Variational Bayesian Image Modelling

Li Cheng LICHENG@CS.UALBERTA.CA
Department of Computing Science, University of Alberta, Canada

Feng Jiao FJIAO@UWATERLOO.CA
School of Computer Science, University of Waterloo, Canada

Dale Schuurmans DALE@CS.UALBERTA.CA
Shaojun Wang SWANG@CS.UALBERTA.CA
Department of Computing Science, University of Alberta, Canada

Abstract

We present a variational Bayesian framework for performing inference, density estimation and model selection in a special class of graphical models—Hidden Markov Random Fields (HMRFs). HMRFs are particularly well suited to image modelling and in this paper, we apply them to the problem of image segmentation. Unfortunately, HMRFs are notoriously hard to train and use because the exact inference problems they create are intractable. Our main contribution is to introduce an efficient variational approach for performing approximate inference of the Bayesian formulation of HMRFs, which we can then apply to the density estimation and model selection problems that arise when learning image models from data. With this variational approach, we can conveniently tackle the problem of image segmentation. We present experimental results which show that our technique outperforms recent HMRF-based segmentation methods on real world images.

1. Introduction

A number of variational algorithms have been developed for performing density estimation and model selection in complex graphical models with hidden variables (Jordan et al., 1999; Attias, 2000; Beal & Ghahramani, 2003). However, these techniques do not exploit all of the structure available in real world models; for example, the structure presented in image data where latent (pixel) variables exhibit strong spatial correlations (Besag, 1986).

Hidden Markov Random Fields (HMRF) are a particularly natural model to apply in domains with numerous, correlated hidden variables of this form, and have been extensively applied in areas such as computational vision (Forbes & Peyrard, 2003; Heitz & Bouthemy, 1993). However, because of the large number of hidden variables and their complex graphical structure, density estimation in these models is computationally hard. The problem is made even harder by the fact that there is an intrinsic "model selection" problem: one also needs to determine how many values (i.e. components) each hidden variable can take on.

Early work on learning HMRFs attempted to use EM for density estimation, but assumed the number of components for each variable was known *a priori* (Zhang, 1992)—hence avoiding the model selection problem altogether. However, because inference in these models is intractable, approximation strategies still had to be devised to implement EM. The main source of difficulty in HMRF inference, as we will see, is the need to compute the normalization constant (the "partition function"). In Zhang (1992), the author employs a simple mean field approximation to achieve tractability. Since then, most authors have adopted a mean field approximation for the inference step in EM. However, a significant amount of effort has been recently devoted to more effectively approximating inference in an HMRF; for example by using loopy belief propagation and convex optimization methods (Yedidia et al., 2003; Opper & Saad, 2001; Wainwright & Jordan, 2003). None of this work however addresses the model selection problem.

Since Zhang (1992), some work has addressed the model selection problem for HMRFs. Cross validation was first investigated in Zhang (1993). More recently, authors have been exploring techniques for explicitly approximating a Bayesian posterior. For example, Stanford and Raftery

Appearing in *Proceedings of the 22^{nd} International Conference on Machine Learning*, Bonn, Germany, 2005.

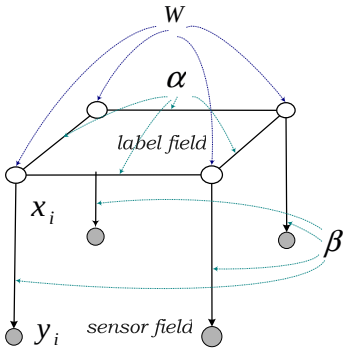

Figure 1. An exemplar 2×2 HMRF model. The white nodes are the latent variables x, and the grey nodes are the observed variables y. $\{\alpha, W, \beta\}$ denote the parameters.

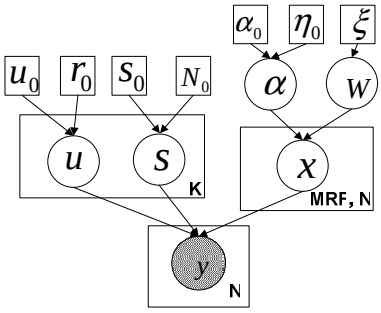

Figure 2. The graphical model view of the proposed the hierarchical Bayesian HMRF model. Fig.1 illustrates how the variables $x = x_i$ and $y = y_i$ are statistically correlated in this model. Notice here $\beta_k = (\mu_k, s_k)$ define a Gaussian distribution for feature $\phi(x_i, y_i)$. See text for details.

(2002) introduced the PLIC criterion for model selection, based on making independence assumptions to render computation of the partition function tractable. More recently a technique based on approximating the Bayesian Information Criterion (BIC) has been proposed by Forbes and Peyrard (2003), referred to as BIC^{GBF}, which we consider below. Although these techniques are increasingly effective at choosing the right number of components, they still retain heuristic elements. For example, none of these existing methods can provide any distributional information about the predicted parameters of an HMRF.

In this paper, we propose a variational Bayesian approach for image modelling that solves the three problems—inference, density estimation, and model selection—within a unified and principled framework. To demonstrate the practical utility of HMRFs and our approximation technique, we conduct a set of experiments on unsupervised image segmentation and obtain favorable results against current methods, such as BIC^{GBF}.

