_{ij} is specified as either a zero or a free entry, this prior was not concentrated around the simpler of these two models in each case. So, an intuitive rationale may be that the signal in the data will be recognised, and hence models, similar and at least as complex as the truth, would be accepted during the MCMC run.

However, although the sparsity structure is being predicted here to some extent, the paradox is visible. That is, models, even simpler than the truth, are being accepted, and hence favoured (shown in frame (a) by the white rectangles darkening or even disappearing) since the non-zero elements of A are not being detected in the data. This is illustrated further by the inaccuracy of S for this `image` plot. In the Metropolis-Hastings sampler, very few proposed models are being accepted, hence leading to the lack of complexity shown here. Indeed, across iterations, the overall acceptance rate of proposals was merely 2%. We realise that a comparable, but less severe scenario is displayed in frame (b).

In contrast, consider frame (f), which resulted from specifying a highly informative prior. Correct edges are now only selected infrequently, and more uncertainty has arisen about the truth. When examining the matrix $\hat{\Pi}$ itself, all incorrect edges are chosen more often than in the other cases, although to an insufficient level so as to register on the `image`

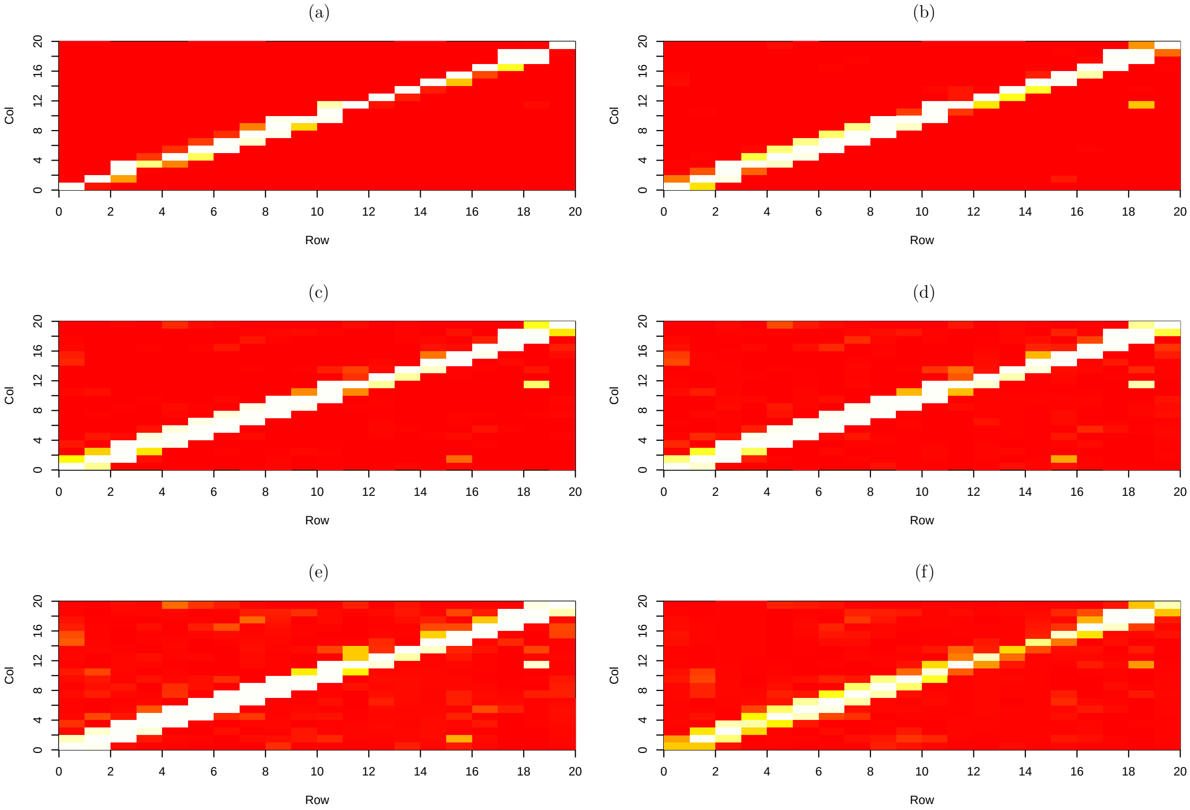


Figure 4.9: Plots of $\hat{\Pi}$ for different specifications of c: (a) $c = 10,000$, (b) $c = 10$, (c) $c = 0.5$, (d) $c = 0.1$, (e) $c = 0.01$, (f) $c = 0.001$

Specification of c	S
$10,000$	21.244
10	10.018
0.5	6.439
0.1	6.513
0.01	7.078
0.001	11.731

Table 4.1: Comparing the accuracy of $\hat{\Pi}$ for each choice of c

plot. Resultantly, the S-value dictates that accuracy has been lost. So, we intimate a slight tendency to approve denser models, the converse of the paradox, noted by the more routine acceptance of proposals (23% acceptance rate).

The three foremost specifications of c appear to be $c = 0.5$, 0.1, 0.01. When studying frame (e), 'non-existent' edges are occasionally identified, and hence there is a preference for models with more complexity. Yet, there is little difference between frames (c) and (d), whilst the values of S, when $c = 0.5$ and $c = 0.1$, are almost identical. Thus, it is evident that, in each case, the truth is well represented. We notice that elements $a_{16,\,2}$ and $a_{19,\,12}$ are incorrectly recognised as non-zeroes in both of these plots, although more recurrently when $c = 0.1$. Therefore, it is suggested that $c = 0.5$ is a sensible choice for this example to compromise between continuous acceptance of simpler or more complex models.

As a consequence of this sensitivity analysis, the prior parameter specifications of $\alpha = 1$, $\beta = 0.001$ and $c = 0.5$ are henceforth maintained for further studies. We now wish to establish the extent to which results are affected by varying the choice of p. For this purpose, we retained $d = 20$, and utilised the same dataset as simulated above. Moreover, for each choice of p, the MCMC sampler was also run in an identical fashion. The specifications given for p were namely $p = 0.95$, 0.855, 0.5, 0.3, 0.1. `Image` plots and their accuracy to the truth are displayed in Figure 4.10 and Table 4.2 respectively.

When $A = \text{tridiag}(0.2,\, 0.4,\, 0.2)$, recall that specifying $p = 0.855$ will induce the correct level of sparsity for models accepted during the scheme. Thus, it follows that frame (b) is the best portrayal of the truth. If p is assigned above this level, we would then expect those models considered too sparse to be in favour. A slight indication of this is revealed in frame (a), and explains why the value of S has now become more discrepant. In addition, the acceptance rate of proposals here is only 3%. Similarly, when $p < 0.855$, more dense models are preferred. As p approaches zero, this bias is more conspicuous, and the acceptance rate rises dramatically; for instance, when $p = 0.1$, the rate is 58%.

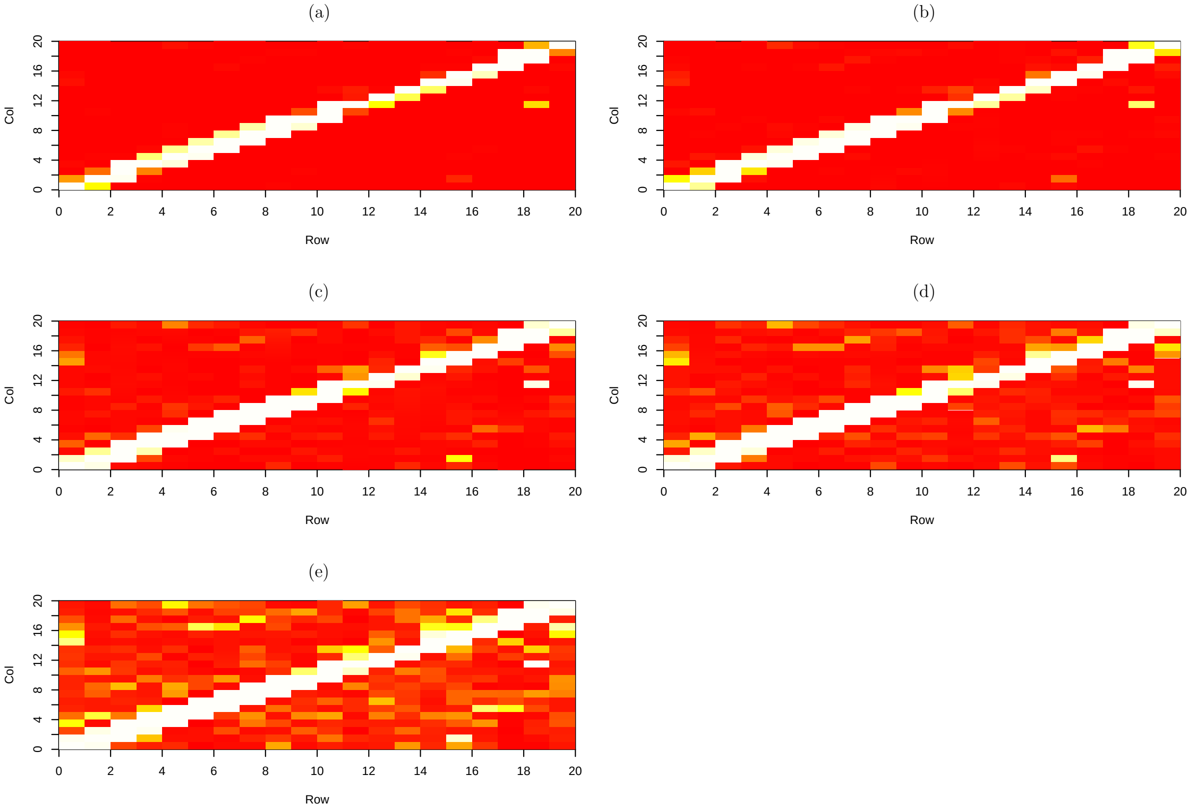


Figure 4.10: Plots of $\hat{\Pi}$ for different specifications of p: (a) $p = 0.95$, (b) $p = 0.855$, (c) $p = 0.5$, (d) $p = 0.3$, (e) $p = 0.1$

Specification of p	S
0.95	9.164
0.855	6.439
0.5	8.312
0.3	16.123
0.1	61.563

Table 4.2: Comparing the accuracy of $\hat{\Pi}$ for each choice of p

Hence, in frame (e), although most true links are routinely discovered, it is unsurprising that a plethora of false edges are now common, implying an inflated S value. We conclude that altering the specification of p can have an extreme influence upon subsequent results.

4.3.4 Small sample size

We explore another study, namely to establish how well the true A is represented over a Metropolis-Hastings run as the sample size N is changed. The choices for d, A and σ^2 are retained from the above section, whereas the prior specifications are set with $\alpha = 1$, $\beta = 0.001$, $c = 0.5$ and $p = 0.855$. Five separate datasets were simulated for differing values of N, in particular, $N = 100, 80, 50, 20, 10$. This was performed in such a way that the matrix Y, of dimension $N \times d$, would form the first N rows of the new Y for the next highest selection of N. This ensured consistency between the data-sets. Algorithm 5 was run for each N, whereby initialising graph, length of MCMC run, burn-in period and thinning ratio were maintained as above.

The outcome of the investigation is summarised in Figure 4.11 and Table 4.3. Thus, it is apparent that, as N decreases, the algorithm struggles to locate the truth and, as such, the accuracy of $\hat{\Pi}$ deteriorates. This is to be expected since, in this case, the signal in the data will be weak. In the figure, for the higher choices of N, frames (a), (b) and (c) do show that some correct edges are being continually recognised. However, once $N = 20$, the signal disappears completely, and no obvious pattern emerges on the `image` plots. As a consequence, values of S increase significantly.

One final analysis was administered by altering the value of the noise variance, σ^2. For this purpose, we let $N = 250$, and all other specifications stated above remained the same. The choices of σ^2 considered were $\sigma^2 = 0.1, 0.5, 1, 5, 10$. Upon running the MCMC algorithm and constructing `image` plots in each case, negligible difference was seen between Figure 4.9(c) (when $\sigma^2 = 0.1$) and all other plots. Hence, all S-values were

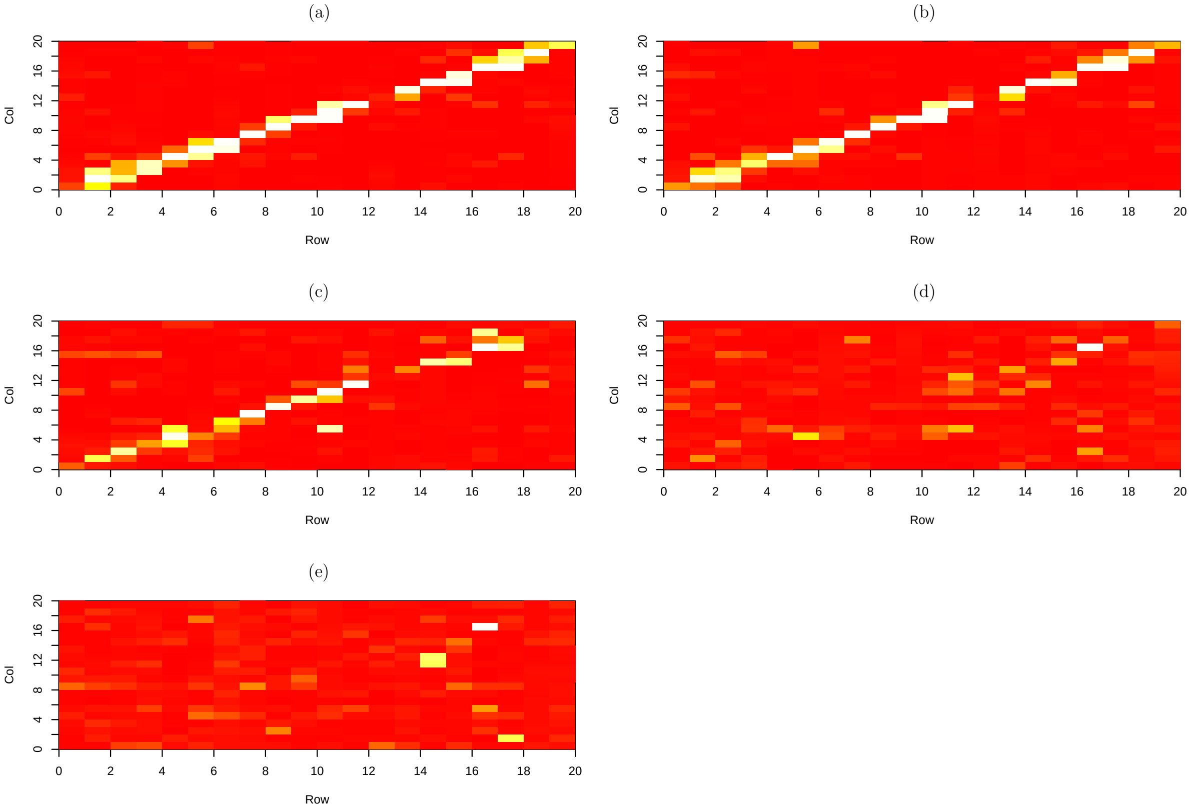


Figure 4.11: Plots of $\hat{\Pi}$ for different specifications of N: (a) $N = 100$, (b) $N = 80$, (c) $N = 50$, (d) $N = 20$, (e) $N = 10$

Specification of N	S
100	19.437
80	23.477
50	35.035
20	47.113
10	51.729

Table 4.3: Comparing the accuracy of $\hat{\Pi}$ for each choice of N

akin. So, analogous to varying the prior specification on σ^2, it follows that adequate observations are present here to estimate the noise in the data. Thus, whilst the true noise is changed, the correct sparsity structure can be predicted accurately.

4.3.5 A further example

Hitherto, the Metropolis-Hastings algorithm has been applied to find high scoring models in graphical spaces of small to moderate dimension. Here, a more challenging example is inspected whereby a dataset is simulated of size $N = 250$, with noise variance specified as $\sigma^2 = 0.1$ and $A = \text{tridiag}(0.2,\, 0.4,\, 0.2)$ as erstwhile, but now dimension $d = 100$. The parameters for the prior distributions are maintained as $\alpha = 1$, $\beta = 0.001$ and $c = 0.5$, but we now let $p = 0.9702$, the proportion of zeroes in the true A.

It has been customary thus far to run the MCMC sampler for $10,000,000$ iterations. However, on this occasion, a problem may be encountered since the amount of computational time to complete the run can become too great. A simple solution to this is to make several, shorter, parallel runs of the scheme, each independent and initialised from a varying starting value (known as a seed). An advantage of this technique is that, by comparing MCMC output, it is simple to identify any salient differences between several seemingly converged chains, if stationarity has yet to be reached.

Thus, five Markov chains were simulated from different, user-specified seeds in `C`, each of length $2,000,000$ iterations, and initialised from the same choice of G_0 as before. Each chain had the first $100,000$ iterations dropped as burn-in, and was then thinned by 1000. Hence, in total, 9500 iterations remained for analysis purposes. Figure 4.12 reveals the plots of the output from the algorithm. Notice that the histogram displays the pooled lower bound values for all five chains. Here, `CODA` has been applied to produce an overall trace plot, whereby the traces of the five, individual chains are overlaid on top of each other. Convergence can be realised when all chains possess similar behaviour, and hence are independent of each initial choice of seed. By examining the plot, the mean and variance of each chain are similar since the chains all overlap. Thus, there is reasonable evidence for convergence.

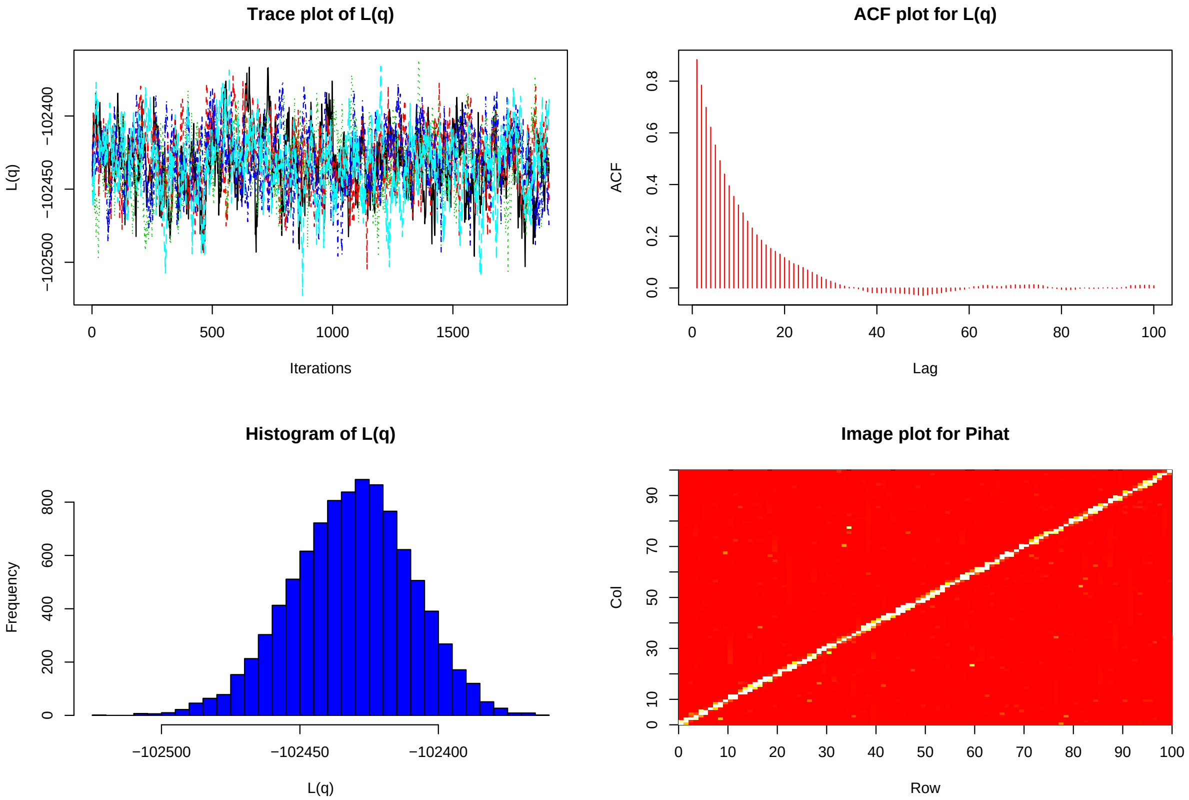


Figure 4.12: Plots for the analysis of the MCMC output in Example 4.3.5

The ACF plot shows the average autocorrelation across all five chains for a set of lags, calculated simply by application of the function `autocorr.diag`. Here, the ACF declines steadily, reaching zero at around lag 35, whereby the values in the chain are recognised as being approximately independent. In hindsight, the chain could be thinned by a higher factor than 1000 to reduce this initial dependence between consecutive values. Nevertheless, the trace and ACF plots still show acceptably quick mixing, and thus good graphical space coverage.

At this stage, it would be beneficial to amplify this view formally by the use of a convergence diagnostic. However, both tests discussed previously can be utilised when only one chain is simulated. Fortunately, in the case of parallel chains, Gelman and Rubin (1992) formulated such a diagnostic. For a set of $m > 1$ multiple chains, each containing n iterations with the first $\frac{n}{2}$ then discarded as burn-in, the statistic is based upon comparing the variance within each chain, and the variance between chains. Thus, to estimate the variance, κ^2, of the stationary distribution, we can compute both W, the mean of the m within-chain variances, and $\hat{\kappa}^2$, the variance of the mn values from all chains combined.

A variance ratio (often referred to as the potential scale reduction factor), denoted as $\hat{R}$, is then computed, dependent upon these two estimates. Consequently, if $\hat{R}$ is approximately equal to 1, the variances within and between chains are coinciding, and so there is evidence that all chains have converged to the stationary distribution, indicated on a trace plot by overlapping. That is, each run is viewed as being independent of its initial seed choice. In practice, Gelman et al. (1995) suggest a value of $\hat{R} \approx 1.2$ should be sufficient for this purpose, otherwise further iterations will be required to improve the estimates, W and $\hat{\kappa}^2$.

As with previous diagnostics, CODA can be used to apply the Gelman and Rubin test. So, in the current example, the corresponding output was

```
Iterations = 1:1900
Thinning interval = 1
Number of chains = 5
Sample size per chain = 1900

Potential scale reduction factor:

     Point est. 97.5% quantile
[1,]       1.02           1.04
```

Reported here are the estimated value and 97.5% quantile of $\hat{R}$, the latter, an upper limit, derived from its approximate sampling distribution — see Gelman and Rubin for additional details. We realise that the first $n = 950$ iterations of each chain are dropped as burn-in automatically, and so the above values are computed using the lower bounds in the second half of the chains. As $\hat{R}$ and its upper limit are both close to 1, we can conclude that 950 iterations were enough to enable convergence, and samples $950 - 1900$ from all chains are assumed to be draws from the stationary distribution.

We now return to Figure 4.12, and examine the plot of $\hat{\Pi}$. This was produced by summing the counting matrices from each of the five chains, where recall that the total number of iterations was $n = 9500$. Upon visual inspection, most true edges are recognised regularly, whilst, due to the specification of p, false links are seldom in favour. In this case, S is computed as 82.611, a higher value than has been noted for previous examples. Yet, this is hardly surprising when we realise that the size of the graphical space is now $2^{10,000} - 1$. If the size of the dataset were increased, we would expect S to be reduced, since the algorithm should be able to locate the true sparsity structure with a much stronger signal now in the data.

Finally, consider Figure 4.13, providing approximate, marginal posterior summaries for a set of a_{ij}. In accordance with the initial specification, the estimated posterior probabilities that a_{13} and a_{31} are zero are very high. Moreover, the plots of $p_{\text{var}}(a_{ij} \,|\, a_{ij} \neq 0, D)$ have predicted the true coefficient values with impressive accuracy. Here, we realise that it is important to specify both parts of the approximate posterior distribution, as noted by Scott and Berger (2006). When studying both a_{21} and a_{32}, it is seen that $\mathrm{P}_{\text{var}}(a_{ij} = 0 \,|\, D) \approx 0.4$ in each case. Yet, the density portion is concentrated around non-zero values, as we would expect. Again, our uncertainty about these coefficients would have been reduced if a larger dataset had been simulated.

4.3.6 Application to ERP data

A fresh example of our Metropolis-Hastings algorithm is now presented, where it is now applied to real time series data, as opposed to the simulated datasets used previously. This data has been analysed formerly by Delorme et al. (2004), and is freely available at

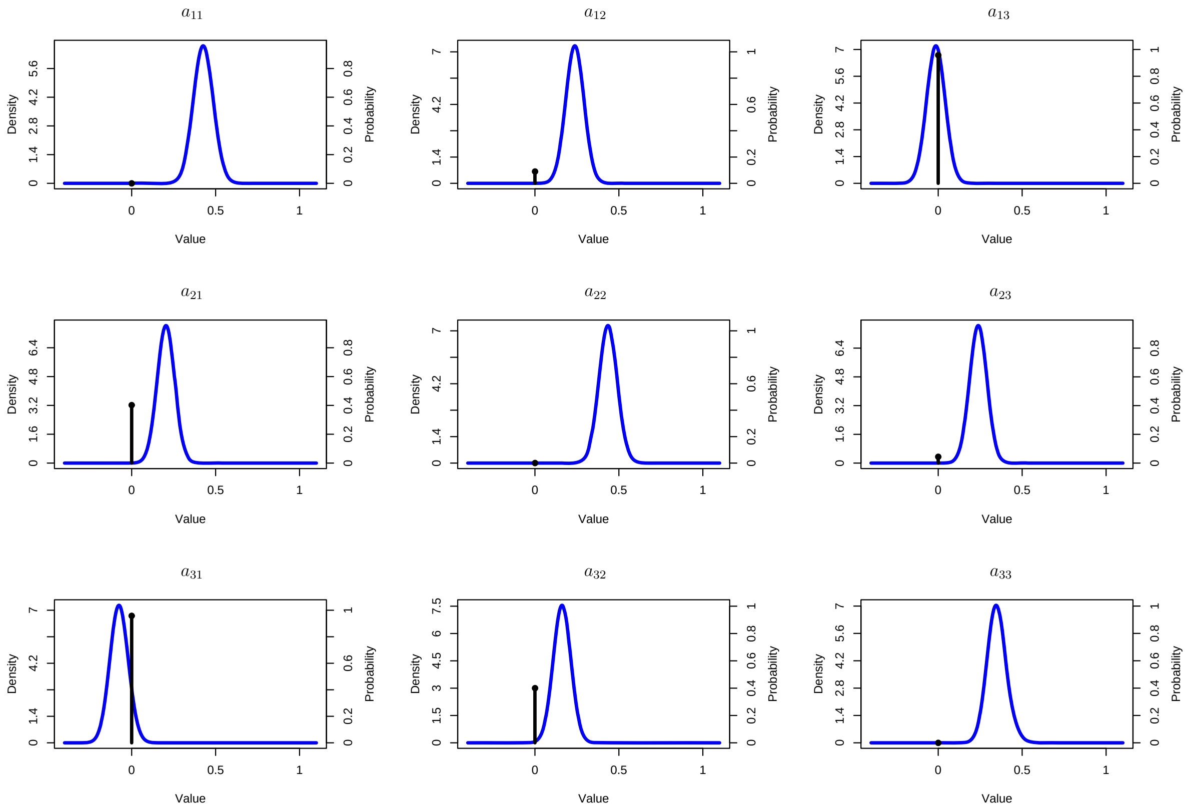


Figure 4.13: Plots showing estimated, marginal posterior distributions for a_{ij}, i, j = 1, 2, 3 in Example 4.3.5

http://sccn.ucsd.edu/∼arno/fam2data/publicly_available_EEG_data.html. The authors of the paper also provide details about the experiment that gave rise to the data, and this is now summarised.

Several human subjects participated in an animal categorisation task, involving 100 different photographs being shown at random to each volunteer for 20ms, with a small, random time between each display. Of the images exhibited, half were target photographs, featuring an animal, and the remaining 50 were non-targets, each known as a distractor. The subjects had to respond within 1s, by releasing a button whenever the picture was an animal, or keep the button depressed if a distractor was identified.

During the experiment, for each participant, the electrical activity produced by the brain (measured in microvolts, denoted μV) was recorded via $d = 32$ electrodes placed on the scalp, hence giving rise to a set of electro-encephalographic (EEG) data. In every case, this was split into two different datasets, representing the display and correct identification of either a target or non-target image. Moreover, the animal and distractor datasets were further separated into a series of time periods or epochs, each lasting for approximately 3s, in which time one image was displayed. In what follows, $N = 250$ corresponding, independent time points were examined in each epoch, a number sufficient for decent analysis. To reduce the amount of data further to a more, manageable level, the EEG measurements were averaged at each time point across epochs, thus forming a set of event-related potential (ERP) data. This procedure is also performed in practice, as it makes the brain response to a particular stimulus more visible graphically.

For our analysis, we want to compare the fast, cerebral processing involved, by taking ERP measurements, when correctly identifying either a target or non-target image. By modelling the two sets of 32-electrode data by a zero mean VAR(1) process with unknown A and σ^2, our primary focus is to infer the sparsity structure of A. Moreover, an A-graph can be constructed, with each node an electrode, determining the neural dynamics of the information processing. That is, our interest is to discover which electrodes were significant by activating further responses elsewhere, and whether this was consistent between viewing an animal or a distractor photograph.

We consider the animal and distractor ERP data for one particular subject in the study, shown respectively in Figures 4.14 and 4.15. Here, ERP value (in μV) is shown on the vertical axis, time (in ms) the horizontal axis. Each figure displays the position of every electrode on the scalp and, moreover, ERP measurements at each electrode. In all cases, the stimulus was administered at time 0ms, hence the scale on the time axis. It is noticeable that these two figures reveal similar ERPs at corresponding electrodes. It will be seen in due course whether the same processing pathways are also involved upon observation of the two distinct stimuli.

In each case, a data matrix Y of dimension 250×32 was formed such that each row was the measurement for a single time point across all electrodes. Moreover, as we have assumed that the mean of the process is zero, the data can be centred by subtracting the sample mean vector of the ERP values at each electrode from every time point. This is a standard practice to estimate the mean, and is performed to ensure that results will not be affected if the true mean is significantly different from zero. Prior specifications for $\mathbf{a}$ and σ^2 were maintained from the previous section.

A valid question to ask at this stage is what might be a sensible choice for p. In general, we know that each A-matrix in a candidate set has dimension $d \times d$, and so contains d^2 elements. Typically, by definition, the number of non-zeroes in any sparse matrix of such size will be of order $\mathcal{O}(d)$. Suppose that we set $p = 1 - \frac{1}{d}$. Then, using (4.4), the number of non-zeroes present will follow a binomial distribution with parameters d^2, the number of trials, and $1 - p = \frac{1}{d}$, the success probability. Moreover, it follows from Johnson et al. (1992) that this distribution has expected value equal to d, which is clearly of the required order. Hence, $p = 1 - \frac{1}{d}$ is a general specification that induces the correct level of sparsity for every A-matrix *a priori*.

