Stat Comput (2009) 19: 329–340
DOI 10.1007/s11222-008-9095-6

Variational Bayes for estimating the parameters of a hidden Potts model

C.A. McGrory · D.M. Titterington · R. Reeves · A.N. Pettitt

Received: 3 October 2006 / Accepted: 18 July 2007 / Published online: 12 September 2008

Abstract Hidden Markov random field models provide an appealing representation of images and other spatial problems. The drawback is that inference is not straightforward for these models as the normalisation constant for the likelihood is generally intractable except for very small observation sets. Variational methods are an emerging tool for Bayesian inference and they have already been successfully applied in other contexts. Focusing on the particular case of a hidden Potts model with Gaussian noise, we show how variational Bayesian methods can be applied to hidden Markov random field inference. To tackle the obstacle of the intractable normalising constant for the likelihood, we explore alternative estimation approaches for incorporation into the variational Bayes algorithm. We consider a pseudo-likelihood approach as well as the more recent reduced dependence approximation of the normalisation constant. To illustrate the effectiveness of these approaches we present empirical results from the analysis of simulated datasets. We also analyse a real dataset and compare results with those of previous analyses as well as those obtained from the recently developed auxiliary variable MCMC method and the recursive MCMC method. Our results show that the variational Bayesian analyses can be carried out much faster than the MCMC analyses and produce good estimates of model parameters. We also found that the reduced dependence approximation of the normalisation constant outperformed the pseudo-likelihood approximation in our analysis of real and synthetic datasets.

Keywords Potts/Ising model · Hidden Markov random field · Variational approximation · Bayesian inference · Pseudo-likelihood · Reduced dependence approximation

C.A. McGrory (✉) · R. Reeves · A.N. Pettitt
School of Mathematical Sciences, Queensland University of Technology, P.O. Box 2434, Brisbane, Queensland, 4001, Australia
e-mail: c.mcgrory@qut.edu.au

D.M. Titterington
University of Glasgow, Glasgow, UK

1 Introduction

Markov random fields (MRFs) are spatial models whose spatial locations or sites generally follow some sort of lattice structure. In a discrete Markov random field, the observation at each of these sites belongs to one of K, say, possible states. Each site on the lattice has a set of neighbouring sites and the attractiveness of such models lies in the fact that the conditional probability at each site is dependent only upon the values of its neighbours. This structure is useful in a variety of research areas where there is an interest in representing the spatial association between data. Some examples are medical imaging, environmental statistics and genetics. Usually in practical applications it is not clear to which state a given observation belongs, and in some cases even the number of states is unknown. In this setting, the hidden Markov random field (HMRF) representation is an appropriate one.

HMRFs have been used in various areas where interest lies in modelling spatial dependency between regions or objects which are geographically close to one another or which are related in some other way. Image analysis is an important application area for HMRFs, the work by Geman and Geman (1984) and Besag (1986) having been influential in this. Recent application areas include micro-array data analysis (Gottardo et al. 2006), brain imaging (Smith and Fahrmeir 2007), disease mapping (Green and Richardson 2002) and

agricultural field experiments (Besag and Higdon 1999). The Ising model for representing binary lattice data and its generalisation to categorical data, the Potts model, are two well-known and widely applied examples of an MRF-type model. These models originated in statistical physics but have been useful in many other modelling areas. This paper focuses on Bayesian inference in the case where the hidden data are represented by the Ising/Potts HMRF model.

A difficulty with HMRF models, and hence the hidden Ising and Potts models, is that the normalising constant associated with the likelihood is generally intractable. This of course presents a problem in Bayesian inference as the computation of the likelihood is integral to the approach. A very recent development for Markov chain Monte Carlo (MCMC) algorithms for analysing distributions such as HMRFs with intractable normalising constants is presented in Møller et al. (2006). Møller et al. (2006) avoid the problem of estimating the normalising constant altogether by developing an auxiliary variable MCMC scheme in which the troublesome constant is cancelled out. This work is developed further by Murray et al. (2006). This method produces accurate results but the drawback is that the computing time involved can be considerable.

Many modern applications involve large data sets. Therefore, methods of analysis that are computationally efficient as well as accurate are very valuable. Approximate methods can often provide a time efficient solution to inference problems with only a small reduction in accuracy. In this paper we show how variational approximations, or variational Bayes, can be used to carry out Bayesian inference for estimating the parameters of a hidden Ising/Potts model. We use the auxiliary variable MCMC approach as a reference method with which we compare our results. Variational methods are non-simulation based and computationally efficient and their usefulness has already been demonstrated in other scenarios such as mixture model analysis (McGrory and Titterington 2007; Corduneanu and Bishop 2001; Attias 1999) and hidden Markov chain analysis (McGrory and Titterington 2008; MacKay 1997).

Calculation of the posterior distribution of model parameters given observed data is a key objective in Bayesian analysis. This is not always straightforward as in many situations posterior distributions are intractable as is the case with HMRF models. The variational Bayes method provides a way of approximating such posteriors through the introduction of a simpler approximating function that is chosen to minimise the Kullback-Leibler divergence between the true and approximating function. However, as we have already mentioned, a difficulty which arises in the analysis of the hidden Potts model and other HMRF models, and hence in the variational Bayes scheme, is that the normalisation constant associated with the likelihood is usually intractable. This presents an obstacle for inference and consequently the investigation of normalising constant estimation techniques is an area of active research. Approaches include approximate methods such as thermodynamic integration, or path sampling (Gelman and Meng 1998; Pettitt et al. 2003). Reeves and Pettitt (2004) present an exact recursive method which enables calculation of the normalising constant for small lattices and this is extended in Friel et al. (2008) for application to larger lattices through an approximate method called the Reduced Dependence Approximation or RDA.

In our variational Bayes scheme we employ the RDA method of Friel et al. (2008) to estimate the normalising constant and compare results with those obtained by taking a pseudo-likelihood approach (Besag 1974, 1975, 1986) at the relevant point in the variational algorithm. Although the pseudo-likelihood may be a crude approximation, it is simplistic and can be computed quickly. For this reason we considered it an option worth exploring. In our analysis of the soil phosphate dataset, discussed in Sect. 5.2, we also compare our results with those obtained using the auxiliary variable MCMC approach of Møller et al. (2006) and the recursive method of Reeves and Pettitt (2004).

In Sect. 2 we outline the fundamental principles of the variational Bayes approach. In Sect. 3 we describe the Ising and Potts models and discuss some methods for approximating the intractable normalising constant of the associated likelihood. In Sect. 4 we show how variational Bayes can be used to infer the parameters of a hidden Potts model in conjunction with the normalising constant approximations. Results of our analyses of synthetic and real data are presented in Sect. 5 with conclusions given in Sect. 6.