2. The HMRF model

An HMRF is a graphical model where the random variables are partitioned into an observed set $Y = \{y_i, i \in V\}$ and an unobserved set $X = \{x_i, i \in V\}$, such that each observed variable y_i is connected only to a corresponding hidden variable x_i; there are no direct links between observed variables; and hidden variables can be directly linked to each other, usually in a regular spatial or temporal pattern. In this paper, we concentrate on the nearest-neighbor MRFs which are particularly suitable for vision applications (Besag, 1986). We call X the label field which defines a MRF, and Y the sensor field. The model also possesses certain parameters. For example, Fig. 1 illustrates a small HMRF model with 2×2 hidden and visible variables.

More formally, an HMRF is specified by a graph $G = (V, E)$, where V denotes the set of nodes and E denotes the set of edges. For each index i, the hidden variable x_i, which takes value $k \in \{1, ..., K\}$, is linked with the visible variable $y_i \in \mathbf{R}$ by some edge $\hat{e}$. Thus, we let $E = \{\mathcal{E}, \hat{\mathcal{E}}\}$ where $\hat{\mathcal{E}}$ is the set of edges between X and Y, and $\mathcal{E}$ is the set of edges within X. Let the edge $e(i, j) \in \mathcal{E}$ connect a pair of adjacent hidden nodes, and the edge $\hat{e}(i, i) \in \hat{\mathcal{E}}$ link the hidden node x_i and the corresponding observation y_i. Let N be the number of nodes in X (and Y), and let x_e denote the local configuration (x_i, x_j) over an edge $e(i, j) \in \mathcal{E}$. Associated with any edge e connecting two neighbouring nodes (x_i, x_j), there is the potential function represented as $\phi(x_e)$. Similarly, for any (x_i, y_i) pair there is an associated potential function represented as $\phi(x_i, y_i)$. In this paper, we assume Potts model (Besag, 1986) for feature function $\phi(x_e) = \delta(x_i \neq x_j), e = (i \sim j)$ associated with parameter α, and Gaussian distribution $\beta_{i,k} = \beta_k$ for feature $\phi(x_i, y_i)$ with x_i taking value k.

To determine the probability distribution specified by an HMRF, we first need a "model", m, which specifies, besides the priors, how many values each hidden variable $x_i \in X$ can take (assume K), and hence how many parameters are associated with each local potential. For a given model m, let $\theta = (\alpha, \beta, W)$ denote the set of parameters specifying the local potentials, where $\alpha \in \mathbb{R}$ for the Potts model, $\beta = \{\beta_k\}_{k=1}^K$, and $W = \{w_k\}_{k=1}^K$. A model m and parameters θ define a probability distribution over X, Y by:

$$p(X, Y|m, \theta) = p(Y|X, \beta)p(X|W, \alpha)$$

with the components defined as follows. First, the sensor conditional likelihood is given by:

$$\begin{aligned} p(Y|X, \beta) &= \prod_{i=1}^{N} p(y_i|x_i, \beta_{x_i}), \\ p(y_i|x_i = k, \beta_k) &\sim \mathcal{N}(y_i; \mu_k, s_k^{-1}). \end{aligned}$$

hence $\beta_k = (\mu_k, s_k)$ is associated with $\hat{\mathcal{E}}$, as mean μ_k and precision s_k. Second, denote $W = \{w_k\}_{k=1}^K$ as the mixing

weights for each node x_i. Also, the distribution over X is given by a Markov random field (MRF), for which we adopt the K-class Potts model. The Gibbs function that respects the Hammersley-Clifford theorem (Jordan et al., 1999), is given by:

$$p(X|W,\alpha) = \exp\{\sum_i \log w_{i,k} + \alpha \sum_e \phi(x_e) - Z(\alpha, W)\},$$

$$Z(\alpha, W) = \log \sum_X \exp\{\sum_i \log w_{i,k} + \alpha \sum_e \phi(x_e)\} \quad (1)$$

where α is associated with $\mathcal{E}$, and $w_{i,k} = w_k$ for node x_i taking value k.

In the HMRF model, β controls the contributions from sensors; α determines the interaction strength among neighbor nodes in the label field X; and W takes care of the distribution of nodes $\{x_i\}$, assigning values in $\{1, \cdots, K\}$, by the global constraint $\sum_k w_k = 1$. Our goals thus turn out to be: (i) decide the model m. That is, decide the number of components (classes) K which also decides the dimensionality of the parameters (θ); (ii) estimate the parameters θ; and (iii) infer the X configuration.