Figure 4.14: Animal ERP data for one subject

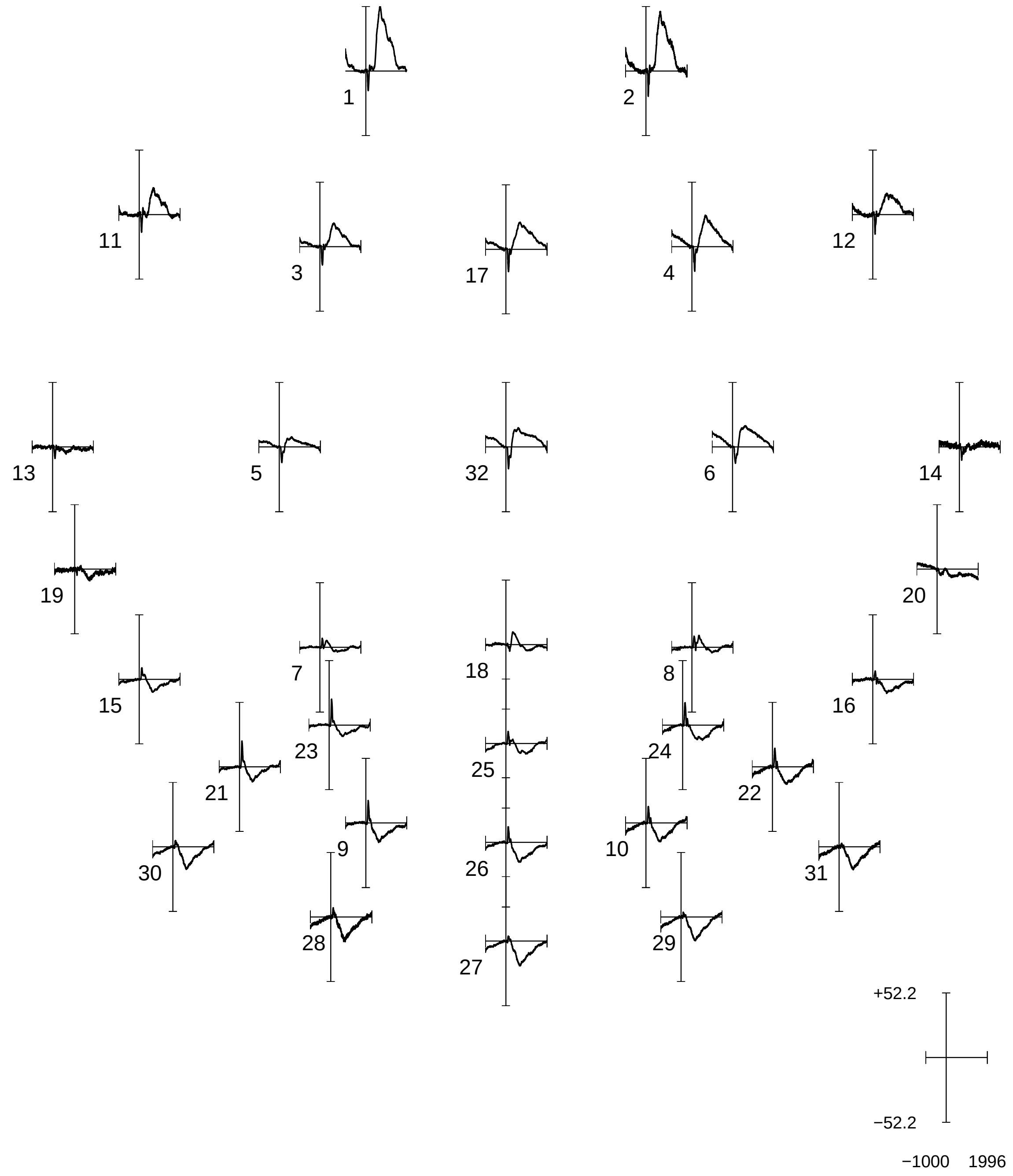
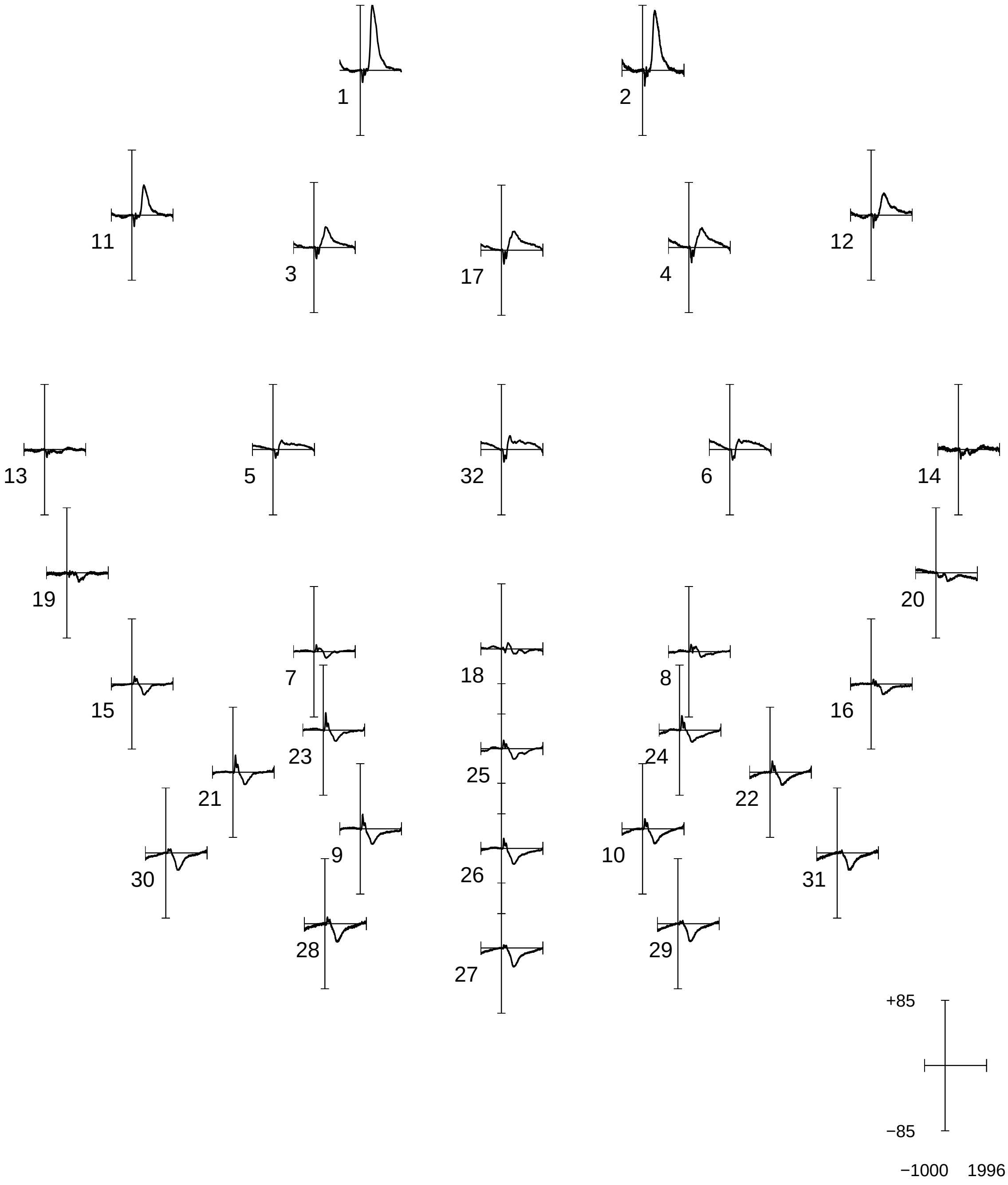


Figure 4.15: Distractor ERP data for one subject



So, in the current scenario, we let $p = \frac{31}{32}$. Then, the MCMC scheme was run for both datasets for $10,000,000$ iterations, with the same burn-in period, thinning interval and choice of G_0 as used previously. The output for the animal and distractor data is shown in Figures 4.16 and 4.17 respectively. In each case, the trace plot indicates that the chain is mixing reasonably well, although slower than has been seen in previous examples. This is a consequence of fewer proposed moves being accepted. Moreover, the ACF plots reveal that the autocorrelation only reduces to approximately zero by lag 100. Of course, it is now beneficial to apply our convergence diagnostics. For the animal data, using CODA, the test of Raftery-Lewis yielded

```
Iterations = 1:9900
Thinning interval = 1
Number of chains = 1
Sample size per chain = 9900

Quantile (q) = 0.025
Accuracy (r) = +/- 0.005
Probability (s) = 0.95

    Burn-in  Total Lower bound  Dependence
    (M)      (N)   (Nmin)       factor (I)
Lq   6       7800  3746         2.08
```

Moreover, Heidelberger-Welch produced the output

```
     Stationarity start     p-value
     test         iteration
Lq   passed       1         0.558

     Halfwidth Mean   Halfwidth
     test
Lq   passed    -31934 0.798
```

Likewise, the Raftery-Lewis diagnostic for the chain that arose from the distractor data supplied the following:

```
Iterations = 1:9900
Thinning interval = 1
Number of chains = 1
```

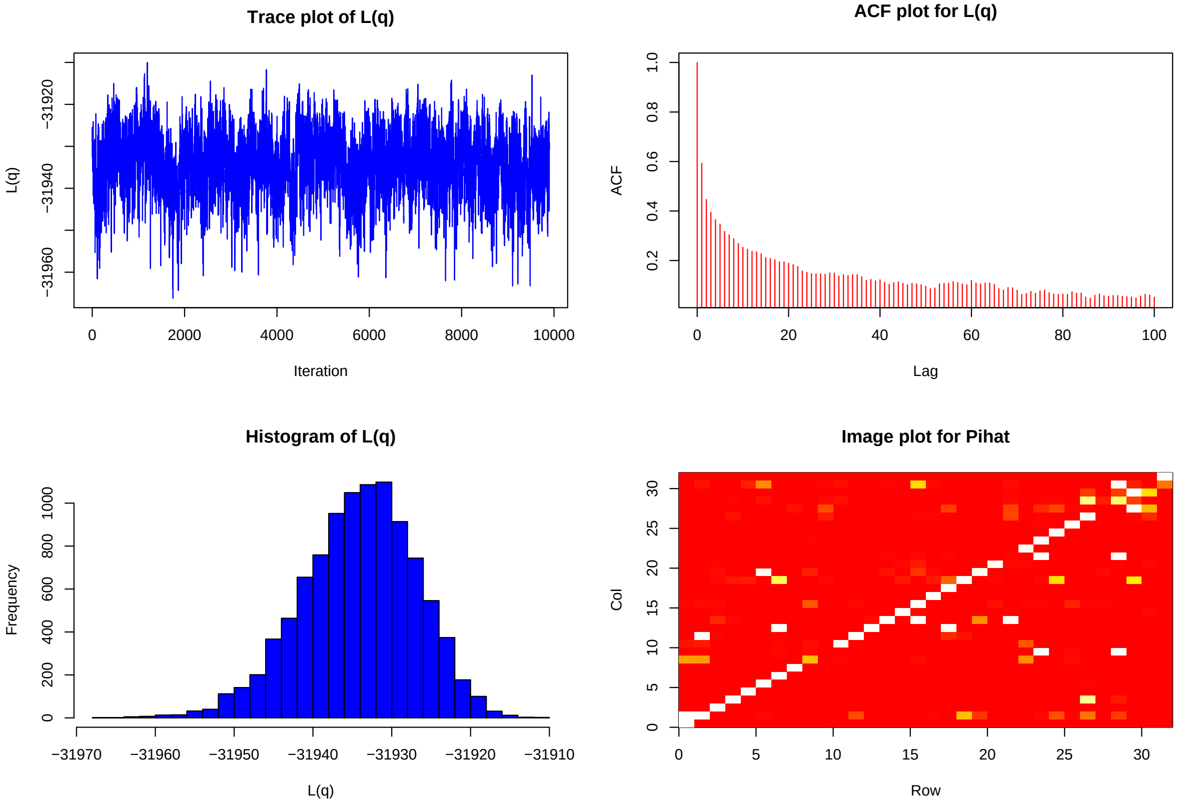


Figure 4.16: Plots for the analysis of the MCMC output for the animal ERP data

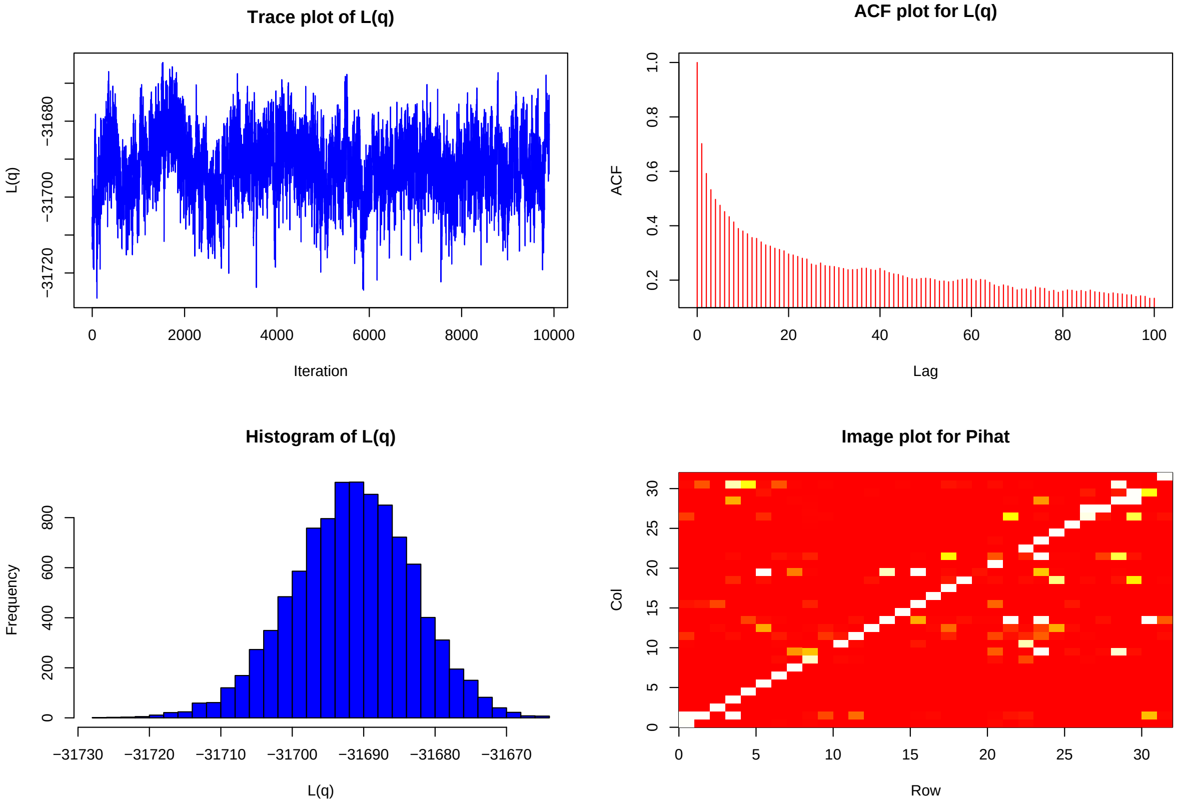


Figure 4.17: Plots for the analysis of the MCMC output for the distractor ERP data

```
Sample size per chain = 9900

Quantile (q) = 0.025
Accuracy (r) = +/- 0.005
Probability (s) = 0.95

    Burn-in  Total Lower bound  Dependence
    (M)      (N)   (Nmin)       factor (I)
Lq   10      10996  3746         2.94
```

Meanwhile, the results of Heidelberger-Welch in this case were

```
     Stationarity start     p-value
     test         iteration
Lq   passed       1         0.143

     Halfwidth Mean   Halfwidth
     test
Lq   passed    -31692 1.23
```

So, these diagnostics still offer good evidence of convergence to the respective stationary distributions in each case. We note however that, due to the slower movement of these chains than before, a longer MCMC run must be used to produce better coverage of the graphical space, and hence more independent chain values. Along with additional thinning, this would improve the trace and ACF plots in Figures 4.16 and 4.17.

We now compare the two `image` plots for the datasets. Here, the axes on these plots correspond to the respective, numbered electrodes, as pictured on Figures 4.14 and 4.15. It is clear that these graphs are quite similar. For instance, when examining the main diagonals, it follows that many nodes have analogous self-loops, and moreover, several common links are discovered elsewhere. An intriguing feature seen in both cases is however that not all nodes possess a self-loop. From Chapter 3, this implies that, on the time series graph, an electrode i at time $t-1$ will not cause a reaction in the same electrode at the next time point t, but instead, stimulate other electrodes on the scalp.

In both cases, it is interesting that many electrodes regularly activate a response in electrodes 24 and 29 that are, independently, in reasonable proximity to these nodes. In fact, the converse is true of electrodes 2 and 14 that act as influencing nodes. Unsurprisingly, differences can be spotted. For the distractor data, numerous edges are widespread between medium and high-numbered electrodes in both directions (as displayed above and below the main diagonal). Additionally, in the animal case, electrode 4 is stimulated less frequently whereas electrode 20 has reduced affect over other electrodes. However, in general, we conclude that, upon display of either a target or non-target image, the cerebral processing of the stimulus is reasonably consistent.

Finally, we can contrast the approximate posterior distributions for a_{ij}, where again $i,\ j = 1,\ 2,\ 3$, for the animal data (Figure 4.18) and distractor data (Figure 4.19). In both cases, we can suggest confidently that a_{13}, a_{21}, a_{23} and a_{31} are all zero in the true specification of A, since $\mathrm{P}_{\text{var}}(a_{ij} = 0 \,|\, D)$ is extremely high for these coefficients. The most captivating plot is that for a_{32}. At first glance, we realise that $\mathrm{P}_{\text{var}}(a_{32} = 0 \,|\, D) \approx 1$ for the two datasets. Yet, despite this, $p_{\text{var}}(a_{32} \,|\, a_{32} \neq 0,\ D)$ takes values that are somewhat distinct from zero. A simple explanation for this is that very few non-zero values of a_{32} were discovered during the MCMC run (as we are extremely certain that this coefficient is zero). Hence, little weight is given to this density, and so a slight perturbation in the variational approximation will produce this imprecision. All other coefficients plotted would appear to be non-zero signals, and the mode values of the corresponding non-zero densities between datasets are most alike.

It is also noticeable that, upon close inspection, some of these densities for each dataset consist of a mixture of components, and thus are multimodal. We have established previously that any node y_j causes a fixed y_i on an A-graph at each MCMC iteration if the element a_{ij}, for $j = 1, \ldots, d$, is non-zero. That is, our attention is on the i-th row of A. Then, the models accepted throughout the sampler will reveal that each y_i can be affected by several nodes to varying degrees; the cumulative effect of this is shown in each `image` plot.

In fact, during the scheme, many different combinations of edges from nodes y_j can influence y_i for numerous j, hence affecting the non-zero value of a particular a_{ik}. So, an intuitive explanation for these multimodal plots is that each peak is a consequence of one such combination. Moreover, the highest peak corresponds to the most likely combination, *i.e.* that which occurs most regularly in the MCMC run. Similarly, the next highest peak will arise from the second most plausible combination, and so on. This feature is intriguing since it was not observed in any previous examples that used simulated data. To understand this, we realise that the ERP datasets possess a more complicated structure whereby, judging from each `image` plot in Figures 4.16 and 4.17, many edges exist between different nodes.

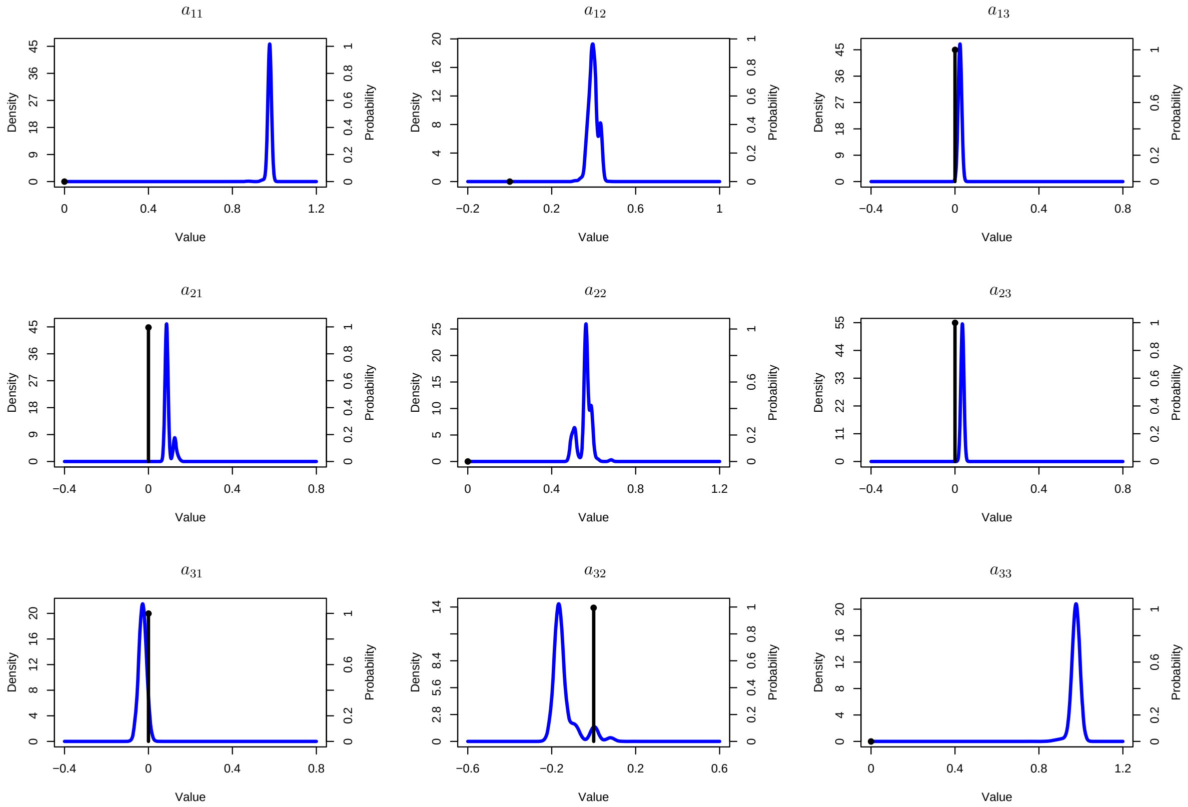


Figure 4.18: Plots showing estimated, marginal posterior distributions for a_{ij}, i, j = 1, 2, 3, for the animal ERP data

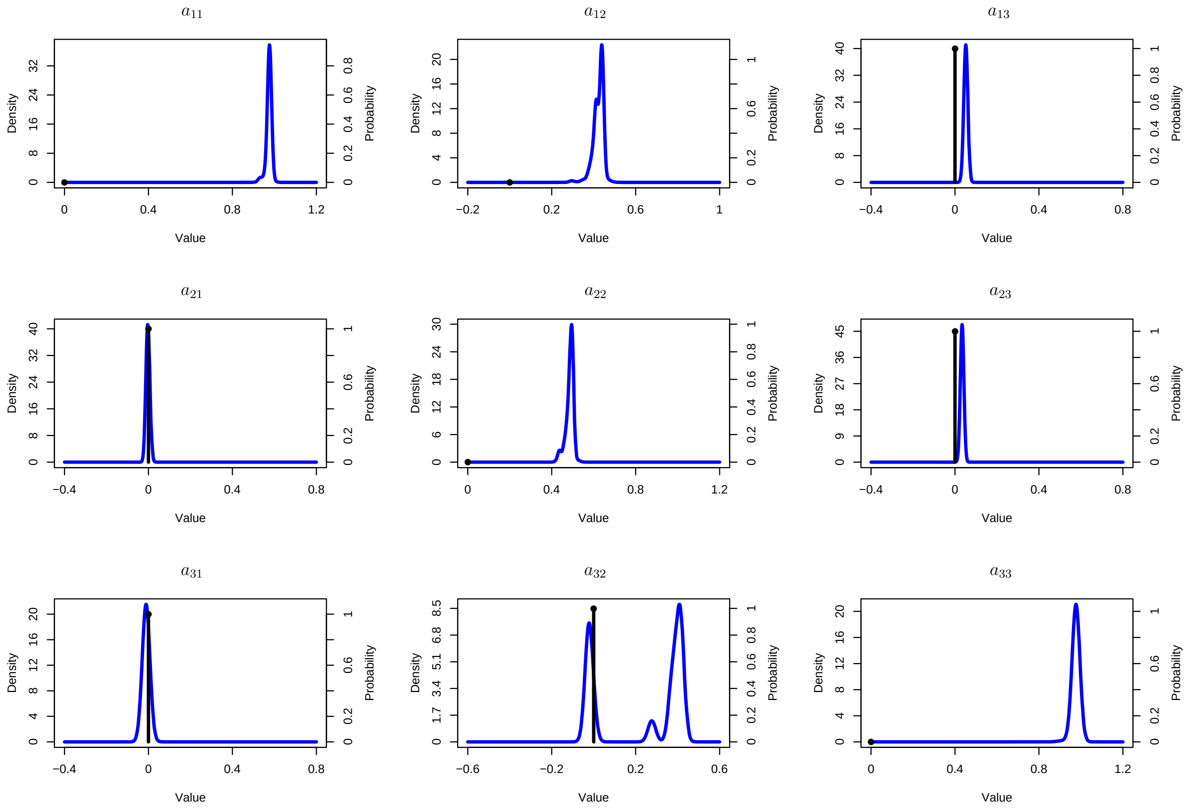


Figure 4.19: Plots showing estimated, marginal posterior distributions for a_{ij}, i, j = 1, 2, 3, for the distractor ERP data

4.3.7 Application to microarray data

Finally, we wish to determine a graphical structure for a set of microarray time series data. A study of gene expression was conducted in the gram-positive bacterium *Bacillus subtilis* whereby $d = 9$ genes are believed to affect the organism's decision on whether to sporulate. Hence, the levels of mRNA are measured for each gene at $N = 40$ time points. For the purposes of understanding the subsequent `image` plot, a number is assigned to every gene, as shown in Table 4.4. As before, the data was centred by subtracting the sample mean at every time point, and modelled with a zero mean VAR(1) process with unknown A and σ^2. The optimum sparsity level is now given as $p = \frac{8}{9}$, whilst all other specifications are preserved.

Number	1	2	3	4	5	6	7	8	9
Gene	spo0A	sda	kinA	lexA	dnaA	spoIIAA	clpP	spo0F	spo0B

Table 4.4: Genes examined in the microarray experiment

Figure 4.20 reveals the results of the MCMC run for this dataset.

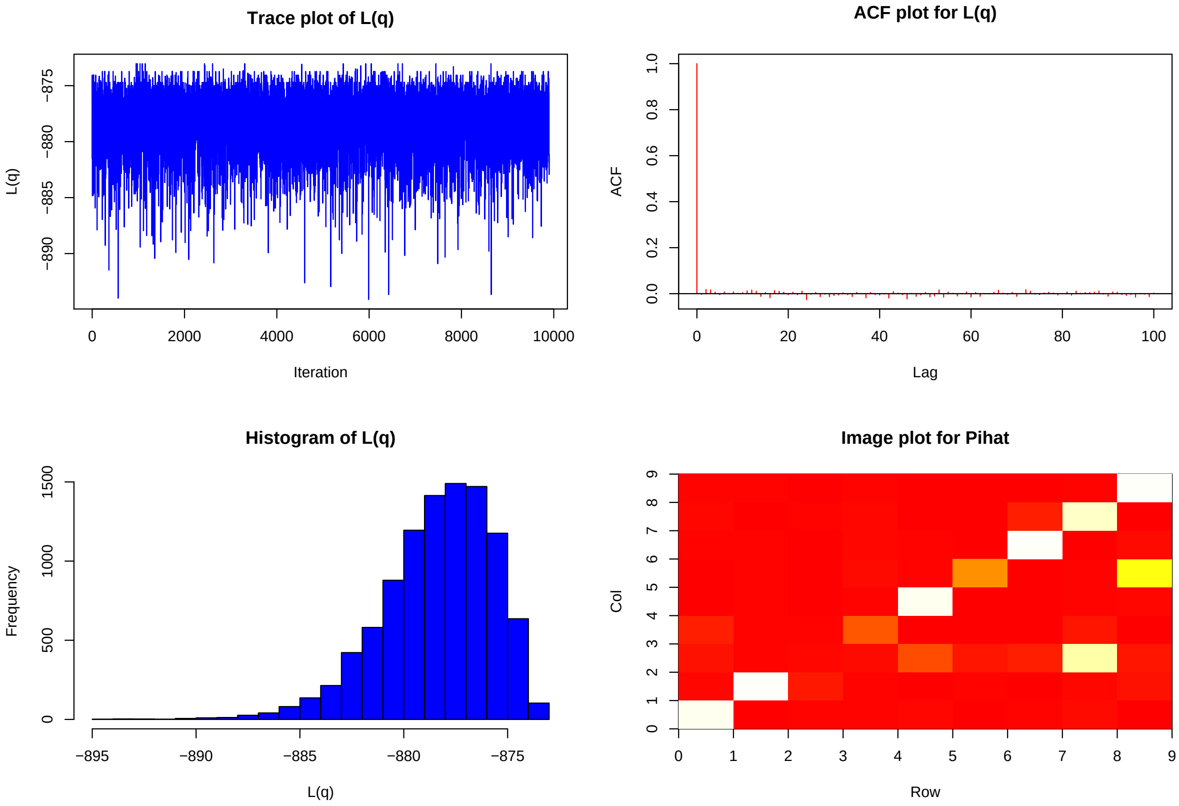


Figure 4.20: Plots for the analysis of the MCMC output for the microarray data

On this occasion, excellent evidence for convergence is displayed by both the trace and ACF plots. This is amplified by the output of the usual diagnostics. For completeness, Raftery-Lewis returned

```
Iterations = 1:9900
Thinning interval = 1
Number of chains = 1
Sample size per chain = 9900
```

```
Quantile (q) = 0.025
Accuracy (r) = +/- 0.005
Probability (s) = 0.95

    Burn-in  Total Lower bound  Dependence
    (M)      (N)   (Nmin)       factor (I)
Lq   1       3748  3746         1
```

whereas Heidelberger-Welch issued

```
     Stationarity start     p-value
     test         iteration
Lq   passed       1         0.215

     Halfwidth Mean   Halfwidth
     test
Lq   passed    -878   0.05
```

To interpret the `image` plot, we refer back to Table 4.4. It follows that many genes stimulate a response in the same gene at consecutive time points, although this is not the case for gene kinA. Moreover, lexA seems to be rather insignificant in the decision making process. Two clear links are spotted between different genes in the study, namely the influence of spoIIAA over spo0B and the reaction caused by kinA in spo0F.

Variational posterior summaries for a set of a_{ij} are provided in Figure 4.21. It appears that only a_{11} and a_{22} are true signals since the estimate of $P(a_{ij} = 0 \,|\, D)$ for all other coefficients is approximately equal to 1. Yet, despite this, these latter entries often possess a plot of $p_{\text{var}}(a_{ij} \,|\, a_{ij} \neq 0,\, D)$ that is peaked away from zero. This is merely a consequence of the size of the dataset — if N was increased, we would anticipate, from previous experience, greater precision in these densities. It is also observed that the likely values of a_{11} and a_{22} are in close proximity.

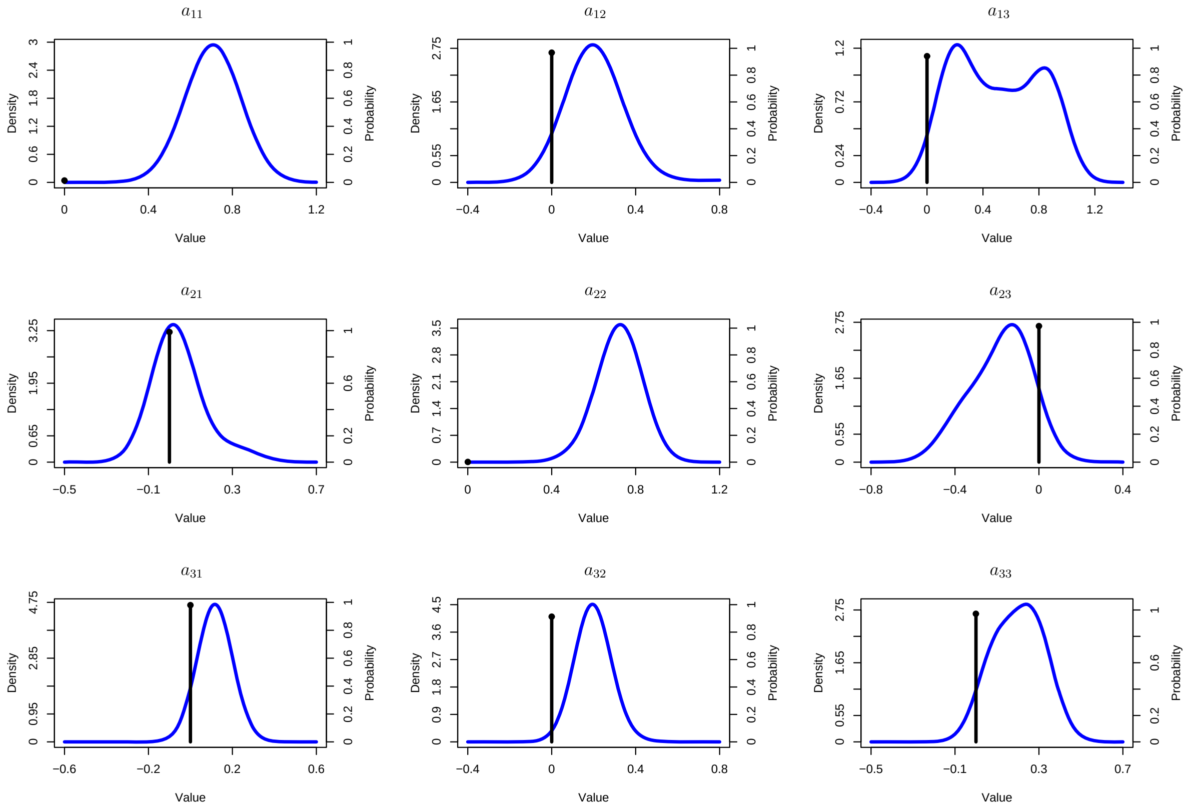


Figure 4.21: Plots showing estimated, marginal posterior distributions for a_{ij}, i, $j = 1, 2, 3$, for the microarray data

4.4 Summary

In this chapter, our main focus was to develop a procedure by which graphical spaces of increasing dimension could be searched, rapidly and effectively, to locate high ranking models. Moreover, we wished to explore the posterior distribution on model space. A Metropolis-Hastings algorithm was constructed whereby, at each iteration, a new model was proposed uniformly, and then accepted on the basis of an acceptance probability. This probability, $\alpha(\cdot, \cdot)$, was itself dependent on the lower bound approximation to the logarithm of the marginal likelihood for each model, derived by variational Bayes. A matrix $\hat{\Pi}$ was constructed to count the number of occasions that an edge between any pair of nodes was included in the accepted model at each iteration. The algorithm was applied to a variety of simulated examples where, *inter alia*, the prior probability $p = P(a_{ij} = 0)$ was altered. Throughout, the true sparsity structure of A was identified with impressive accuracy.

It was also possible to determine an approximate, marginal posterior distribution for each a_{ij} across models. That is, we were able to display graphically the probability that a particular a_{ij} was zero, and the probable magnitude of a_{ij}, given that it was non-zero. For the simulated data, it was feasible to ascertain which a_{ij} were actually signals, and, in that case, estimate their true values precisely. Finally, the MCMC scheme was applied to two real cases, namely a set of ERP and microarray data. For the former, we concluded that there are many similarities in processing the visual information when correctly recognising either an animal or distractor photograph, a decision made in equivalent brain areas. Moreover, in the latter, a gene network was identified, regarding the decision of a bacterium to sporulate. In the next chapter, the ideas presented here, and in Chapter 3, are extended, so that the variational Bayesian treatment is provided to VAR(1) models that are no-longer assumed to have zero mean.

Chapter 5

Generalisation to non-zero mean VAR(1) models

5.1 Introduction

In Chapter 3, a candidate set of VAR(1) graphical models was constructed, each reliant upon the differing sparsity structure of the autoregressive matrix A. The evidence for each model, given by the corresponding marginal likelihood, was approximated using the variational Bayesian algorithm. Hence, it was possible to determine the more plausible models in the set.

Recall that, in general, a VAR(1) model of dimension d is specified as

$$\mathbf{y}_t = \mathbf{y}_{t-1} A + \mathbf{e}_t,$$

with the noise vector $\mathbf{e}_t \sim \mathcal{N}(\mathbf{0}, \Gamma)$. However, alternatively, the model can be rewritten in mean-adjusted form (Lütkepohl, 2005) such that

$$\mathbf{y}_t - \boldsymbol{\mu} = (\mathbf{y}_{t-1} - \boldsymbol{\mu})A + \mathbf{e}_t, \tag{5.1}$$

whereby $\boldsymbol{\mu} = (\mu_1,\, \mu_2,\, \ldots,\, \mu_d) = \mathrm{E}(\mathbf{y}_t)$ for all t, a $(1 \times d)$ vector, and the noise is distributed as before. All other dimensions are as previous. (5.1) is now classified as the *non-zero mean* VAR(1) model. Of course, as Lütkepohl informs us, this generalisation can be extended further to the VAR(p) process.

Therefore, using again the graphical representation of the sparse matrix A, we can conduct a similar procedure to that of Chapter 3 by applying the variational algorithm to derive a lower bound, $\mathcal{L}_{M_i}(q_i)$, for each graphical model M_i, an approximation to the logarithm of the corresponding marginal likelihood for each model, $p(\{\mathbf{y}_t\} \,|\, M_i)$. Consequently, by ranking models, we seek primarily to learn the sparsity structure of A. However, analysis is now complicated since, by model (5.1), the parameter set is now defined as $\boldsymbol{\theta} = \{A,\, \sigma^2,\, \boldsymbol{\mu}\}$. Here, as erstwhile, we let the covariance matrix of the noise vector be $\Gamma = \sigma^2 \mathrm{I}_d$ for ease of prior specification. Thus, by the definition of the lower bound (2.6), we extend previous work by now placing a further prior on $\boldsymbol{\mu}$, and subsequently deducing an additional variational distribution for this parameter. Henceforth, conditioning and dependence on M_i is assumed throughout, although not expressed explicitly.

5.2 Scoring non-zero mean VAR(1) models

Suppose that $t = 1, \ldots, N$ independent samples of the time series have been collected. Initially, by rewriting (5.1) as a matrix equation, we derive an expression for the data likelihood. Define Y, X and E, with the same dimension, as in Section 3.3, whereby $\mathbf{x}_t = [\mathbf{y}_{t-1}]$ for all t. Moreover, let M be a $N \times d$ matrix, with each row of M given by the vector $\boldsymbol{\mu}$. Thus, we designate (5.1) in matrix form as

$$Y - M = [X - M]A + E. \tag{5.2}$$

At the end-points, we again allow $\mathbf{x}_N = \mathbf{y}_{N-1}$. Moreover, we assume the process is initialised such that $\mathbf{x}_1 = \mathbf{y}_0 = \boldsymbol{\mu}$. As $\mathrm{E}(\mathbf{y}_t) = \boldsymbol{\mu}$ for all t, this is, by definition, the stationary mean. Recall that this value exists if all eigenvalues of A have modulus less than 1, hence the VAR(1) process is referred to as stable.