2 The variational Bayes approach

Suppose we have a missing data model with parameters θ and latent variables z. Calculation of the posterior distribution of the parameters θ, given the observed data y, is central to carrying out a Bayesian analysis. However, in many practical situations, posterior distributions cannot be computed exactly. In these situations, the variational Bayes method provides a means of approximating the intractable posterior. This is done by introducing a variational function $q(\theta, z)$ as an approximation to the joint posterior $p(\theta, z|y)$ from which the required posterior over the parameters can be found. To find an approximating function that is as close as possible to the true posterior, $q(\theta, z)$ is chosen to be the minimiser of the Kullback-Leibler (KL) divergence between $q(\theta, z)$ and the joint posterior $p(\theta, z|y)$. The idea behind this is that minimising this KL divergence is the same as finding a rigorous lower bound on the log marginal likelihood. To see this, note that, by Jensen's inequality, the log-likelihood can be lower bounded in the following way.

$$\log p(y) = \log \int \sum_{\{z\}} p(y, z, \theta) d\theta$$

$$= \log \int \sum_{\{z\}} q(\theta, z) \frac{p(y, z, \theta)}{q(\theta, z)} d\theta$$

$$\geq \int \sum_{\{z\}} q(\theta, z) \log \frac{p(y, z, \theta)}{q(\theta, z)} d\theta. \quad (1)$$

Since the difference between the two sides of the inequality (1) is given by the KL divergence,

$$KL(q|p) = \int \sum_{\{z\}} q(\theta, z) \log \frac{q(\theta, z)}{p(\theta, z|y)} d\theta,$$

it is evident that maximising the lower bound, (1), is equivalent to minimising the KL divergence between the two quantities. It is well known that the KL divergence is minimised by taking $q(\theta, z) = p(\theta, z|y)$, but, in order to make any progress towards a solution, $q(\theta, z)$ must be computable. Therefore, to simplify matters we assume that $q(\theta, z)$ factorises over the parameters and latent values, leading to the distributional form $q(\theta, z) = q_\theta(\theta) q_z(z)$. The variational Bayes algorithm then proceeds to iteratively maximise (1) with respect to $q_\theta(\theta)$ and $q_z(z)$. At each iteration the lower bound is increased provided it is not already at a maximum. As a result of this similarity with the expectation maximisation (EM) algorithm, the algorithm is sometimes referred to as the variational Bayes expectation maximisation (VBEM) algorithm. Another outcome is that, when variational Bayes is applied to exponential-family models with the appropriate conjugate priors selected, the variational posterior for the model parameters will also belong to that conjugate family (for more detail on this point see, for example, Beal and Gharamani 2003 or McGrory 2005).

3 Hidden Markov random field modelling

Consider a grid or lattice with regularly spaced sites $i = 1, \ldots, n$; in the context of image analysis, these sites correspond to pixels. Each site belongs to one of K states. For example, these states might correspond to pixel colour when the lattice represents an image or, in a remote sensing problem, the states might correspond to land-use types. The true states $z = \{z_1, \ldots, z_n\}$ are hidden and what we observe are noisy observations $y = \{y_1, \ldots, y_n\}$. The conditional probability at a given site depends only upon the values of neighbouring sites. In a first-order neighbourhood, the neighbours of a site are the sites above, below, to the left and to the right. The notation δ_i denotes the sites that are the neighbours of site i. We also use the notation $i \sim j$ to indicate that i and j are neighbours of one another. It is possible to consider higher-order neighbourhood systems but we do not do so in this paper. Increasing the order of the neighbourhood considered increases the amount of information used to estimate the true value of each pixel. This could improve the performance of inference techniques in some cases but would also increase computation time.

The Hammersley-Clifford Theorem identifies MRFs with Gibbs distributions (see Besag 1974, 1975, for example). This is a valuable result because it provides a straightforward way of specifying the distribution of a MRF. For this reason MRFs are usually defined through their representations as Gibbs distributions. Two well-studied examples of Gibbsian models that are often used to represent images are the Ising model and its generalisation, the Potts model.

3.1 The Ising model and the Potts model

The idea of the Ising model translates in a natural way to the binary image setting, representing the two colours present by $+/-1$ and assuming that nearby pixels are likely to have similar colour values. Suppose that the parameter of the MRF is β and that the variables z_i, which represent the state or colour at a given pixel, take values in $\{-1, +1\}$. The Ising model is of the form

$$p(z|\beta) = \frac{\exp\{\beta \sum_{i \sim j} z_i z_j\}}{G(\beta)},$$

where $G(\beta)$ is a normalising constant which is usually not computable. The parameter β measures the strength of association between neighbouring pixels. Large positive values of β encourage neighbouring pixels to be of the same colour so that increasing β leads to bigger patches of like-coloured pixels in the image. In 2 dimensions the Ising model undergoes a continuous phase transition at the critical value $\beta_c \approx 0.44$ and moves from a disordered to an ordered phase. The development of ordering occurs gradually as β increases above the critical value and, in our context, this means that when β is sufficiently above the critical value, the entire image will be made up of only one colour. For this reason, values of β that are too large would not be of interest here. This transition phenomenon is very important in statistical physics where the Ising model has its origins as a model for magnetic materials. For further detail see, for example, Newman and Barkema (1999).

The natural extension of the Ising model to more than two states is given by the Potts model. In the Potts model each site can take more than two discrete values. For instance, a K-state Potts model is one in which each site can take values in states $1, \ldots, K$. The variables z corresponding to the states are K-vectors, $z_i = (z_{i1}, \ldots, z_{iK})$, the elements of which take values in $\{0, 1\}$ such that $z_{il} = 1$ if and only if the observation at site i belongs to state l. Then

$$p(z|\beta) = \frac{\exp\{\beta \sum_{i \sim j} \delta(z_i, z_j)\}}{G(\beta)}, \quad (2)$$

and

$$\delta(z_i, z_j) = 1, \quad \text{if } {z_i}^T z_j = \sum_{l=1}^{K} z_{il} z_{jl} = 1$$
$$= -1, \quad \text{otherwise, i.e. if } {z_i}^T z_j = \sum_{l=1}^{K} z_{il} z_{jl} = 0.$$

We can write

$$\delta(z_i, z_j) = 2 \sum_{l=1}^{K} z_{il} z_{jl} - 1.$$

For $K = 2$ states, the Potts model is equivalent to the Ising model up to an additive constant.

3.2 The hidden Potts model

This paper focuses on the case where our data are modelled by a K-state hidden Potts model with independent Gaussian noise. The joint probability distribution of y, z and θ is

$$p(y, z, \theta) = \left\{ \prod_{i=1}^{n} p(y_i|z_i, \phi) \right\} p(z|\beta) \left\{ \prod_{l=1}^{K} p(\phi_l) \right\} p(\beta),$$

where $\theta = (\phi, \beta)$. Here $\phi = (\phi_1, \ldots, \phi_K)$ where the ϕ_l's are parameters within the lth noise model and

$$p(y, z, \theta) = \left[\prod_{i=1}^{n} \prod_{l=1}^{K} \{p(y_i|\phi_l)\}^{z_{il}} \right] p(z|\beta) \times \left\{ \prod_{l=1}^{K} p(\phi_l) \right\} p(\beta). \tag{3}$$

We assume that the lth noise model is $N(\mu_l, \tau_l^{-1})$, with mean μ_l and where τ_l is the precision ($\tau_l^{-1} = \sigma_l^2$) so that $\phi_l = (\mu_l, \tau_l)$.