2.1. A Hierarchical Bayesian framework

Following the Bayesian methodology, consider a specific HMRF model $m \in \mathcal{M}$ where $\mathcal{M} = \{1, \cdots, M\}$ denotes the model space. The joint density function for m is:

$$p(Y, X, \theta|m) = p(Y|X, \beta)p(X|W, \alpha)p(\theta|m).$$

where $p(\theta|m)$ contains the set of hyper-priors for the parameters θ, as:

$$p(\theta|m) = \prod_{k=1}^{K} p(\mu_k|\mu_0, \gamma_0) \prod_{k=1}^{K} p(s_k|s_0, N_0) p(\alpha|\alpha_0, \eta_0) p(W|\xi).$$

By taking conjugate priors, the priors for μ, s are Gaussian and Gamma distributions respectively, as:

$$p(\mu|\mu_0, \gamma_0) \sim \mathcal{N}(\mu; \mu_0, \gamma_0^{-1}),$$

$$p(s|s_0, N_0) \sim \Gamma(s; s_0, N_0).$$

Similarly, the priors on the mixing weights W and Potts weight α are Gaussian and Dirichlet, as:

$$p(\alpha|\alpha_0, \eta) \sim \mathcal{N}(\alpha; \alpha_0, \eta^{-1}),$$

$$p(W|\xi) \sim \mathcal{D}(W; \xi_1, ..., \xi_K).$$

Fig.2 graphically illustrates the hierarchical Bayesian formulation. Here all unknown quantities (parameter nodes, latent variable nodes) are treated as random variables and are denoted as white nodes. In contrast, observed data are denoted as grey nodes. Round nodes represent latent variables and square nodes represent pre-fixed prior variables which are either determined empirically from the data or set as uninformative priors (O'Ruanaidh & Fitzgerald, 1996). The plates with label "K" and "N" denote K and N *iid* copies of such variables. The plate with label "MRF, N" denotes the latent N copies of variables that form a MRF.

2.2. Related Models

The proposed HMRF model has close connections with several existing graphical models, such as Boltzmann machines (BMs) and hidden Markov models (HMMs). When we omit the mixing weight W in the fixed model m, the HMRF can be regarded as a variant of the BM (Ackley et al., 1985) with both latent and observed nodes. Given a graphical model over variables $\overline{x}$, the Boltzmann Machine can be defined as

$$p(\overline{x}|\Lambda) = \exp\Big\{ \sum_{i \neq j} \Lambda_{i,j} \overline{x}_i \overline{x}_j - Z \Big\}$$

where Λ is a symmetric matrix with zero-diagonal elements, and Z is the log partition function. The HMRF model could thus be modelled as a Boltzmann Machine with a sparse and highly structured $2N \times 2N$ Λ matrix.

The HMRF can also be viewed as a 2D generalization of HMM (Jordan et al., 1999), where the HMRF is defined over a general nearest-neighbor graph rather than a 1D sequence for HMM. This difference, however, leads to the intractability of inference in HMRFs due to the global partition function $Z(\alpha, W)$.

3. Variational approximation

The hierarchical Bayesian framework poses a significant challenge for optimization. There are, in general, two approaches for coping this problem: one is to use Monte Carlo sampling techniques; the other, as we adopt here, is to employ deterministic approximation methods. In addition to a speed-up, the adoption of conjugate priors in the modelling phase allows analytical solutions of integrals to be computed practically. In this paper, we apply the variational approximation method, and call this formulation variational Bayesian HMRF (VB-HMRF) for convenience.

Also, for ease of notation, we denote three log-evidence functions $L(\theta; Y) = \log p(Y|\theta)$, $L(m; Y) = \log p(Y|m)$, and $L(Y) = \log p(Y)$. Following MacKay (1991), we can write, by applying Jensen's Inequality (Jordan et al., 1999),

$$\begin{aligned} L(\theta; Y) &\geq E_{q(X)}\left[\log p(Y|X, \beta)\right] - KL(q(X)||p(X)) \\ &= \mathcal{L}(q(X), \theta), \end{aligned} \quad (2)$$

$$\begin{aligned} L(m; Y) &\geq E_{q(\theta)}\left[L(\theta; Y)\right] - KL(q(\theta)||p(\theta)) \\ &\geq \mathcal{L}(q(X), q(\theta)), \end{aligned} \quad (3)$$

$$\begin{aligned} L(Y) &\geq E_{q(m)}\left[L(m;Y)\right] - KL(q(m)||p(m)) \\ &\geq \mathcal{L}(q(m)). \end{aligned} \tag{4}$$

where $KL(\cdot||\cdot)$ denotes the relative entropy; $\mathcal{L}(q(X),\theta)$, $\mathcal{L}(q(X),q(\theta))$ and $\mathcal{L}(q(m))$ are the lower bounds of the respective log evidence functions; and $q(X)$, $q(\theta)$ are the approximating distribution functions. Interestingly, the second inequality of Eq.(3) is obtained by substituting the inequality in Eq.(2) of $L(\theta;Y)$ into the first inequality; similarly, we derive the second inequality of Eq.(4). These lower bounds derivations, although not necessarily concave, provides intuitions for iteratively ascending schemes. It is known (Attias, 2000; Neal & Hinton, 1998) that the EM algorithm maximizes the first lower bound $\mathcal{L}(q(X),\theta)$ and is equivalent to the ML estimate. The variational Bayes approach, as a generalization of ML, maximizes the second lower bound $\mathcal{L}(q(X),q(\theta))$ and outputs the prior-corrected posterior density estimates of X and θ (MacKay, 1991). Herein, we mainly focus on the second (and refer to it as short-handed $\mathcal{L}$) and third lower bounds in the latter derivations. The technical details of deriving the set of update equations are omitted due to space constraints. The main idea is that, instead of computing the true probability distributions ($p(X|Y,\theta,m)$ and $p(\theta|X,Y,\mathcal{M})$) of a given HMRF,[1] one can approximate them by maximizing $\mathcal{L}(q(X),q(\theta))$, where the approximate distributions $q(X)$ and $q(\theta)$ can be chosen to alleviate the computation burden.