Now, decompose (5.2) into vector form such that

$$\begin{aligned}
\mathrm{vec}(Y - M) &= \mathrm{vec}([X - M]A + E) \\
\implies \mathrm{vec}(Y) - \mathrm{vec}(M) &= \mathrm{vec}(XA) - \mathrm{vec}(MA) + \mathrm{vec}(E) \\
\implies \mathbf{y} - \mathbf{m} &= (\mathrm{I}_d \otimes X)\mathbf{a} - (\mathrm{I}_d \otimes M)\mathbf{a} + \mathbf{e}, \qquad (5.3)
\end{aligned}$$

using the same results as in Section 3.3. Moreover, let $\mathrm{vec}(M) = \mathbf{m}$, a $dN \times 1$ vector, and allow all other definitions as before. Determination of the probability density function for $\mathbf{e}$ is exactly as before since, again, $\mathbf{e}_t \sim \mathcal{N}(\mathbf{0},\, \sigma^2\mathrm{I}_d)$. Consequently, $\mathbf{e} \sim \mathcal{N}(\mathbf{0},\, \mathrm{I}_d \otimes \sigma^2\mathrm{I}_N)$. Hence, we can rearrange (5.3) in terms of $\mathbf{e}$, and substitute into the probability density function of $\mathbf{e}$, (3.6). The exponent of this expression is then

$$\begin{aligned}
&\exp\left\{-\frac{1}{2}\mathbf{e}^T(\mathrm{I}_d \otimes \sigma^{-2}\mathrm{I}_N)\mathbf{e}\right\} \\
&= \exp\biggl\{ -\frac{1}{2}\left[\mathbf{y} - \mathbf{m} - (\mathrm{I}_d \otimes X)\mathbf{a} + (\mathrm{I}_d \otimes M)\mathbf{a}\right]^T (\mathrm{I}_d \otimes \sigma^{-2}\mathrm{I}_N) \\
&\qquad\qquad \times \left[\mathbf{y} - \mathbf{m} - (\mathrm{I}_d \otimes X)\mathbf{a} + (\mathrm{I}_d \otimes M)\mathbf{a}\right] \biggr\} \\
&= \exp\left\{-\frac{1}{2}\left[\mathrm{vec}(Y - M - [X - M]A)\right]^T (\mathrm{I}_d \otimes \sigma^{-2}\mathrm{I}_N)\left[\mathrm{vec}(Y - M - [X - M]A)\right]\right\} \\
&= \exp\left\{-\frac{1}{2}\mathrm{Tr}\left[\sigma^{-2}g(A,\, M)\right]\right\}, \qquad (5.4)
\end{aligned}$$

where $g(A,\, M) = (Y - M - [X - M]A)^T(Y - M - [X - M]A)$, a $d \times d$ matrix and using corresponding results. Resultantly, given a data-set $D = \{X,\, Y\}$, the probability of the data is such that

$$p(D\,|A,\, \sigma^2,\, \boldsymbol{\mu}) = (2\pi\sigma^2)^{-\frac{dN}{2}} \exp\left\{-\frac{1}{2}\mathrm{Tr}\left[\sigma^{-2}g(A,\, M)\right]\right\}. \tag{5.5}$$

5.2.1 Priors

As mentioned in Section 5.1, we must denote priors over, not only $\mathbf{a} = \mathrm{vec}(A)$ and σ^2 as before, but also $\boldsymbol{\mu}$. Therefore, the specifications are

$$p(\mathbf{a}) = \mathcal{N}(\mathbf{a}\,|\,\mathbf{0},\, C^*) \tag{5.6}$$

$$p(\sigma^2) = \mathcal{IG}(\sigma^2\,|\,\alpha,\, \beta) \tag{5.7}$$

$$p(\boldsymbol{\mu}) = \mathcal{N}(\boldsymbol{\mu}\,|\,\mathbf{b},\, \Delta). \tag{5.8}$$

There is no motivation to change the priors on both $\mathbf{a}$ and σ^2, and so these are maintained from Section 3.3.1. Again, we define $C^* = \mathrm{diag}\{\mathrm{vec}(C)\}$, where C is indicated by equation (3.11). We realise that C^* can be rank deficient, hence creating problems in later analysis. Thus, the sparsity structure on A is retained in the prior by constraining C. The prior on $\boldsymbol{\mu}$ was chosen since the multivariate normal distribution is the typical conjugate choice for the unknown mean vector of a normal random sample. Of course, these specifications are assumed to be independent such that $p(\mathbf{a},\, \sigma^2,\, \boldsymbol{\mu}) = p(\mathbf{a})p(\sigma^2)p(\boldsymbol{\mu})$.

5.2.2 Free form method

When considering non-zero mean VAR(1) models, the variational distributions, which approximate each true posterior, are here defined to be $q(\mathbf{a}\,|\,D)$, $q(\sigma^2\,|\,D)$ and $q(\boldsymbol{\mu}\,|\,D)$.

We continue by applying the same procedure as in Chapter 3. That is, a free form variational method is implemented initially to derive the variationals for σ^2, $\boldsymbol{\mu}$ and $\mathbf{a}$ respectively in the dense case. Furthermore, a fixed form technique is then employed to calculate the lower bound and, more importantly, to constrain as a result of a given sparsity structure.

From a free form approach, we only assume independence between the variational distributions, *i.e.* $q(\mathbf{a},\, \sigma^2,\, \boldsymbol{\mu} \,|\, D) = q(\mathbf{a} \,|\, D)\, q(\sigma^2 \,|\, D)\, q(\boldsymbol{\mu} \,|\, D)$. This is an approximation since $\mathbf{a}$, σ^2 and $\boldsymbol{\mu}$ are not *a posteriori* independent. This was seen previously for zero mean VAR(1) models, and hence still holds in the more advanced situation here.

By definition, the lower bound is specified as

$$\mathcal{L}_{\boldsymbol{\mu}}(q) = \iiint q(\mathbf{a}, \sigma^2, \boldsymbol{\mu} \,|\, D)\, \log \left[\frac{p(D \,|\, A,\, \sigma^2,\, \boldsymbol{\mu})\, p(\mathbf{a},\, \sigma^2,\, \boldsymbol{\mu})}{q(\mathbf{a}, \sigma^2, \boldsymbol{\mu} \,|\, D)} \right] \mathrm{d}\mathbf{a}\, \mathrm{d}\sigma^2\, \mathrm{d}\boldsymbol{\mu}, \tag{5.9}$$

where the subscript $\boldsymbol{\mu}$ on $\mathcal{L}(q)$ represents the non-zero mean model. As seen formerly, this equation can be rewritten as a sum of integrals using the independence of the distributions involved, and simplified further by integrating out parameters when necessary. Hence, by recombining integrals, we can denote $\mathcal{L}_{\boldsymbol{\mu}}(q)$ as a functional of $q(\mathbf{a} \,|\, D)$, $q(\sigma^2 \,|\, D)$ and $q(\boldsymbol{\mu} \,|\, D)$. These are given respectively by equations (5.10), (5.11) and (5.12) below.

$$\begin{aligned}\mathcal{L}_{\boldsymbol{\mu}}(q) = \int q(\mathbf{a} \,|\, D) \bigg[\iint & q(\sigma^2 \,|\, D)\, q(\boldsymbol{\mu} \,|\, D)\, \log p(D \,|\, A,\, \sigma^2,\, \boldsymbol{\mu})\, \mathrm{d}\sigma^2\, \mathrm{d}\boldsymbol{\mu} \\ & + \log p(\mathbf{a}) - \log q(\mathbf{a} \,|\, D) \bigg] \mathrm{d}\mathbf{a} \,+\, \text{const.}\end{aligned} \tag{5.10}$$

$$\begin{aligned}\mathcal{L}_{\boldsymbol{\mu}}(q) = \int q(\sigma^2 \,|\, D) \bigg[\iint & q(\mathbf{a} \,|\, D)\, q(\boldsymbol{\mu} \,|\, D)\, \log p(D \,|\, A,\, \sigma^2,\, \boldsymbol{\mu})\, \mathrm{d}\mathbf{a}\, \mathrm{d}\boldsymbol{\mu} \\ & + \log p(\sigma^2) - \log q(\sigma^2 \,|\, D) \bigg] \mathrm{d}\sigma^2 \,+\, \text{const.}\end{aligned} \tag{5.11}$$

$$\begin{aligned}\mathcal{L}_{\boldsymbol{\mu}}(q) = \int q(\boldsymbol{\mu} \,|\, D) \bigg[\iint & q(\mathbf{a} \,|\, D)\, q(\sigma^2 \,|\, D)\, \log p(D \,|\, A,\, \sigma^2,\, \boldsymbol{\mu})\, \mathrm{d}\mathbf{a}\, \mathrm{d}\sigma^2 \\ & + \log p(\boldsymbol{\mu}) - \log q(\boldsymbol{\mu} \,|\, D) \bigg] \mathrm{d}\boldsymbol{\mu} \,+\, \text{const.}\end{aligned} \tag{5.12}$$

We shall tackle each of these integrals. Examine first (5.11). In this equation, we can substitute for both $\log p(D \mid A,\, \sigma^2,\, \boldsymbol{\mu})$ and $\log p(\sigma^2)$. Upon comparison, we notice the similarity between (3.15) and (5.11), the functionals of $q(\sigma^2 \mid D)$ in both the zero and non-zero mean cases. Moreover, as the prior on σ^2 is identical and the data likelihood of similar form to the zero mean circumstance, by dropping terms independent of σ^2 which disappear upon differentiation with respect to $q(\sigma^2 \mid D)$, we will arrive at an expression that is akin to (3.16). Thus, we acquire

$$\begin{aligned}\mathcal{L}_{\boldsymbol{\mu}}(q) = \int q(\sigma^2 \mid D) \Bigg[&- \frac{dN}{2} \log \sigma^2 - \frac{(\sigma^2)^{-1}}{2} \iint q(\mathbf{a} \mid D)\, q(\boldsymbol{\mu} \mid D) \left\{ \mathrm{Tr}\ [g(A,\, M)] \right\}\, \mathrm{d}\mathbf{a}\, \mathrm{d}\boldsymbol{\mu} \\ &- (\alpha + 1)\, \log \sigma^2 - \beta (\sigma^2)^{-1} - \log q(\sigma^2 \mid D) \Bigg] \mathrm{d}\sigma^2 + \mathrm{const.}' \end{aligned} \tag{5.13}$$

Now notice, by the definition of $g(A,\, M)$ and the derivation of (5.3), that,

$$\begin{aligned}\mathrm{Tr}\ [g(A,\, M)] &= [\mathrm{vec}(Y - M - [X - M]A)]^T\, [\mathrm{vec}(Y - M - [X - M]A)] \\ &= [\mathbf{y} - \mathbf{m} - (\mathrm{I}_d \otimes X)\mathbf{a} + (\mathrm{I}_d \otimes M)\mathbf{a}]^T\, [\mathbf{y} - \mathbf{m} - (\mathrm{I}_d \otimes X)\mathbf{a} + (\mathrm{I}_d \otimes M)\mathbf{a}] \\ &\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad (5.14) \\ &=: h(\mathbf{a},\, M),\end{aligned}$$

again using identity (3.18), and defining the function h to ease the algebra. Hence, we take the expectation of this expression with respect to the variational distributions for both $\mathbf{a}$ and $\boldsymbol{\mu}$. Thus, we compute

$$\begin{aligned}&\mathrm{E}_{q(\mathbf{a} \mid D)} \left\{ \mathrm{E}_{q(\boldsymbol{\mu} \mid D)} \left\{ \mathrm{Tr}\ [g(A,\, M)] \right\} \right\} \\ &= \mathbf{y}^T\mathbf{y} - [\mathrm{vec}(\Omega)]^T\, \mathbf{y} - \boldsymbol{\rho}^T (\mathrm{I}_d \otimes X^T)\mathbf{y} + \boldsymbol{\rho}^T (\mathrm{I}_d \otimes \Omega^T)\mathbf{y} - \mathbf{y}^T \mathrm{vec}(\Omega) \\ &\qquad + \mathrm{E}_{q(\boldsymbol{\mu} \mid D)}\{\mathbf{m}^T\} \mathrm{E}_{q(\boldsymbol{\mu} \mid D)}\{\mathbf{m}\} + \mathrm{Tr}[\mathrm{Var}_{q(\boldsymbol{\mu} \mid D)}\{\mathbf{m}\}] + \boldsymbol{\rho}^T (\mathrm{I}_d \otimes X^T) \mathrm{vec}(\Omega) \\ &\qquad - \boldsymbol{\rho}^T \mathrm{E}_{q(\boldsymbol{\mu} \mid D)} \left\{ \mathrm{vec}(M^T M) \right\} - \mathbf{y}^T (\mathrm{I}_d \otimes X)\boldsymbol{\rho} + [\mathrm{vec}(\Omega)]^T\, (\mathrm{I}_d \otimes X)\boldsymbol{\rho} \\ &\qquad + \mathrm{E}_{q(\mathbf{a} \mid D)}\{\mathbf{a}^T\} (\mathrm{I}_d \otimes X^T X) \mathrm{E}_{q(\mathbf{a} \mid D)}\{\mathbf{a}\} + \mathrm{Tr}[(\mathrm{I}_d \otimes X^T X) \mathrm{Var}_{q(\mathbf{a} \mid D)}\{\mathbf{a}\}]\end{aligned}$$

$$
\begin{aligned}
&- \mathrm{E}_{q(\mathbf{a}\,|\,D)}\{\mathbf{a}^T\}(\mathrm{I}_d \otimes \Omega^T X)\mathrm{E}_{q(\mathbf{a}\,|\,D)}\{\mathbf{a}\} - \mathrm{Tr}[(\mathrm{I}_d \otimes \Omega^T X)\mathrm{Var}_{q(\mathbf{a}\,|\,D)}\{\mathbf{a}\}] + \mathbf{y}^T(\mathrm{I}_d \otimes \Omega)\boldsymbol{\rho} \\
&- \mathrm{E}_{q(\boldsymbol{\mu}\,|\,D)}\left\{\left[\mathrm{vec}(M^T M)\right]^T\right\}\boldsymbol{\rho} - \mathrm{E}_{q(\mathbf{a}\,|\,D)}\{\mathbf{a}^T\}(\mathrm{I}_d \otimes X^T\Omega)\mathrm{E}_{q(\mathbf{a}\,|\,D)}\{\mathbf{a}\} \\
&\quad - \mathrm{Tr}[(\mathrm{I}_d \otimes X^T\Omega)\mathrm{Var}_{q(\mathbf{a}\,|\,D)}\{\mathbf{a}\}] + \mathrm{E}_{q(\mathbf{a}\,|\,D)}\left\{\mathbf{a}^T\left(\mathrm{I}_d \otimes \mathrm{E}_{q(\boldsymbol{\mu}\,|\,D)}\{M^T M\}\right)\mathbf{a}\right\}. \qquad (5.15)
\end{aligned}
$$

Here, we define $\boldsymbol{\rho} = \mathrm{E}_{q(\mathbf{a}\,|\,D)}\{\mathbf{a}\}$ as before. In addition, let $\mathrm{E}_{q(\boldsymbol{\mu}\,|\,D)}\{\boldsymbol{\mu}\} = \boldsymbol{\omega}$ and, moreover, $\mathrm{E}_{q(\boldsymbol{\mu}\,|\,D)}\{M\} = \Omega$. That is, by construction of M, Ω is the $N \times d$ matrix with each row equal to the d-vector $\boldsymbol{\omega}$. Furthermore, $\mathrm{E}_{q(\boldsymbol{\mu}\,|\,D)}\{\mathbf{m}\} = \mathrm{vec}(\mathrm{E}_{q(\boldsymbol{\mu}\,|\,D)}\{M\}) = \mathrm{vec}(\Omega)$. To derive (5.15), the identities (3.20) and (2.51), used at the corresponding stage in Chapter 3, have been applied, and moreover (3.5) (by setting $P = M^T$, $Q = M$). We also realise that $\mathrm{E}_{q(\boldsymbol{\mu}\,|\,D)}\{\mathrm{I}_d \otimes M\} = \mathrm{I}_d \otimes \mathrm{E}_{q(\boldsymbol{\mu}\,|\,D)}\{M\}$ since $\mathrm{I}_d \otimes M$ is merely the block diagonal matrix where each block is equivalent to M. Taking expectations here is simplified by the variational *a posteriori* independence between $\mathbf{a}$ and $\boldsymbol{\mu}$.

We now account for the other terms in (5.15). As erstwhile, let $\tau = \mathrm{Var}_{q(\mathbf{a}\,|\,D)}\{\mathbf{a}\}$. Moreover, define $\mathrm{Var}_{q(\boldsymbol{\mu}\,|\,D)}\{\boldsymbol{\mu}\} = \Lambda$ and $\mathrm{Var}_{q(\boldsymbol{\mu}\,|\,D)}\{\mathbf{m}\} = \Xi$. In the above expression, we must find $\mathrm{Tr}[\Xi]$. However, as $\mathbf{m} = (\mu_1, \mu_1, \ldots, \mu_1, \mu_2, \mu_2, \ldots, \mu_2, \ldots, \mu_d, \mu_d, \ldots \mu_d)^T$, with each component repeated N times by definition, it follows that

$$
\begin{aligned}
\mathrm{Tr}[\Xi] &= N(\mathrm{Var}_{q(\boldsymbol{\mu}\,|\,D)}\{\mu_1\} + \mathrm{Var}_{q(\boldsymbol{\mu}\,|\,D)}\{\mu_2\} + \cdots + \mathrm{Var}_{q(\boldsymbol{\mu}\,|\,D)}\{\mu_d\}) \\
&= N\,\mathrm{Tr}[\mathrm{Var}_{q(\boldsymbol{\mu}\,|\,D)}\{\boldsymbol{\mu}\}] \\
&= N\,\mathrm{Tr}[\Lambda]. \qquad (5.16)
\end{aligned}
$$

We now endeavour to find $\mathrm{E}_{q(\boldsymbol{\mu}\,|\,D)}\left\{\mathrm{vec}(M^T M)\right\} = \mathrm{vec}\left(\mathrm{E}_{q(\boldsymbol{\mu}\,|\,D)}\left\{M^T M\right\}\right)$. It helps to inspect this problem in component form. So, let $M = (m_{ij})$ such that $M^T = (m_{ji})$. Then, by definition of matrix multiplication,

$$
[M^T M]_{ij} = \sum_{k=1}^{N} m_{ki} m_{kj} = \sum_{k=1}^{N} \mu_i \mu_j = N\mu_i\mu_j,
$$

since each element of the j-th column of M is μ_j. The expectation of this expression with respect to $q(\boldsymbol{\mu}\,|\,D)$ is now taken. Thus,

$$\begin{aligned}
\mathrm{E}_{q(\boldsymbol{\mu}\,|\,D)}\{[M^TM]_{ij}\} &= N\,\mathrm{E}_{q(\boldsymbol{\mu}\,|\,D)}\{\mu_i\mu_j\} \\
&= N\,\mathrm{E}_{q(\boldsymbol{\mu}\,|\,D)}\{\mu_i\}\mathrm{E}_{q(\boldsymbol{\mu}\,|\,D)}\{\mu_j\} + N\mathrm{Cov}_{q(\boldsymbol{\mu}\,|\,D)}\{\mu_i,\,\mu_j\} \\
&= \sum_{k=1}^{N}\mathrm{E}_{q(\boldsymbol{\mu}\,|\,D)}\{m_{ki}\}\mathrm{E}_{q(\boldsymbol{\mu}\,|\,D)}\{m_{kj}\} + N\left[\mathrm{Var}_{q(\boldsymbol{\mu}\,|\,D)}\{\boldsymbol{\mu}\}\right]_{ij} \\
&= \left[\mathrm{E}_{q(\boldsymbol{\mu}\,|\,D)}\{M^T\}\mathrm{E}_{q(\boldsymbol{\mu}\,|\,D)}\{M\}\right]_{ij} + N\left[\mathrm{Var}_{q(\boldsymbol{\mu}\,|\,D)}\{\boldsymbol{\mu}\}\right]_{ij},
\end{aligned}$$

using the definition of covariance. Consequently, by removing subscripts and returning to matrix form, we obtain the identity

$$\mathrm{E}_{q(\boldsymbol{\mu}\,|\,D)}\{M^TM\} = \Omega^T\Omega + N\Lambda, \tag{5.17}$$

and hence $\mathrm{E}_{q(\boldsymbol{\mu}\,|\,D)}\left\{\mathrm{vec}(M^TM)\right\} = \mathrm{vec}(\Omega^T\Omega) + N\mathrm{vec}(\Lambda)$. As a result, the final term of (5.15) is now equal to

$$\begin{aligned}
&\mathrm{E}_{q(\mathbf{a}\,|\,D)}\left\{\mathbf{a}^T\left(\mathrm{I}_d\otimes\mathrm{E}_{q(\boldsymbol{\mu}\,|\,D)}\{M^TM\}\right)\mathbf{a}\right\} \\
&= \mathrm{E}_{q(\mathbf{a}\,|\,D)}\left\{\mathbf{a}^T\right\}\left(\mathrm{I}_d\otimes[\Omega^T\Omega+N\Lambda]\right)\mathrm{E}_{q(\mathbf{a}\,|\,D)}\left\{\mathbf{a}\right\} + \mathrm{Tr}[\left(\mathrm{I}_d\otimes[\Omega^T\Omega+N\Lambda]\right)\mathrm{Var}_{q(\mathbf{a}\,|\,D)}\{\mathbf{a}\}] \\
&= \boldsymbol{\rho}^T\left(\mathrm{I}_d\otimes[\Omega^T\Omega+N\Lambda]\right)\boldsymbol{\rho} + \mathrm{Tr}[\left(\mathrm{I}_d\otimes[\Omega^T\Omega+N\Lambda]\right)\tau].
\end{aligned}$$

By (5.14), the expression (5.15) can be simplified somewhat. In fact, we reach

$$\begin{aligned}
&\mathrm{E}_{q(\mathbf{a}\,|\,D)}\left\{\mathrm{E}_{q(\boldsymbol{\mu}\,|\,D)}\left\{\mathrm{Tr}\ [g(A,\,M)]\right\}\right\} \\
&= h(\boldsymbol{\rho},\,\Omega) + N\,\mathrm{Tr}[\Lambda] - N\boldsymbol{\rho}^T\mathrm{vec}(\Lambda) + \mathrm{Tr}[(\mathrm{I}_d\otimes X^TX)\tau] - \mathrm{Tr}[(\mathrm{I}_d\otimes\Omega^TX)\tau] \\
&\quad - N[\mathrm{vec}(\Lambda)]^T\boldsymbol{\rho} - \mathrm{Tr}[(\mathrm{I}_d\otimes X^T\Omega)\tau] + N\boldsymbol{\rho}^T(\mathrm{I}_d\otimes\Lambda)\boldsymbol{\rho} \\
&\quad + \mathrm{Tr}[(\mathrm{I}_d\otimes\Omega^T\Omega)\tau] + N\,\mathrm{Tr}[(\mathrm{I}_d\otimes\Lambda)\tau]
\end{aligned}$$

$$
\begin{aligned}
&= h(\boldsymbol{\rho},\, \Omega) + N\,\mathrm{Tr}[\Lambda] + N\boldsymbol{\rho}^T\left[(\mathrm{I}_d \otimes \Lambda)\boldsymbol{\rho} - 2\mathrm{vec}(\Lambda)\right] \\
&\qquad + \mathrm{Tr}\left[\left\{\mathrm{I}_d \otimes \left[(X-\Omega)^T(X-\Omega) + N\Lambda\right]\right\}\tau\right] \\
&=: h(\boldsymbol{\rho},\, \Omega) + j(\boldsymbol{\rho},\, \tau,\, \Omega,\, \Lambda), \qquad (5.18)
\end{aligned}
$$

introducing a further function j to simplify notation. Now, substitute (5.18) into (5.13) so that the latter is now independent of $\mathbf{a}$ and $\boldsymbol{\mu}$. As in Section 3.3.2, we use the Lagrange multiplier ν_{σ^2} to construct the functional $\tilde{\mathcal{L}}_{\boldsymbol{\mu}}(q)$ (*c.f.* (3.21)). Maximising this functional with respect to $q(\sigma^2 \,|\, D)$ then provides

$$
\begin{aligned}
\frac{\partial \tilde{\mathcal{L}}_{\boldsymbol{\mu}}(q)}{\partial q(\sigma^2 \,|\, D)} = -\frac{dN}{2}\log\sigma^2 - \frac{(\sigma^2)^{-1}}{2}\left[h(\boldsymbol{\rho},\, \Omega) + j(\boldsymbol{\rho},\, \tau,\, \Omega,\, \Lambda)\right] \\
- (\alpha+1)\,\log\sigma^2 - \beta(\sigma^2)^{-1} - \log q(\sigma^2 \,|\, D) - 1 + \nu_{\sigma^2} = 0.
\end{aligned}
$$

When comparing this expression with the zero mean case, it is clear that

$$
q(\sigma^2 \,|\, D) \propto (\sigma^2)^{-(\alpha+\frac{dN}{2}+1)}\,\exp\left\{-(\sigma^2)^{-1}\left(\beta + \frac{1}{2}\left[h(\boldsymbol{\rho},\, \Omega) + j(\boldsymbol{\rho},\, \tau,\, \Omega,\, \Lambda)\right]\right)\right\}.
$$

Therefore, using the same notation as before, we can easily see that, as in Chapter 3,

$$
q(\sigma^2 \,|\, D) = \mathcal{IG}(\sigma^2 \,|\, \gamma,\, \delta), \qquad (5.19)
$$

but now with variational parameters specified as

$$
\begin{aligned}
\gamma &= \alpha + \frac{dN}{2} && (5.20) \\
\delta &= \beta + \frac{1}{2}\left[h(\boldsymbol{\rho},\, \Omega) + j(\boldsymbol{\rho},\, \tau,\, \Omega,\, \Lambda)\right], && (5.21)
\end{aligned}
$$

whereby

$$\begin{aligned} h(\boldsymbol{\rho},\,\Omega) &= \left[\mathbf{y} - \mathrm{vec}(\Omega) - (\mathrm{I}_d \otimes X)\boldsymbol{\rho} + (\mathrm{I}_d \otimes \Omega)\boldsymbol{\rho}\right]^T \\ &\qquad \times \left[\mathbf{y} - \mathrm{vec}(\Omega) - (\mathrm{I}_d \otimes X)\boldsymbol{\rho} + (\mathrm{I}_d \otimes \Omega)\boldsymbol{\rho}\right] \qquad (5.22) \\ j(\boldsymbol{\rho},\,\tau,\,\Omega,\,\Lambda) &= N\,\mathrm{Tr}[\Lambda] + N\boldsymbol{\rho}^T\left[(\mathrm{I}_d \otimes \Lambda)\boldsymbol{\rho} - 2\mathrm{vec}(\Lambda)\right] \\ &\qquad + \mathrm{Tr}\left[\left\{\mathrm{I}_d \otimes \left[(X-\Omega)^T(X-\Omega) + N\Lambda\right]\right\}\tau\right]. \qquad (5.23) \end{aligned}$$

It is worth mentioning that the expression for γ is the same as that in the zero mean case. That for δ, however, is more complicated; yet, if the mean is equated to zero, then (5.21) is equivalent to (3.24). We now reconsider (5.12), and find the variational distribution for $\boldsymbol{\mu}$. By substituting the likelihood and prior for $\boldsymbol{\mu}$, given by (5.8), this functional becomes

$$\begin{aligned} \mathcal{L}_{\boldsymbol{\mu}}(q) = \int q(\boldsymbol{\mu}\,|\,D)\Bigg[\iint q(\mathbf{a}\,|\,D)\,q(\sigma^2\,|\,D)\Bigg\{ -\frac{dN}{2}\log 2\pi\sigma^2 \\ -\frac{1}{2}\mathrm{Tr}\left[(\sigma^2)^{-1}g(A,\,M)\right]\Bigg\}\,\mathrm{d}\mathbf{a}\,\mathrm{d}\sigma^2 - \frac{d}{2}\log 2\pi - \frac{1}{2}\log|\Delta| \\ -\frac{1}{2}(\boldsymbol{\mu}-\mathbf{b})\Delta^{-1}(\boldsymbol{\mu}-\mathbf{b})^T - \log q(\boldsymbol{\mu}\,|\,D)\Bigg]\,\mathrm{d}\boldsymbol{\mu} + \mathrm{const.} \end{aligned}$$

Here, we show care when specifying the prior density since $\boldsymbol{\mu}$ is a *row* vector. By dropping terms independent of $\boldsymbol{\mu}$, we then arrive at

$$\begin{aligned} \mathcal{L}_{\boldsymbol{\mu}}(q) = \int q(\boldsymbol{\mu}\,|\,D)\Bigg[-\frac{1}{2}\int q(\sigma^2\,|\,D)(\sigma^2)^{-1}\,\mathrm{d}\sigma^2\int q(\mathbf{a}\,|\,D)\mathrm{Tr}\left[g(A,\,M)\right]\,\mathrm{d}\mathbf{a} \\ -\frac{1}{2}(\boldsymbol{\mu}-\mathbf{b})\Delta^{-1}(\boldsymbol{\mu}-\mathbf{b})^T - \log q(\boldsymbol{\mu}\,|\,D)\Bigg]\,\mathrm{d}\boldsymbol{\mu} + \mathrm{const.}' \qquad (5.24) \end{aligned}$$

This expression can be simplified by the substitution of result (A.5). We now concentrate on computing the expectation of $\mathrm{Tr}\left[g(A,\,M)\right]$ with respect to the variational distribution $q(\mathbf{a}\,|\,D)$. To this end, $\mathrm{Tr}\left[g(A,\,M)\right]$ is expanded in a slightly different way to that seen previously. That is, we can rewrite (5.14) as

$$\begin{aligned}\mathrm{Tr}\ [g(A,\, M)] = \big[\mathbf{y} - \mathbf{m} - (\mathrm{I}_d \otimes X)\mathbf{a} + (A^T \otimes \mathrm{I}_N)\mathbf{m}\big]^T \\ \times \big[\mathbf{y} - \mathbf{m} - (\mathrm{I}_d \otimes X)\mathbf{a} + (A^T \otimes \mathrm{I}_N)\mathbf{m}\big]\,,\end{aligned}$$

using the necessary identities of (3.5). This will aid greatly in later algebra simplification. Taking the afore-mentioned expectation of the above provides

$$\begin{aligned}&\mathrm{E}_{q(\mathbf{a}\,|\,D)}\{\mathrm{Tr}\ [g(A,\, M)]\} \\ &= \mathbf{y}^T\mathbf{y} - \mathbf{m}^T\mathbf{y} - \boldsymbol{\rho}^T(\mathrm{I}_d \otimes X^T)\mathbf{y} + \mathbf{m}^T(\hat{A} \otimes \mathrm{I}_N)\mathbf{y} - \mathbf{y}^T\mathbf{m} + \mathbf{m}^T\mathbf{m} + \boldsymbol{\rho}^T(\mathrm{I}_d \otimes X^T)\mathbf{m} \\ &\quad - \mathbf{m}^T(\hat{A} \otimes \mathrm{I}_N)\mathbf{m} - \mathbf{y}^T(\mathrm{I}_d \otimes X)\boldsymbol{\rho} + \mathbf{m}^T(\mathrm{I}_d \otimes X)\boldsymbol{\rho} \\ &\quad + \mathrm{E}_{q(\mathbf{a}\,|\,D)}\{\mathbf{a}^T\}(\mathrm{I}_d \otimes X^TX)\mathrm{E}_{q(\mathbf{a}\,|\,D)}\{\mathbf{a}\} + \mathrm{Tr}[(\mathrm{I}_d \otimes X^TX)\mathrm{Var}_{q(\mathbf{a}\,|\,D)}\{\mathbf{a}\}] \\ &\quad - \mathrm{E}_{q(\mathbf{a}\,|\,D)}\{\mathbf{m}^T\mathrm{vec}(XAA^T)\} + \mathbf{y}^T(\hat{A}^T \otimes \mathrm{I}_N)\mathbf{m} - \mathbf{m}^T(\hat{A}^T \otimes \mathrm{I}_N)\mathbf{m} \\ &\qquad - \mathrm{E}_{q(\mathbf{a}\,|\,D)}\left\{\big[\mathrm{vec}(XAA^T)\big]^T\mathbf{m}\right\} + \mathbf{m}^T\left(\mathrm{E}_{q(\mathbf{a}\,|\,D)}\{AA^T\} \otimes \mathrm{I}_N\right)\mathbf{m}. \qquad (5.25)\end{aligned}$$

In the above, we have made use of (3.20) and (2.51). Further, it is realised that

$$\mathrm{vec}(PQR) = (R^T \otimes P)\mathrm{vec}(Q) \qquad (5.26)$$

for compatible matrices P, Q, R (Henderson and Searle, 1979). Moreover, we have defined $\mathrm{E}_{q(\mathbf{a}\,|\,D)}\{A\} = \hat{A}$. As mentioned in Section 3.5, the $d \times d$ matrix $\hat{A}$ is created by unstacking the d^2-vector $\boldsymbol{\rho}$. So, if $\boldsymbol{\rho} = (\rho_1,\, \rho_2, \ldots, \rho_{d^2})^T$, then

$$\hat{A} = \begin{pmatrix} \rho_1 & \rho_{d+1} & \cdots & \rho_{d(d-1)+1} \\ \rho_2 & \rho_{d+2} & \cdots & \rho_{d(d-1)+2} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_d & \rho_{2d} & \cdots & \rho_{d^2} \end{pmatrix}.$$

To enable this, we introduce the notation vec^{-1} to be the *inverse vec operator*, and hence write $\hat{A} = \mathrm{vec}_d^{-1}(\boldsymbol{\rho})$, where the subscript d represents the number of rows in the new matrix. Thus, the operator is defined such that this subscript must be a divisor of the length of the vector, and can produce non-square matrices. Notice that, if P is any matrix with r rows, it follows that $\mathrm{vec}_r^{-1}[\mathrm{vec}(P)] = P$.