3.3 Approximating the normalising constant

The normalising constant, also referred to as the partition function, for the likelihood of a HMRF cannot be evaluated exactly unless the observation set is very small. In this paper we consider two approaches to dealing with the issue of approximating the normalising constant of the hidden Potts/Ising model within the variational framework. The first of these is the reduced dependence approximation (RDA) to the normalising constant and the second is the pseudo-likelihood approach.

The reduced dependence approximation

The RDA (Friel et al. 2008) extends the recursion method for normalising constant calculation, presented in Reeves and Pettitt (2004), to problems involving larger lattices. This results in an approximation to the true normalising constant. The recursive method (Reeves and Pettitt 2004) enables exact computation of the normalising constant for an unnormalised joint likelihood that can be expressed as a product of factors. As factorisation of the joint likelihood is applicable to any discrete Markov random field, the recursion method is suitable for these models. However, it is only feasible when the lattice size is not too large, that is up to approximately 20 rows, with computing time increasing proportionally with the number of columns.

The recursion method can be used to compute the normalising constant for the unnormalised likelihood, $\varphi(z|\beta)$, of a HMRF as follows. Here $\varphi(z|\beta)$ can be expressed in a factorised form as

$$\varphi(z|\beta) = \varphi_1(z_1, z_2, \ldots, z_{r+1}|\beta)\varphi_2(z_2, z_3, \ldots, z_{r+2}|\beta) \cdots \times \varphi_s(z_s, z_{s+1}, \ldots, z_n|\beta),$$

where $r < n$ and $s = n - r$. This is referred to as a lag-r model. Here r determines the degree of complexity or dependence of the model after optimal indexing of the z's so as to make r as small as possible. Note that the exact grouping of the terms of the HMRF in the factorisation is not unique nor is it of consequence. For an example of a possible factorisation for a HMRF model see Reeves and Pettitt (2004). As a result of the above factorisation, the normalising constant, $G(\beta) = \sum_{\{z_1, \ldots, z_n\}} \varphi(z|\beta)$, can be expressed as

$$G(\beta) = \sum_{z_{s+1} \ldots z_n} \sum_{z_s} \varphi_s(z_s, z_{s+1}, \ldots, z_n|\beta) \times \sum_{z_{s-1}} \varphi_{s-1}(z_{s-1}, z_s, \ldots, z_{n-1}|\beta) \cdots \times \sum_{z_1} \varphi_1(z_1, z_2, \ldots, z_{r+1}|\beta).$$

This expression can then be evaluated through forward recursion with a computation time significantly less than that of a straightforward summation over all possible realisations of the $(z_1, \ldots, z_n)$. Note that, when $r = 1$, the recursion method corresponds to the well-known forward-backward algorithm for hidden Markov models and so the recursion method is a generalisation of this result to a lattice. See Reeves and Pettitt (2004) for further detail on the recursive method. See also Jordan (2004) for an alternative perspective on this approach.

The RDA provides a means of extending this approach to larger lattice sizes. It finds a close approximation to the true normalising constant through the relaxation of certain dependencies within the model. This involves approximating $p(z|\beta)$ so that it can be expressed as a product of factors that are defined on sublattices. The normalising constants of

these sublattices can then be computed using the recursive method described above allowing us to obtain an approximation to the true normalising constant as follows. We denote the vector of states in row ρ by $\mathbf{r}_\rho$ and we denote by u', say, a number of rows which is smaller than the actual number of rows, u in the lattice. Then, making use of the Markov property and an approximation, we can approximate $p(z|\beta)$ as

$$\begin{aligned} p(z|\beta) &= p(\mathbf{r}_{u-u'+1},\ldots,\mathbf{r}_u|\beta)\prod_{\rho=1}^{u-u'} p(\mathbf{r}_\rho|\mathbf{r}_{\rho+1},\beta) \\ &= p(\mathbf{r}_{u-u'+1},\ldots,\mathbf{r}_u|\beta)\prod_{\rho=1}^{u-u'} \frac{p(\mathbf{r}_1,\ldots,\mathbf{r}_\rho|\mathbf{r}_{\rho+1},\beta)}{p(\mathbf{r}_1,\ldots,\mathbf{r}_{\rho-1}|\mathbf{r}_\rho,\beta)} \\ &\approx p(\mathbf{r}_{u-u'+1},\ldots,\mathbf{r}_u|\beta)\prod_{\rho=1}^{u-u'} \frac{p(\mathbf{r}_{p-u'},\ldots,\mathbf{r}_\rho|\beta)}{p(\mathbf{r}_{p-u'},\ldots,\mathbf{r}_{\rho-1}|\beta)}. \end{aligned} \tag{4}$$

The marginal probabilities in (4) can be approximated as

$$p(\mathbf{r}_{p-u'},\ldots,\mathbf{r}_\rho|\beta) \approx \frac{\varphi(\mathbf{r}_{p-u'},\ldots,\mathbf{r}_\rho|\beta)}{G_{(u'+1)\times v}(\beta)}.$$

In the above, v is the number of columns in the lattice, $\varphi(\mathbf{r}_{\rho-u'},\ldots,\mathbf{r}_\rho|\beta)$ represents the un-normalised distribution of the sub-lattice of dimension $(u'+1)\times v$ defined on rows $\rho-u',\ldots,\rho$ of the lattice and $G_{(u'+1)\times v}(\beta)$ is the corresponding normalising constant of the sub-lattice. The above expression is an approximation because it ignores some of the associations between rows within the lattice. See Friel et al. (2008) for more detail on this point. Expressing each of the marginal distributions in (4) similarly leads to the approximation for the normalising constant

$$\widetilde{G(\beta)} = \frac{(G_{(u'+1)\times v}(\beta))^{u-u'}}{(G_{u'\times v}(\beta))^{u-u'-1}}.$$

The normalising constants of these sub-lattices can be computed exactly via the recursive method when u' is chosen to be reasonably small.

The pseudo-likelihood approach

This approach involves replacing the intractable likelihood, at appropriate stages, by the pseudo-likelihood in the spirit of Rydén and Titterington (1998). The pseudo-likelihood is given by

$$\begin{aligned} p_{\mathrm{PL}}(z|\beta) &= \prod_{i=1}^{n} p(z_i|z_{-i},\beta) \quad \text{where } z_{-i}=\{z_j : j\neq i\} \\ &= \prod_{i=1}^{n} p(z_i|z_{\delta_i},\beta), \end{aligned}$$

where z_{δ_i} denotes the z-values of the pixels which are neighbours of pixel i. The normalising constants for the factors of the pseudo-likelihood are computable and so the unobtainable quantity has been replaced by something more amenable.

4 Variational Bayesian analysis of a hidden K-state Potts model

We consider a K-state hidden Potts model with independent Gaussian noise and joint probability distribution given by (3) as described in Sect. 3.2.