To simplify the notation we use the following short-cuts:

$$q(x_i) = \sum_{X_{.\setminus i}} q(X), \qquad q(x_e) = \sum_{X_{.\setminus (i,j)}} q(X)$$

$$q_{ik} = q(x_i = k), \qquad N_k = \sum_i q_{ik}$$

$$\overline{y_k} = \frac{1}{N_k}\sum_i q_{ik} y_i, \qquad \overline{y_k^2} = \frac{1}{N_k}\sum_i q_{ik} y_i^2$$

$$\tilde{\mu_k} = \frac{s_0}{\tilde{s_k}}\mu_0 + \frac{N_k E_q[s_k]}{\tilde{s_k}}\overline{y_k}, \qquad \tilde{s_k} = s_0 + N_k E_q[s_k]$$

$$\hat{N_k} = N_0 + \frac{N_k}{2}, \qquad \hat{s_k}^{-1} = s_0^{-1} + \frac{N_k}{2}\overline{s_{\mu_k}}^{-1}$$

$$\overline{s_{\mu_k}}^{-1} = \overline{y_k^2} - 2\overline{y_k}u_k + \frac{1}{N_k}(\tilde{\mu_k}^2 + \tilde{s_k}^{-1})$$

$$\log\overline{\omega}_k = \int_{\omega_k} \log\omega_k d\omega_k = \Psi(\xi_k + N_k) - \Psi(\textstyle\sum_k \xi_k + N_k)$$

$$\log\overline{s}_k = \int_{s_k} \log s_k ds_k = \Psi(\hat{N_k}) + \log\hat{S_k}$$

$$E_q[S_k] = \int_{s_k} q(s_k) s_k ds_k = \hat{N_k}\hat{S_k}.$$

Here $\Psi(\cdot)$ denotes the digamma function, and $\cdot\setminus i$ represents the entire set except the ith element.

[1] They turn out to be computationally infeasible due to the existence of integrals and the partition function $Z(\alpha, W)$.

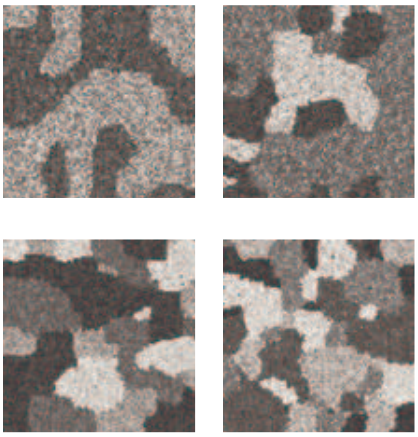

Figure 3. Exemplar synthetic images. In scanline order, a) is a 2-class image with $\alpha = 1.5$ and $\mu = \{80, 60\}$ and all $\sigma = 60$; b) is a 3-class image with $\alpha = 1.5$ and $\mu = \{60, 120, 200\}$ and all $\sigma = 40$; c) is a 4-class image with $\alpha = 1.4$ and $\mu = \{50, 105, 160, 210\}$ and all $\sigma = 30$; d) is a 5-class image with $\alpha = 1.3$ and $\mu = \{40, 85, 130, 175, 215\}$ and all $\sigma = 25$.

Inference By taking the functional derivative of Eq.(3) and equating to zero $\delta\mathcal{L}/\delta q(X) = 0$, we get the update:

$$\log q(X) = E_{q(\beta)}\left[\sum_i \log p(y_i|x_i,\beta_{i,k})\right] + E_{q(W)}\left[\sum_i \log w_{i,k}\right] + E_{q(\alpha)}\left[\alpha \sum_{i\sim j} e(x_i, x_j)\right] + const.$$

Parameter Estimation We evaluate $q(\theta)$ by $\delta\mathcal{L}/\delta q(\theta) = 0$, and get the following update equations

$$q(\mu_k) \sim \mathcal{N}(\mu_k; \tilde{\mu_k}, \tilde{s_k}^{-1}), \tag{5}$$

$$q(s_k) \sim \Gamma(s_k; \hat{s_k}, \hat{N_k}), \tag{6}$$

$$q(\omega_k) \sim \mathcal{D}(\sum_i q_{ik} + \xi_k), \tag{7}$$

$$q(\alpha) \propto \exp\left\{\alpha \sum_X q(X) \sum_e \phi(x_e) - \frac{1}{2}(\alpha - \alpha_0)^2\eta\right\}, \tag{8}$$

where $\tilde{\mu_k}$ and $\tilde{s_k}$ denote the prior-corrected mean and precision parameters for the Gaussian, and $\hat{s_k}$ and $\hat{N_k}$ denote the scale and shape parameters for the gamma distribution.