In (5.25), we now must find $\mathrm{E}_{q(\mathbf{a}\,|\,D)}\{AA^T\}$. In particular, it is apparent that $\mathrm{E}_{q(\mathbf{a}\,|\,D)}\{\mathbf{m}^T\mathrm{vec}(XAA^T)\} = \mathbf{m}^T\mathrm{vec}(X\mathrm{E}_{q(\mathbf{a}\,|\,D)}\{AA^T\})$. Again via component form, the expectation of the $(i,\, j)$-th element of AA^T, using the definition of matrix multiplication and covariance, is seen to be

$$\begin{aligned}
\mathrm{E}_{q(\mathbf{a}\,|\,D)}\{[AA^T]_{ij}\} &= \mathrm{E}_{q(\mathbf{a}\,|\,D)}\left\{\sum_{k=1}^{d} a_{ik}a_{jk}\right\} \\
&= \sum_{k=1}^{d} \mathrm{E}_{q(\mathbf{a}\,|\,D)}\{a_{ik}a_{jk}\} \\
&= \sum_{k=1}^{d} \left(\mathrm{E}_{q(\mathbf{a}\,|\,D)}\{a_{ik}\}\mathrm{E}_{q(\mathbf{a}\,|\,D)}\{a_{jk}\} + \mathrm{Cov}_{q(\mathbf{a}\,|\,D)}\{a_{ik},\, a_{jk}\}\right) \\
&= \left[\mathrm{E}_{q(\mathbf{a}\,|\,D)}\{A\}\mathrm{E}_{q(\mathbf{a}\,|\,D)}\{A^T\}\right]_{ij} + \sum_{k=1}^{d} \left[\mathrm{Var}_{q(\mathbf{a}\,|\,D)}\{A_k\}\right]_{ij}, \qquad (5.27)
\end{aligned}$$

where we define A_k to be the k-th column vector of A. We can simplify this expression yet further by examining $\mathrm{Var}_{q(\mathbf{a}\,|\,D)}\{A_k\}$. Suppose that the k-th column of I_d is denoted by $\mathbf{i}_k$, *i.e.* $\mathrm{I}_d = (\mathbf{i}_1\,|\,\mathbf{i}_2\,|\,\ldots\,|\,\mathbf{i}_d)$. Then, it follows that $A_k = A\mathbf{i}_k$. We realise that $\mathrm{Var}_{q(\mathbf{a}\,|\,D)}\{\mathrm{vec}(A)\} = \tau$ by the definition of $\mathbf{a}$. Consequently,

$$\begin{aligned}
\mathrm{Var}_{q(\mathbf{a}\,|\,D)}\{A_k\} &= \mathrm{Var}_{q(\mathbf{a}\,|\,D)}\{A\mathbf{i}_k\} \\
&= \mathrm{Var}_{q(\mathbf{a}\,|\,D)}\{\mathrm{vec}(A\mathbf{i}_k)\} \\
&= \mathrm{Var}_{q(\mathbf{a}\,|\,D)}\{(\mathbf{i}_k^T \otimes \mathrm{I}_d)\mathrm{vec}(A)\} \\
&= (\mathbf{i}_k^T \otimes \mathrm{I}_d)\mathrm{Var}_{q(\mathbf{a}\,|\,D)}\{\mathrm{vec}(A)\}(\mathbf{i}_k^T \otimes \mathrm{I}_d)^T \\
&= (\mathbf{i}_k^T \otimes \mathrm{I}_d)\tau(\mathbf{i}_k \otimes \mathrm{I}_d)
\end{aligned}$$

by (3.5), and since $\mathrm{Var}\{P\mathbf{w}\} = P\,\mathrm{Var}\{\mathbf{w}\}P^T$ for any random vector $\mathbf{w}$ and fixed matrix P of suitable dimension. Moreover, for any vector $\mathbf{r}$, it is clear that $\mathrm{vec}(\mathbf{r}) = \mathbf{r}$. Hence, substituting back into (5.27) and returning to matrix form provides

$$\begin{aligned}\mathrm{E}_{q(\mathbf{a}\,|\,D)}\{AA^T\} &= \hat{A}\hat{A}^T + \sum_{k=1}^{d}(\mathbf{i}_k^T \otimes \mathrm{I}_d)\tau(\mathbf{i}_k \otimes \mathrm{I}_d) \\ &= \hat{A}\hat{A}^T + (\mathbf{1}_d^T \otimes \mathrm{I}_d)\tau(\mathbf{1}_d \otimes \mathrm{I}_d),\end{aligned} \tag{5.28}$$

where $\mathbf{1}_d$ is defined to be the vector of 1s of length d. Here, we notice the result

$$\sum_{i=1}^{d}\sum_{j=1}^{d}(\mathbf{i}_i^T \otimes \mathrm{I}_d)\tau(\mathbf{i}_j \otimes \mathrm{I}_d) = (\mathbf{1}_d^T \otimes \mathrm{I}_d)\tau(\mathbf{1}_d \otimes \mathrm{I}_d).$$

However, as τ is diagonal, then (5.28) holds. All outstanding terms in (5.25) have now been accounted for, and this expression in turn can be substituted into (5.24). Again, we establish $\tilde{\mathcal{L}}_{\boldsymbol{\mu}}(q) = \mathcal{L}_{\boldsymbol{\mu}}(q) + \nu_{\boldsymbol{\mu}}(\int q(\boldsymbol{\mu}\,|\,D)\,\mathrm{d}\boldsymbol{\mu} - 1)$, and differentiate now with respect to $q(\boldsymbol{\mu}\,|\,D)$. Accordingly, this leads to

$$\begin{aligned}\frac{\partial\tilde{\mathcal{L}}_{\boldsymbol{\mu}}(q)}{\partial q(\boldsymbol{\mu}\,|\,D)} = -\frac{\gamma}{2\delta}\mathrm{E}_{q(\mathbf{a}\,|\,D)}\{\mathrm{Tr}\,[g(A,\,M)]\} - \frac{1}{2}(\boldsymbol{\mu}-\mathbf{b})\Delta^{-1}(\boldsymbol{\mu}-\mathbf{b})^T \\ - \log q(\boldsymbol{\mu}\,|\,D) - 1 + \nu_{\boldsymbol{\mu}} = 0.\end{aligned}$$

Upon rearranging and dropping all terms that are independent of $\boldsymbol{\mu}$, in particular those within $\mathrm{E}_{q(\mathbf{a}\,|\,D)}\{\mathrm{Tr}\,[g(A,\,M)]\}$, we arrive at

$$\begin{aligned}q(\boldsymbol{\mu}\,|\,D) \propto \exp\Bigg\{ -\frac{\gamma}{2\delta}\Bigg[\mathbf{m}^T\Bigg(\mathrm{I}_{dN} - (\hat{A}\otimes\mathrm{I}_N) - (\hat{A}^T\otimes\mathrm{I}_N) + (\hat{A}\hat{A}^T\otimes\mathrm{I}_N) \\ + \Big[\Big\{(\mathbf{1}_d^T\otimes\mathrm{I}_d)\tau(\mathbf{1}_d\otimes\mathrm{I}_d)\Big\}\otimes\mathrm{I}_N\Big]\Bigg)\mathbf{m} \\ - \mathbf{m}^T\Bigg(\mathbf{y} - \mathrm{vec}(Y\hat{A}^T) - \mathrm{vec}(X\hat{A}) + \mathrm{vec}(X\hat{A}\hat{A}^T) + \mathrm{vec}\Big(X(\mathbf{1}_d^T\otimes\mathrm{I}_d)\tau(\mathbf{1}_d\otimes\mathrm{I}_d)\Big)\Bigg)\end{aligned}$$

$$- \Big(\mathbf{y}^T - \Big[\mathrm{vec}(Y\hat{A}^T)\Big]^T - \Big[\mathrm{vec}(X\hat{A})\Big]^T + \Big[\mathrm{vec}(X\hat{A}\hat{A}^T)\Big]^T \\ + \Big[\mathrm{vec}\Big(X(\mathbf{1}_d^T \otimes \mathrm{I}_d)\tau(\mathbf{1}_d \otimes \mathrm{I}_d)\Big)\Big]^T\Big)\mathbf{m}\Big] - \frac{1}{2}(\boldsymbol{\mu} - \mathbf{b})\Delta^{-1}(\boldsymbol{\mu} - \mathbf{b})^T\Big\},$$

using (3.5) and that $(P+Q) \otimes R = (P \otimes R) + (Q \otimes R)$. This expression can be further simplified:

$$q(\boldsymbol{\mu}\,|\,D) \propto \exp\Big\{ - \frac{\gamma}{2\delta}\Big[\mathbf{m}^T\Big(\Big\{\Big[\mathrm{I}_d - \hat{A}\Big]\Big[\mathrm{I}_d - \hat{A}\Big]^T + (\mathbf{1}_d^T \otimes \mathrm{I}_d)\tau(\mathbf{1}_d \otimes \mathrm{I}_d)\Big\} \otimes \mathrm{I}_N\Big)\mathbf{m} \\ - \mathbf{m}^T\Big(\mathrm{vec}\Big(\Big[Y - X\hat{A}\Big]\Big[\mathrm{I}_d - \hat{A}\Big]^T + X(\mathbf{1}_d^T \otimes \mathrm{I}_d)\tau(\mathbf{1}_d \otimes \mathrm{I}_d)\Big)\Big) \\ - \Big[\mathrm{vec}\Big(\Big[Y - X\hat{A}\Big]\Big[\mathrm{I}_d - \hat{A}\Big]^T + X(\mathbf{1}_d^T \otimes \mathrm{I}_d)\tau(\mathbf{1}_d \otimes \mathrm{I}_d)\Big)\Big]^T\mathbf{m}\Big] \\ - \frac{1}{2}(\boldsymbol{\mu} - \mathbf{b})\Delta^{-1}(\boldsymbol{\mu} - \mathbf{b})^T\Big\}, \quad (5.29)$$

whereby $\mathbf{y} = \mathrm{vec}(Y)$ and $\mathrm{I}_d \otimes \mathrm{I}_N = \mathrm{I}_{dN}$ (Harville, 1997). Here, there is an evident problem. We require a variational distribution for $\boldsymbol{\mu}$ whereas the expression above is given predominantly in terms of $\mathbf{m} = \mathrm{vec}(M)$. Recall that M is a matrix with each row equivalent to $\boldsymbol{\mu}$. So, we need to manipulate such terms, in particular, those of the form $\mathbf{m}^T\mathbf{r}$ and $\mathbf{m}^T C\mathbf{m}$ for a given dN-vector $\mathbf{r}$ and $dN \times dN$ matrix C.

With regards to the former, notice that, by definition, $\mathbf{m} = \boldsymbol{\mu}^T \otimes \mathbf{1}_N$, remembering that $\boldsymbol{\mu}$ is a row vector. Thus, by introducing a $N \times d$ matrix R such that $R = \mathrm{vec}_N^{-1}(\mathbf{r})$, it follows that

$$\begin{aligned} \mathbf{m}^T\mathbf{r} = \mathbf{r}^T\mathbf{m} &= \mathbf{m}^T\mathrm{vec}(R) \\ &= (\boldsymbol{\mu}^T \otimes \mathbf{1}_N)^T\mathrm{vec}(R) \\ &= (\boldsymbol{\mu} \otimes {\mathbf{1}_N}^T)\mathrm{vec}(R) \\ &= \mathrm{vec}({\mathbf{1}_N}^T R\,\boldsymbol{\mu}^T) \\ &= {\mathbf{1}_N}^T R\,\boldsymbol{\mu}^T, \end{aligned} \quad (5.30)$$

where (5.26) has been applied. On the other hand, we also realise that

$$\begin{aligned}
\mathbf{m}^T C \mathbf{m} &= (\boldsymbol{\mu}^T \otimes \mathbf{1}_N)^T \, C \, (\boldsymbol{\mu}^T \otimes \mathbf{1}_N) \\
&= (\boldsymbol{\mu} \otimes {\mathbf{1}_N}^T) \, C \, (\boldsymbol{\mu}^T \otimes \mathbf{1}_N) \\
&= (\boldsymbol{\mu} \otimes 1)(\mathrm{I}_d \otimes {\mathbf{1}_N}^T) \, C \, (\mathrm{I}_d \otimes \mathbf{1}_N)(\boldsymbol{\mu}^T \otimes 1) \\
&= \boldsymbol{\mu}(\mathrm{I}_d \otimes {\mathbf{1}_N}^T) \, C \, (\mathrm{I}_d \otimes \mathbf{1}_N)\boldsymbol{\mu}^T, \qquad (5.31)
\end{aligned}$$

noting (3.20) and that $P \otimes z = zP$ for any matrix P and scalar z. Hence, we can utilise (5.30) and (5.31) to ensure that each term of (5.29) is written in terms of $\boldsymbol{\mu}$. To effect this, we must be aware that ${\mathbf{1}_N}^T \mathbf{1}_N = N$. As a result, (5.29) can be expressed in the form

$$\begin{aligned}
q(\boldsymbol{\mu} \,|\, D) &\propto \exp\left\{ -\frac{1}{2} \left[\boldsymbol{\mu}\Upsilon\boldsymbol{\mu}^T - \Theta\boldsymbol{\mu}^T - \boldsymbol{\mu}\Theta^T \right] \right\} \qquad (5.32) \\
&\propto \exp\left\{ -\frac{1}{2} \left[(\boldsymbol{\mu} - \Theta\Upsilon^{-1})\Upsilon(\boldsymbol{\mu} - \Theta\Upsilon^{-1})^T \right] \right\},
\end{aligned}$$

where we have

$$\Upsilon = \frac{N\gamma}{\delta} \left(\left[\mathrm{I}_d - \hat{A} \right] \left[\mathrm{I}_d - \hat{A} \right]^T + (\mathbf{1}_d^T \otimes \mathrm{I}_d)\tau(\mathbf{1}_d \otimes \mathrm{I}_d) \right) + \Delta^{-1} \qquad (5.33)$$

$$\Theta = \frac{\gamma}{\delta} {\mathbf{1}_N}^T \left(\left[Y - X\hat{A} \right] \left[\mathrm{I}_d - \hat{A} \right]^T + X(\mathbf{1}_d^T \otimes \mathrm{I}_d)\tau(\mathbf{1}_d \otimes \mathrm{I}_d) \right) + \mathbf{b}\Delta^{-1}. \qquad (5.34)$$

So ultimately, the variational distribution for $\boldsymbol{\mu}$ is such that

$$q(\boldsymbol{\mu} \,|\, D) = \mathcal{N}(\boldsymbol{\mu} \,|\, \boldsymbol{\omega}, \, \Lambda) \qquad (5.35)$$

whereby

$$\boldsymbol{\omega} = \Theta\Upsilon^{-1} \qquad (5.36)$$

$$\Lambda = \Upsilon^{-1}. \qquad (5.37)$$

Application of the free form method is concluded by seeking a form for $q(\mathbf{a}\,|\,D)$. By substitution of terms into (5.10) and dropping those independent of $\mathbf{a}$, hence, in accordance with (3.25), we acquire

$$\begin{aligned}\mathcal{L}_{\boldsymbol{\mu}}(q) = \int q(\mathbf{a}\,|\,D)\biggl[-\frac{\gamma}{2\delta}\int q(\boldsymbol{\mu}\,|\,D)\mathrm{Tr}\left[g(A,\,M)\right]\,\mathrm{d}\boldsymbol{\mu} \\ -\frac{1}{2}\mathbf{a}^T{C^*}^{-1}\mathbf{a} - \log q(\mathbf{a}\,|\,D)\biggr]\,\mathrm{d}\mathbf{a} + \mathrm{const.}', \end{aligned} \tag{5.38}$$

moreover via (A.5). If $\mathrm{Tr}\left[g(A,\,M)\right]$ is expanded with respect to (5.14), then, in comparison with (5.15), taking expectations implies

$$\begin{aligned}&\mathrm{E}_{q(\boldsymbol{\mu}\,|\,D)}\{\mathrm{Tr}\ [g(A,\,M)]\} \\ &= \mathbf{y}^T\mathbf{y} - \left[\mathrm{vec}(\Omega)\right]^T\mathbf{y} - \mathbf{a}^T(\mathrm{I}_d\otimes X^T)\mathbf{y} + \mathbf{a}^T(\mathrm{I}_d\otimes\Omega^T)\mathbf{y} - \mathbf{y}^T\mathrm{vec}(\Omega) + \left[\mathrm{vec}(\Omega)\right]^T\left[\mathrm{vec}(\Omega)\right] \\ &\quad + N\,\mathrm{Tr}[\Lambda] + \mathbf{a}^T(\mathrm{I}_d\otimes X^T)\mathrm{vec}(\Omega) - \mathbf{a}^T\mathrm{vec}(\Omega^T\Omega) - N\mathbf{a}^T\mathrm{vec}(\Lambda) - \mathbf{y}^T(\mathrm{I}_d\otimes X)\mathbf{a} \\ &\quad + \left[\mathrm{vec}(\Omega)\right]^T(\mathrm{I}_d\otimes X)\mathbf{a} + \mathbf{a}^T(\mathrm{I}_d\otimes X^TX)\mathbf{a} - \mathbf{a}^T(\mathrm{I}_d\otimes\Omega^TX)\mathbf{a} + \mathbf{y}^T(\mathrm{I}_d\otimes\Omega)\mathbf{a} \\ &\quad - \left[\mathrm{vec}(\Omega^T\Omega)\right]^T\mathbf{a} - N\left[\mathrm{vec}(\Lambda)\right]^T\mathbf{a} - \mathbf{a}^T(\mathrm{I}_d\otimes X^T\Omega)\mathbf{a} + \mathbf{a}^T\left(\mathrm{I}_d\otimes\{\Omega^T\Omega + N\Lambda\}\right)\mathbf{a},\end{aligned}$$

in particular, as a consequence of (5.16) and (5.17). We now feed this expression into (5.38), and subsequently form $\tilde{\mathcal{L}}_{\boldsymbol{\mu}}(q)$ as before, using the Lagrange multiplier $\nu_{\mathbf{a}}$. Optimising this functional with respect to $q(\mathbf{a}\,|\,D)$ provides

$$\frac{\partial\tilde{\mathcal{L}}_{\boldsymbol{\mu}}(q)}{\partial q(\mathbf{a}\,|\,D)} \;=\; -\frac{\gamma}{2\delta}\mathrm{E}_{q(\boldsymbol{\mu}\,|\,D)}\{\mathrm{Tr}\ [g(A,\,M)]\} - \frac{1}{2}\mathbf{a}^T{C^*}^{-1}\mathbf{a} - \log q(\mathbf{a}\,|\,D) - 1 + \nu_{\mathbf{a}} \;=\; 0.$$

By rearranging and dropping all terms that do not depend on $\mathbf{a}$, we obtain

$$\begin{aligned}q(\mathbf{a}\,|\,D) \propto \exp\biggl\{ -\frac{1}{2}\biggl[\mathbf{a}^T\Bigl(\frac{\gamma}{\delta}\Bigl\{(\mathrm{I}_d\otimes X^TX) - (\mathrm{I}_d\otimes\Omega^TX) - (\mathrm{I}_d\otimes X^T\Omega) + (\mathrm{I}_d\otimes\Omega^T\Omega) \\ + N(\mathrm{I}_d\otimes\Lambda)\Bigr\} + {C^*}^{-1}\Bigr)\mathbf{a}\end{aligned}$$

$$
\begin{aligned}
&-\frac{\gamma}{\delta}\mathbf{a}^T\Big((\mathrm{I}_d \otimes X^T)\mathbf{y} - (\mathrm{I}_d \otimes \Omega^T)\mathbf{y} - (\mathrm{I}_d \otimes X^T)\mathrm{vec}(\Omega) + \mathrm{vec}(\Omega^T\Omega) + N\mathrm{vec}(\Lambda)\Big) \\
&-\frac{\gamma}{\delta}\Big(\mathbf{y}^T(\mathrm{I}_d \otimes X) - \mathbf{y}^T(\mathrm{I}_d \otimes \Omega) - [\mathrm{vec}(\Omega)]^T(\mathrm{I}_d \otimes X) + \big[\mathrm{vec}(\Omega^T\Omega)\big]^T + N\,[\mathrm{vec}(\Lambda)]^T\Big)\mathbf{a}\Big]\Big\}.
\end{aligned}
$$

Since $\mathrm{vec}(\Omega^T\Omega) = (\mathrm{I}_d \otimes \Omega^T)\mathrm{vec}(\Omega)$, this equation can be easily rewritten as

$$
\begin{aligned}
q(\mathbf{a}\,|\,D) \propto \exp\Big\{ -\frac{1}{2}\Big[&\mathbf{a}^T\Big(\frac{\gamma}{\delta}\Big\{\mathrm{I}_d \otimes \big[(X-\Omega)^T(X-\Omega) + N\Lambda\big]\Big\} + C^{*^{-1}}\Big)\mathbf{a} \\
&- \mathbf{a}^T\Big(\frac{\gamma}{\delta}\Big\{\Big(\mathrm{I}_d \otimes [X-\Omega]^T\Big)(\mathbf{y} - \mathrm{vec}(\Omega)) + N\mathrm{vec}(\Lambda)\Big\}\Big) \\
&\qquad - \Big(\frac{\gamma}{\delta}\Big\{[\mathbf{y} - \mathrm{vec}(\Omega)]^T(\mathrm{I}_d \otimes [X-\Omega]) + N\,[\mathrm{vec}(\Lambda)]^T\Big\}\Big)\mathbf{a}\Big]\Big\}.
\end{aligned}
$$

It is apparent that this is now in the form as given by (5.32), and consequently

$$
q(\mathbf{a}\,|\,D) = \mathcal{N}(\mathbf{a}\,|\,\boldsymbol{\rho},\,\tau) \tag{5.39}
$$

as in the zero-mean case, but now with $\boldsymbol{\rho}$ and τ equivalent to

$$
\begin{aligned}
\boldsymbol{\rho} = &\Big[\frac{\gamma}{\delta}\Big\{\mathrm{I}_d \otimes \big[(X-\Omega)^T(X-\Omega) + N\Lambda\big]\Big\} + C^{*^{-1}}\Big]^{-1} \\
&\times \Big[\frac{\gamma}{\delta}\Big\{\Big(\mathrm{I}_d \otimes [X-\Omega]^T\Big)(\mathbf{y} - \mathrm{vec}(\Omega)) + N\mathrm{vec}(\Lambda)\Big\}\Big]
\end{aligned} \tag{5.40}
$$

$$
\tau = \Big[\frac{\gamma}{\delta}\Big\{\mathrm{I}_d \otimes \big[(X-\Omega)^T(X-\Omega) + N\Lambda\big]\Big\} + C^{*^{-1}}\Big]^{-1}. \tag{5.41}
$$

Thus, the expressions for the variational parameters $\{\gamma,\,\delta,\,\boldsymbol{\omega},\,\Lambda,\,\boldsymbol{\rho},\,\tau\}$ are (5.20), (5.21), (5.36), (5.37), (5.40) and (5.41) respectively. Once again, these are update equations, which must be solved iteratively.

5.2.3 Fixed form method

The variational distributions for both σ^2, $\boldsymbol{\mu}$ and $\mathbf{a}$, when this vector is dense, have now been accounted for. However, as in Chapter 3, we must now derive the approximation for $\mathbf{a}$ in the sparse case, *i.e.* we constrain elements of $\boldsymbol{\rho}$ and τ to zero relative to the sparsity structure of a given A-matrix. Consequently, the fixed form procedure is again most beneficial.

Therefore, we assume fixed forms for each variational, as given by (5.19), (5.35) and (5.39) in the free form approach. Recall that an inherent component of this method is to derive the lower bound initially. So, using again the independence of both prior and variational distributions, (5.9) can be split up into a sum of integrals:

$$\begin{aligned}\mathcal{L}_{\boldsymbol{\mu}}(q) = &\iiint q(\mathbf{a}\,|\,D)q(\sigma^2\,|\,D)q(\boldsymbol{\mu}\,|\,D)\log p(D\,|\,A,\,\sigma^2,\,\boldsymbol{\mu})\,\mathrm{d}\mathbf{a}\,\mathrm{d}\sigma^2\,\mathrm{d}\boldsymbol{\mu} + \int q(\mathbf{a}\,|\,D)\log p(\mathbf{a})\,\mathrm{d}\mathbf{a} \\ &+ \int q(\sigma^2\,|\,D)\,\log p(\sigma^2)\,\mathrm{d}\sigma^2 + \int q(\boldsymbol{\mu}\,|\,D)\log p(\boldsymbol{\mu})\,\mathrm{d}\boldsymbol{\mu} - \int q(\mathbf{a}\,|\,D)\log q(\mathbf{a}\,|\,D)\,\mathrm{d}\mathbf{a} \\ &- \int q(\sigma^2\,|\,D)\log q(\sigma^2\,|\,D)\,\mathrm{d}\sigma^2 - \int q(\boldsymbol{\mu}\,|\,D)\log q(\boldsymbol{\mu}\,|\,D)\,\mathrm{d}\boldsymbol{\mu}. \qquad (5.42)\end{aligned}$$

Each integral is subsequently tackled in turn. Thus, initially, it follows that

$$\begin{aligned}&\iiint q(\mathbf{a}\,|\,D)q(\sigma^2\,|\,D)q(\boldsymbol{\mu}\,|\,D)\log p(D\,|\,A,\,\sigma^2,\,\boldsymbol{\mu})\,\mathrm{d}\mathbf{a}\,\mathrm{d}\sigma^2\,\mathrm{d}\boldsymbol{\mu} \\ &= \iiint q(\mathbf{a}\,|\,D)q(\sigma^2\,|\,D)q(\boldsymbol{\mu}\,|\,D)\left[-\frac{dN}{2}\log 2\pi\sigma^2 - \frac{1}{2}\mathrm{Tr}\left[(\sigma^2)^{-1}g(A,\,M)\right]\right]\mathrm{d}\mathbf{a}\,\mathrm{d}\sigma^2\,\mathrm{d}\boldsymbol{\mu} \\ &= -\frac{dN}{2}\log 2\pi - \frac{dN}{2}\int q(\sigma^2\,|\,D)\,\log\sigma^2\,\mathrm{d}\sigma^2 \\ &\quad -\frac{1}{2}\iint q(\mathbf{a}\,|\,D)q(\boldsymbol{\mu}\,|\,D)\mathrm{Tr}\left[g(A,\,M)\right]\,\mathrm{d}\mathbf{a}\,\mathrm{d}\boldsymbol{\mu}\int q(\sigma^2\,|\,D)(\sigma^2)^{-1}\,\mathrm{d}\sigma^2 \\ &= -\frac{dN}{2}\log 2\pi - \frac{dN}{2}\left[\log\delta - \psi(\gamma)\right] - \frac{\gamma}{2\delta}\left[h(\boldsymbol{\rho},\,\Omega) + j(\boldsymbol{\rho},\,\tau,\,\Omega,\,\Lambda)\right], \qquad (5.43)\end{aligned}$$

via equations (A.5), (A.6) and (5.18). In addition, using (2.51) and the variational distribution for $\boldsymbol{\mu}$,

$$\begin{aligned}
&\int q(\boldsymbol{\mu}\,|\,D) \log p(\boldsymbol{\mu})\,\mathrm{d}\boldsymbol{\mu} \\
&= \int q(\boldsymbol{\mu}\,|\,D) \left[-\frac{d}{2}\log 2\pi - \frac{1}{2}\log|\Delta| - \frac{1}{2}\left(\boldsymbol{\mu}-\mathbf{b}\right)\Delta^{-1}\left(\boldsymbol{\mu}-\mathbf{b}\right)^T\right]\,\mathrm{d}\boldsymbol{\mu} \\
&= -\frac{d}{2}\log 2\pi - \frac{1}{2}\log|\Delta| - \frac{1}{2}\mathrm{E}_{q(\boldsymbol{\mu}\,|\,D)}\left\{\boldsymbol{\mu}-\mathbf{b}\right\}\Delta^{-1}\mathrm{E}_{q(\boldsymbol{\mu}\,|\,D)}\left\{\left[\boldsymbol{\mu}-\mathbf{b}\right]^T\right\} \\
&\quad - \frac{1}{2}\mathrm{Tr}\left[\Delta^{-1}\mathrm{Var}_{q(\boldsymbol{\mu}\,|\,D)}\left\{\left[\boldsymbol{\mu}-\mathbf{b}\right]^T\right\}\right] \\
&= -\frac{d}{2}\log 2\pi - \frac{1}{2}\log|\Delta| - \frac{1}{2}\left(\boldsymbol{\omega}-\mathbf{b}\right)\Delta^{-1}\left(\boldsymbol{\omega}-\mathbf{b}\right)^T - \frac{1}{2}\mathrm{Tr}\left[\Delta^{-1}\Lambda\right].
\end{aligned}$$

Furthermore, using the above computation,

$$\begin{aligned}
&\int q(\boldsymbol{\mu}\,|\,D) \log q(\boldsymbol{\mu}\,|\,D)\,\mathrm{d}\boldsymbol{\mu} \\
&= -\frac{d}{2}\log 2\pi - \frac{1}{2}\log|\Lambda| - \frac{1}{2}\mathrm{E}_{q(\boldsymbol{\mu}\,|\,D)}\left\{\boldsymbol{\mu}-\boldsymbol{\omega}\right\}\Lambda^{-1}\mathrm{E}_{q(\boldsymbol{\mu}\,|\,D)}\left\{\left[\boldsymbol{\mu}-\boldsymbol{\omega}\right]^T\right\} \\
&\quad - \frac{1}{2}\mathrm{Tr}\left[\Lambda^{-1}\mathrm{Var}_{q(\boldsymbol{\mu}\,|\,D)}\left\{\left[\boldsymbol{\mu}-\boldsymbol{\omega}\right]^T\right\}\right] \\
&= -\frac{d}{2}\log 2\pi - \frac{1}{2}\log|\Lambda| - \frac{d}{2},
\end{aligned}$$

as $\mathrm{Tr}\left[\Lambda^{-1}\Lambda\right] = \mathrm{Tr}\left[\mathrm{I}_d\right] = d$. At this moment, we notice that, as the prior and variational distributions for both $\mathbf{a}$ and σ^2 are identical for both the zero and non-zero mean cases (albeit with differing update equations for the variational parameters), all other integrals in (5.42) have been previously evaluated in Section 3.3.3. Thus, to recap, we have

$$\begin{aligned}
\int q(\mathbf{a}\,|\,D)\,\log p(\mathbf{a})\,\mathrm{d}\mathbf{a} &= -\frac{d^2}{2}\log 2\pi - \frac{1}{2}\log|C^*| - \frac{1}{2}\boldsymbol{\rho}^T {C^*}^{-1}\boldsymbol{\rho} - \frac{1}{2}\mathrm{Tr}\left[{C^*}^{-1}\tau\right]. \\
\int q(\sigma^2\,|\,D)\,\log p(\sigma^2)\,\mathrm{d}\sigma^2 &= \alpha\log\beta\, -\, \log\Gamma(\alpha) - (\alpha+1)[\log\delta - \psi(\gamma)] - \frac{\beta\gamma}{\delta}. \\
\int q(\mathbf{a}\,|\,D)\,\log q(\mathbf{a}\,|\,D)\,\mathrm{d}\mathbf{a} &= -\frac{d^2}{2}\log 2\pi - \frac{1}{2}\log|\tau| - \frac{d^2}{2}. \\
\int q(\sigma^2\,|\,D)\,\log q(\sigma^2\,|\,D)\,\mathrm{d}\sigma^2 &= -\log\Gamma(\gamma) - \log\delta + (\gamma+1)\psi(\gamma) - \gamma.
\end{aligned}$$

Ultimately, by substituting back into (5.42), the lower bound is given by

$$\begin{aligned}\mathcal{L}_{\boldsymbol{\mu}}(q) = -\frac{dN}{2}\log 2\pi - \frac{dN}{2}\log\delta + \frac{dN}{2}\psi(\gamma) - \frac{\gamma}{2\delta}\left[h(\boldsymbol{\rho},\,\Omega) + j(\boldsymbol{\rho},\,\tau,\,\Omega,\,\Lambda)\right] - \frac{1}{2}\log|C^*| \\ - \frac{1}{2}\boldsymbol{\rho}^T C^{*^{-1}}\boldsymbol{\rho} - \frac{1}{2}\mathrm{Tr}\left[C^{*^{-1}}\tau\right] + \alpha\log\beta - \log\Gamma(\alpha) - \alpha\log\delta + \alpha\psi(\gamma) \\ - \frac{\beta\gamma}{\delta} - \frac{1}{2}\log|\Delta| - \frac{1}{2}\left(\boldsymbol{\omega}-\mathbf{b}\right)\Delta^{-1}\left(\boldsymbol{\omega}-\mathbf{b}\right)^T - \frac{1}{2}\mathrm{Tr}\left[\Delta^{-1}\Lambda\right] \\ + \frac{1}{2}\log|\tau| + \frac{d^2}{2} + \log\Gamma(\gamma) - \gamma\psi(\gamma) + \gamma + \frac{1}{2}\log|\Lambda| + \frac{d}{2}. \qquad (5.44)\end{aligned}$$

As before, we realise that (3.40) can be applied to compute the logarithm of the determinant for any singular matrix in (5.44), in particular, for C^* and τ. If we knew $\boldsymbol{\mu} = \mathbf{0}$, then, of course, this expression would simplify down to the lower bound in the zero mean case, given by (3.30). We now consider the maximisation of $\mathcal{L}_{\boldsymbol{\mu}}(q)$ with respect to $\boldsymbol{\rho}$ and τ, and hence subsequently, enforce the sparsity constraint. By differentiating with respect to $\boldsymbol{\rho}$, we have

$$\begin{aligned}\frac{\partial\mathcal{L}_{\boldsymbol{\mu}}(q)}{\partial\boldsymbol{\rho}} &= \frac{\partial}{\partial\boldsymbol{\rho}}\left\{-\frac{\gamma}{2\delta}\left[h(\boldsymbol{\rho},\,\Omega) + j(\boldsymbol{\rho},\,\tau,\,\Omega,\,\Lambda)\right] - \frac{1}{2}\boldsymbol{\rho}^T C^{*^{-1}}\boldsymbol{\rho}\right\} \\ &= \frac{\partial}{\partial\boldsymbol{\rho}}\Bigg\{ - \frac{\gamma}{2\delta}\left[\mathbf{y} - \mathrm{vec}(\Omega) - (\mathrm{I}_d\otimes[X-\Omega])\,\boldsymbol{\rho}\right]^T\left[\mathbf{y} - \mathrm{vec}(\Omega) - (\mathrm{I}_d\otimes[X-\Omega])\,\boldsymbol{\rho}\right] \\ &\qquad - \frac{\gamma}{2\delta}\left(N\boldsymbol{\rho}^T\left[(\mathrm{I}_d\otimes\Lambda)\boldsymbol{\rho} - 2\mathrm{vec}(\Lambda)\right]\right) - \frac{1}{2}\boldsymbol{\rho}^T C^{*^{-1}}\boldsymbol{\rho}\Bigg\} \\ &= \frac{\partial}{\partial\boldsymbol{\rho}}\left\{\boldsymbol{\rho}^T H\boldsymbol{\rho} + \mathbf{c}^T\boldsymbol{\rho}\right\},\end{aligned}$$

whereby

$$H = -\frac{\gamma}{2\delta}\Big\{\mathrm{I}_d\otimes\left[(X-\Omega)^T(X-\Omega) + N\Lambda\right]\Big\} - \frac{1}{2}C^{*^{-1}} \qquad (5.45)$$

$$\mathbf{c}^T = \frac{\gamma}{\delta}\Big\{\left(\mathbf{y} - \mathrm{vec}(\Omega)\right)^T\left(\mathrm{I}_d\otimes[X-\Omega]\right) + N\left[\mathrm{vec}(\Lambda)\right]^T\Big\}, \qquad (5.46)$$

and $\left[(\mathrm{I}_d\otimes[X-\Omega])\,\boldsymbol{\rho}\right]^T\left[\mathbf{y} - \mathrm{vec}(\Omega)\right] = \left[\mathbf{y} - \mathrm{vec}(\Omega)\right]^T\left[(\mathrm{I}_d\otimes[X-\Omega])\,\boldsymbol{\rho}\right]$. Therefore, we have the quadratic programming problem (3.31) that was faced in Section 3.3.3, with different specifications of H and $\mathbf{c}^T$. By again defining $\boldsymbol{\rho}_1$ as the non-zero elements of a given A, we must thus maximise $\boldsymbol{\rho}_1^T \mathbf{H}_{11} \boldsymbol{\rho}_1 + \mathbf{c}_1^T \boldsymbol{\rho}_1$ with respect to $\boldsymbol{\rho}_1$ where $\mathbf{H}_{11}$ and $\mathbf{c}_1$ are defined as previous. Resultantly, upon optimisation and solving for $\boldsymbol{\rho}_1$, it is evident that

$$\begin{aligned}\boldsymbol{\rho}_1 = &\left(\left[\frac{\gamma}{\delta} \left\{ \mathrm{I}_d \otimes \left[(X - \Omega)^T (X - \Omega) + N\Lambda \right] \right\} + C^{*^{-1}} \right]_{11} \right)^{-1} \\ &\times \left[\frac{\gamma}{\delta} \left\{ \left(\mathrm{I}_d \otimes [X - \Omega]^T \right) (\mathbf{y} - \mathrm{vec}(\Omega)) + N \mathrm{vec}(\Lambda) \right\} \right]_1 . \qquad (5.47)\end{aligned}$$

This definition of $\boldsymbol{\rho}_1$ is again used to reconstruct $\boldsymbol{\rho}$, according to the prescribed sparsity structure. Recall that the subscript notation refers to choosing the correct submatrix $\mathbf{H}_{11}$ and subvector $\mathbf{c}_1$, following row and column permutation.