Assigning the prior distributions

We assign independent Gaussian priors to the means $\mu=(\mu_1,\ldots,\mu_K)$, conditional on the precisions $\tau=(\tau_1,\ldots,\tau_K)$, so that

$$p(\mu|\tau)=\prod_{l=1}^{K} N(\mu_l; m_l^{(0)}, (\lambda_l^{(0)}\tau_l)^{-1}).$$

The precisions are given independent Gamma prior distributions:

$$p(\tau)=\prod_{l=1}^{K} Ga\left(\tau_l; \frac{1}{2}\gamma_l^{(0)}, \frac{1}{2}\xi_l^{(0)}\right).$$

In the above, the $\{m_l^{(0)}\}$, $\{\lambda_l^{(0)}\}$, $\{\gamma_l^{(0)}\}$ and $\{\xi_l^{(0)}\}$ are chosen hyperparameters and $N(.;.,.)$ and $Ga(.;.,.)$ denote the Gaussian and Gamma probability density functions, respectively.

4.1 Form of the variational posterior distributions for the noise model parameters

We wish to maximise (1). We assume that

$$q(z,\theta)=q_z(z)\left\{\prod_{l=1}^{K} q_l(\phi_l)\right\} q_\beta(\beta).$$

This results in an optimal $q_l(\phi_l)$ of the form

$$q_l(\phi_l) \propto \prod_{i=1}^{n} \left\{p(y_i|\phi_l)^{q_{il}}\right\} p(\phi_l), \tag{5}$$

where $q_{il}=\mathbf{E}_z(z_{il})=P_{q_z}(i\text{th data point belongs to state } l)$. (The expectations here and throughout are with respect to the variational approximation.) See the Appendix for a derivation of the above formula. This results in variational posteriors of the form

$$q(\mu_l|\tau_l)=N(\mu_l; m_l, (\lambda_l\tau_l)^{-1}),$$

$$q(\tau_l)=Ga\left(\tau_l; \frac{1}{2}\lambda_l, \frac{1}{2}\xi_l\right),$$

with hyperparameters given by

$$\lambda_l = \lambda_l^{(0)} + \sum_{i=1}^{n} q_{il},$$

$$\gamma_l = \gamma_l^{(0)} + \sum_{i=1}^{n} q_{il},$$

$$m_l = \frac{\lambda_l^{(0)} m_l^{(0)} + \sum_{i=1}^{n} q_{il} y_i}{\lambda_l},$$

$$\xi_l = \xi_l^{(0)} + \sum_{i=1}^{n} q_{il} y_i^2 + \lambda_l^{(0)} m_l^{(0)^2} - \lambda_l m_l^2.$$

4.2 Optimisation of $q_z(z)$

The optimal $q_z(z)$ (see Appendix) is

$$q_z(z) \propto \exp\left\{ \sum_{i=1}^{n} \sum_{l=1}^{K} z_{il} \mathbf{E}_{\phi_l} \log p(y_i|\phi_l) + \mathbf{E}_\beta \log p(z|\beta) \right\}.$$

We are unable to compute the optimal $q_z(z)$ explicitly because of the complexity of $p(z|\beta)$. The simplest proposal is to assume a fully factorised form for $q_z(z)$, i.e. $q_z(z) = \prod_{i=1}^{n} q_{z_i}(z_i)$. This gives

$$q_{z_i}(z_i) \propto \exp\left\{ \sum_{l=1}^{K} z_{il} \mathbf{E}_{\phi_l} \log p(y_i|\phi_l) + \mathbf{E}_\beta \mathbf{E}_{z_{-i}} \log p_i(z_i|\beta) \right\}, \tag{6}$$

where z_{-i} represents all z_j's except for z_i. Here $\log p_i(z_i|\beta)$ is the part of $\log p(z|\beta)$ that depends on z_i. By the definition of $\delta(z_i, z_j)$, we can write

$$p(z|\beta) = \frac{\exp\{2\beta \sum_{i\sim j} \sum_{l=1}^{K} z_{il} z_{jl}\}}{G^*(\beta)}$$

where $G^*(\beta) = G(\beta)e^\beta$. Thus, $\mathbf{E}_\beta \mathbf{E}_{z_{-i}}[\log p_i(z_i|\beta)]$ is given by

$$\mathbf{E}_\beta \left\{ 2\beta \sum_{l=1}^{K} z_{il} \sum_{j\epsilon\delta_i} \mathbf{E}_{z_j} z_{jl} \right\} = 2\mathbf{E}_\beta(\beta) \sum_{l=1}^{K} z_{il} \left(\sum_{j\epsilon\delta_i} q_{jl} \right).$$

Substituting the above expression into (6) gives

$$q_{z_i}(z_i) \propto \exp\left\{ \sum_{l=1}^{K} z_{il} (\mathbf{E}_{\phi_l} \log p(y_i|\phi_l) + 2\mathbf{E}_\beta(\beta) \sum_{j\epsilon\delta_i} q_{jl}) \right\},$$

and therefore

$$q_{il} \propto \exp\left\{ \mathbf{E}_{\phi_l}[\log p(y_i|\phi_l)] + 2\mathbf{E}_\beta(\beta) \sum_{j\epsilon\delta_i} q_{jl} \right\},$$

$$l = 1, \ldots, K, \tag{7}$$

normalised so that $\sum_{l=1}^{K} q_{il} = 1$. We perform five iterations of the calculation of the above equation to obtain a result. In the above, apart from an additive constant,

$$\mathbf{E}_{\phi_l}[\log p(y_i|\phi_l)] = \frac{1}{2}\mathbf{E}_{\phi_l}[\log(\tau_l)] - \frac{1}{2}\mathbf{E}_{\phi_l}(\tau_l)(y_i - m_l)^2 - \frac{1}{2\xi_l},$$

with expectations given by

$$\mathbf{E}_{\phi_l}[\log(\tau_l)] = \Psi\left(\frac{\gamma_l}{2}\right) - \log\left(\frac{\xi_l}{2}\right),$$

$$\mathbf{E}_{\phi_l}(\tau_l) = \frac{\gamma_l}{\xi_l},$$

where Ψ is the digamma function.

Therefore, given the q_{il}'s, the $q_l(\phi_l)$'s can be updated through (5). Given $\mathbf{E}_\beta(\beta)$, the q_{il}'s can be calculated/updated through (7).

4.3 Optimisation of $q_\beta(\beta)$

The optimal $q_\beta(\beta)$ (see Appendix) is

$$q_\beta(\beta) \propto \exp\{\mathbf{E}_z \log p(z|\beta)\}\, p(\beta). \tag{8}$$

There is no conjugate set-up for the prior for β and so we use a uniform prior, i.e. we set $p(\beta) =$ constant over the range of permitted values. The complexity of $p(z|\beta)$ also prevents us from calculating (8) explicitly. By the assumption of a factorised form for $q_z(z)$,

$$\mathbf{E}_z \log p(z|\beta) = 2\beta \sum_{i\sim j} \sum_{l=1}^{K} q_{il} q_{jl} - \log G^*(\beta),$$

so that

$$q_\beta(\beta) \propto \frac{\exp\{2\beta \sum_{i\sim j} \sum_{l=1}^{K} q_{il} q_{jl}\} p(\beta)}{G^*(\beta)}.$$

In principle $q_\beta(\beta)$ can be updated from the q_{il}'s. However, this is computationally infeasible because of the intractability of the normalising constant and so we consider the two aforementioned approaches to dealing with this problem.