Model Selection The VB framework is known (Attias, 2000) to be equivalent to the BIC framework as the sample size goes to infinity, and the posterior density $p(\theta)$ peaks around the ML estimation θ_{ML}:

$$BIC(m) = L(\theta_{ML}; Y) - \frac{|\theta|}{2}\log N \tag{9}$$

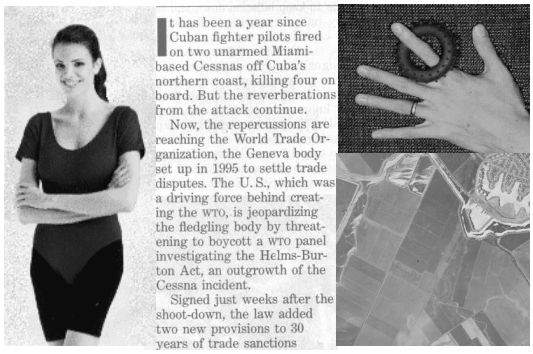

Figure 4. Some exemplar real world images.

where $|\theta|$ denotes the size of the parameter space.

If we evaluate at $\delta\mathcal{L}(q(m))/\delta q(m) = 0$, and assume uniform distribution over $\mathcal{M}$, we get:

$$q(m) = \frac{\exp\{\mathcal{L}_m\}}{\sum_{m'} \exp\{\mathcal{L}_{m'}\}} \tag{10}$$

where $\mathcal{L}_m$ denotes $\mathcal{L}(q(X), q(\theta))$ for model m.

3.1. The Mean Field Approximation

Thus far, we have been executing an exact inference step. In practice, the computational complexity of evaluating $q(X)$ grows exponentially with the number of nodes in graph G. Therefore, we must resort to approximation methods which include, for example, the family of mean field algorithms. Given a graph G, we define a subgraph $G_{\mathcal{H}} = (V, E_{\mathcal{H}})$, where $E_{\mathcal{H}} \subseteq E$. The essence of the mean field algorithms is as follows: instead of computing the exact probability $q(X)$, a factorability assumption is imposed by deliberately removing some edges from the original graph G.[2] Thus, one approximates $q(X)$ by $q_{\mathcal{H}}(X) = \prod_{h\in\mathcal{H}} q_h(x_h)$, resulting in the variational E step:

$$q_{\mathcal{H}}(X) \quad = \quad \arg\max_{q_{\mathcal{H}}(X)} \mathcal{L}(q_{\mathcal{H}}(X), q(\theta)). \tag{11}$$

Naive Mean Field Algorithm Consider the simplest case when all edges are removed from G, forming a subgraph $G_{\mathcal{H}_0} = (V, E_{\mathcal{H}_0})$ with $E_{\mathcal{H}_0} = \varnothing$ all nodes $\{x_i \; i \in V\}$ within X are independent of each other. This is the Naive Mean Field (NMF) setting, with the probability function being factorized as $q(X) = \prod_{i\in V} q(x_i)$.

In this NMF approximation, since $E_{\mathcal{H}_0} = \varnothing$, for any node i and its neighbouring node j,[3] the correlation term $x_i x_j$ is replaced by $x_i\mu_j$ where μ_j is the estimated mean value of x_j. Thus we compute in the E step:

$$q_{ik} = \overline{\omega}_k \overline{s}_k^{\frac{1}{2}} \exp\{-\frac{1}{2} E_q[S_k]((y_i - \tilde{\mu_k})^2 + \tilde{S_k}^{-1}) \tag{12}$$

$$+ \frac{1}{2} E_{q(\alpha)}[\alpha] \sum_{j\in\partial i} e(x_i = k, \mu_j)\}, \tag{13}$$

where $\log\overline{\omega}_k$, $\log\overline{s}_k$ and $E_q[S_k]$ denote the weight, precision and expected precision estimates, respectively.

In the same manner, $q(\alpha)$ is given as:

$$q(\alpha) \propto \exp\left\{\frac{1}{2}\alpha \sum_i \sum_{j\in\partial i} \mu_i\mu_j - \frac{1}{2}(\alpha - \alpha_0)^2\eta\right\} \tag{14}$$

4. Experiments

The BIC^{GBF} Method We conduct experiments on unsupervised segmentation of both synthetic images and real world images, and compare the results with the BIC^{GBF} method. The BIC^{GBF} method is a first-order approximation of Eq.(9) on HMRF provided simulated field approximation[4]. However for the sake of simplicity, here we solely consider the naive mean field approximation. Define the log-partition function as:

$$Z(\theta, Y) = \tag{15}$$
$$\log\sum_X \exp\left\{\sum_i \log w_{i,k} + \alpha\sum_e \phi(x_e) + \log p(Y|X,\beta)\right\}$$

It has been shown that both partition functions in Eq.(1) and Eq.(15) can be approximated via the GBF bound (Forbes & Peyrard, 2003), provided the mean field approximation. For example,

$$Z(\alpha, W) \approx Z^{mf}(\alpha, W) + \triangle Z^{mf}(\alpha, W) \doteq Z^{GBF}(\alpha, W)$$

Since we can further write:

$$L(\theta; Y) = Z(\theta; Y) - Z(\alpha, W)$$

we can plug these approximations into the BIC criteria in Eq.(9), to arrive at Eq.(23) of (Forbes & Peyrard, 2003):

$$BIC^{GBF} = Z^{GBF}(\theta; Y) - Z^{GBF}(\alpha, W) - \frac{|\theta|}{2}\log N.$$

Experiments On Synthetic Images We first compare the BIC^{GBF} method with the proposed VB-HMRF method on synthetic images (Forbes & Peyrard, 2003) sampled

[2] Here we only consider the removal of the edges within X.