When maximising $\mathcal{L}_{\boldsymbol{\mu}}(q)$ with respect to τ, as in the zero mean case, it is apparent that the sparsity constraint cannot be enforced using quadratic programming. Thus, we derive an expression for those elements of τ with non-zero variational posterior variance in component form as before. The prior distributions for σ^2 and $\boldsymbol{\mu}$, parameters unaffected by sparsity, remain specified by equations (5.7) and (5.8), whereas that for $\mathbf{a}$ is denoted by (3.33), a product of univariate Gaussians. Using a fixed form variational procedure, the variational distribution for $\mathbf{a}$ is again given, in component form, by (3.34).

We thus strive to re-calculate (5.42) at the component level. By using the identity (3.35), the probability of the data can be rewritten as

$$\begin{aligned}&p(D \,|\, \{a_{ij}\}, \, \sigma^2, \, \boldsymbol{\mu}) \\ &= (2\pi\sigma^2)^{-\frac{dN}{2}} \exp\left\{ -\frac{(\sigma^2)^{-1}}{2} \mathrm{Tr} \left[(Y - M - [X - M]A)^T (Y - M - [X - M]A) \right] \right\} \\ &= (2\pi\sigma^2)^{-\frac{dN}{2}} \exp\left\{ -\frac{(\sigma^2)^{-1}}{2} \sum_{j=1}^{N} \sum_{k=1}^{d} \left([Y - M - [X - M]A]_{jk} \right)^2 \right\} \\ &= (2\pi\sigma^2)^{-\frac{dN}{2}} \exp\left\{ -\frac{(\sigma^2)^{-1}}{2} \sum_{j=1}^{N} \sum_{k=1}^{d} \left(y_{jk} - m_{jk} - \sum_{i=1}^{d} x_{ji} a_{ik} + \sum_{i=1}^{d} m_{ji} a_{ik} \right)^2 \right\} .\end{aligned}$$

The definition of matrix multiplication is again noted. Hence, the first integral of (5.42) is computed as

$$
\begin{aligned}
&\iiint q(\mathbf{a}\,|\,D)q(\sigma^2\,|\,D)q(\boldsymbol{\mu}\,|\,D)\log p(D\,|\,\{a_{ij}\},\,\sigma^2,\,\boldsymbol{\mu})\,\mathrm{d}\mathbf{a}\,\mathrm{d}\sigma^2\,\mathrm{d}\boldsymbol{\mu}\\
&= \iiint q(\mathbf{a}\,|\,D)q(\sigma^2\,|\,D)q(\boldsymbol{\mu}\,|\,D)\Bigg[-\frac{dN}{2}\log 2\pi\sigma^2\\
&\quad -\frac{(\sigma^2)^{-1}}{2}\sum_{j=1}^{N}\sum_{k=1}^{d}\left(y_{jk}-m_{jk}-\sum_{i=1}^{d}x_{ji}a_{ik}+\sum_{i=1}^{d}m_{ji}a_{ik}\right)^2\Bigg]\,\mathrm{d}\mathbf{a}\,\mathrm{d}\sigma^2\,\mathrm{d}\boldsymbol{\mu}\\
&= -\frac{dN}{2}\log 2\pi-\frac{dN}{2}\left[\log\delta-\psi(\gamma)\right]\\
&\quad -\frac{\gamma}{2\delta}\sum_{j=1}^{N}\sum_{k=1}^{d}\iint q(\mathbf{a}\,|\,D)q(\boldsymbol{\mu}\,|\,D)\left(y_{jk}-m_{jk}-\sum_{i=1}^{d}x_{ji}a_{ik}+\sum_{i=1}^{d}m_{ji}a_{ik}\right)^2\,\mathrm{d}\mathbf{a}\,\mathrm{d}\boldsymbol{\mu},
\end{aligned}
$$

via (5.43). Moreover, the double integral in the above expression can be calculated as

$$
\begin{aligned}
&\mathrm{E}_{q(\mathbf{a}\,|\,D)}\left\{\mathrm{E}_{q(\boldsymbol{\mu}\,|\,D)}\left\{\left(y_{jk}-m_{jk}-\sum_{i=1}^{d}x_{ji}a_{ik}+\sum_{i=1}^{d}m_{ji}a_{ik}\right)^2\right\}\right\}\\
&= y_{jk}^2+\left[\mathrm{E}_{q(\boldsymbol{\mu}\,|\,D)}\{m_{jk}\}\right]^2+\mathrm{Var}_{q(\boldsymbol{\mu}\,|\,D)}\{m_{jk}\}+\left[\sum_{i=1}^{d}x_{ji}\mathrm{E}_{q(\mathbf{a}\,|\,D)}\{a_{ik}\}\right]^2\\
&\quad +\sum_{i=1}^{d}x_{ji}^2\mathrm{Var}_{q(\mathbf{a}\,|\,D)}\{a_{ik}\}+\mathrm{E}_{q(\boldsymbol{\mu}\,|\,D)}\left\{\left[\sum_{i=1}^{d}m_{ji}\mathrm{E}_{q(\mathbf{a}\,|\,D)}\{a_{ik}\}\right]^2\right\}\\
&\quad +\mathrm{E}_{q(\boldsymbol{\mu}\,|\,D)}\left\{\sum_{i=1}^{d}m_{ji}^2\mathrm{Var}_{q(\mathbf{a}\,|\,D)}\{a_{ik}\}\right\}-2y_{jk}\Omega_{jk}-2y_{jk}\sum_{i=1}^{d}x_{ji}\rho_{(i,k)}\\
&\quad +2y_{jk}\sum_{i=1}^{d}\Omega_{ji}\,\rho_{(i,k)}+2\Omega_{jk}\sum_{i=1}^{d}x_{ji}\,\rho_{(i,k)}-2\mathrm{E}_{q(\boldsymbol{\mu}\,|\,D)}\left\{m_{jk}\sum_{i=1}^{d}m_{ji}\,\rho_{(i,k)}\right\}\\
&\qquad -2\mathrm{E}_{q(\mathbf{a}\,|\,D)}\left\{\sum_{i=1}^{d}x_{ji}a_{ik}\sum_{u=1}^{d}\Omega_{ju}a_{uk}\right\},
\end{aligned}
\tag{5.48}
$$

using (2.51) and previous definitions of variational parameters. Recall that $\rho_{(i,k)}$ and $\tau_{(i,k)}$ are the variational posterior mean and variance corresponding to element a_{ik} respectively.

Notice that, in the final line, the indices i and u are used, merely to distinguish the two summations. Since we only require to maximise $\mathcal{L}_{\boldsymbol{\mu}}(q)$ with respect to τ, a little extra work can be saved by computing only those outstanding terms in (5.48) that will depend upon τ. Clearly, $\sum_{i=1}^{d} x_{ji}^2 \mathrm{Var}_{q(\mathbf{a}\,|\,D)}\{a_{ik}\} = \sum_{i=1}^{d} x_{ji}^2 \tau_{(i,k)}$. Moreover,

$$\begin{aligned}
&\mathrm{E}_{q(\boldsymbol{\mu}\,|\,D)}\left\{\sum_{i=1}^{d} m_{ji}^2 \mathrm{Var}_{q(\mathbf{a}\,|\,D)}\{a_{ik}\}\right\} \\
&= \sum_{i=1}^{d} \tau_{(i,k)} \left(\left[\mathrm{E}_{q(\boldsymbol{\mu}\,|\,D)}\{m_{ji}\}\right]^2 + \mathrm{Var}_{q(\boldsymbol{\mu}\,|\,D)}\{m_{ji}\}\right) \\
&= \sum_{i=1}^{d} \tau_{(i,k)} \left(\Omega_{ji}^2 + \Lambda_{ii}\right).
\end{aligned}$$

Here, we realise that $\mathrm{Var}_{q(\boldsymbol{\mu}\,|\,D)}\{m_{ji}\} = \mathrm{Var}_{q(\boldsymbol{\mu}\,|\,D)}\{\mu_i\} = \Lambda_{ii}$, by construction of $M = (m_{ij})$. Furthermore,

$$\begin{aligned}
&\mathrm{E}_{q(\mathbf{a}\,|\,D)}\left\{\sum_{i=1}^{d} x_{ji} a_{ik} \sum_{u=1}^{d} \Omega_{ju} a_{uk}\right\} \\
&= \mathrm{E}_{q(\mathbf{a}\,|\,D)}\left\{\sum_{i=1}^{d} x_{ji} a_{ik}\right\} \mathrm{E}_{q(\mathbf{a}\,|\,D)}\left\{\sum_{u=1}^{d} \Omega_{ju} a_{uk}\right\} + \sum_{i=1}^{d}\sum_{u=1}^{d} x_{ji}\Omega_{ju} \mathrm{Cov}_{q(\mathbf{a}\,|\,D)}\{a_{ik},\, a_{uk}\} \\
&= \sum_{i=1}^{d} x_{ji}\rho_{(i,k)} \sum_{u=1}^{d} \Omega_{ju}\rho_{(u,k)} + \sum_{i=1}^{d} x_{ji}\Omega_{ji}\tau_{(i,k)}.
\end{aligned}$$

Notice that $\mathrm{Cov}_{q(\mathbf{a}\,|\,D)}\{a_{ik},\, a_{uk}\} \neq 0$ only if $i = u$ as τ is diagonal. The above computations can then be substituted back into (5.48). Consequently, by dropping all terms independent of τ, we obtain

$$\begin{aligned}
&\iiint q(\mathbf{a}\,|\,D) q(\sigma^2\,|\,D) q(\boldsymbol{\mu}\,|\,D) \log p(D\,|\,\{a_{ij}\},\, \sigma^2,\, \boldsymbol{\mu})\, \mathrm{d}\mathbf{a}\, \mathrm{d}\sigma^2\, \mathrm{d}\boldsymbol{\mu} \\
&\propto -\frac{\gamma}{2\delta} \sum_{j=1}^{N}\sum_{k=1}^{d}\sum_{i=1}^{d} \tau_{(i,k)} \left[(x_{ji} - \Omega_{ji})^2 + \Lambda_{ii}\right].
\end{aligned}$$

Calculating the additional integrals in (5.42) is straightforward since the prior and variational posterior for $\mathbf{a}$ are, in effect, identical to those in the zero mean case. Hence, to recap the results from Section 3.3.3,

$$\int q(\mathbf{a}\,|\,D)\log p(\mathbf{a})\,\mathrm{d}\mathbf{a} = -\sum_{(p,q)\in I}\left[\frac{1}{2}\log 2\pi + \frac{1}{2}\log C^*_{(p,q)} + \frac{1}{2}\frac{\tau_{(p,q)}}{C^*_{(p,q)}} + \frac{1}{2}\frac{\rho^2_{(p,q)}}{C^*_{(p,q)}}\right]$$

$$\int q(\mathbf{a}\,|\,D)\log q(\mathbf{a}\,|\,D)\,\mathrm{d}\mathbf{a} = -\sum_{(p,q)\in I}\left[\frac{1}{2}\log 2\pi + \frac{1}{2}\log \tau_{(p,q)} + \frac{1}{2}\right],$$

where $(p,q)\in I$ if and only if element $a_{pq}\neq 0$. All other terms in (5.42) are independent of τ. Thus, in component form and as a function of τ, the lower bound is now such that

$$\mathcal{L}_{\boldsymbol{\mu}}(q) \propto -\frac{\gamma}{2\delta}\sum_{j=1}^{N}\sum_{k=1}^{d}\sum_{i=1}^{d}\tau_{(i,k)}\left[(x_{ji}-\Omega_{ji})^2+\Lambda_{ii}\right] - \frac{1}{2}\sum_{(p,q)\in I}\left[\frac{\tau_{(p,q)}}{C^*_{(p,q)}} - \log\tau_{(p,q)}\right]. \quad (5.49)$$

Maximising (5.49) with respect to the element $\tau_{(p,q)}$ thus provides

$$\frac{\partial\mathcal{L}_{\boldsymbol{\mu}}(q)}{\partial\tau_{(p,q)}} = -\frac{\gamma}{2\delta}\left(\sum_{j=1}^{N}\left[(x_{jp}-\Omega_{jp})^2\right] + N\Lambda_{pp}\right) - \frac{1}{2}\left[\frac{1}{C^*_{(p,q)}} - \frac{1}{\tau_{(p,q)}}\right].$$

Upon equating to zero, this equation can be quickly solved for the non-zero $\tau_{(p,q)}$. Hence, the diagonal elements of τ are declared as

$$\tau_{(p,q)} = \begin{cases} \left(\dfrac{1}{C^*_{(p,q)}} + \dfrac{\gamma}{\delta}\displaystyle\sum_{j=1}^{N}(x_{jp}-\Omega_{jp})^2 + \dfrac{N\gamma}{\delta}\Lambda_{pp}\right)^{-1} & \text{if } a_{pq}\neq 0 \\ 0 & \text{if } a_{pq}=0 \end{cases}.$$

To conclude, update equations have been derived for the variational parameters of σ^2, namely γ and δ, and, moreover, for those of $\boldsymbol{\mu}$, that is $\boldsymbol{\omega}$ and Λ. Furthermore, we can use $\boldsymbol{\rho}_1$ and $\tau_{(p,q)}$ to construct $\boldsymbol{\rho}$ and τ, the parameters of $\mathbf{a}$. By running until convergence, parameter values for $q(\sigma^2\,|\,D)$, $q(\boldsymbol{\mu}\,|\,D)$ and $q(\mathbf{a}\,|\,D)$ are acquired. Of course, the converged value of $\mathcal{L}_{\boldsymbol{\mu}}(q)$ will provide evidence for each model. As before, we can rewrite the lower bound (5.44) so that constant terms across models are disregarded:

$$\begin{aligned}\mathcal{L}_{\boldsymbol{\mu}}(q) \propto -\frac{dN}{2}\log\delta - \frac{\gamma}{2\delta}\left[h(\boldsymbol{\rho},\,\Omega) + j(\boldsymbol{\rho},\,\tau,\,\Omega,\,\Lambda)\right] - \frac{1}{2}\log|C^*| - \frac{1}{2}\boldsymbol{\rho}^T {C^*}^+ \boldsymbol{\rho} \\ - \frac{1}{2}\mathrm{Tr}\left[{C^*}^+\tau\right] - \alpha\log\delta - \frac{\beta\gamma}{\delta} - \frac{1}{2}\left(\boldsymbol{\omega} - \mathbf{b}\right)\Delta^{-1}\left(\boldsymbol{\omega} - \mathbf{b}\right)^T \\ - \frac{1}{2}\mathrm{Tr}\left[\Delta^{-1}\Lambda\right] + \frac{1}{2}\log|\tau| + \frac{1}{2}\log|\Lambda|. \qquad (5.50)\end{aligned}$$

Here, C^* is now inverted using the Moore-Penrose inverse, ${C^*}^+$, as explained in Section 3.4.1.

5.3 Toy example

The methodology discussed thus far in this chapter is elucidated via a straightforward example. Here, we examine an arbitrary non-zero mean VAR(1) model. In fact, the true model is chosen with specifications identical to those given in the corresponding example for the zero mean case in Section 3.5. Moreover here, the mean is specified as $\boldsymbol{\mu} = (1,\,1)$, a row vector. Consequently, a dataset was generated from the model (5.1), where A is represented graphically by Figure 3.3. We let $\mathbf{x}_1 = \boldsymbol{\mu}$ and $\mathbf{x}_{250} = \mathbf{y}_{249}$. Again, a candidate set of 15 A-graphs, ignoring the null graph, is constructed, each of which is scored using $\mathcal{L}_{\boldsymbol{\mu}}(q)$.

The choice of prior parameter values is made as before for both $\mathbf{a}$ and σ^2. That is, we avoid Lindley's paradox, namely that a simpler model will be favoured as the prior is made to be more diffuse, by choosing an informative prior of the form $\mathcal{N}(0,\,C^*)$ on $\mathbf{a}$, where $c_{ij} \in \{0,\,0.5\}$. Moreover, a vague $\mathcal{IG}(1,\,0.001)$ prior is specified on σ^2. For further details, refer back to Section 3.4.2. Furthermore, a new prior is required for $\boldsymbol{\mu}$. This is denoted as

$$p(\boldsymbol{\mu}) = \mathcal{N}(\boldsymbol{\mu}\,|\,0,\,10{,}000\mathrm{I}_d).$$

In this specification, it seems sensible to centre the distribution at the zero mean case. Each element of the vector $\boldsymbol{\mu}$ is then assigned prior variance equal to $10,000$. Hence, prior ignorance is represented since the prior is not concentrated around any particular value. By running all update equations until convergence, we sought to find values for the variational parameters for each variational distribution, given by (5.19), (5.35) and (5.39). As erstwhile, convergence of update equations and lower bound values took 4 iterations. The expressions for $\boldsymbol{\omega}$ and Λ were run first so that, to start the algorithm, initial, arbitrary choices were made such that $\gamma = \delta = 1$ and $\rho_{(i,j)} = \tau_{(i,j)} = 1$, whenever $a_{ij} \neq 0$.

The results obtained in this case are revealed in Table 5.1. Consider firstly the values of the lower bound, $\mathcal{L}_{\boldsymbol{\mu},\, M_i}(q_i)$. As in the zero mean case, the true A was deemed to be the most plausible model in the candidate set. Moreover, those models, containing at least the two, correct free elements, were again ranked highly. Hence, the more complex models were penalised sufficiently by the choice of informative prior variance on $\mathbf{a}$. Candidates with neither of the true non-zero elements of A unsurprisingly fared poorly, thus indicating a very weak signal in the data for the two, true zero elements being non-zero.

In addition, the posterior means of A, σ^2 and $\boldsymbol{\mu}$ are now inspected and compared to the truth. With a dataset of size $N = 250$, $\hat{A}$ and $\mathrm{E}_{q(\boldsymbol{\mu}\,|\,D)}\{\boldsymbol{\mu}\}$ are reasonably close to the truth for each candidate A-matrix. In fact, in both cases, the estimates are extremely akin to each other. In particular, if we misspecify the model, the estimates for $\boldsymbol{\mu}$, the point about which the data fluctuates, are unaffected. However, those for σ^2 tend to show more discrepancy, a scenario also seen in the zero mean case. If a candidate model was specified with at least the correct free elements seen in the truth, the afore-mentioned estimates were extremely accurate. Yet, if the wrong model was chosen, *i.e.* an incorrect sparsity structure of A, the resulting error provided inaccuracy in $\mathrm{E}_{q(\sigma^2\,|\,D)}(\sigma^2)$.

Finally, we compare Tables 3.1 and 5.1, the zero and non-zero mean models respectively. It is seen in these tables that the estimates of σ^2 for each candidate are almost identical.

Specification	Posterior means			$\mathcal{L}_{\boldsymbol{\mu},\,M_i}(q_i)$
A-matrix	$\hat{A}$-matrix	$\mathrm{E}_{q(\sigma^2\,\vert\,D)}\{\sigma^2\}$	$\mathrm{E}_{q(\boldsymbol{\mu}\,\vert\,D)}\{\boldsymbol{\mu}\}$	
$\begin{pmatrix} * & 0 \\ 0 & 0 \end{pmatrix}$	$\begin{pmatrix} -0.069 & 0 \\ 0 & 0 \end{pmatrix}$	0.134	(0.973, 0.951)	−1142.389
$\begin{pmatrix} 0 & 0 \\ 0 & * \end{pmatrix}$	$\begin{pmatrix} 0 & 0 \\ 0 & -0.055 \end{pmatrix}$	0.134	(0.973, 0.951)	−1142.610
$\begin{pmatrix} 0 & * \\ 0 & 0 \end{pmatrix}$	$\begin{pmatrix} 0 & 0.667 \\ 0 & 0 \end{pmatrix}$	0.109	(0.973, 0.951)	−1091.392
$\begin{pmatrix} 0 & 0 \\ * & 0 \end{pmatrix}$	$\begin{pmatrix} 0 & 0 \\ 0.338 & 0 \end{pmatrix}$	0.125	(0.974, 0.951)	−1126.006
$\begin{pmatrix} * & * \\ 0 & 0 \end{pmatrix}$	$\begin{pmatrix} -0.070 & 0.662 \\ 0 & 0 \end{pmatrix}$	0.109	(0.973, 0.951)	−1093.702
$\begin{pmatrix} * & 0 \\ * & 0 \end{pmatrix}$	$\begin{pmatrix} -0.072 & 0 \\ 0.338 & 0 \end{pmatrix}$	0.125	(0.973, 0.951)	−1128.282
$\begin{pmatrix} 0 & * \\ 0 & * \end{pmatrix}$	$\begin{pmatrix} 0 & 0.668 \\ 0 & -0.060 \end{pmatrix}$	0.109	(0.973, 0.951)	−1093.840
$\begin{pmatrix} 0 & 0 \\ * & * \end{pmatrix}$	$\begin{pmatrix} 0 & 0 \\ 0.338 & -0.055 \end{pmatrix}$	0.125	(0.973, 0.951)	−1128.561
$\begin{pmatrix} * & 0 \\ 0 & * \end{pmatrix}$	$\begin{pmatrix} -0.069 & 0 \\ 0 & -0.055 \end{pmatrix}$	0.134	(0.973, 0.951)	−1144.945
$\begin{pmatrix} 0 & * \\ * & 0 \end{pmatrix}$	$\begin{pmatrix} 0 & 0.669 \\ 0.340 & 0 \end{pmatrix}$	0.100	(0.974, 0.952)	−1073.091
$\begin{pmatrix} * & * \\ 0 & * \end{pmatrix}$	$\begin{pmatrix} -0.070 & 0.663 \\ 0 & -0.060 \end{pmatrix}$	0.109	(0.973, 0.951)	−1096.150
$\begin{pmatrix} * & 0 \\ * & * \end{pmatrix}$	$\begin{pmatrix} -0.072 & 0 \\ 0.338 & -0.055 \end{pmatrix}$	0.125	(0.973, 0.951)	−1130.837
$\begin{pmatrix} * & * \\ * & 0 \end{pmatrix}$	$\begin{pmatrix} -0.073 & 0.664 \\ 0.341 & 0 \end{pmatrix}$	0.100	(0.974, 0.952)	−1075.335
$\begin{pmatrix} 0 & * \\ * & * \end{pmatrix}$	$\begin{pmatrix} 0 & 0.670 \\ 0.340 & -0.060 \end{pmatrix}$	0.100	(0.974, 0.952)	−1075.536
$\begin{pmatrix} * & * \\ * & * \end{pmatrix}$	$\begin{pmatrix} -0.073 & 0.665 \\ 0.340 & -0.060 \end{pmatrix}$	0.100	(0.974, 0.952)	−1077.778

Table 5.1: Lower bounds and posterior means for each non-zero mean VAR(1) model

Here, there is a correspondence between the two cases whereby, if we select an A-matrix with an erroneous sparsity pattern, the variational posterior mean of the noise variance suffers, regardless of our beliefs about $\boldsymbol{\mu}$. The most notable difference stems from the values of $\mathcal{L}_{M_i}(q_i)$ and $\mathcal{L}_{\boldsymbol{\mu}, M_i}(q_i)$. Although the correct model is selected in each case, the lower bound values are higher when $\boldsymbol{\mu} = \mathbf{0}$. When the mean is unknown, the resulting uncertainty in the problem implies that the approximate evidence for each model, denoted by $\mathcal{L}_{\boldsymbol{\mu}, M_i}(q_i)$, is reduced. This may also account for the slight discrepancy between the corresponding $\hat{A}$-matrices. These estimates are very similar, although marginally more inaccurate to the truth in the non-zero mean case.

5.4 Taking a random walk

Hitherto in this chapter, variational Bayesian methods have been utilised to derive an approximation, $\mathcal{L}_{\boldsymbol{\mu}, M_i}(q_i)$, to the logarithm of the marginal likelihood, $p(D \mid M_i)$. Hence, we were able to score non-zero mean VAR(1) models, in particular for graphs with a small number of nodes. However, we can apply the methods of Chapter 4 to find the most plausible models in graphical spaces of higher dimension. In particular, the variational algorithm, presented in this chapter, can again be embedded within the Metropolis-Hastings scheme, given by Algorithm 5, so that a random walk can be made across the space. The principles behind the MCMC algorithm remain the same — a new model is proposed by the addition or deletion of a randomly selected edge from the current model, and is accepted on the basis of a log acceptance probability.

For analysis, trace and ACF plots can be used to test for the convergence of the chain, as well as the more formal diagnostics previously described. Furthermore, `image` plots of the counting matrix $\hat{\Pi}$ are produced, which will be dependent on the choice of $p = \mathrm{P}(a_{ij} = 0)$. When using simulated data, $\hat{\Pi}$ can then be normalised, and hence compared to the true adjacency matrix Π by computing the residual sum of squares, S (*c.f.* (4.11)).

Moreover, approximate posterior summaries can be produced for the coefficients a_{ij} of the matrix A. We realise that of additional inferential interest here are the components of the mean vector $\boldsymbol{\mu}$ across models. As the prior specification for $\boldsymbol{\mu}$, given by (5.8), is an equivalent choice for every model, its conditioning on M_i (although not stated explicitly) can be dropped. Thus, we wish to update the prior

$$p(\mu_j) = \mathcal{N}(\mu_j \,|\, b_j,\, \Delta_{jj}),$$

and subsequently infer the marginal posterior $p(\mu_j \,|\, D)$, where $j = 1, \ldots, d$. As before, Bayesian model averaging can be applied to estimate this true density, *i.e.* for a converged chain of length n, we average all variational densities for μ_j (*c.f.* (5.35)) that are associated with the models accepted across the scheme. So, akin to (4.12) and using the corresponding notation, we aim to compute

$$p_{\text{var}}(\mu_j \,|\, D) = \frac{1}{n} \sum_{k=1}^{n} \mathcal{N}\left(\mu_j \,|\, \omega_j^{(k)},\, \Lambda_{jj}^{(k)}\right). \tag{5.51}$$

5.4.1 Examples

In the following, the same specifications were maintained from Section 4.3.2, *i.e.* $d = 10$, $N = 250$ and $\sigma^2 = 0.1$. The prior on $\mathbf{a}$ was such that $c_{ij} \in \{0,\, 0.5\}$ and for that on σ^2, $\alpha = 1$, $\beta = 0.001$. Moreover, a $\mathcal{N}(0,\, 10,000\mathrm{I}_d)$ prior was allowed for $\boldsymbol{\mu}$ as in Section 5.3. By now simulating data from the non-zero mean VAR(1) model (5.1), only A, p and also now $\boldsymbol{\mu}$ were changed between examples. The MCMC algorithm was initialised from the graph with only one self-loop on node y_1, and run in `C` for $10,000,000$ iterations, of which the first $100,000$ were discarded as burn-in and the remainder thinned by 1000.

Example 1

We allow direct comparison between this and the corresponding first example in Section 4.3.2 by specifying $A = \text{diag}(0.8)$ and $p = 0.5$, but, furthermore, $\boldsymbol{\mu} = (1, \ldots, 1)$, a 10-vector. The output of the scheme is displayed graphically in Figure 5.1.

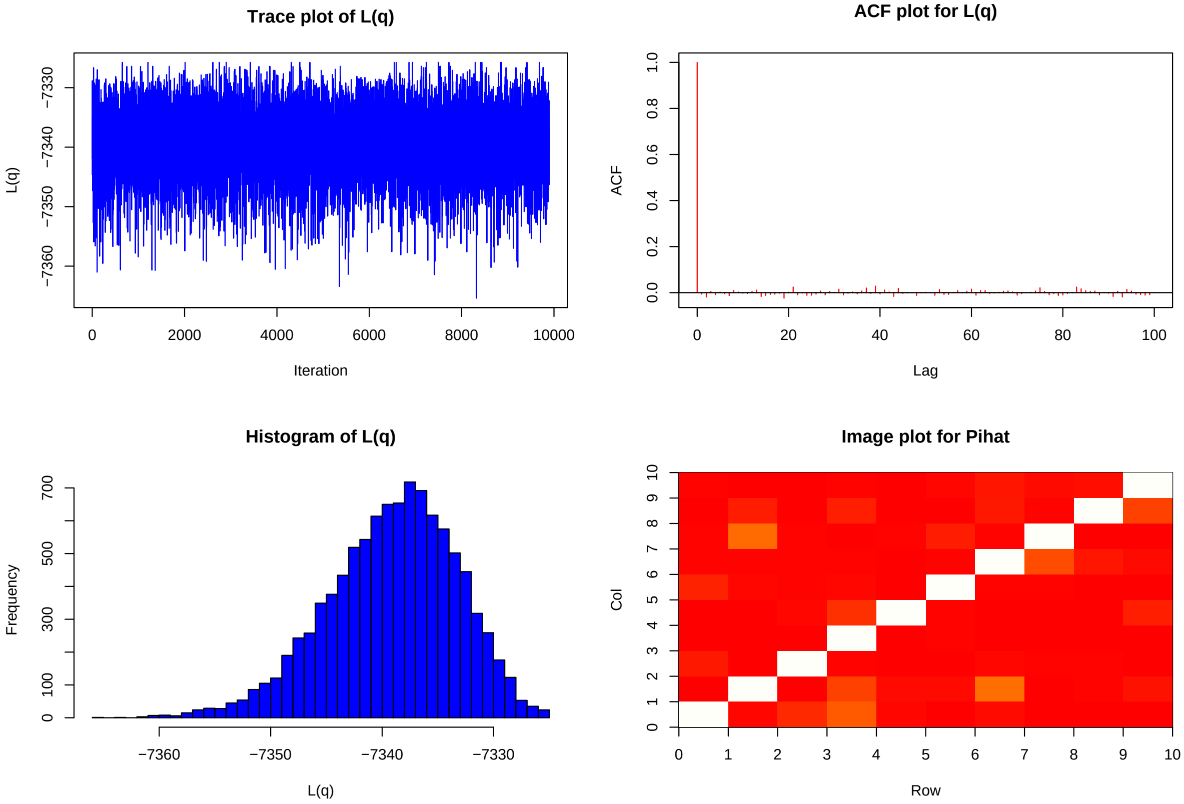


Figure 5.1: Plots for the analysis of the MCMC output in Example 1 (non-zero mean)

The trace and ACF plots are extremely similar to those in the corresponding zero mean example, indicating a well-mixing and independent chain. The only significant different is that the mean value of the lower bound, about which the values of the chain fluctuate, is greater in this case due to the additional uncertainty about $\boldsymbol{\mu}$. Moreover, the `effectiveSize` of the chain is equal to 9900, indicating that the chain is fully independent. When applied to the variational scores, the Raftery-Lewis test yielded the following:

```
Iterations = 1:9900
Thinning interval = 1
Number of chains = 1
Sample size per chain = 9900

Quantile (q) = 0.025
Accuracy (r) = +/- 0.005
Probability (s) = 0.95

    Burn-in  Total Lower bound  Dependence
    (M)      (N)   (Nmin)       factor (I)
Lq   2       3794  3746         1.01
```

The Heidelberger-Welch diagnostic reached a similar conclusion:

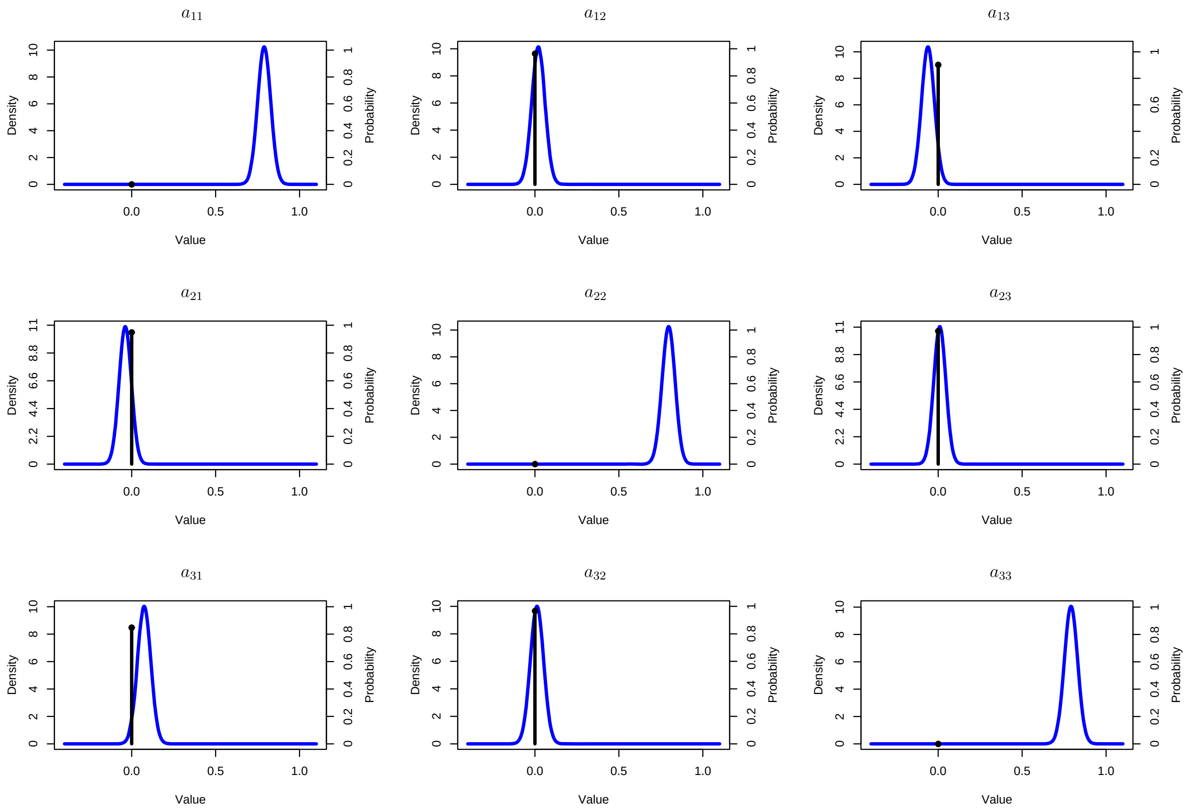
```
     Stationarity start     p-value
     test         iteration
Lq   passed       1         0.0708

     Halfwidth Mean  Halfwidth
     test
Lq   passed    -7339 0.114
```

Each diagnostic has produced overwhelming evidence in favour of the chain having converged. In addition, when employing these tests for components of $\boldsymbol{\rho}$, τ, $\boldsymbol{\omega}$ and Λ, stored at each iteration, the results produced were concurrent with those above.