The reduced dependence approximation

Using the RDA, the true normalising constant for the likelihood is approximated by

$$\widetilde{G(\beta)} = \frac{(G_{(u'+1)\times v}(\beta))^{u-u'}}{(G_{u'\times v}(\beta))^{u-u'-1}},$$

giving

$$q_\beta^{\mathrm{RDA}}(\beta) \propto \frac{\exp\{2\beta \sum_{i\sim j} \sum_{l=1}^{K} q_{il} q_{jl}\} p(\beta) e^{-\beta}}{\widetilde{G(\beta)}}. \tag{9}$$

A pseudo-likelihood approach

Here we replace $p(z|\beta)$ by

$$p_{\mathrm{PL}}(z|\beta) := \prod_{i=1}^{n} p(z_i|z_{-i},\beta) = \prod_{i=1}^{n} p(z_i|z_{\delta_i},\beta),$$

where

$$p(z_i|z_{\delta_i},\beta) = \frac{\exp\{2\beta\sum_{l=1}^{K} z_{il}\sum_{j\epsilon\delta_i} z_{jl}\}}{\sum_{z_{i'}}\exp\{2\beta\sum_{l=1}^{K} z_{i'l}\sum_{j\epsilon\delta_{i'}} z_{jl}\}} = \frac{\exp\{2\beta\sum_{l=1}^{K} z_{il}\sum_{j\epsilon\delta_i} z_{jl}\}}{\sum_{l=1}^{K}\exp\{2\beta\sum_{j\epsilon\delta_i} z_{jl}\}}.$$

Thus

$$p_{\mathrm{PL}}(z|\beta) = \prod_{i=1}^{n} \frac{\exp\{2\beta\sum_{l=1}^{K} z_{il}\sum_{j\epsilon\delta_i} z_{jl}\}}{\sum_{l=1}^{K}\exp\{2\beta\sum_{j\epsilon\delta_i} z_{jl}\}}.$$

If we replace $p(z|\beta)$ by $p_{\mathrm{PL}}(z|\beta)$, then the optimum $q_\beta(\beta)$ has the form

$$q_\beta^{\mathrm{PL}}(\beta) \propto \exp\{\mathbf{E}_z \log p_{\mathrm{PL}}(z|\beta)\}p(\beta), \tag{10}$$

where

$$\mathbf{E}_z \log p_{\mathrm{PL}}(z|\beta) = 2\beta\sum_{i=1}^{n}\sum_{j\in\delta_i}\sum_{l=1}^{K} q_{il}q_{jl} - \sum_{i=1}^{n}\mathbf{E}_{z_{\delta_i}}\left[\log\left\{\sum_{l=1}^{K}\exp\left(2\beta\sum_{j\in\delta_i} z_{jl}\right)\right\}\right],$$

and so we have

$$\exp\{\mathbf{E}_z \log p_{\mathrm{PL}}(z|\beta)\} = \frac{\exp\{2\beta\sum_{i=1}^{n}\sum_{j\in\delta_i}\sum_{l=1}^{K} q_{il}q_{jl}\}}{\exp(\sum_{i=1}^{n}\mathbf{E}_{z_{\delta_i}}[\log\{\sum_{l=1}^{K}\exp(2\beta\sum_{j\in\delta_i} z_{jl})\}])}. \tag{11}$$

In principle this can be used in (10), but in the case of a first-order hidden Markov random field, each $\mathbf{E}_{z_{\delta_i}}$ in the denominator contains K^4 terms, making computation impractical.

Here we suggest tackling this obstacle by approximating the denominator of (11) by

$$\exp\left[\sum_{i=1}^{n}\left[\log\left\{\sum_{l=1}^{K}\exp\left(2\beta\sum_{j\epsilon\delta_i} q_{jl}\right)\right\}\right]\right] = \prod_{i=1}^{n}\left\{\sum_{l=1}^{K}\exp\left(2\beta\sum_{j\epsilon\delta_i} q_{jl}\right)\right\},$$

so that

$$q_\beta^{\mathrm{PL}}(\beta) \propto \prod_{i=1}^{n}\frac{\exp\{2\beta\sum_{j\epsilon\delta_i}\sum_{l=1}^{K} q_{il}q_{jl}\}p(\beta)}{\{\sum_{l=1}^{K}\exp(2\beta\sum_{j\epsilon\delta_i} q_{jl})\}}. \tag{12}$$

Note that the use of the pseudo-likelihood approximation in addition to the approximation to the denominator of (11) amounts to taking a mean field-like approach to approximating the likelihood of the HMRF.

Approximating the expected value of β

Calculation of $\mathbf{E}_\beta(\beta)$ for use in (7) also requires us to normalise (9) and (12). We have

$$q_\beta(\beta) = C\,Q(\beta),$$

where $Q(\beta)$ equals the right-hand side of (9) or (12) for the RDA and PL, respectively and C is a normalising constant. We use numerical quadrature to approximate the normalising constant and first moment for $Q(\beta)$. The prior for β was taken as being uniform on $(0, 0.6)$ and so $p(\beta)$ was constant in the calculation. The prior for β allows values slightly above the critical value for phase transition. Values larger than 0.6 do not seem worthy of consideration as we are interested in modelling images which of course comprise more than one colour.

5 Results

5.1 Simulated images

Using simulated datasets we compared results obtained from the variational analysis using a pseudo-likelihood approximation to those obtained from the variational analysis using the reduced dependence approximation to the normalisation constant. We considered two images of size 40 by 40 simulated from the Ising model with β equal to 0.3 and 0.4, respectively. We added independent Gaussian noise to the two images with means equal to 0 and standard deviations of 0.6, 0.7, 1 and 1.25, respectively. We repeated the simulation of each image and the subsequent analysis 20 times. The results presented in Table 1 are an average of the estimates obtained over the 20 simulations. To give a comparison of our variational method with an alternative MCMC approach, we compared results with those obtained using the auxiliary variable MCMC method. Results from the MCMC analysis are given in Table 2 and these are also an average of 20 replications. Note that the recursive method was unavailable to us here as the lattice size is too large.