[3] j is the neighbouring node of i, when $(i, j) \in E$. Sometimes we also use the notation $j \in \partial i$.

[4] The simulated field approximation (Forbes & Peyrard, 2003) is an simulated annealing variant of the naive mean field algorithm, which samples the conditional distribution with a Gibbs sampler. It was shown to have more accurate results than the naive mean field algorithm.

$K=2, \alpha=1.5$	
Estimated BIC^{GBF}	$K=2$
	50
VB-HMRF	50

$K=3, \alpha=1.5$			
Estimated BIC^{GBF}	$K<3$	$K=3$	$K>3$
	0	49	1
VB-HMRF	2	47	1

Table 1. Comparison of BIC^{GBF} and VB-HMRF on datasets of fifty synthetic images of 2-class (3-class).

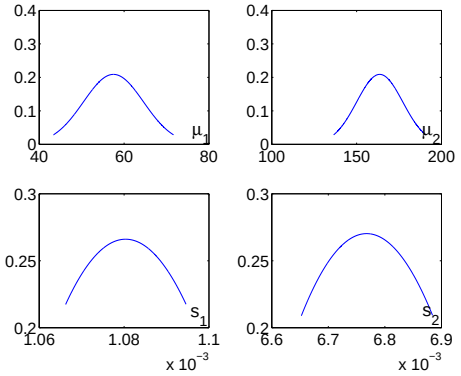

Figure 5. The predictive posterior distributions of means (μ_1, μ_2) and precisions (s_1, s_2) of the hand image, by applying the VB-HMRF method, and the predicted expectation of α value is 0.8. The model dimension is $K=2$ as in Fig.7.

from a known model (class number) and parameters (θ) using a Gibbs sampler. Fig.3 shows some representative images from this dataset. Due to space constraints, we only compare the model selection results here.

Table 1 considers the model estimation characteristic on datasets of 2-class or 3-class synthetic images (each dataset contains fifty images sampled from fixed model and parameters). The BIC^{GBF} method performs quite well on these synthetic datasets. The VB-HMRF method performs generally on par with BIC^{GBF}. However, since we essentially consider a simpler image model with fixed parameters, the synthesized images tend to be noisier than the real ones, this results in the slightly inferior performance of VB-HMRF. In these experiments, the VB-HMRF method favors simpler models (smaller K values which incurs larger regions for fixed image.), while the BIC^{GBF} method doesn't have such tendency.

Experiments On Real World Images We have observed that the BIC^{GBF} method and the proposed VB-HMRF method give similar results on synthetic images. However, while dealing with real world images (Some exemplar images are shown in Fig.4), the VB-HMRF delivers more reasonable results than the BIC^{GBF} method.

μ	32.3	66.4	120.5	159.2	174.0	178.9
s	0.008	0.005	0.003	0.078	1.294	0.024

Table 2. The mean/variance parameters estimate of applying the BIC^{GBF} method to the hand image, with the model is picked $K=6$ in Fig.7.

Fig.6 and Fig.7 present experimental results on document and hand images, respectively. In the first row of each figure, the left side shows the BIC^{GBF} values as a function of K, the feasible models, while the right side shows the predicted distribution for the model space where $K \in \{2 \dots 6\}$. The second row presents the inference results for the BIC^{GBF} and the VB-HMRF methods, respectively, after picking the model with the highest BIC^{GBF} values or the MAP model. Note that the presented segmentation results are the MAP pixel classification, with different colors representing different classes in the models. In the third row, the density estimations of both methods are plotting against the image histogram. For the remaining woman and remote sensing images in Fig.4, the VB-HMRF gives 3 classes for both images, while the BIC^{GBF} reports 4 classes for the woman image and 5 classes for the remote sensing image.

In these experiments, the estimated parameters of BIC^{GBF} are point estimates in the parameter space, while the counterparts of VB-HMRF are predicted posterior distributions. As an example, Table 2 shows the resultant mean-variance parameters point estimates for the hand image, from application of the BIC^{GBF} method; Fig.5 shows the corresponding predictive distributions estimate with the proposed VB-HMRF. Notice that the inference [5] results of VB-HMRF method tend to be smoother compared to the BIC^{GBF} method. In addition, the VB-HMRF method favors the model with fewer components, compared to the BIC^{GBF} method. In particular, the proposed approach cleanly segments the textured background of the hand image, while the BIC^{GBF} produces noisy results, when the raw colors are used in both cases. By adopting more salient features (such as wavelet coefficients) with the proposed approach, we believe that the two rings on the fingers can be further preserved.