Moreover, it is evident from Figures 4.3 and 5.1 that the `image` plots of $\hat{\Pi}$ in both the zero and non-zero mean cases are well-matched and, after normalising, will be close to the truth Π. In this case, the residual sum of squares is computed as $S = 0.757$, a value only marginally bigger than that in the zero mean case. This implies that there is sufficient data available here to learn the unknown mean, and hence the zero and non-zero mean cases subsequently become most alike. So, the new variational algorithm, derived in this chapter, is able to accurately predict the sparsity structure of the true A from the simulated data.

Finally, Figures 5.2 and 5.3 display approximate posterior information for numerous coefficients of both A and $\boldsymbol{\mu}$.



a_{11} a_{12} a_{13}

a_{21} a_{22} a_{23}

a_{31} a_{32} a_{33}

Figure 5.2: Plots showing estimated, marginal posterior distributions for a_{ij}, $i,\ j = 1,\ 2,\ 3$, in Example 1 (non-zero mean)

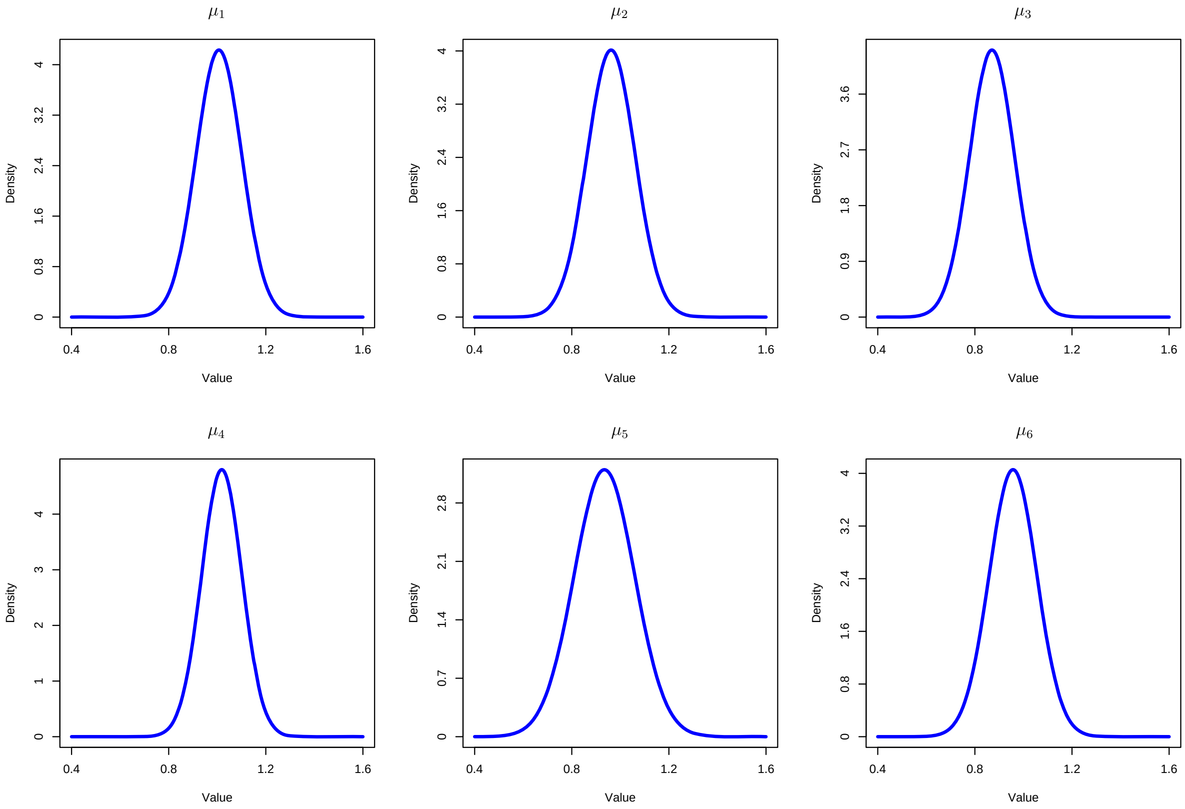


Figure 5.3: Plots showing estimated, marginal posterior distributions for μ_j, $j = 1, \ldots, 6$, in Example 1

As in the zero mean case, the approximate posterior probability that $a_{ij} = 0$ is large for all off-diagonal elements. Moreover, the density plots, given that $a_{ij} \neq 0$, are peaked at around the true specification for the diagonal entries. We realise that Figures 4.4 and 5.2 are almost identical. As mentioned above, this is a consequence of the mean being estimated accurately, a fact borne out by Figure 5.3. Here, as expected, we see that the approximate posterior mode of each μ_j is close to 1.

Example 2

We now choose $A = \text{tridiag}(0.2, 0.4, 0.2)$, as in the corresponding zero mean case. However, on this occasion, the prior probability p is assigned such that $p = 0.9$. In addition, the mean is specified as $\boldsymbol{\mu} = (2, \ldots, 2)$. Figure 5.4 reveals the results of the algorithm.

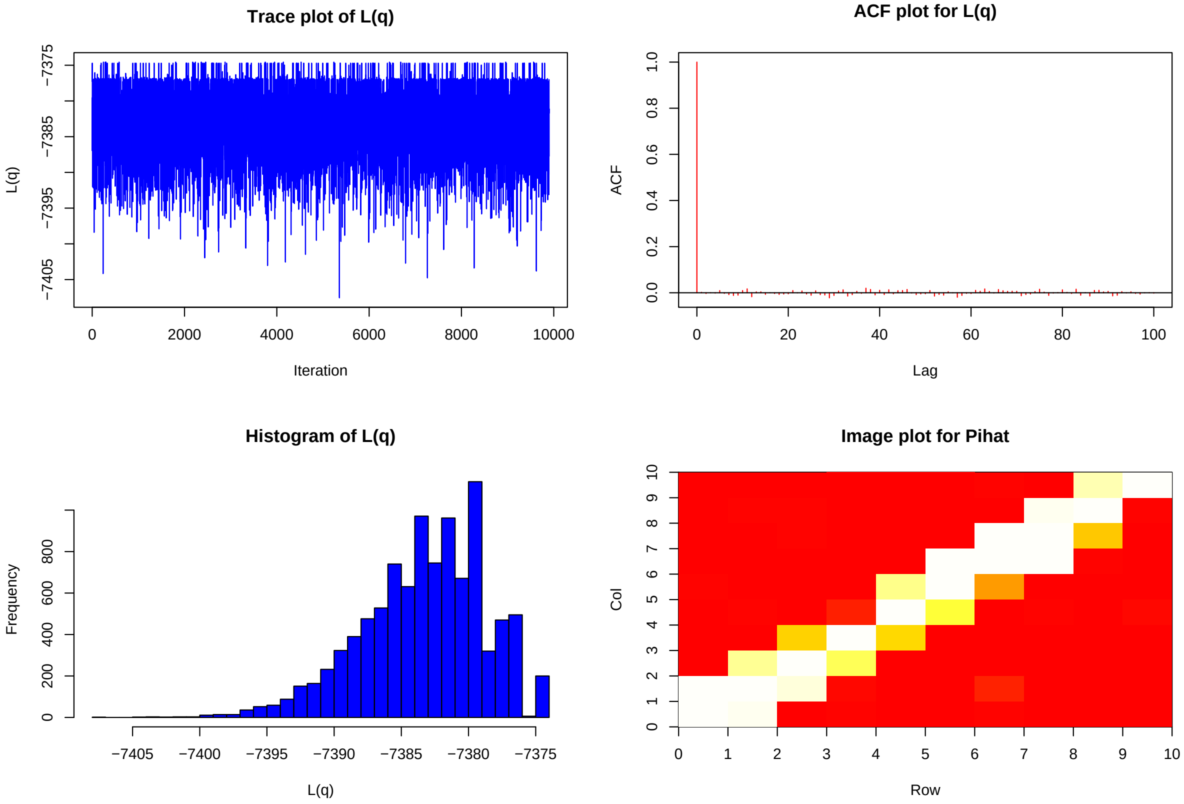


Figure 5.4: Plots for the analysis of the MCMC output in Example 2 (non-zero mean)

From the plots, it is seen that the chain is moving freely and quickly in the graphical space, as well as the autocorrelation dropping to zero immediately. Moreover, the `effectiveSize` of the chain, once again computed as 9900, intimates full independence of the values. Both convergence diagnostics produce favourable results, whereby the Raftery-Lewis output is

```
Iterations = 1:9900
Thinning interval = 1
Number of chains = 1
Sample size per chain = 9900

Quantile (q) = 0.025
Accuracy (r) = +/- 0.005
Probability (s) = 0.95
```

```
        Burn-in  Total Lower bound  Dependence
        (M)      (N)   (Nmin)       factor (I)
   Lq    2       3768  3746         1.01
```

and the Heidelberger-Welch test gives the results

```
         Stationarity start     p-value
         test         iteration
   Lq    passed       1         0.657

         Halfwidth Mean  Halfwidth
         test
   Lq    passed    -7383 0.0875
```

Again, upon application of these diagnostics to components of $\boldsymbol{\rho}$, τ, $\boldsymbol{\omega}$ and Λ, the same outcome was provided. Henceforth, only the chain of lower bound is thus analysed.

Noticeable differences are apparent upon comparison of the `image` plots in the two cases due to the change in specification of p. In this case, $p = 0.9$ was chosen to be higher than the 'true' value (computed as $p = 0.72$). Consequently, a slight preference has been given to the acceptance of models considered too sparse, as displayed in Figure 5.4. That is, many true edges are identified from the data with less regularity than seen in Figure 4.5, whilst the link from y_5 to y_4 is no-longer recognised. This trait is reflected by the calculation of $S = 2.619$, a value relatively less accurate than in the zero mean circumstance, where $p = 0.5$.

By studying Figure 5.5, it follows that the true specifications of a_{ij} are being well represented in these graphical summaries. Moreover, it is clear that there is much similarity between these plots and those in the zero mean case, shown in Figure 4.6. This is despite the choice of $p = \mathrm{P}(a_{ij} = 0)$ being increased here. Although this implies a bias for the selection of more sparse models, the variational algorithm is still able to predict accurately those values of a_{ij} that are not constrained to zero in all models accepted across the scheme. In addition, it is obvious from Figure 5.6 that all plots of $p_{\text{var}}(\mu_j \,|\, D)$ are peaked near to the true value. At each iteration, accurate estimates of all μ_j have been

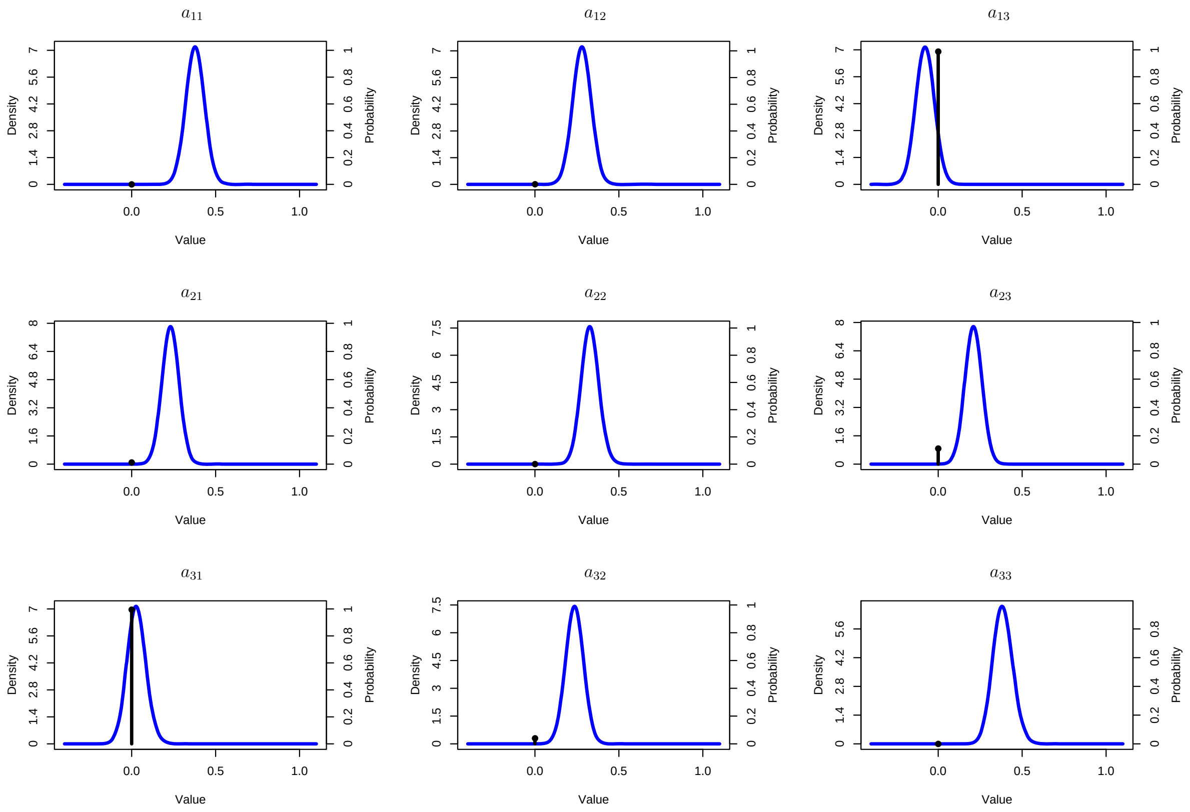


Figure 5.5: Plots showing estimated, marginal posterior distributions for a_{ij}, $i,\ j = 1,\ 2,\ 3$, in Example 2 (non-zero mean)

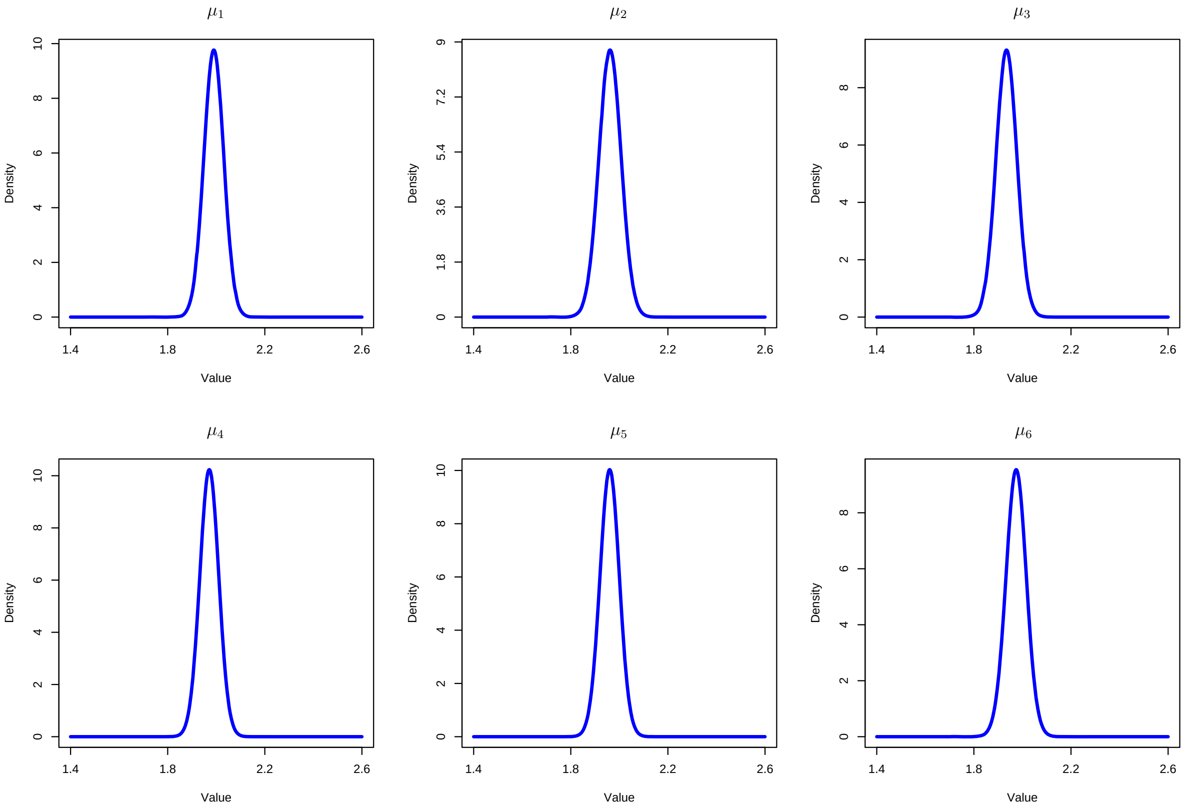


Figure 5.6: Plots showing estimated, marginal posterior distributions for μ_j, $j = 1, \ldots, 6$, in Example 2

provided, and uncertainty about every component reduced. Hence, as in the previous example, the density plots are alike in shape.

Example 3

The third example in Section 4.3.2 was also repeated with A = tridiag(0.4, 0, 0.4) and $p = 0.9$ as before, but now $\boldsymbol{\mu} = (3, \ldots, 3)$. The output was analysed and is displayed graphically below.

When calculating convergence diagnostics, the Raftery-Lewis test yielded

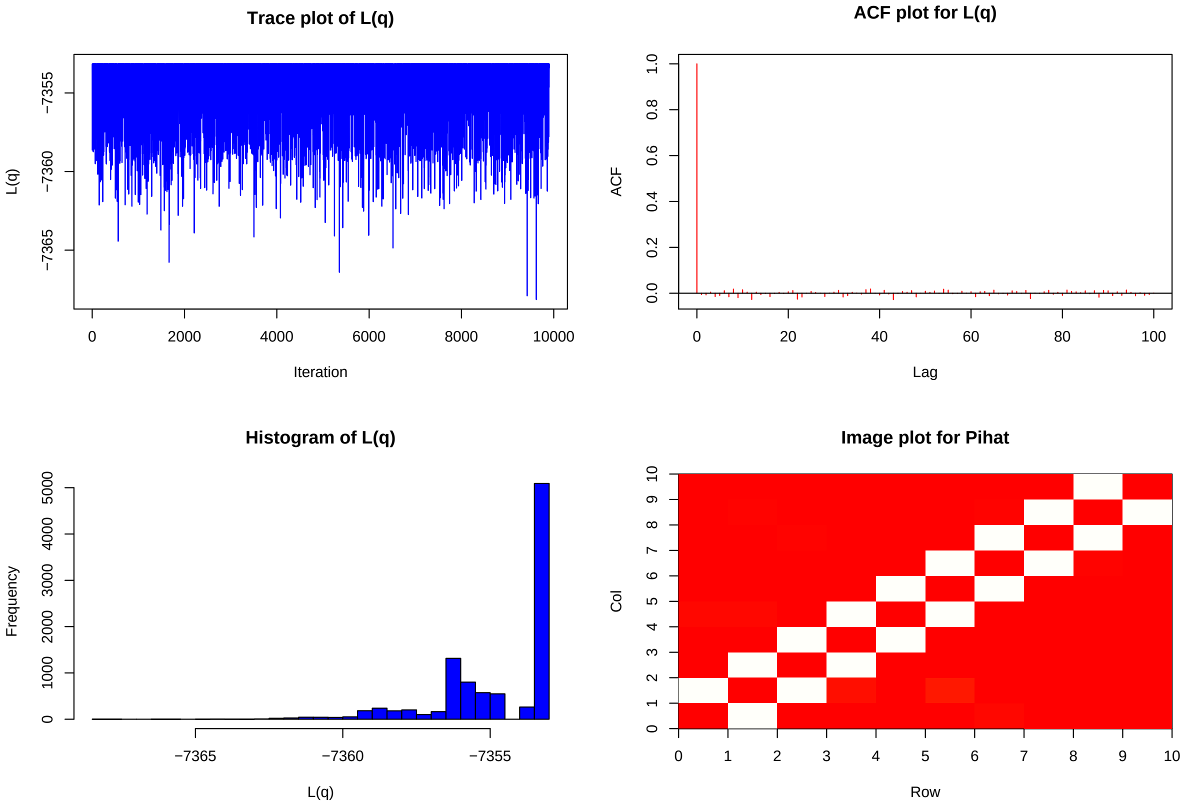


Figure 5.7: Plots for the analysis of the MCMC output in Example 3 (non-zero mean)

```
Iterations = 1:9900
Thinning interval = 1
Number of chains = 1
Sample size per chain = 9900

Quantile (q) = 0.025
Accuracy (r) = +/- 0.005
Probability (s) = 0.95

    Burn-in  Total Lower bound  Dependence
    (M)      (N)   (Nmin)       factor (I)
Lq   2       3856  3746         1.03
```

Moreover, application of Heidelberger and Welch resulted in

```
     Stationarity start     p-value
     test         iteration
Lq   passed       1         0.113
```

```
          Halfwidth Mean  Halfwidth
          test
     Lq   passed    -7355 0.0386
```

Therefore, the plots and diagnostics are all concurrent with chain convergence and independence of values (`effectiveSize` $= 9900$). Notice that the histograms produced in Figures 4.7 and 5.7 are almost identical due to the larger specification of p in each case. Similarly, the `image` plots also overlap significantly, revealing an obvious tendency to select models that are not dense. For completeness, we note that $S = 0.015$ here. To conclude, we again realise that the approximate marginal posterior summaries for both a_{ij} (Figure 5.8) and μ_j (Figure 5.9) are a strong reflection of the truth.

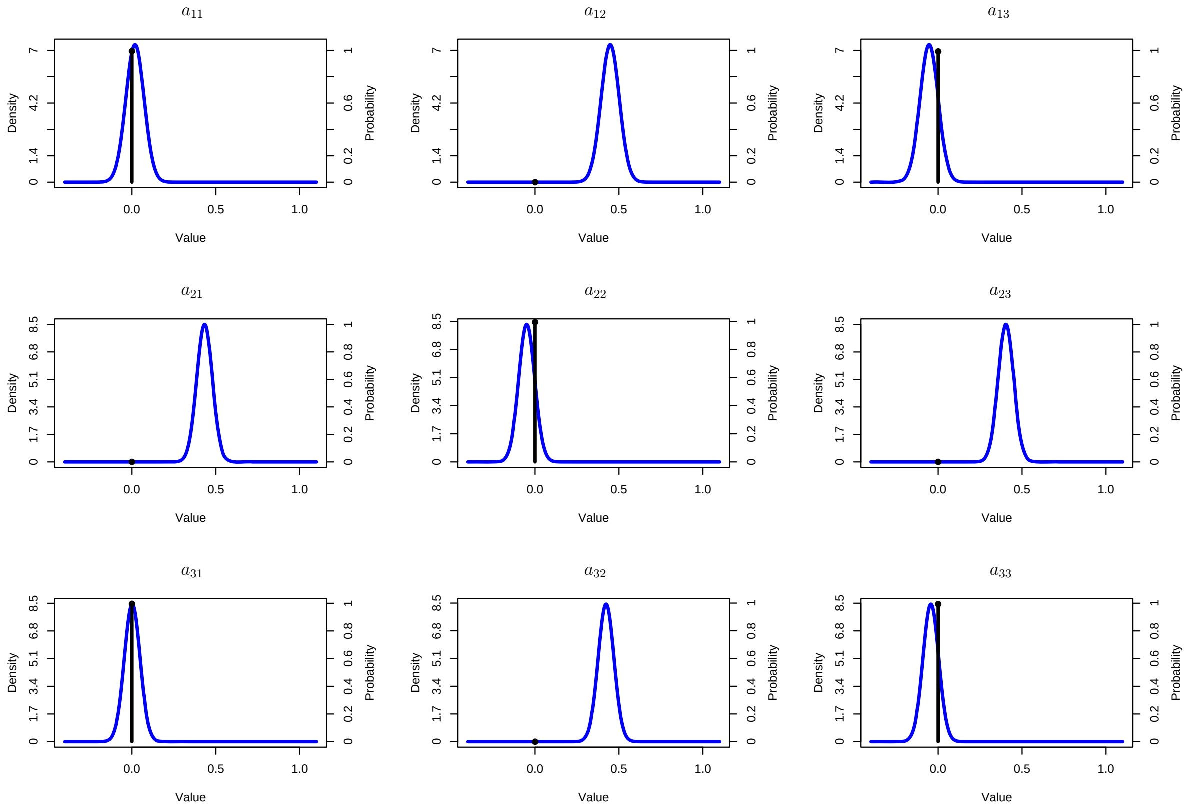


Figure 5.8: Plots showing estimated, marginal posterior distributions for a_{ij}, i, $j = 1$, 2, 3, in Example 3 (non-zero mean)

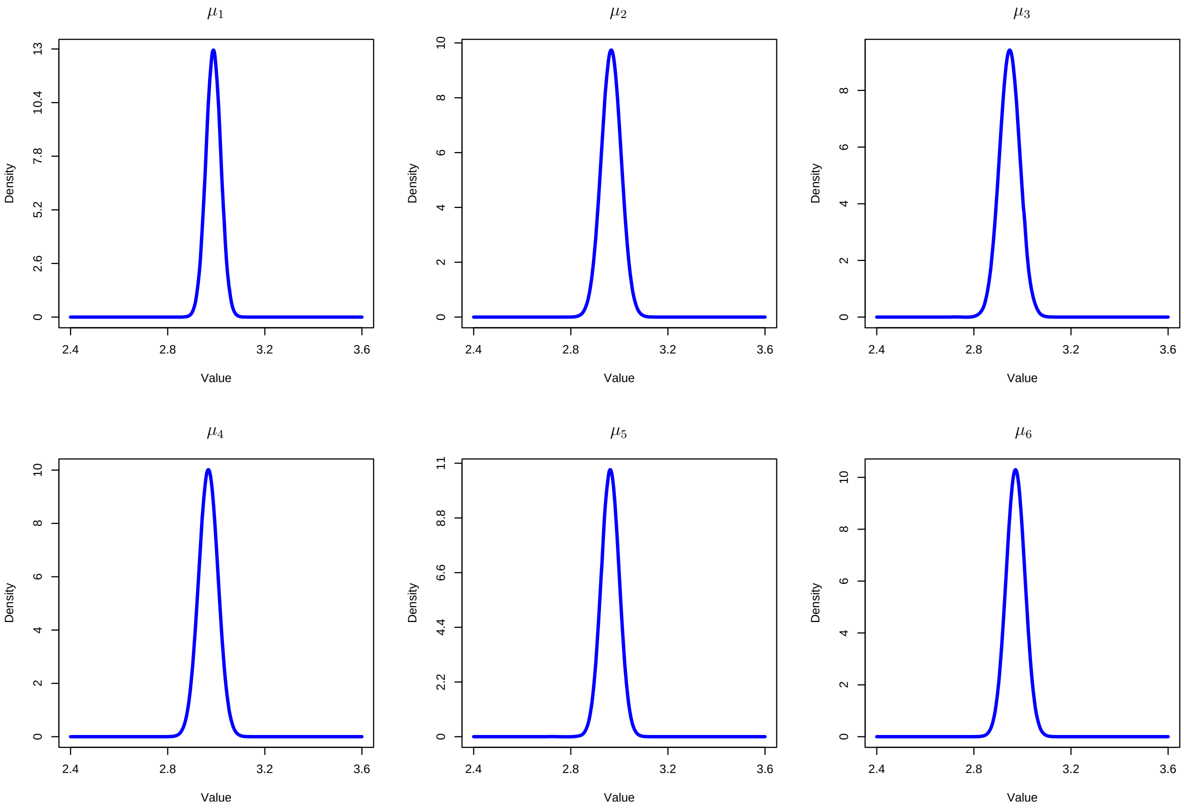


Figure 5.9: Plots showing estimated, marginal posterior distributions for μ_j, $j = 1, \ldots, 6$, in Example 3

5.4.2 Application to ERP data

We now reconsider the ERP data, introduced previously in Section 4.3.6. It is recalled that an animal and distractor dataset, each of size $N = 250$, were obtained by monitoring the cerebral activity produced at $d = 32$ electrodes for a particular volunteer in the study (shown in Figures 4.14 and 4.15). On this occasion, each dataset was fitted to a non-zero mean VAR(1) model. The sparsity structure of A and the likely values of the coefficients a_{ij}, μ_j are of inferential interest.

Here, the sample mean of ERP values was not subtracted from the data since the true mean is itself estimated during the algorithm. Thus, the Metropolis-Hastings scheme was run twice in an identical fashion to that described in Section 5.4.1. The prior distributions for $\mathbf{a}$, σ^2 and $\boldsymbol{\mu}$ were also chosen as here, whereas we fixed $p = \frac{31}{32}$ in accordance with the zero mean case. Figures 5.10 and 5.11 display the graphical summaries of the sampler for the animal and distractor datasets respectively.

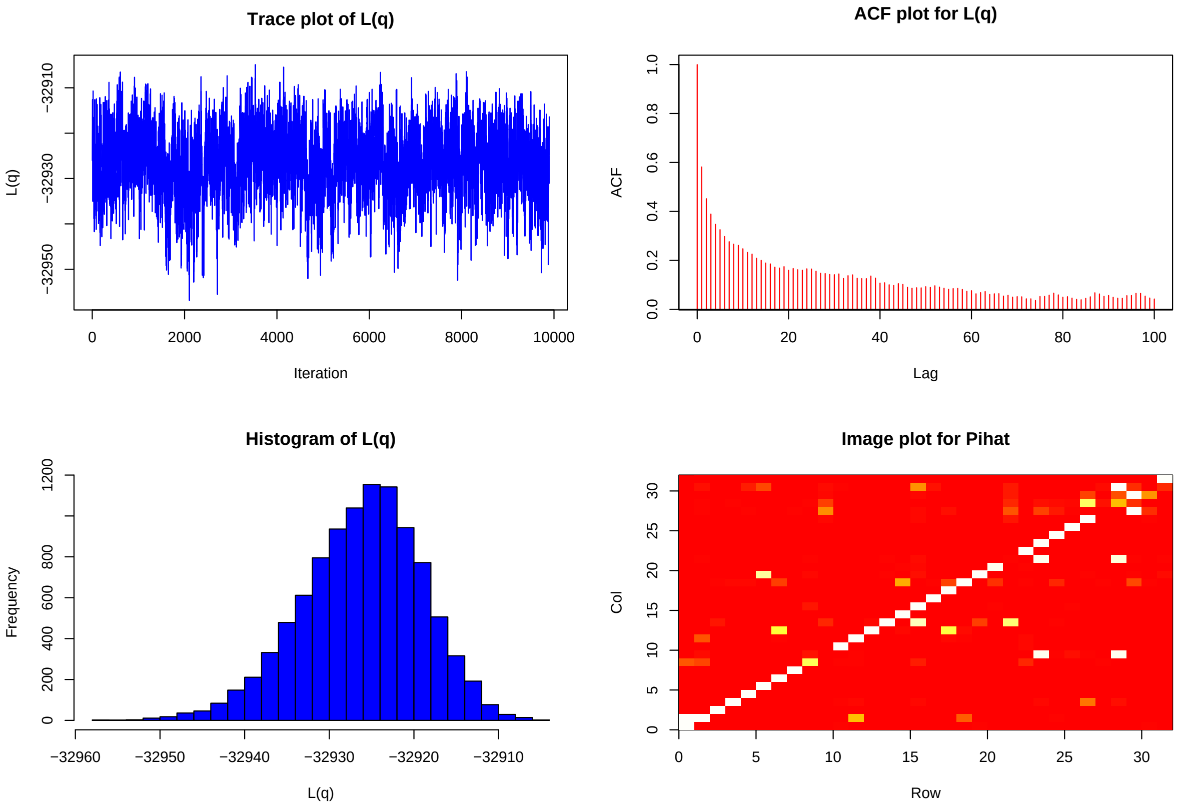


Figure 5.10: Plots for the analysis of the MCMC output for the animal ERP data (non-zero mean)

Formal diagnostics can now be administered to test for convergence of the chains. For the animal dataset, Raftery-Lewis offered the results

```
Iterations = 1:9900
Thinning interval = 1
Number of chains = 1
Sample size per chain = 9900

Quantile (q) = 0.025
Accuracy (r) = +/- 0.005
Probability (s) = 0.95
```

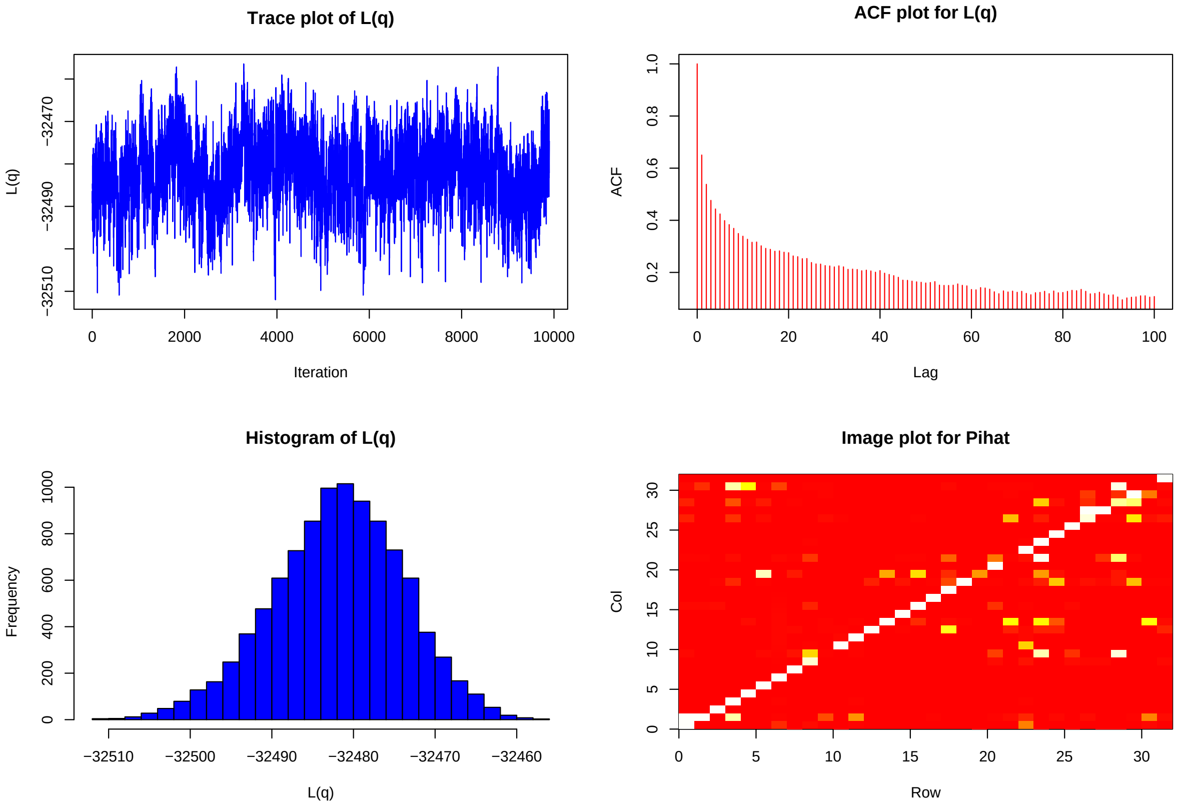


Figure 5.11: Plots for the analysis of the MCMC output for the distractor ERP data (non-zero mean)

```
     Burn-in  Total Lower bound  Dependence
     (M)      (N)   (Nmin)       factor (I)
Lq    7       8315  3746         2.22
```

whereas the Heidelberger-Welch diagnostic revealed

```
      Stationarity start     p-value
      test         iteration
Lq    passed       1         0.484

      Halfwidth Mean   Halfwidth
      test
Lq    passed    -32926 0.757
```

Moreover, in the distractor case, the output of the Raftery-Lewis test was

```
Iterations = 1:9900
Thinning interval = 1
Number of chains = 1
Sample size per chain = 9900

Quantile (q) = 0.025
Accuracy (r) = +/- 0.005
Probability (s) = 0.95

    Burn-in  Total Lower bound  Dependence
    (M)      (N)   (Nmin)       factor (I)
Lq   8       10428  3746         2.78
```

Additionally, Heidelberger-Welch returned

```
     Stationarity start     p-value
     test         iteration
Lq   passed       1         0.246

     Halfwidth Mean   Halfwidth
     test
Lq   passed    -32482 1.04
```

So, as was noted in the zero mean case due to the length of the run, the graphical output and diagnostics suggest that these chains have converged, but without rapid exploration of the space, and thus with fewer independent values.