In our implementation, the hyperparameters $m_1^{(0)}$ and $m_2^{(0)}$ were both chosen to be 0 and $\lambda_1^{(0)}$ and $\lambda_2^{(0)}$ were set

Table 1 Estimation of parameters in the simulation study using the variational Bayes scheme: averages from 20 replications

True distribution			Variational post. means using PL			Variational post. means using RDA 10		
$\hat{\beta}$	$\hat{\mu_1}$	$\hat{\sigma_1}$	$\hat{\beta}$	$\hat{\mu_1}$	$\hat{\sigma_1}$	$\hat{\beta}$	$\hat{\mu_1}$	$\hat{\sigma_1}$
	$\hat{\mu_2}$	$\hat{\sigma_2}$		$\hat{\mu_2}$	$\hat{\sigma_2}$		$\hat{\mu_2}$	$\hat{\sigma_2}$
0.30	−1.00	0.60	0.315	−0.998	0.598	0.300	−0.999	0.596
	1.00	0.60		1.010	0.599		1.001	0.598
0.30	−1.00	0.70	0.331	−1.050	0.669	0.303	−1.042	0.697
	1.00	0.70		0.989	0.731		0.985	0.711
0.30	−1.00	1.00	0.389	−1.044	0.985	0.288	−1.032	0.936
	1.00	1.00		0.899	1.044		1.014	0.938
0.30	−1.00	1.25	0.424	−1.117	1.326	0.269	−1.105	1.085
	1.00	1.25		0.882	1.295		1.126	1.116
0.40	−1.00	0.60	0.412	−1.005	0.596	0.400	−1.002	0.597
	1.00	0.60		1.007	0.587		1.001	0.596
0.40	−1.00	0.70	0.439	−0.981	0.730	0.403	−0.987	0.678
	1.00	0.70		1.081	0.679		1.124	0.727
0.40	−1.00	1.00	0.445	−0.958	1.010	0.398	−0.975	0.967
	1.00	1.00		1.120	0.978		1.122	0.972
0.40	−1.00	1.25	0.491	−0.902	1.289	0.391	−0.980	1.185
	1.00	1.25		1.288	1.255		1.047	1.192

Table 2 Estimation of parameters in the simulation study using the auxiliary variable method: averages from 20 replications

True distribution			Post. mean using auxiliary variable MCMC		
$\hat{\beta}$	$\hat{\mu_1}$	$\hat{\sigma_1}$	$\hat{\beta}$	$\hat{\mu_1}$	$\hat{\sigma_1}$
	$\hat{\mu_2}$	$\hat{\sigma_2}$		$\hat{\mu_2}$	$\hat{\sigma_2}$
0.30	−1.00	0.60	0.297	−1.010	0.594
	1.00	0.60		0.997	0.597
0.30	−1.00	0.70	0.298	−1.012	0.693
	1.00	0.70		0.995	0.696
0.30	−1.00	1.00	0.299	−1.013	0.989
	1.00	1.00		1.000	0.991
0.30	−1.00	1.25	0.299	−1.014	1.233
	1.00	1.25		1.003	1.288
0.40	−1.00	0.60	0.399	−1.007	0.594
	1.00	0.60		0.984	0.603
0.40	−1.00	0.70	0.398	−1.009	0.693
	1.00	0.70		0.982	0.702
0.40	−1.00	1.00	0.399	−1.012	0.986
	1.00	1.00		0.977	0.999
0.40	−1.00	1.25	0.403	−1.036	1.233
	1.00	1.25		0.963	1.253

Table 3 Estimates of model parameters resulting from alternative analyses of the soil phosphate dataset

	Variational post. means using PL	Variational post. means using RDA 10	Recursive method post. means	Auxiliary variable MCMC post. means
$\hat{\beta}$	0.599	0.445	0.442	0.442
$\hat{\mu}_1$	3.871	3.851	3.897	3.895
$\hat{\sigma}_1$	0.345	0.337	0.361	0.359
$\hat{\mu}_2$	4.349	4.349	4.360	4.359
$\hat{\sigma}_2$	0.278	0.275	0.239	0.241

as 0.05. The Gamma prior distribution for the precision was also chosen to be uninformative.

Using the RDA, our estimates of the association parameter β seem not to be biased. However, the RDA approach slightly underestimates β when the image is very noisy in the case where the true value is 0.3. We observed this in the majority of the replications of the experiment. The PL, on the other hand, has a tendency to overestimate β when the data are noisy. We found this to be the case in all of the replications of the experiment. Both approximations produced good variational posterior estimates of parameters. The more accurate estimate of β that resulted from the RDA led to only slightly improved noise model parameter estimates.

The auxiliary variable MCMC analysis produces quality estimates of the model parameters and is less sensitive to noise than the other approaches. However, the drawback of the method is the length of time required to implement it. From a computational time perspective, variational Bayes is much more efficient than the auxiliary method. Images with a significant amount of noise result in a longer computation time than those that are less noisy. On a 3.2 GHz Pentium 4 desktop PC with 1 GB of RAM, variational Bayes with PL took approximately between 10 and 19 minutes to analyse a dataset of this size. Variational Bayes with RDA took roughly 12 to 21 minutes. The auxiliary variable MCMC method with 10,000 iterations required approximately 2 hours to analyse the datasets with association parameter β equal to 0.3. For β equal to 0.4 the computing time required for auxiliary variable MCMC increased to around 10 hours per image. If β were to be increased further we would expect the computational time to be lengthened also. It should be noted however that the auxiliary method involves perfect sampling and much of the computing time for the method can be attributed to that. The computing time for the auxiliary method could be reduced by improving the efficiency of the perfect sampler.

5.2 Soil phosphate measurements at an archaeological survey site

The data in this application comprise a set of soil phosphate measurements taken during the 1987 year of the Laconia Archaeological Survey in Greece. Soil phosphate concentration, resulting from the decomposition of organic material, is often higher in areas where archaeological activity is known to have taken place. Therefore, locating regions of high and low phosphorus content across a survey area can be helpful in identifying sites where archaeological activity has occurred. The measurements in this dataset were taken over a 16 by 16 grid at 10 m intervals at the Greek archaeological site and there are missing data at 9 of the sites. These data were first analysed by Buck et al. (1988), using Bayesian change-point analysis, and later by Besag et al. (1991) using Bayesian image analysis techniques with an Ising model representation. We model the data using the Ising model and we assume that the means and variances of the Gaussian noise distributions are unknown and estimate them along with the association parameter β. Following the previous analyses, we assume Gaussian distributions for the log phosphate concentrations in mg P/100 g of soil. However, unlike the previous analyses of the data we do not assume equal variances for the noise models. We used the same uninformative priors for the means and precisions as in the previous section.

Table 3 displays the posterior means resulting from four different analyses of this dataset. The first two methods are the variational approach presented in this paper using the PL approach and the RDA approach conditioning on 10 rows (no advantage was achieved by conditioning on more than 10 rows) to estimate the normalising constant. Results are also given from two MCMC analyses of the data using the forward recursion method and the auxiliary variable MCMC method, respectively, for comparison. Figures 1 and 2 represent the areas which were identified as being high or low in phosphorous concentration. In these figures the posterior probability of high concentration at each grid site is plotted using a grey scale in which high concentration corresponds to black and low concentration corresponds to white.