Implementation Details As shown in Fig.2, several priors have to be set in the implementation in line with the empirical Bayesian methods (O'Ruanaidh & Fitzgerald, 1996); We adopt the same settings throughout the experiments presented in this paper, as follows. For the normal prior of the mean μ, the mean μ_0 is set as the mean of the image data, and the precision γ_0 is set to a real value that is closely related to the sensor variance, here we use

[5] Inference in image segmentation domain is the pixel classification problem.

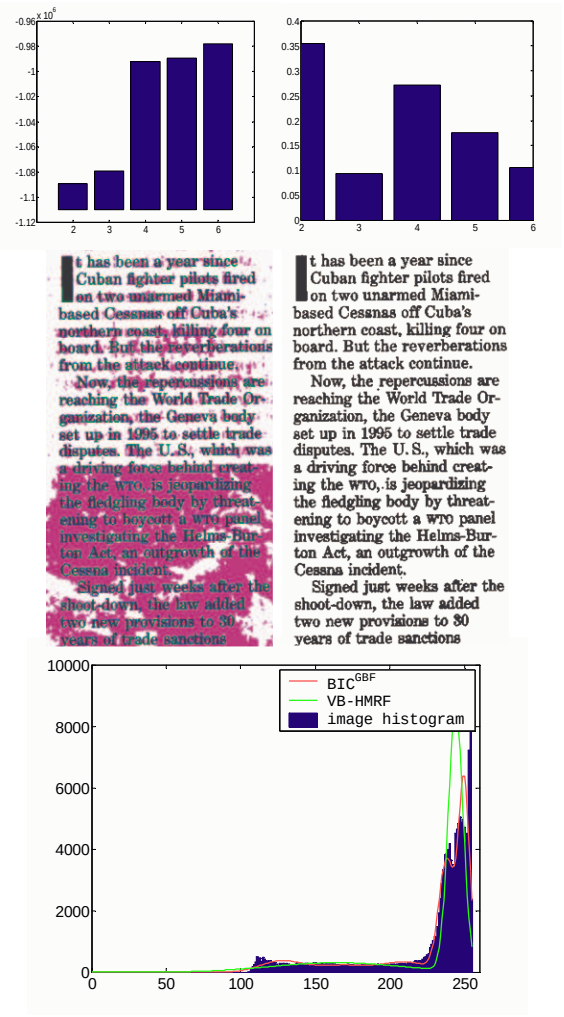

Figure 6. Top-left shows the BIC^{GBF} values as function of model space ($K \in \{2...6\}$), and middle-left is the resultant 6-class segmentation; Similarly, top-right shows the posterior distribution over the model space, and middle-right is the resultant 2-class segmentation, by applying the VB-HMRF method. In the bottom panel, the blue spikes show the image histogram; the red curve is the BIC^{GBF} predicted distribution estimated from the parameters (θ) with $K = 6$; the green curve is the VB-HMRF predicted posterior distribution estimated from the parameters (θ) with $K = 2$.

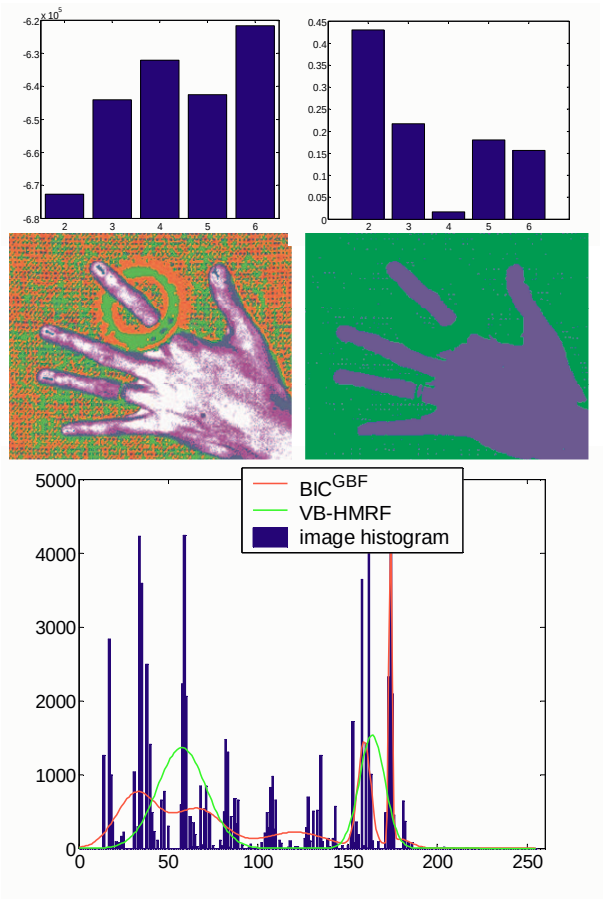

Figure 7. Top-left shows the BIC^{GBF} values as function of model space ($K \in \{2...6\}$), and middle-left is the resultant 6-class segmentation; Similarly, top-right shows the posterior distribution over the model space, and middle-right is the resultant 2-class segmentation, by applying the VB-HMRF method. In the bottom panel, the blue spikes show the image histogram; the red curve is the BIC^{GBF} predicted distribution estimated from the parameters (θ) with $K = 6$; the green curve is the VB-HMRF predicted posterior distribution estimated from the parameters (θ) with $K = 2$, and the predicted expectation of α value is 1.5.