Upon comparison of the two `image` plots for each dataset, all of the conclusions seen when $\boldsymbol{\mu} = \mathbf{0}$ can again be reached. For further discussion, the reader is referred back to the corresponding stage in Section 4.3.6. So, when the mean is non-zero, we can again surmise that the decision needed to categorise both animal and distractor images is made along similar neural pathways. Moreover, the analogous animal and distractor `image` plots are closely related, independent of the value of $\boldsymbol{\mu}$. This is particularly true along the main diagonals, but many edges between different electrodes are also regularly recognised.

Figures 5.12 and 5.13 provide variational posterior summaries for a set of a_{ij} in the animal and distractor cases respectively.

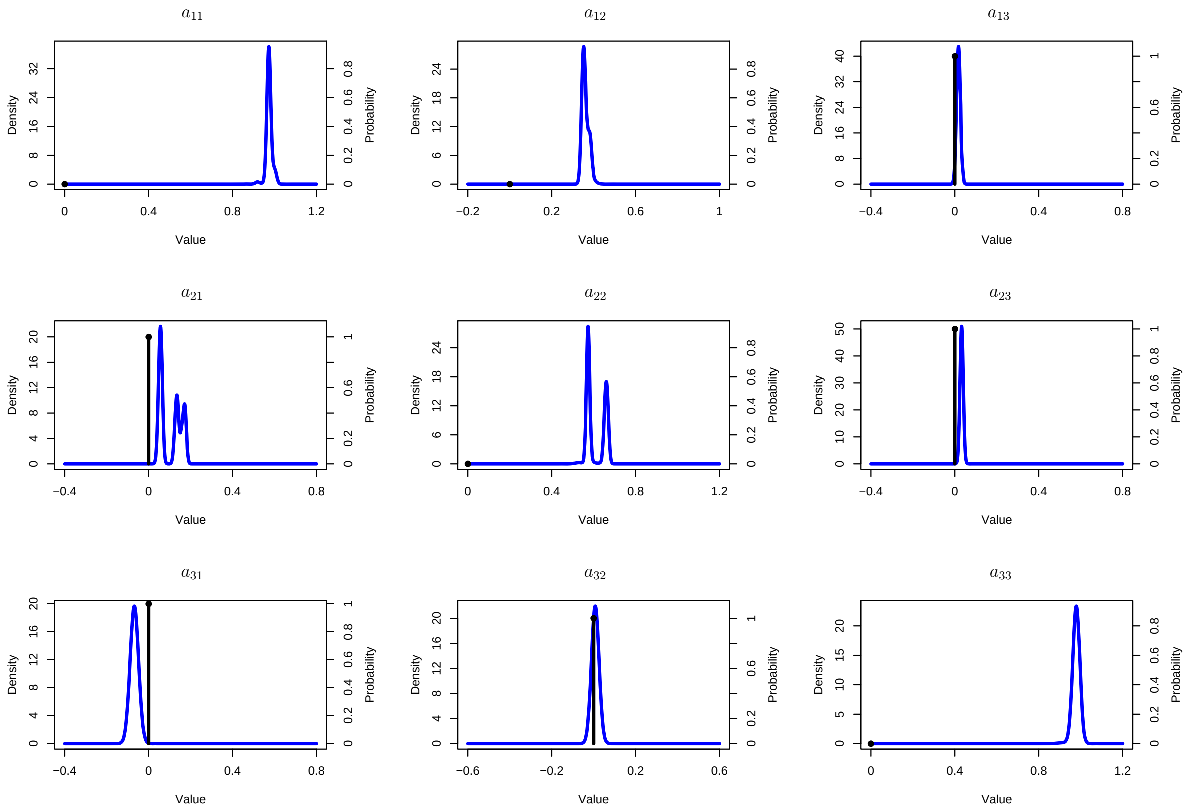


Figure 5.12: Plots showing estimated, marginal posterior distributions for a_{ij}, i, j = 1, 2, 3, for the animal ERP data (non-zero mean)

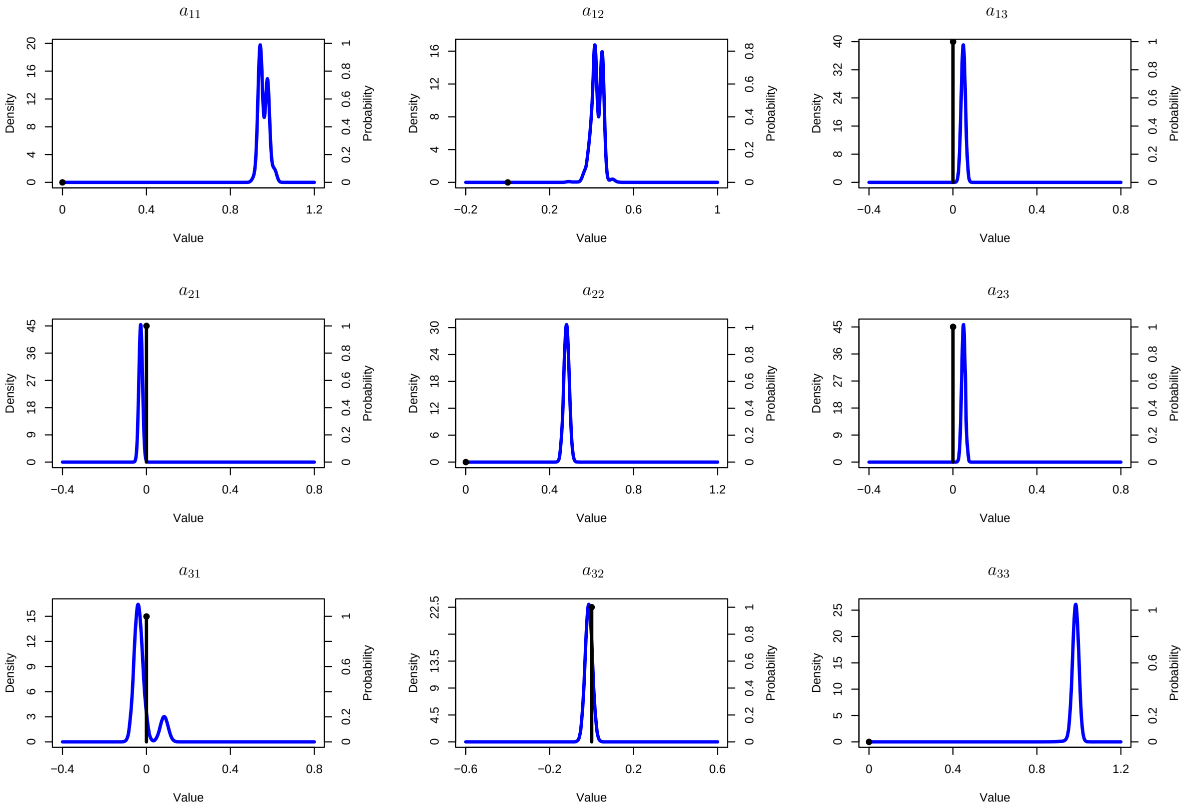


Figure 5.13: Plots showing estimated, marginal posterior distributions for a_{ij}, i, $j = 1, 2, 3$, for the distractor ERP data (non-zero mean)

It follows that a_{13}, a_{21}, a_{23} and a_{31} appear all to be zero in these figures as before. In fact, the respective densities for these coefficients are similar both to each other and the corresponding plots in the zero mean case, with the possible exception of $p_{\mathrm{var}}(a_{21} \,|\, a_{21} \neq 0,\, D)$. It is additionally evident that $a_{32} = 0$; here, we note that, unlike in Figures 4.18 and 4.19, the densities for this coefficient are tightly peaked around zero. All of the remaining elements considered are suggested to be non-zero, with the likely values of each akin for the two datasets. Moreover, it is realised that some of the densities are multimodal, as discussed erstwhile.

For completeness, plots of $p_{\mathrm{var}}(\mu_j \,|\, D)$ are provided for the two datasets in Figures 5.14 and 5.15. Upon comparison to each other, the densities are peaked around quite distinct values for all coefficients, apart from that for μ_5 and especially μ_6. Thus, we can suggest that the mean level of electrical activity varies regularly at corresponding electrodes for the two datasets. In such cases, the response is greater upon recognition of an animal despite the use of comparable circuits in each case to process the information.

A valid question to ask at this stage is whether there is much gain in applying the more complex non-zero mean approach as opposed to comparing zero mean VAR(1) models with centralised input data (*i.e.* by subtracting the sample mean). For instance, in the current scenario, results are similar between the two approaches. We have learnt additionally about the likely values of components of $\boldsymbol{\mu}$ here, but this required a large quantity of theoretical and computational work. Yet, in poor datasets, differences may exist if we assume either zero or non-zero mean models. In fact, evidence of this is provided in the final example below.

5.4.3 Application to microarray data

To conclude, our Metropolis-Hastings algorithm is run again for the microarray data, introduced in the previous chapter, and now modelled via a non-zero mean VAR(1) process

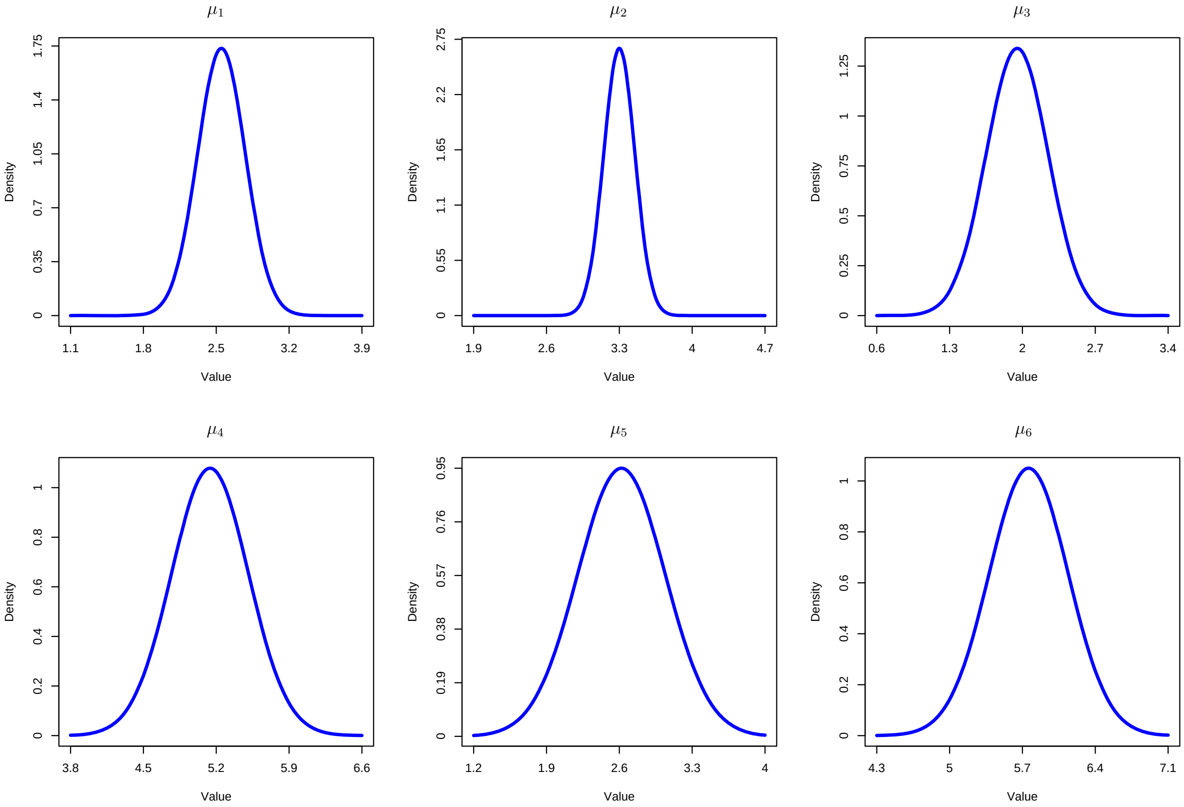


Figure 5.14: Plots showing estimated, marginal posterior distributions for μ_j, $j = 1, \ldots, 6$, for the animal ERP data

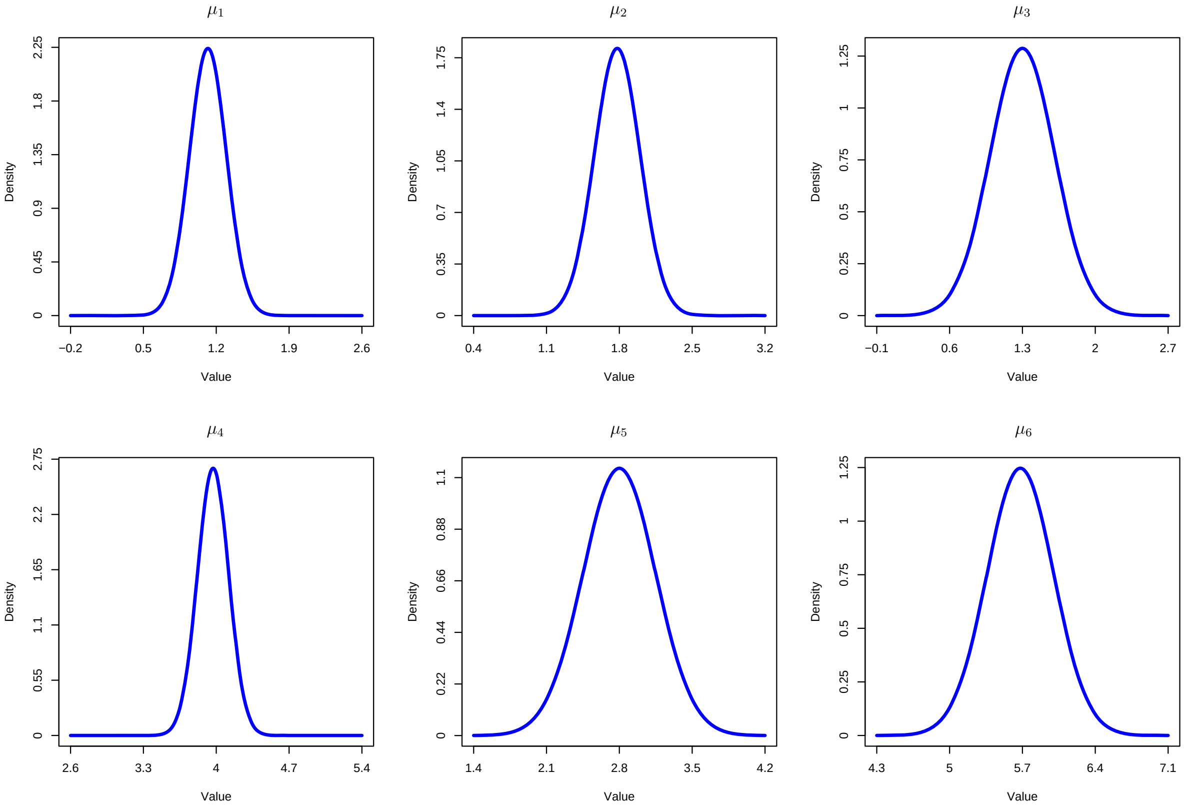


Figure 5.15: Plots showing estimated, marginal posterior distributions for μ_j, $j = 1, \ldots, 6$, for the distractor ERP data

with unknown A, σ^2 and $\boldsymbol{\mu}$. Recall that, for this dataset, we have $d = 9$ and $N = 40$. Moreover, we let $p = \frac{8}{9}$ and retain the remaining specifications from before. The genes considered in the study are displayed in Table 4.4.

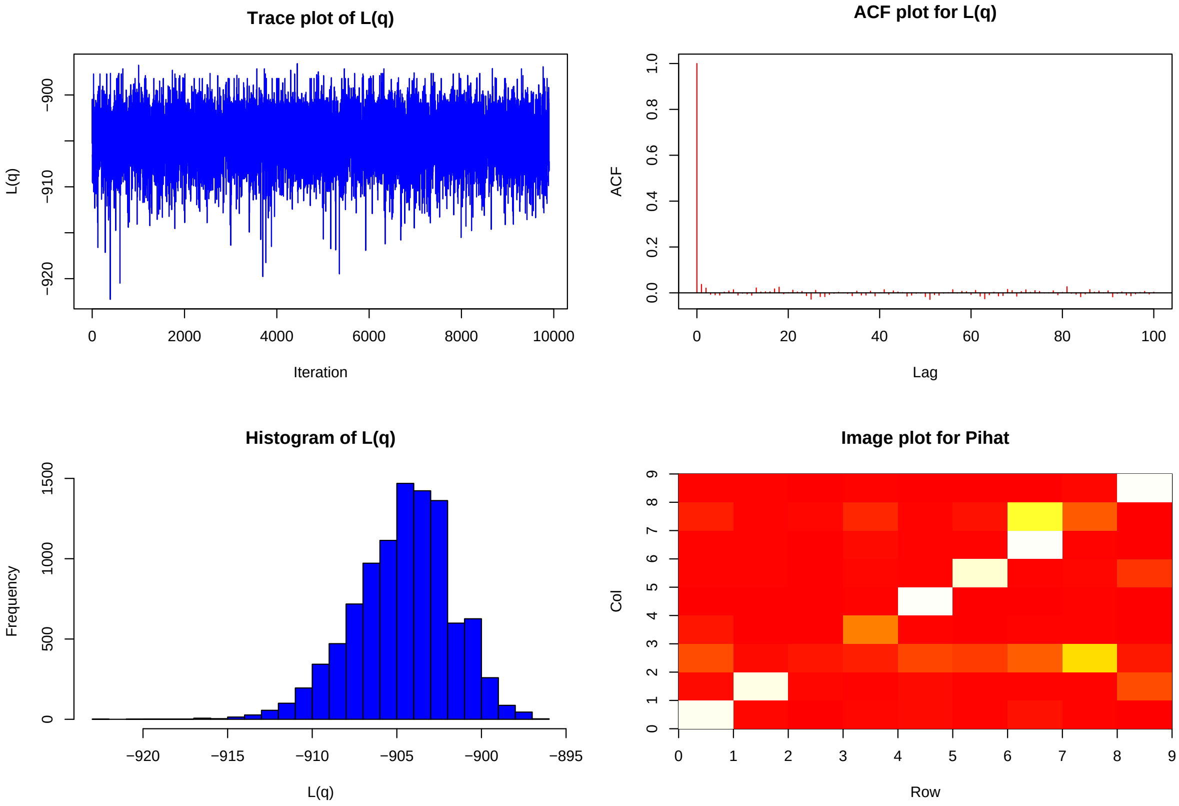


Figure 5.16: Plots for the analysis of the MCMC output for the microarray data (non-zero mean)

The output of the scheme, displayed in Figure 5.16, is now analysed. It is clear from the trace and ACF plots that the chain is moving rapidly through the space and contains many independent values. Furthermore, the Raftery-Lewis test yielded

```
Iterations = 1:9900
Thinning interval = 1
Number of chains = 1
Sample size per chain = 9900

Quantile (q) = 0.025
Accuracy (r) = +/- 0.005
Probability (s) = 0.95
```

```
      Burn-in  Total Lower bound  Dependence
      (M)      (N)   (Nmin)       factor (I)
 Lq    2       3812  3746         1.02
```

whilst Heidelberger-Welch produced

```
       Stationarity start     p-value
       test         iteration
 Lq    passed       1         0.528

       Halfwidth Mean   Halfwidth
       test
 Lq    passed    -905   0.0525
```

Thus, it can be suggested confidently that the stationary distribution of the chain has been reached. When examining the `image` plot, we see that there are similarities with that in the zero mean case (*c.f.* Figure 4.20). For instance, kinA causes a reaction in a distinct gene, namely spo0F, as opposed to influencing itself at the next time point. However, on this occasion, a new link is determined from spoOF to clpP whereas the affect of spoIIAA over spo0B is scarcely recognised. In fact, there may exist other such edges between different genes, although these associations seem rather weak.

Finally, the approximate posterior information for the coefficients a_{ij} and μ_j is revealed in Figures 5.17 and 5.18 respectively. As in the zero mean case, the only non-zero coefficients of A shown are a_{11} and a_{22}, although the possible value of a_{22} appears to be marginally smaller than before. On the other hand, most of the densities $p_{\text{var}}(\mu_j \mid D)$ have negative modal values. Again, for a larger dataset, we would expect these densities to be more tightly peaked, and hence the same links would be suggested in the `image` plots for when the mean was both zero or otherwise.

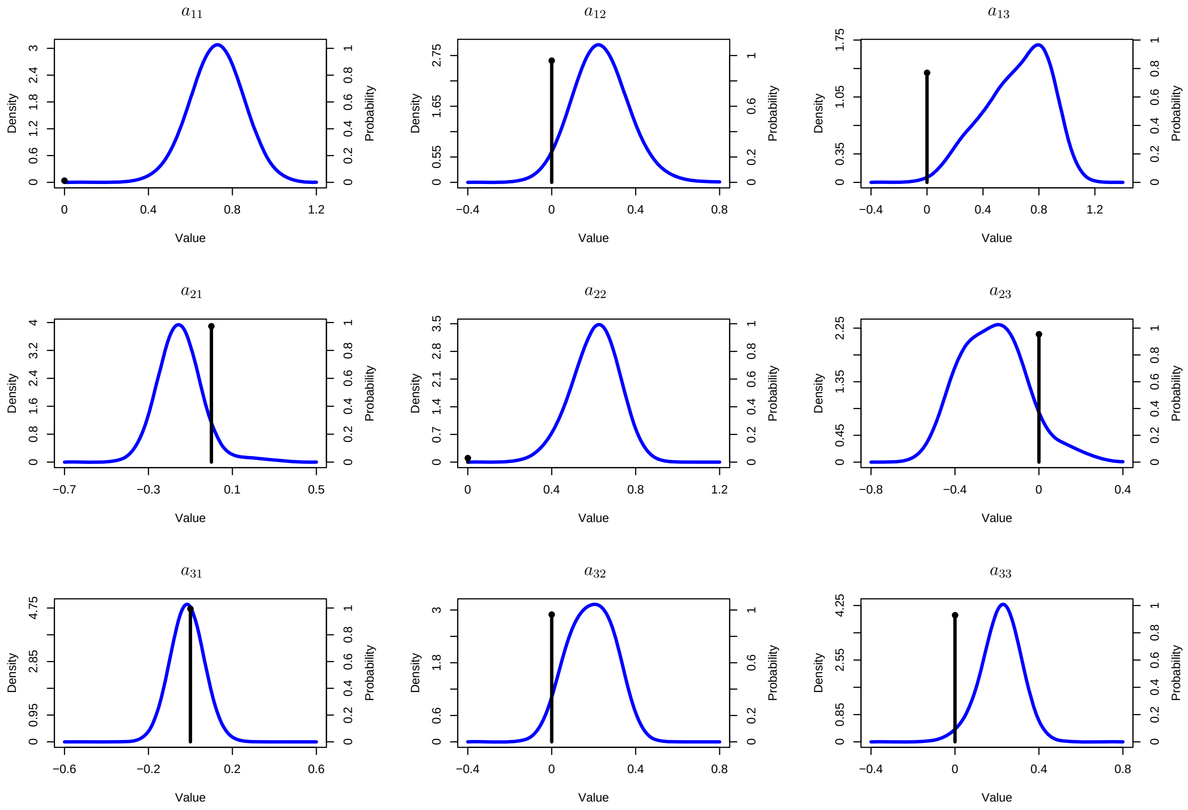


Figure 5.17: Plots showing estimated, marginal posterior distributions for a_{ij}, i, j = 1, 2, 3, for the microarray data (non-zero mean)

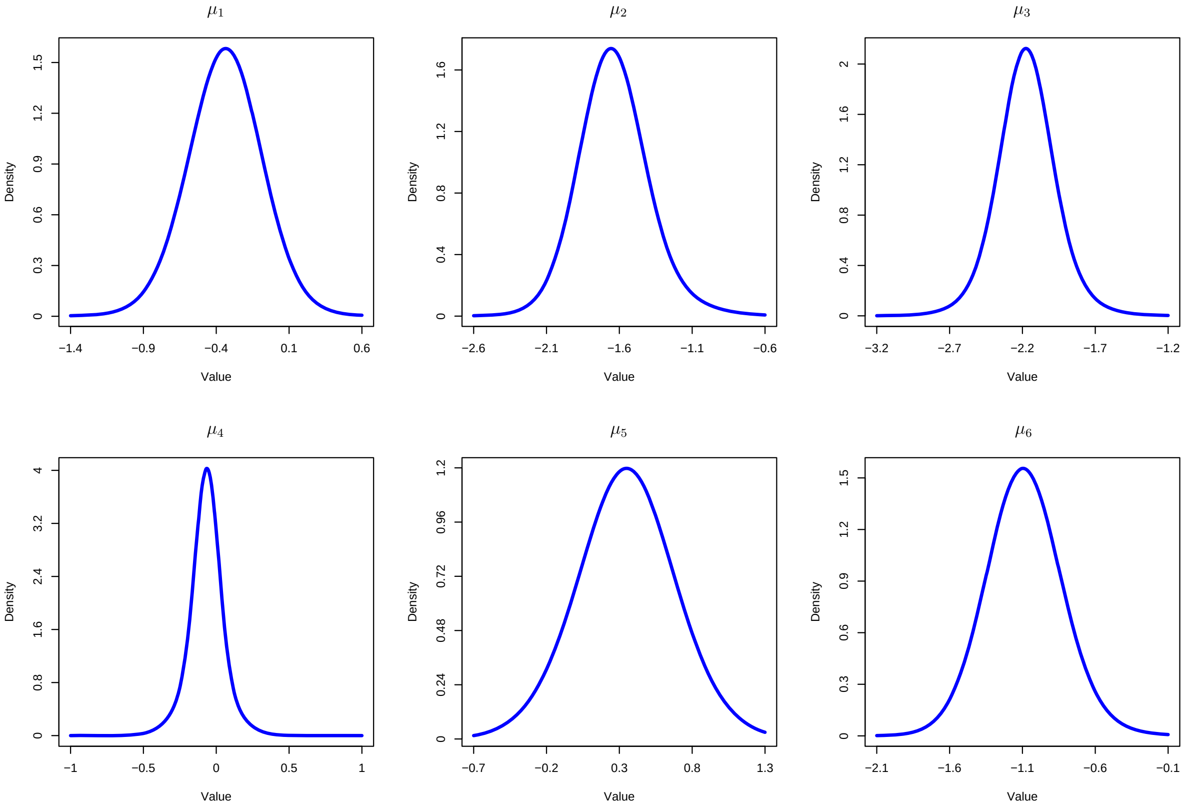


Figure 5.18: Plots showing estimated, marginal posterior distributions for μ_j, $j = 1, \ldots, 6$, for the microarray data

Chapter 6

Conclusions and further work

6.1 Conclusions

In this thesis, the predominant focus has been to illustrate how variational Bayesian methods can be applied so that sparse VAR(1) graphical models may be scored. As noted, this approximation has been utilised previously by Penny and Roberts (2002) in zero mean VAR(p) models for the purpose of model-order selection, itself a significant, inferential problem. Here however, our wish was to estimate the unknown sparsity structure of the autoregressive matrix A in both zero and non-zero mean processes, as seen respectively in Chapters 3 and 5.

To rank models, we realise that an inherent feature of variational Bayes methodology is that a lower bound is formed on the logarithm of the marginal likelihood, an essential statistic for Bayesian model comparison. At the same time, an attractive benefit of the approach is that a global approximation can also be made to each parameter posterior by minimising the KL divergence between the true and variational distribution. The optimality of every estimate is ensured by iterating update equations that are derived for the corresponding set of variational parameters until convergence. It was shown via example in Chapter 2 that such an approximation competes favourably with that of the EM algorithm and Gibbs' sampling.

An additional advantage here is that variational distributions may be determined by either a free form or fixed form approach. Of course, this proved to be of importance in both Chapters 3 and 5. If A was dense, application of the free form method was possible to find $q(\mathbf{a} \,|\, D)$. However, in the sparse case, the most natural and straightforward way to proceed was to assume fixed forms for the variational posteriors, namely those suggested by the free form approach. A toy example based upon simulated data was considered in both the zero and non-zero mean cases, and positive results were produced. In particular, the model possessing the true sparsity structure was ranked highest in each case and, when unknown, the mean of the process was accurately estimated.

In Chapter 4, an MCMC algorithm was constructed to traverse quickly graphical spaces of higher dimensions. Any move to a neighbouring graph was proposed by the addition or deletion of a single edge from the current graph, and accepted in accordance with a Metropolis-Hastings acceptance probability. Throughout the scheme, a Markov chain of accepted lower bounds was formed, enabling exploration of an approximation to the model posterior distribution. For analysis purposes, `image` plots could be produced to display which edges were accepted most frequently during the run. Moreover, the probable values of the coefficients a_{ij} could be determined by estimating both $P(a_{ij} = 0 \,|\, D)$ and the marginal density $p(a_{ij} \,|\, a_{ij} \neq 0,\, D)$. Similarly, in Chapter 5, an approximate posterior summary could also be provided for each coefficient of the mean vector.

The algorithm was tested on several datasets of varying dimension, simulated from both zero and non-zero mean VAR(1) models. In this case, the results produced throughout accurately predicted the true specifications. Moreover, two sets of real time series data were also considered. Initially, for the 32-electrode ERP datasets, we concluded that the networks required to process the information of target and non-target photographs were similar, independently of whether the mean was equal to zero or otherwise. Then, for the microarray data, it was possible to discover which genes were influential in determining whether an organism should sporulate. So, in summary, upon modelling a real dataset by a zero or non-zero mean VAR(1) process, we can use our algorithm to locate high scoring models with computational efficiency in graphical spaces of potentially huge dimension.

6.2 Further work

We consider briefly how the methodology that is comprised within this thesis could be extended. Recall that, in Chapter 3, the VAR(1) model was specified such that the noise vector $\mathbf{e}_t$ was distributed with covariance matrix $\Gamma = \sigma^2 \mathrm{I}_d$. Thus, a simple direction to take would be to implement the variational Bayesian approach for ranking sparse VAR(1) models when Γ was no-longer constrained. In this case, the most natural way to proceed would be to place an inverse Wishart prior on Γ. Alternatively, we could examine the scenario when the noise is modelled as a mixture of Gaussian distributions, as opposed to the standard single Gaussian. This has been tackled previously to identify the optimal model order by Roberts and Penny (2002).

However, an additional area of research that is perhaps most clearly motivated here is to compare sparse VAR(p) models. Now, our task would be to determine the sparsity structure of all p autoregressive matrices in the process, where each $A(i)$ is of dimension $d \times d$. By defining $\mathbf{x}_t = [\mathbf{y}_{t-1}, \mathbf{y}_{t-2}, \ldots, \mathbf{y}_{t-p}]$ where $t = 1 \ldots, N$, we could follow Penny and Roberts (2002) and rewrite (3.1) as

$$\mathbf{y}_t = \mathbf{x}_t W + \mathbf{e}_t.$$

Here, W is a $pd \times d$ matrix, formed by stacking the $A(i)$-matrices. Thus, by specifying a prior on vec(W) that imposes the correct sparsity structure for each model, the variational algorithm could proceed as before. In particular, it would be interesting to see how effectively our Metropolis-Hastings algorithm could handle moving through graphical spaces of such extreme dimension.

Of course, we are not restricted to model time series data using just VAR processes. Hence finally, it is noted that the variational Bayes treatment could be given to such alternatives. For instance, one possibility is the VARMA(p, q) (*vector autoregressive moving average*) process, defined as

$$\mathbf{y}_t = \sum_{i=1}^{p} \mathbf{y}_{t-i} A(i) + \mathbf{e}_t + \sum_{j=1}^{q} \mathbf{e}_{t-j} \phi(j),$$

where again $\mathbf{e}_t \sim \mathcal{N}(\mathbf{0}, \Gamma)$. Thus, for $i = 1, \ldots, p$ and $j = 1, \ldots, q$, our parameter set would be $\{A(i), \phi(j), \Gamma\}$, where each $\phi(j)$ has dimension $d \times d$. For further information on this and other related models, the reader is referred to Lütkepohl (2005).

Appendix A

Probability distributions

In this appendix, some standard, continuous probability distributions are documented. In each case, the probability density function is defined, together with any salient expectations, taken with respect to this density.