We used a uniform prior for β over the range 0 to 0.6. The PL approximation has led to an overly high estimate of β. This is apparent in Fig. 1 as the evidence of human presence in the lower right section of the grid is suppressed and the map of high and low areas we obtained when using the PL shows less similarity to that obtained using the other methods. The Gibbs sampling analysis of this dataset based

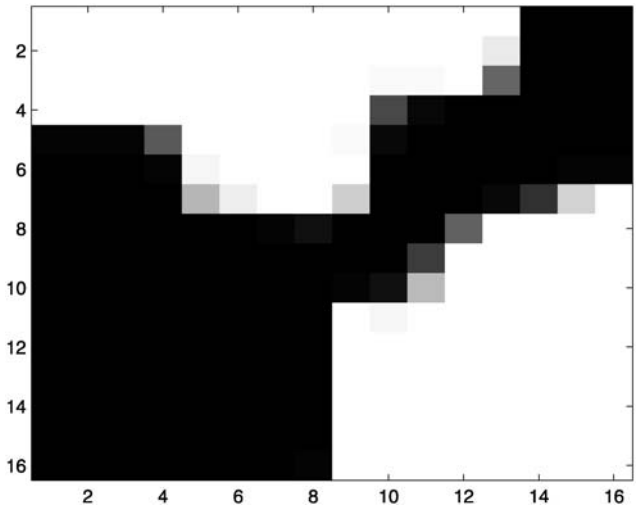

Fig. 1 High and low phosphorous areas as identified by variational Bayes with PL estimation

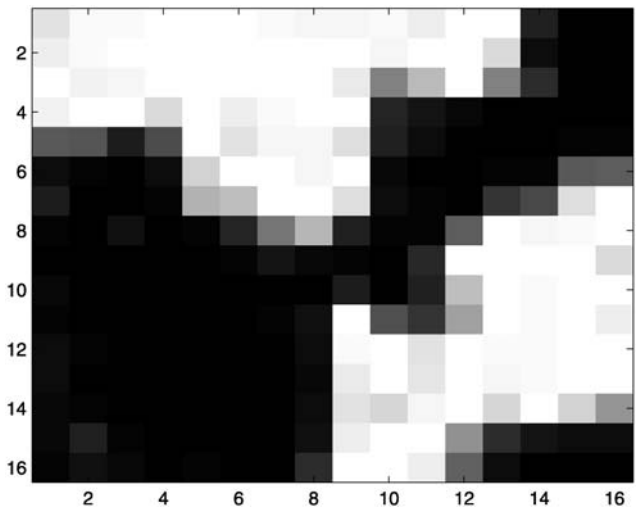

Fig. 2 High and low phosphorous areas as identified by variational Bayes with RDA

on the PL performed in the paper by Besag et al. (1991) also resulted in an overly high final estimate β. However, despite the overestimation of β that is obtained from the PL approach, we can see from Table 3 that the resulting variational posterior parameter estimates still compare favourably with those of the other analyses.

The estimates of β resulting from the variational Bayes with RDA approach, the auxiliary variable MCMC method and recursive method are in close agreement. The regions identified by variational Bayes with the RDA approximation as having high posterior probability of human archaeological activity (Fig. 2) are comparable to those found by an auxiliary variable MCMC analysis, by the recursive method and those found in Besag et al. (1991). All of these methods find two main activity areas and a smaller area in the lower right part of the grid. The analysis by Buck et al. (1988) did not provide posterior probabilities for the phosphorous concentration levels, but it also picked out two larger regions of likely activity and one smaller area.

The computing time required to perform both variational analyses was considerably shorter than that needed for the recursive and auxiliary variable approaches. On a 3.2 GHz Pentium 4 desktop PC with 1 GB of RAM, the computing time for the recursive analysis was approximately 3 days and the time for the auxiliary variable analysis was approximately 24 hours while variational Bayes with PL took approximately 5 minutes and variational Bayes with RDA took around 7 minutes.

6 Conclusions

We have shown how a variational Bayes scheme can be constructed to analyse a hidden Potts model and demonstrated the results on binary images. In doing so we have demonstrated that variational Bayes is an effective means of inference in the challenging situation where noise model parameters and the association parameter in the Gibbsian model are unknown. We have also compared the quality of results obtained when using two alternative approaches to overcome the normalising constant estimation problem. Both approaches resulted in a time efficient algorithm and good approximations to the noise model parameters. However, the use of the RDA to approximate the normalising constant led to more accurate results on the whole than did the PL approach.

In addition, we compared the variational scheme with two recent MCMC approaches in our analysis of a real data set and synthetic datasets. The computational time for the variational analyses was significantly shorter than that required for the MCMC analyses. Results from the variational Bayes analysis were comparable with the MCMC results lending further support to the usefulness of variational approximate methods for HMRF inference in the particular case where the Ising model is used to represent the data.

Extension of the variational Bayes technique to more general HMRF-type models such as the autologistic model is possible and is a topic of current research. The variational technique can be straightforwardly applied to models belonging to the exponential family, therefore we can easily replace the noise model with another from that family. Using a second-order neighbourhood structure would not excessively increase the computation time for a variational analysis, therefore future work will include investigation of what improvements in inference the use of a higher order neighbourhood might bring.

Acknowledgements The authors would like to thank the Associate Editor and two anonymous reviewers for their constructive suggestions which led to improvements in this paper. The research of the authors C.A.M., R.R. and A.N.P. was partially supported by an Australian Research Council Grant DP0558199.

Appendix

To derive the variational posterior distributions we have to maximise the lower bound (1). We can express the lower bound as follows:

$$\int \sum_{\{z\}} q(\theta,z)\log\frac{p(y,z,\theta)}{q(\theta,z)}d\theta = \int \sum_{\{z\}} q_z(z)\left\{\prod_{l=1}^{K} q_l(\phi_l)\right\} q_\beta(\beta) \times \log\frac{\{\prod_{i=1}^{n}\prod_{l=1}^{K} p(y_i|\phi_l)^{z_{il}}\}p(z|\beta)\{\prod_{l=1}^{K} p(\phi_l)\}p(\beta)}{q_z(z)\{\prod_{l=1}^{K} q_l(\phi_l)\}q_\beta(\beta)}d\phi d\beta. \tag{13}$$

To find the form of the posterior for $q_z(z)$ we concentrate on the parts of (13) that involve z to obtain the following:

$$\begin{aligned}
&\int \sum_{\{z\}} q_z(z)\left\{\prod_{l=1}^{K} q_l(\phi_l)\right\} q_\beta(\beta)\log\frac{\{\prod_{i=1}^{n}\prod_{l=1}^{K} p(y_i|\phi_l)^{z_{il}}\}p(z|\beta)}{q_z(z)}d\phi d\beta \\
&\quad + \text{terms not involving } q_z(z) \\
&= \sum_{\{z\}} q_z(z)\log\frac{\exp\{\int\{\prod_{l=1}^{K} q_l(\phi_l)\}q_\beta(\beta)\sum_{i=1}^{n}\sum_{l=1}^{K} z_{il}\log p(y_i|\phi_l)+\log p(z|\beta)d\phi d\beta\}}{q_z(z)} \\
&\quad + \text{terms not involving } q_z(z) \\
&= \sum_{\{z\}} q_z(z)\log\frac{\exp\{\sum_{i=1}^{n}\sum_{l=1}^{K} z_{il}\mathbf{E}_{\phi_l}\log p(y_i|\phi_l)+\mathbf{E}_\beta\log p(z|\beta)\}}{q_z(z)} \\
&\quad + \text{terms not involving } q_z(z).
\end{aligned}$$