$\gamma_0 = \frac{(\sup y - \inf y)^2}{4}$. For the gamma prior of the precision s, the scale s_0 is set to 10^4 and the shape parameter N_0 to 3×10^{-4}. For the normal prior of the Potts weight α, we set the mean α_0 to zero and the precision η_0 to $1/100$. Finally, the Dirichlet priors of the mixing weights W are all set as $\xi = 5$.

In the implementation, an iterative procedure is employed as follows:

1. Choose a model m.
   (a) Set $t = 0$. Initialize the latent variables shown in Fig.2 by a standard K-means algorithm.
   (b) iteration t: inference of the latent variables X, μ, s, α and W by Eq.(12),(5),(6),(7) and Eq.(14), respectively.
   (c) $t \leftarrow t + 1$, goto 2
2. Compute the lower bound $\mathcal{L}$ for the model m. Take the MAP of the $q(X)$ distribution as the pixel labelling of this model.

In practice, the inner loop iterates 60 times to guarantee the convergence. After exploring all models $m \in \mathcal{M}$, once we have the predictive model space distribution Eq.(10), we can decide the MAP model (eg. K value) accordingly. The implementation is a mixture of matlab and c codes. In the experiments, it takes less than one minute for processing a 100 by 100 image with two classes, for both BIC^{GBF} and VB-HMRF methods, on a Pentium 4 PC.

5. Conclusion

We presented a Bayesian framework for image modelling and applied it to the problem of unsupervised image segmentation. We obtained favorable results relative to recent HMRF-based segmentation methods based on BIC. The improvement is due to the fact that BIC, since it originates from large sample theory, relies on large sample size for asymptotically stable behavior. By contrast, due to the model averaging effect, the variational Bayesian approach has reasonable performance at small sample size, as in segmentation problems on limited size images.

The framework is quite generic and could be applied to a broad class of image modelling problems, including image content retrieval and video data segmentation. To improve the performance upon image segmentation problem, future work includes exploring task specific feature functions; incorporating more domain-specific knowledge into the priors for both the model and the parameter spaces, as for example, the data-driven approach of Tu and Zhu (2002).

Acknowledgments

Research supported by the Alberta Ingenuity Centre for Machine Learning, NSERC, MITACS, CRC and CFI.

References

Ackley, D., Hinton, G., & Sejnowski, T. (1985). A learning algorithm for Boltzmann machine. *Cognitive Science, 9*, 147–169.

Attias, H. (2000). A variational Bayesian framework for graphical models. *Advances in Neural Information Processing Systems 12* (pp. 209–215).

Beal, M., & Ghahramani, Z. (2003). The variational Bayesian EM algorithm for incomplete data: with application to scoring graphical model structures. *Bayesian Statistics 7* (pp. 453–464).

Besag, J. (1986). On the statistical analysis of dirty pictures (with discussions). *Journal of the Royal Statistical Society, Series B, 48*, 259–302.

Forbes, F., & Peyrard, N. (2003). Hidden Markov random field model selection criteria based on mean field-like approximation. *IEEE Transaction on Pattern Recognition and Machine Intelligence, 25*, 1089–1101.

Heitz, F., & Bouthemy, P. (1993). Multimodal estimation of discontinuous optical flow using Markov random fields. *IEEE Transaction on Pattern Recognition and Machine Intelligence, 15*, 1217–1232.

Jordan, M., Ghahramani, Z., Jaakkola, T., & Saul, L. (1999). An introduction to variational methods for graphical models. *Machine Learning, 37*, 183–233.

MacKay, D. (1991). *Bayesian methods for adaptive models*. Doctoral dissertation, Computation and Neural Systems, California Institute of Technology.

Neal, R., & Hinton, G. (1998). A view of the EM algorithm that justifies incremental, sparse, and other variants. *Learning in Graphical Models*. Kluwer.

Opper, M., & Saad, D. (Eds.). (2001). *Advanced mean field methods: Theory and practice*. the MIT press.

O'Ruanaidh, J., & Fitzgerald, W. (1996). *Numerical Bayesian methods applied to signal processing*. Statistics and Computing. Springer.

Stanford, D., & Raftery, A. (2002). Approximate Bayes factors for image segmentation: the pseudolikelihood information criterion (plic). *IEEE Transaction on Pattern Analysis and Machine Intelligence, 24*, 1517–1520.

Tu, Z., & Zhu, S. (2002). Image segmentation by data-driven Markov chain Monte Carlo. *IEEE Trans on Pattern Analysis and Machine Intelligence, 24*, 657–673.

Wainwright, M. J., & Jordan, M. I. (2003). *Graphical models, exponential families, and variational inference* (Technical Report 649). UC Berkeley, Dept. of Statistics.

Yedidia, J., Freeman, W. T., & Weiss, Y. (2003). Understanding belief propagation and its generalizations. *Exploring Artificial Intelligence in the New Millennium* (pp. 239–269).

Zhang, J. (1992). The mean field theory in EM procedures for Markov random fields. *IEEE Transaction on Signal Processing, 40*, 2570–2583.

Zhang, P. (1993). Model selection via multifold cross validation. *Annals of Statistics, 21*, 299–313.