A.1 Gaussian distribution

The Gaussian (normal) distribution with mean m and variance $v > 0$ is denoted as

$$\begin{aligned} p(x \,|\, m,\, v) &= \mathcal{N}(x \,|\, m,\, v) \\ &= \frac{1}{\sqrt{2\pi v}} \exp\left\{-\frac{(x-m)^2}{2v}\right\}. \end{aligned} \tag{A.1}$$

An important result is that

$$\mathrm{E}\left\{X^2\right\} = m^2 + v. \tag{A.2}$$

A.2 Inverse gamma distribution

With support wherever $x > 0$, the density of the inverse gamma distribution is

$$\begin{aligned} p(x \mid a,\, b) &= \mathcal{IG}(x \mid a,\, b) \\ &= \frac{b^a}{\Gamma(a)} x^{-(a+1)} \exp\left\{-bx^{-1}\right\}, \end{aligned} \tag{A.3}$$

with parameters $a,\, b > 0$. We realise three pertinent identities for this distribution.

$$\mathrm{E}\left\{X\right\} = \frac{b}{a-1} \qquad \text{for } a > 1 \tag{A.4}$$

$$\mathrm{E}\left\{X^{-1}\right\} = \frac{a}{b} \tag{A.5}$$

$$\mathrm{E}\left\{\log X\right\} = \log b - \psi(a), \tag{A.6}$$

by both Beal (2003) and Nicolas (2002). Here, for $z \in \mathbb{R}$, we define $\psi(z)$ to be the *digamma* function (Johnson et al., 1992), *i.e.* the logarithmic derivative of the gamma function, given by

$$\psi(z) = \frac{\mathrm{d}}{\mathrm{d}z} \log \Gamma(z) = \frac{\Gamma'(z)}{\Gamma(z)}.$$

A.3 Multivariate Gaussian distribution

The univariate Gaussian can be generalised to d dimensions with density

$$\begin{aligned} p(\mathbf{x} \mid \mathbf{m},\, V) &= \mathcal{N}(\mathbf{x} \mid \mathbf{m},\, V) \\ &= (2\pi)^{-d/2} \left|V\right|^{-1/2} \exp\left\{-\frac{1}{2}(\mathbf{x}-\mathbf{m})^T V^{-1} (\mathbf{x}-\mathbf{m})\right\}, \end{aligned} \tag{A.7}$$

where $\mathbf{m} = (m_1, \ldots, m_d)$ is the mean vector and V is the symmetric, positive-definite,

$d \times d$ covariance matrix. Akin to the univariate case, we have

$$\mathrm{E}\left\{\mathbf{X}\mathbf{X}^T\right\} = \mathbf{m}\mathbf{m}^T + V. \tag{A.8}$$

A.4 Inverse Wishart distribution

A multivariate generalisation of the inverse gamma distribution is the inverse Wishart, with density function for $d \times d$ matrix X given by O'Hagan et al. (1994) as

$$\begin{aligned} p(X \mid B, r) &= \mathcal{IW}(X \mid B, r) \\ &= k^{-1}|B|^{r/2}|X|^{-(r+d+1)/2} \exp\left\{-\mathrm{Tr}\left[X^{-1}B\right]/2\right\}, \end{aligned} \tag{A.9}$$

where the normalising constant is

$$k = 2^{rd/2}\pi^{d(d-1)/4}\prod_{i=1}^{d}\Gamma\left\{(r+1-i)/2\right\}. \tag{A.10}$$

The parameters of this distribution are B, a symmetric, positive definite, $d \times d$ matrix and a scalar $r > d$. The afore-mentioned authors also indicate that

$$\begin{aligned} \mathrm{E}\left\{X\right\} &= \frac{B}{r-d-1} \qquad \text{for } r > d+1 \\ \mathrm{E}\left\{X^{-1}\right\} &= rB^{-1}. \end{aligned} \tag{A.11}$$

Appendix B

Graphical Models

The main focus of this appendix is to introduce the concept of a graphical model. This will lead us to examine briefly both graphical Gaussian models and Bayesian networks. Initially, we define conditional independence, the key notion that characterises a graphical model.

B.1 Conditional Independence

Suppose we have two random variables, X_1 and X_2, possessing a joint probability density function $p_{X_1,\,X_2}$. Then, these random variables are independent, written $X_1 \perp\!\!\!\perp X_2$, if $p_{X_1,\,X_2}(x_1,\,x_2) = p_{X_1}(x_1)p_{X_2}(x_2)$. An equivalent formulation is $p_{X_1\,|\,X_2}(x_1\,|\,x_2) = p_{X_1}(x_1)$, *i.e.* the conditional density of X_1, given $X_2 = x_2$, is not a function of x_2, and that $p_{X_2\,|\,X_1}(x_2\,|\,x_1) = p_{X_2}(x_2)$.

Now, introduce a third variable, X_3. We say X_1 and X_2 are *conditionally independent* given X_3, written $X_1 \perp\!\!\!\perp X_2\,|\,X_3$, if X_1 and X_2 are independent in their conditional distribution given $X_3 = x_3$, for any value of x_3. In other words, given knowledge of X_3, subsequent understanding of X_2 will not provide any new information about X_1.

Conditional independence can be characterised in terms of density functions as follows:

$$X_1 \perp\!\!\!\perp X_2 \,|\, X_3 \iff p_{X_1,\, X_2 \,|\, X_3}(x_1,\, x_2 \,|\, x_3) = p_{X_1 \,|\, X_3}(x_1 \,|\, x_3) p_{X_2 \,|\, X_3}(x_2 \,|\, x_3) \tag{B.1}$$

$$\iff p_{X_1 \,|\, X_2,\, X_3}(x_1 \,|\, x_2,\, x_3) = p_{X_1 \,|\, X_3}(x_1 \,|\, x_3) \tag{B.2}$$

$$\iff p_{X_2 \,|\, X_1,\, X_3}(x_2 \,|\, x_1,\, x_3) = p_{X_2 \,|\, X_3}(x_2 \,|\, x_3). \tag{B.3}$$

B.2 Graph theory

Some standard notation and terminology for graphs is recalled. For further discussion, the reader is referred to Cowell et al. (1999).

By definition, a graph is a pair $G = (V,\, E)$, whereby V is a finite set of nodes (or vertices) and E a set of edges of ordered pairs of nodes. If, for two nodes a and b, $(a,\, b) \in E$ and $(b,\, a) \in E$, the edge between them is described as *undirected*, written $a \sim b$ (represented on a graph by a line between the two nodes). Thus, a and b are described as *neighbours*.

On the contrary, if $(a,\, b) \in E$, but $(b,\, a) \notin E$, the edge is called *directed*, written $a \to b$ (represented on a graph by an arrow from a to b). In this case, a is termed as a *parent* of b, and b one of the *children* of a. We denote $\mathrm{pa}(b)$ to be the set of parents of the node b, similarly $\mathrm{ch}(a)$ the set of children of a. The *boundary*, $\mathrm{bd}(a)$, of $a \in V$ is the set of parents and neighbours of this node. Moreover, the *closure*, $\mathrm{cl}(a)$, is the set $a \cup \mathrm{bd}(a)$. If a graph possesses only directed edges, it is referred to as a *directed graph*, similarly an *undirected graph*. If an edge exists between every pair of nodes, the graph is *complete*.

A sequence of distinct nodes $a = a_0, \ldots, a_n = b$, such that $a_{j-1} \sim a_j$ for all $j = 1, \ldots, n$, forms a *path* from a to b of length n. If the path is such that $a = b$, *i.e.* the end-points coincide, it is referred to as an *n-cycle*. A path from a to b, given by the same set of nodes as above, is described as *directed* if it contains at least one directed edge $a_{j-1} \to a_j$ for any j. In this case, a is an *ancestor* of b and b one of the *descendants* of a. Denote $\mathrm{an}(b)$

to be the set of ancestors of b, similarly $\mathrm{de}(a)$ the set of descendants of a. The definition of a *directed n-cycle* follows immediately. A graph without any cycles is called *acyclic.*

Finally, suppose that A, B and C are subsets of V. If all paths from A to B intersect C, then C is deemed to *separate* A and B. The theory presented here is important for what ensues in this appendix.

B.3 Undirected graphical models

Let $G = (V, E)$ be an undirected graph and $\mathbf{X} = (X_1, \ldots, X_p)^T$ a p-dimensional random vector. If the graph has p nodes, then a random variable X_a is associated to each node for all $a \in V$ where, of course, $V = \{1, \ldots, p\}$. In general, note that, on any graph, a circle is used to represent a continuous random variable, a dot for a discrete variable. Here, we are concerned with the former case. Now, suppose a subset $A \subseteq V$. We thus denote $\mathbf{X}_A = (X_a : a \in A)$ to be a collection of random variables.

Furthermore, introduce P, a probability distribution for $\mathbf{X}$. If $A \subseteq V$, then let P_A denote the marginal distribution for X_A. Thus, an important definition is realised.

Definition 1 *Assume that A, B, C are disjoint subsets of V. If $X_A \perp\!\!\!\perp X_B \,|\, X_C$ whenever C separates A and B in the graph G, the distribution P is said to be Markov with respect to G.*

This is known as the *global Markov property.* We stress that, on the graph, if two nodes are conditionally independent, no edge exists between them. It is worth mentioning that other such Markov properties exist over graphs.

Definition 2 *If $X_a \perp\!\!\!\perp X_{V\setminus \mathrm{cl}(a)} \,|\, X_{\mathrm{bd}(a)}$ for any $a \in V$, a distribution P obeys the local Markov property with respect to a graph G.*

Moreover, if $X_a \perp\!\!\!\perp X_b \,|\, X_{V \setminus \{a,\, b\}}$ for any pair $(a,\, b) \notin E$, then, relative to a graph, the *pairwise Markov property* is satisfied. These properties are important since they show that any conditional independencies that can be determined from the graph also hold in the corresponding probability distribution. Further analysis of these Markov properties is provided by Lauritzen (1996). Finally, an *undirected graphical model* (also termed a *Markov network*) for $\mathbf{X}$ is a joint probability distribution for $\mathbf{X}$, that is Markov (obeys the global Markov property) with respect to an undirected graph G.

If the distribution is multivariate Gaussian, say $\mathcal{N}(\mathbf{x} \,|\, \boldsymbol{\mu},\, \Sigma)$, then a *graphical Gaussian model* is so defined. We now examine the conditional independencies between random variables, inherent in such a model. Thus, let $K = \Sigma^{-1}$ be the *concentration (precision) matrix* for such a multivariate Gaussian. Speed and Kiiveri (1986) illustrate that K determines the conditional independence structure of a graphical Gaussian model as follows.

Proposition 1 *Let $a,\, b \in \{1, \ldots, p\}$ be distinct nodes on an undirected graph G, giving rise to a graphical Gaussian model, parameterised by mean vector $\boldsymbol{\mu}$ and covariance matrix Σ. Defining the corresponding concentration matrix as $K = (k_{ab})$, then $X_a \perp\!\!\!\perp X_b \,|\, X_{V \setminus \{a,\, b\}}$ (pairwise Markov property) if and only if $k_{ab} = 0$.*

So, for a given K, a graph can be associated and its independencies identified. Moreover, a given graph determines a sparse matrix K. Of course, as the graph is undirected and with K symmetric, if $k_{ab} = 0$, then $k_{ba} = 0$, implying $X_b \perp\!\!\!\perp X_a \,|\, X_{V \setminus \{a,\, b\}}$. Thus, in this case, no edge would exist between nodes a and b. To clarify, consider this simple example.

B.3.1 Example

Suppose that the random vector $\mathbf{X} = (X_1,\, X_2,\, X_3,\, X_4)^T$ is modelled via a multivariate Gaussian distribution, with concentration matrix K specified as

$$K = \begin{pmatrix} * & * & * & * \\ * & * & 0 & 0 \\ * & 0 & * & 0 \\ * & 0 & 0 & * \end{pmatrix},$$

where $*$ refers to an unspecified, non-zero element. Then, a graphical Gaussian model is defined, with respect to the graph below.

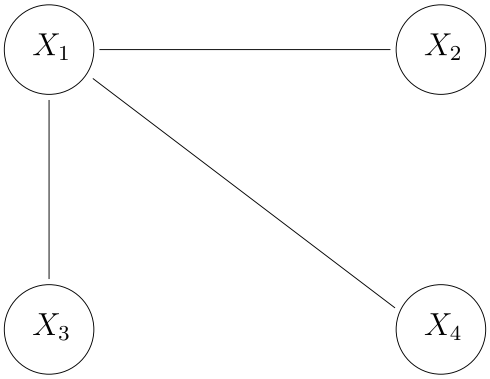


Figure B.1: 4-node graph, corresponding to the choice of K

Using Proposition 1, the subsequent conditional independencies are apparent: $X_2 \perp\!\!\!\perp X_3 \,|\, \{X_1,\, X_4\}$, $X_2 \perp\!\!\!\perp X_4 \,|\, \{X_1,\, X_3\}$ and $X_3 \perp\!\!\!\perp X_4 \,|\, \{X_1,\, X_2\}$. Similarly, the reverse independencies, such as $X_3 \perp\!\!\!\perp X_2 \,|\, \{X_1,\, X_4\}$, also hold.

B.4 Bayesian networks

Let $G = (V, E)$ now be a directed graph and suppose that we have again the random vector $\mathbf{X} = (X_1, \ldots, X_p)^T$. Here, a directed edge implies a causal dependence between a pair of nodes. So if $X_a \rightarrow X_b$ for $a, b \in \{1, \ldots, p\}$, then we say that X_a *causes* (or influences) X_b. Moreover, whenever G contains no directed cycles, then it is referred to as a *directed acyclic graph* (DAG). On any DAG, a (non-unique) ordering of the nodes can be found such that $X_a \rightarrow X_b$ only when $a < b$, *i.e.* every node follows its parents in the ordering.

Introducing a distribution P for $\mathbf{X}$, we now present an analogue to Definition 2 for the directed case.

Definition 3 *Let* $\mathrm{nd}(X_a)$ *represent the set of non-descendants of* X_a *for* $a \in V$. *Then, if* $X_a \perp\!\!\!\perp \mathrm{nd}(X_a) \,|\, \mathrm{pa}(X_a)$, *the distribution* P *obeys the directed local Markov property with respect to a directed acyclic graph* G.

Although not treated here, there are also directed counterparts to the (undirected) global and pairwise Markov properties. In contrast to the undirected case, the directed local and global Markov properties are equivalent over a DAG (Lauritzen 1996, pg. 33, 51). Thus, if either of these two properties is satisfied, P is termed a *directed Markov distribution* (*c.f.* Definition 1). Finally, a *Bayesian network* (also termed a *belief network*) for $\mathbf{X}$ is a joint probability distribution for $\mathbf{X}$, that is directed Markov (obeys the directed global Markov property) with respect to G, a DAG.

Essentially, a Bayesian network is merely a directed, acyclic graphical model, containing an ordering of the nodes. We note that this ordering is consistent with the DAG, but is otherwise arbitrary. Of course, any conditional independencies between variables can be simply read off the graph. Moreover, the probability distribution for $\mathbf{X}$ can be factorised according to the DAG (Cowell, 1998). The condition $X_a \perp\!\!\!\perp \mathrm{nd}(X_a) \,|\, \mathrm{pa}(X_a)$, determining the directed local Markov property, can be re-expressed, in general, as $X_a \perp\!\!\!\perp X_1, \ldots, X_{a-1} \,|\, \mathrm{pa}(X_a)$. This is because $X_b \notin \mathrm{de}(X_a)$ if $b < a$. So, in terms of densities and dropping subscripts on p, we have

$$p(x_a \,|\, x_1, \ldots, x_{a-1}) = p(x_a \,|\, \mathrm{pa}(x_a)),$$

by (B.2). Hence, the full distribution can be factorised with density

$$\begin{aligned} p(x_1, \ldots, x_p) &= p(x_1)p(x_2 \,|\, x_1)p(x_3 \,|\, x_1,\, x_2) \times \cdots \times p(x_p \,|\, x_1, \ldots, x_{p-1}) \\ &= \prod_{a=1}^{p} p(x_a \,|\, x_1, \ldots, x_{a-1}) \\ &= \prod_{a=1}^{p} p(x_a \,|\, \mathrm{pa}(x_a)). \end{aligned} \tag{B.4}$$

In other words, the joint density, represented by the graph, consists of a product of marginal densities for each node, conditioned on the parents of that node. It is evident that this final factorisation is independent of the (arbitrary) choice of ordering.

B.4.1 Example

As a straightforward illustration, Figure B.2 shows a simple, directed acyclic graph, which defines a Bayesian network for $\mathbf{X} = (X_1, \ldots, X_6)^T$, possessing a joint density $p(x_1, \ldots, x_6)$.

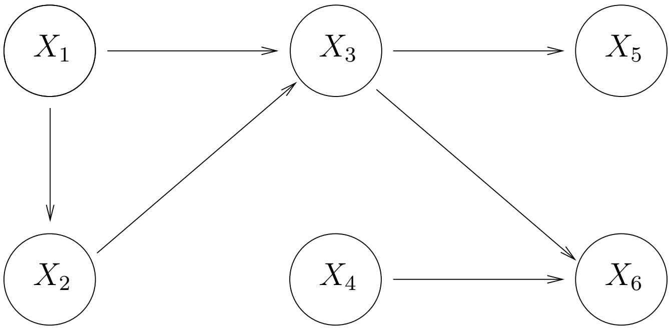


Figure B.2: 6-node DAG

Then, by (B.4), it is evident from the graph that

$$p(x_1, \ldots, x_6) = p(x_1)p(x_2 \,|\, x_1)p(x_3 \,|\, x_2,\, x_1)p(x_4)p(x_5 \,|\, x_3)p(x_6 \,|\, x_3,\, x_4).$$

Hence, for instance, it follows that $X_6 \perp\!\!\!\perp X_1 \,|\, \{X_3,\, X_4\}$, $X_5 \perp\!\!\!\perp X_1 \,|\, X_3$, *etc.*

Appendix C

Generalised inverses

Every non-singular, square matrix P possesses a unique inverse, denoted by P^{-1}, whereby

$$PP^{-1} = P^{-1}P = I. \tag{C.1}$$

The inverse matrix itself has many properties, for instance, $(P^{-1})^{-1} = P$, $(P^T)^{-1} = (P^{-1})^T$ and $(aP)^{-1} = a^{-1}P^{-1}$ for all non-zero $a \in \mathbb{R}$ *etc.* However, as we have seen in both Chapters 3 and 5, it can be the case that we want to find an inverse matrix when P is singular or even not square. To effect this, we search for a generalised inverse (termed by some authors as a pseudoinverse), with similar properties to the standard inverse of a square, non-singular matrix.

Initially, we have the following definition.

Definition 4 *A generalised inverse of a $m \times p$ matrix P is any $p \times m$ matrix G such that*

$$PGP = P. \tag{C.2}$$

Let P^- denote an arbitrary generalised inverse of such a matrix P. Hence, $PP^-P = P$. In the specific case of P being square and non-singular, then this matrix has a unique generalised inverse, the standard inverse P^{-1}. Clearly, when $G = P^{-1}$, the above generalised inverse condition is satisfied. Moreover, if G is a generalised inverse of P, then, by definition, $G = P^{-1}PGPP^{-1} = P^{-1}PP^{-1} = P^{-1}$ (Harville, 1997).

Some of the properties of the standard inverse can correspond to an arbitrary generalised inverse, proven by direct substitution into (C.2). For instance, if P is $m \times p$, then one choice of $(P^T)^-$ is $(P^-)^T$. In this case, $P^T(P^-)^TP^T = (PP^-P)^T = P^T$, hence (C.2) is satisfied. In addition, $a^{-1}P^-$ is a generalised inverse of aP where $a \in \mathbb{R}$ and is non-zero. However, it is not necessarily true that one choice of $(P^-)^-$ is P.

The generalised inverse as defined above exists for any matrix, but is not unique. In fact, for a $m \times p$ matrix P of rank r, there are an infinite number of generalised inverses (Harville, 1997). So, an alternative, unique generalised inverse has been considered, initially by Moore (1920) and then independently by Penrose (1955), which now must hold for several constraints.

Definition 5 *The Moore-Penrose inverse of any $m \times p$ matrix P is the unique $p \times m$ matrix G that holds for the following conditions:*

$$\begin{aligned} PGP &= P && \text{(C.3)} \\ GPG &= G && \text{(C.4)} \\ (PG)^T &= PG && \text{(C.5)} \\ (GP)^T &= GP. && \text{(C.6)} \end{aligned}$$

Let P^+ denote the Moore-Penrose inverse (often referred to as *the* generalised inverse) of such a matrix P. We realise that other generalised inverses exist that meet (C.3) and a combination of the properties (C.4)–(C.6). See Ben-Israel and Greville (1974) or Harville

(1997) for more details.

When P is square and non-singular, then, similar to the case as that of any generalised inverse, $P^+ = P^{-1}$. We realise this since P has a unique generalised inverse, as mentioned earlier, and that $G = P^{-1}$ holds for conditions (C.3)–(C.6).

The Moore-Penrose inverse possesses some properties that are in common with both an arbitrary generalised inverse, P^-, and the standard inverse, P^{-1}. For instance, analogous to previous, $(P^T)^+ = (P^+)^T$ and $(aP)^+ = a^{-1}P^+$ for all non-zero $a \in \mathbb{R}$. However, unlike P^-, we now have $(P^+)^+ = P$. Such results are proven by direct substitution into (C.3)–(C.6) (Harville, 1997). Other such properties of the Moore-Penrose inverse do not hold for any generalised inverse. For a complete list, see Rao (1966).

A simple way to compute P^+ is to use matrix decomposition. Here, we examine one of the more popular methods, used in this context by Rao (1962).

Definition 6 *The singular value decomposition of any $m \times p$ matrix P of rank r is defined to be*

$$P = U \begin{bmatrix} \mathbf{S} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} V^T, \tag{C.7}$$

where U and V are $m \times m$ and $p \times p$ orthogonal matrices respectively (i.e. $U^T U = UU^T = I_m$, similarly for V) and $S = \mathrm{diag}(\sigma_1, \sigma_2, \ldots, \sigma_r)$, an $r \times r$ matrix with strictly positive diagonal elements.

In this definition, the σ_i, $i = 1, \ldots, r$ are the singular values of P and are unique. Note that P is of rank r since it has r non-zero singular values.

Harville (1997) shows that the Moore-Penrose inverse of P with this singular value decomposition is given by

$$P^+ = V \begin{bmatrix} \mathbf{S} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}^+ U^T, \tag{C.8}$$

where

$$\begin{bmatrix} \mathbf{S} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}^{+} = \begin{bmatrix} \mathbf{S}^{+} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}.$$

At this stage, it can be realised that for any $D = \text{diag}(d_1, d_2, \ldots, d_r)$, an $r \times r$ diagonal matrix, then $D^{+} = \text{diag}(d_1^{+}, d_2^{+}, \ldots, d_r^{+})$, whereby, for all i,

$$d_i^{+} = \begin{cases} d_i^{-1} & \text{if } d_i \neq 0 \\ 0 & \text{if } d_i = 0 \end{cases}. \tag{C.9}$$

Thus, $S^{+} = \text{diag}(\sigma_1^{-1}, \sigma_2^{-1}, \ldots, \sigma_r^{-1})$.

The proof is quite straightforward. We note that, as S is diagonal and the product of two diagonal matrices is merely the product of each pair of diagonal entries, S^{+}, as defined above, holds for the conditions (C.3)–(C.6), and hence is the Moore-Penrose inverse of S. Moreover, by substituting (C.7) and (C.8) directly into conditions (C.3)–(C.6), then it is easy to see that (C.8) is the Moore-Penrose inverse of (C.7), as U and V are both orthogonal.

Bibliography

Aitkin, M. (1991). Posterior Bayes factors (with discussion). *Journal of the Royal Statistical Society, Series B 53*, 111–142.

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

Beal, M. (2003). *Variational algorithms for approximate Bayesian inference*. Ph. D. thesis, Gatsby Computational Neuroscience Unit, University College London.

Beal, M., F. Falciani, Z. Ghahramani, C. Rangel, and D. Wild (2005). A Bayesian approach to reconstructing genetic regulatory networks with hidden factors. *Bioinformatics 21*(3), 349–356.

Ben-Israel, A. and T. Greville (1974). *Generalized Inverses: Theory and Applications*. New York, Wiley.

Berger, J. and L. Pericchi (1996). The Intrinsic Bayes factor for model selection and prediction. *Journal of the American Statistical Association 91*(433), 109–122.

Berger, J. and T. Sellke (1987). Testing a point null hypothesis: the irreconcilability of p-values and evidence. *Journal of the American Statistical Association 82*(397), 112–122.

Box, G. and G. Tiao (1992). *Bayesian Inference in Statistical Analysis*. New York, Wiley.

Brooks, S. (2002). Discussion on Bayesian measures of model complexity and fit (by D.J. Spiegelhalter *et al*). *Journal of the Royal Statistical Society, Series B 64*(4), 616–618.

Burnham, K. and D. Anderson (2004). Multimodel inference: Understanding AIC and BIC in model selection. *Sociological Methods and Research 33*(2), 261–304.

Chipman, H., E. George, and R. McCulloch (2001). The practical implementation of Bayesian model selection. In P. Lahiri (Ed.), *Model Selection*, Volume 38, pp. 67–116. IMS, Beachwood, OH.

Cowell, R. (1998). Introduction to inference in Bayesian networks. In M. Jordan (Ed.), *Learning in Graphical Models*, pp. 9–26. Kluwer.

Cowell, R., A. Dawid, S. Lauritzen, and D. Spiegelhalter (1999). *Probabilistic Networks and Expert Systems*. Springer-Verlag, New York.

Dahlhaus, R. and M. Eichler (2003). Causality and graphical models in time series analysis. In P. Green, N. Hjort, and S. Richardson (Eds.), *Highly Structured Stochastic Systems*, pp. 115–137. Oxford University Press.

Delorme, A., G. Rousselet, M.-M. Macé, and M. Fabre-Thorpe (2004). Interaction of top-down and bottom-up processing in the fast visual analysis of natural scenes. *Cognitive Brain Research 19*(2), 103–113.

Dempster, A., N. Laird, and D. Rubin (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B 39*(1), 1–38.

Eichler, M. (2001). Markov properties for graphical time series models. Technical report, Department of Statistics, University of Heidelberg.

Fletcher, R. (2000). *Practical Methods of Optimization* (2nd ed.). Wiley.

Friedman, N., K. Murphy, and S. Russell (1998). Learning the structure of dynamic probabilistic networks. In *Fourteenth Conference on Uncertainty in Artificial Intelligence*, pp. 139–147.

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis 1*(3), 515–533.

Gelman, A., J. Carlin, H. Stern, and D. Rubin (1995). *Bayesian Data Analysis*. Chapman and Hall.

Gelman, A. and D. Rubin (1992). Inference from iterative simulation using multiple sequences. *Statistical Science 7*(4), 457–511.

Geman, S. and D. Geman (1984). Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence 6*, 721–741.

Ghahramani, Z. (1997). Learning Dynamic Bayesian Networks. In C. Giles and M. Gori (Eds.), *Adaptive Processing of Temporal Information*, Lecture Notes in Artificial Intelligence, pp. 168–197. Berlin, Springer-Verlag.

Ghahramani, Z. (2004). Unsupervised Learning. In O. Bousquet, U. von Luxburg, and G. Raetsch (Eds.), *Advanced Lectures in Machine Learning*, pp. 72–112. Berlin, Springer-Verlag.

Giudici, P. and P. Green (1999). Decomposable graphical Gaussian model determination. *Biometrika 86*(4), 785–801.

Häggström, O. (2002). *Finite Markov Chains and Algorithmic Applications*. Cambridge University Press.

Harville, D. (1997). *Matrix Algebra from a Statistician's Perspective*. New York, Springer.

Hastings, W. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika 57*(1), 97–109.

Heckerman, D., D. Geiger, and D. Chickering (1995). Learning Bayesian networks: The combination of knowledge and statistical data. *Machine Learning 20*, 197–243.

Heidelberger, P. and P. Welch (1983). Simulation run length control in the presence of an initial transient. *Operations Research 31*, 1109–1144.

Henderson, H. and S. Searle (1979). Vec and vech operators for matrices, with some uses in Jacobian and multivariate statistics. *The Canadian Journal of Statistics 7*(1), 65–81.

Henderson, H. and S. Searle (1981). The vec-permutation matrix, the vec operator and Kronecker products: a review. *Linear and multilinear algebra 9*, 271–288.

Husmeier, D. (2003). Sensitivity and specificity of inferring genetic regulatory interactions from microarray experiments with dynamic Bayesian networks. *Bioinformatics 19*(17), 2271–2282.

Johnson, N., S. Kotz, and A. Kemp (1992). *Univariate Discrete Distributions* (2nd ed.). New York, Wiley.

Jones, B., C. Carvalho, A. Dobra, C. Hans, C. Carter, and M. West (2005). Experiments in stochastic computation for high-dimensional graphical models. *Statistical Science 20*(4), 388–400.

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

Kullback, S. and R. Leibler (1951). On information and sufficiency. *The Annals of Mathematical Statistics 22*(1), 79–86.

Lappalainen, H. and J. Miskin (2000). Ensemble learning. In M. Girolami (Ed.), *Advances in Independent Component Analysis*, pp. 76–92. Springer-Verlag.

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

Lindley, D. (1991). Discussion on Posterior Bayes factors (by M. Aitkin). *Journal of the Royal Statistical Society, Series B 53*, 130–131.

Lovász, L. (1993). Random walks on graphs: A survey. *Combinatorics, Paul Erdös is Eighty 2*, 353–397.

Lucas, J., C. Carvalho, Q. Wang, A. Bild, J. Nevins, and M. West (2006). Sparse statistical modelling in gene expression genomics. In K. Do, P. Mueller, and M. Vannucci (Eds.), *Bayesian Inference for Gene Expression and Proteomics*, pp. 155–176. Cambridge University Press.

Lütkepohl, H. (2005). *New Introduction to Multiple Time Series.* Springer-Verlag.

MacKay, D. (1995a). Developments in probabilistic modelling with neural networks – ensemble learning. In *Neural Networks: Artificial Intelligence and Industrial Applications. Proceedings of the 3rd Annual Symposium on Neural Networks, Nijmegen, Netherlands*, pp. 191–198. Springer.

MacKay, D. (1995b). Probable networks and plausible predictions - a review of practical Bayesian methods for supervised neural networks. *Network: Computation in Neural Systems 6*, 469–505.

Mihajlovic, V. and M. Petkovic (2001). Dynamic Bayesian Networks: A State of the Art. Technical report, Computer Science Department, University of Twente.

Miskin, J. (2000). *Ensemble learning for independent component analysis.* Ph. D. thesis, University of Cambridge.

Moore, E. (1920). On the reciprocal of the general algebraic matrix. *Bulletin of the American Mathematical Society 26*, 394–395.

Muirhead, R. (1982). *Aspects of Multivariate Statistical Theory.* New York, Wiley.

Murphy, K. (2002). *Dynamic Bayesian Networks: Representation, Inference and Learning.* Ph. D. thesis, UC Berkeley.

Neudecker, H. (1995). Mathematical properties of the variance of the multinomial distribution. *Journal of Mathematical Analysis and Applications 189*, 757–762.

Newton, M. and A. Raftery (1994). Approximate Bayesian inference with the weighted likelihood bootstrap (with discussion). *Journal of the Royal Statistical Society, Series B 56*(1), 3–48.

Nicolas, J. (2002). Introduction to second kind statistics: application of log-moments and log-cumulants to SAR image law analysis. *Traitement du signal 19*(3), 139–167.

O'Hagan, A. (1991). Discussion on Posterior Bayes factors (by M. Aitkin). *Journal of the Royal Statistical Society, Series B 53*, 136.

O'Hagan, A. (1995). Fractional Bayes factors for model comparison. *Journal of the Royal Statistical Society, Series B 57*(1), 99–138.

O'Hagan, A., A. Stuart, J. Ord, and M. Kendall (1994). *Kendall's Advanced Theory of Statistics: Bayesian Inference*, Volume 2B. Edward Arnold.

Penny, W., S. Kiebel, and K. Friston (2006). Variational Bayes. In K. Friston, J. Ashburner, S. Kiebel, T. Nichols, and W. Penny (Eds.), *Statistical Parametric Mapping: The analysis of functional brain images*. Elsevier, London.

Penny, W. and S. Roberts (2000). Bayesian methods for autoregressive models. In *IEEE Workshop on Neural Networks for Signal Processing*, Sydney, Australia.

Penny, W. and S. Roberts (2002, February). Bayesian multivariate autoregressive models with structured priors. In *IEE Proceedings - Vision, Image, and Signal Processing*, Volume 149, pp. 33–41.

Penrose, R. (1955). A generalised inverse for matrices. *Proceedings of the Cambridge Philosophical Society 51*, 406–413.

Petersen, K. and M. Pedersen (2007, September). The Matrix Cookbook. *http://matrixcookbook.com.*

Plummer, M., N. Best, K. Cowles, and K. Vines (2006, March). CODA: Convergence diagnosis and output analysis for MCMC. *R News 6*(1), 7–11.

Raftery, A. and S. Lewis (1992). How many iterations in the Gibbs sampler? In *Bayesian Statistics 4*, pp. 763–773. Oxford University Press.

Rao, C. (1962). A note on a generalized inverse of a matrix with applications to problems in mathematical statistics. *Journal of the Royal Statistical Society, Series B 24*(1), 152–158.

Rao, C. (1966). Generalized inverse for matrices and its applications in mathematical statistics. In F. David (Ed.), *Festschrift for J. Neyman: Research Papers in Statistics*, pp. 263–279. London, Wiley.

Rao, C. and S. Mitra (1972). Generalized inverse of a matrix and its applications. In *Proceedings of the Berkeley Symposium on Mathematical Statistics and Probability*, Volume 1, pp. 601–620. University of California Press.

Rice, J. (1995). *Mathematical Statistics and Data Analysis*. Duxbury Press.

Robert, C. (1993). A note on Jeffreys-Lindley paradox. *Statistica Sinica 3*, 601–608.

Robert, C. and D. Titterington (2002). Discussion on Bayesian measures of model complexity and fit (by D.J. Spiegelhalter *et al*). *Journal of the Royal Statistical Society, Series B 64*(4), 621–622.

Roberts, S. and W. Penny (2002). Variational Bayes for generalized autoregressive models. *IEEE Transactions on Signal Processing 50*(9), 2245–2257.

Russell, S. and P. Norvig (2003). *Artificial Intelligence: A Modern Approach* (2nd ed.). Prentice Hall.

Saad, Y. (2003). *Iterative Methods for Sparse Linear Systems* (2nd ed.). SIAM.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics 6*(2), 461–464.

Scott, J. and J. Berger (2006). An exploration of aspects of Bayesian multiple testing. *Journal of Statistical Planning and Inference 136*, 2144–2162.

Shafer, G. (1982). Lindley's paradox. *Journal of the American Statistical Association 77*(378), 325–334.

Speed, T. and H. Kiiveri (1986). Gaussian Markov distributions over finite graphs. *The Annals of Statistics 14*(1), 138–150.

Spiegelhalter, D., N. Best, B. Carlin, and A. van der Linde (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society, Series B 64*(4), 583–639.

Spiegelhalter, D., A. Thomas, N. Best, and W. Gilks (1995). BUGS: Bayesian inference using Gibbs sampling, Version 0.50. Technical report, Medical Research Council Biostatistics Unit, Institute of Public Health, Cambridge University.

Stoica, P. and Y. Selén (2004). Model-order selection: A review of information criterion rules. *IEEE Signal Processing Magazine 21*(4), 36–47.

Winn, J. (2003). *Variational message passing and its applications.* Ph. D. thesis, Department of Physics, University of Cambridge.