The above expression is optimised when

$$q_z(z)\propto\exp\left\{\sum_{i=1}^{n}\sum_{l=1}^{K} z_{il}\mathbf{E}_{\phi_l}\log p(y_i|\phi_l)+\mathbf{E}_\beta\log p(z|\beta)\right\}.$$

Similarly, to optimise $q_\phi(\phi)$, we concentrate on the parts of (13) that involve ϕ as follows

$$\begin{aligned}
&\int \sum_{\{z\}} q_z(z)\left\{\prod_{l=1}^{K} q_l(\phi_l)\right\} \\
&\quad \times\log\frac{\{\prod_{i=1}^{n}\prod_{l=1}^{K} p(y_i|\phi_l)^{z_{il}}\}\{\prod_{l=1}^{K} p(\phi_l)\}}{\{\prod_{l=1}^{K} q_l(\phi_l)\}}d\phi \\
&\quad + \text{terms not involving } q_\phi(\phi) \\
&= \int\left\{\prod_{l=1}^{K} q_l(\phi_l)\right\} \\
&\quad \times\log\frac{\{\prod_{i=1}^{n}\prod_{l=1}^{K} p(y_i|\phi_l)^{q_{il}}\}\{\prod_{l=1}^{K} p(\phi_l)\}}{\{\prod_{l=1}^{K} q_l(\phi_l)\}}d\phi \\
&\quad + \text{terms not involving } q_\phi(\phi).
\end{aligned}$$

This expression is maximised when

$$q_l(\phi_l)\propto\prod_{i=1}^{n}\left\{p(y_i|\phi_l)^{q_{il}}\right\}p(\phi_l).$$

Similarly, to optimise with respect to $q_\beta(\beta)$, we obtain that (13) is

$$\begin{aligned}
&\int \sum_{\{z\}} q_z(z)q_\beta(\beta)\log\frac{p(z|\beta)p(\beta)}{q_\beta(\beta)}d\beta \\
&\quad + \text{terms not involving } q_\beta(\beta) \\
&= \int q_\beta(\beta)\log\frac{\exp\{\mathbf{E}_z\log p(z|\beta)\}p(\beta)}{q_\beta(\beta)}d\beta \\
&\quad + \text{terms not involving } q_\beta(\beta).
\end{aligned}$$

Thus, the optimal $q_\beta(\beta)$ is

$$q_\beta(\beta)\propto\exp\{\mathbf{E}_z\log p(z|\beta)\}p(\beta).$$

References

Attias, H.: Inferring parameters and structure of latent variable models by variational Bayes. In: Proceedings of 15th Conference on Uncertainty in Artificial Intelligence (1999)

Beal, M.J., Ghahramani, Z.: The variational Bayesian EM algorithm for incomplete data: with application to scoring graphical model structures. In: Bernardo, J.M., Bayarri, M.J., Berger, J.O., David, A.P., Heckerman, D., Smith, A.F.M., West, M. (eds.) Proceedings of the Seventh Valencia International Meeting. Bayesian Statistics, vol. 7, pp. 453–464. Oxford University Press, London (2003)

Besag, J.: Spatial interaction and the statistical analysis of lattice systems (with discussion). J. R. Stat. Soc. Ser. B **36**, 192–236 (1974)

Besag, J.: Statistical analysis of non-lattice data. Statistician **24**, 179–195 (1975)

Besag, J.: On the statistical analysis of dirty pictures (with discussion). J. R. Stat. Soc. Ser. B **48**, 259–302 (1986)

Besag, J., York, J., Mollie, A.: Bayesian image restoration, with two applications in spatial statistics (with discussion). Ann. Inst. Stat. Math. **43**, 1–59 (1991)

Besag, J., Higdon, D.: Bayesian analysis of agricultural field experiments. J. R. Stat. Soc. Ser. B **61**, 691–746 (1999)

Buck, C.E., Cavanagh, W.G., Litton, C.D.: The spatial analysis of site phosphate data. In: Rhatz, S.P.Q. (ed.) Computer Applications and Quantitive Methods in Archeology. British Archaeological Reports, International Series, vol. 446. BAR, Oxford (1988)

Corduneanu, A., Bishop, C.M.: Variational Bayesian model selection for mixture distributions. In: Jaakkola, T., Richardson, T. (eds.) Artificial Intelligence and Statistics, pp. 27–34. Morgan Kaufmann, San Mateo (2001)

Friel, N., Pettitt, A.N., Reeves, R., Wit, E.: Bayesian inference in hidden Markov random fields for binary data defined on large lattices (2008, submitted)

Gelman, A., Meng, X.L.: Simulating normalizing constants: from importance sampling to bridge sampling to path sampling. Stat. Sci. **13**, 163–185 (1998)

Geman, S., Geman, D.: Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. IEEE Trans. Pattern Anal. Mach. Intell. **6**, 721–741 (1984)

Gottardo, R., Besag, J., Stephens, M., Murua, A.: Probabilistic segmentation and intensity estimation for microarray images. Biostatistics **7**, 85–89 (2006)

Green, P.J., Richardson, S.: Hidden Markov models and disease mapping. J. Am. Stat. Assoc. **97**, 1055–1070 (2002)

Jordan, M.: Graphical models. Stat. Sci. **19**, 140–155 (2004)

MacKay, D.J.C.: Ensemble learning for hidden Markov models. Technical Report, Cavendish Laboratory, University of Cambridge (1997)

McGrory, C.A.: Variational Approximations in Bayesian Model Selection. Ph.D. Thesis, University of Glasgow, UK (2005)

McGrory, C.A., Titterington, D.M.: Variational approximations in Bayesian model selection for finite mixture distributions. Comput. Stat. Data Anal. **51**, 5352–5367 (2007)

McGrory, C.A., Titterington, D.M.: Bayesian analysis of hidden Markov models using variational approximations. Aust. N. Z. J. Stat. (2008, to appear)

Møller, J., Pettitt, A.N., Reeves, R., Berthelsen, K.K.: An efficient Markov chain Monte Carlo method for distributions with intractable normalising constants. Biometrika **93**, 451–458 (2006)

Murray, I., Ghahramani, Z., MacKay, D.J.C.: MCMC for doubly-intractable distributions. In: Proceedings of 22nd Conference on Uncertainty in Artificial Intelligence (2006)

Newman, M.E.J., Barkema, G.T.: Monte Carlo Methods in Statistical Physics. Oxford University Press, London (1999)

Pettitt, A.N., Friel, N., Reeves, R.: Efficient calculation of the normalising constant of the autologistic and related models on the cylinder and lattice. J. R. Stat. Soc. Ser. B **65**, 235–247 (2003)

Reeves, R., Pettitt, A.N.: Efficient recursions for general factorisable models. Biometrika **91**, 751–757 (2004)

Rydén, T., Titterington, D.M.: Computational Bayesian analysis of hidden Markov models. J. Comput. Graph. Stat. **7**, 194–211 (1998)

Smith, M., Fahrmeir, L.: Spatial Bayesian variable selection with application to functional magnetic resonance imaging. J. Am. Stat. Assoc. **102**, 417–431 (2007)