Variational Bayesian Image Modelling

Li Chen
Feng Jiao
Dale Schuurmans
Shaojun Wang
Wright State University - Main Campus, shaojun.wang@wright.edu

Follow this and additional works at: http://corescholar.libraries.wright.edu/knoesis
Part of the Bioinformatics Commons, Communication Technology and New Media Commons, Databases and Information Systems Commons, OS and Networks Commons, and the Science and Technology Studies Commons

Repository Citation

This Conference Proceeding is brought to you for free and open access by the The Ohio Center of Excellence in Knowledge-Enabled Computing (Kno.e.sis) at CORE Scholar. It has been accepted for inclusion in Kno.e.sis Publications by an authorized administrator of CORE Scholar. For more information, please contact corescholar@www.libraries.wright.edu.
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 22nd International Conference on Machine Learning, Bonn, Germany, 2005. Copyright 2005 by the author(s)/owner(s).
(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—in 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 $2 \times 2$ 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 $k \in \{1, \ldots, K\}$, is linked with the visible variable $y_i \in \mathbb{R}$ by some edge $\tilde{e}$. Thus, we let $E = \{\tilde{e}, \tilde{e}'\}$ where $\tilde{e}$ is the set of edges between $X$ and $Y$, and $\tilde{e}'$ is the set of edges within $X$. Let the edge $e(i, j) \in \tilde{e}$ connect a pair of adjacent hidden nodes, and the edge $\tilde{e}(i, j) \in \tilde{e}'$ 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 $x_i$ denote the local configuration $(x_i, x_j)$ over an edge $e(i, j) \in \tilde{e}$. 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 $\alpha$, 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 $\theta$ 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:

$$p(Y | X, \beta) = \prod_{i=1}^N p(y_i | x_i, \beta_{x_i}),$$

$$p(y_i | x_i = k, \beta_k) \sim \mathcal{N}(y_i; \mu_k, \nu_k^{-1}).$$

hence $\beta_k = (\mu_k, \nu_k)$ is associated with $\tilde{e}$, as mean $\mu_k$ and precision $\nu_k$. 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:

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

\[
Z(\alpha, W) = \log \sum_X \exp\left\{ \sum_i \log w_{i,k} + \alpha \sum_e \phi(x_e) \right\}
\]

where \( \alpha \) is associated with \( \varepsilon \), and \( w_{i,k} = w_k \) for node \( x_i \) taking value \( k \).

In the HMRF model, \( \beta \) controls the contributions from sensors; \( \alpha \) 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 \( \{1, \ldots, K\} \), 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 \( \theta \); (ii) estimate the parameters \( \theta \); and (iii) infer the \( X \) configuration.

### 2.1. A Hierarchical Bayesian framework

Following the Bayesian methodology, consider a specific HMRF model \( m \in M \) where \( M = \{1, \ldots, M\} \) 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 \( \theta \), 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 \( \mu, s \) 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 \( \alpha \) are Gaussian and Dirichlet, as:

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

\[
p(W|\xi) \sim D(W; \xi_1, \ldots, \xi_K).
\]

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 \( \pi \), the Boltzmann Machine can be defined as:

\[
p(\pi|\Lambda) = \exp\left\{ \sum_{i\neq j} \Lambda_{i,j}\pi_i\pi_j - Z \right\}
\]

where \( \Lambda \) 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 \( 2N \times 2N \) 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),

\[
L(\theta; Y) \geq E_{q(\theta)}[\log p(Y|X, \beta)] - KL(q(\theta)||p(\theta)) = \mathcal{L}(q(\theta), \theta),
\]

\[
L(m; Y) \geq E_{q(\theta)}[L(\theta; Y)] - KL(q(\theta)||p(\theta)) \geq \mathcal{L}(q(\theta), q(\theta)),
\]
\[ L(Y) \geq E_{q(m)} [L(m; Y)] - KL(q(m)||p(m)) \]
\[ \geq \mathcal{L}(q(m)). \]  

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 \( \mathcal{L}(\theta; Y) \) 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.

The VB framework 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 \( \theta \) (MacKay, 1991).

Here, 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 \( p(\theta|X, Y, M) \) of a given HRMF,\(^1\) 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:

\[ q(x_i) = \sum_{X \setminus i} q(X), \quad q(x_e) = \sum_{X \setminus \{i,e\}} q(X) \]
\[ q_{ik} = q(x_i = k), \quad N_k = \sum_i q_{ik} \]
\[ \overline{y}_k = \frac{1}{N_k} \sum_i q_{ik} y_i, \quad \overline{\overline{y}}_k = \frac{1}{N_k} \sum_i q_{ik} y_i^2 \]
\[ \hat{\mu}_k = \frac{s_0}{s_k} \mu_0 + \frac{N_k E[q(s_k)]}{s_k} \overline{y}_k, \quad \hat{s}_k = s_0 + N_k E[q(s_k)] \]
\[ \hat{\hat{N}}_k = \hat{N}_k + \frac{N_k}{2}, \quad \hat{\hat{s}}_k^{-1} = s_0^{-1} + \frac{N_k}{2 s_{\mu_k}} s_{\mu_k}^{-1} \]
\[ \log \gamma_k = \frac{1}{\hat{\hat{N}}_k} - 2 \frac{\hat{\hat{s}}_k}{\hat{\hat{s}}_k} \log \hat{\mu}_k + \frac{1}{\hat{\hat{N}}_k} (\hat{\mu}_k^2 + \hat{s}_k^{-1}) \]
\[ \log \varphi_k = \int_{\xi_k} \log \omega_k d\omega_k = \Psi(\xi_k + N_k) - \Psi(\sum_k \xi_k + N_k) \]
\[ \log \sigma_k = \int_{s_k} \log s_k ds_k = \Psi(\hat{\hat{N}}_k) + \log \hat{\hat{s}}_k \]
\[ E_q[S_k] = \int_{s_k} q(s_k) s_k ds_k = \hat{\hat{N}}_k \hat{\hat{s}}_k. \]

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

\(^1\)They turn out to be computationally infeasible due to the existence of integrals and the partition function \( Z(\alpha, W) \).
In this NMF approximation, since

\[
E_{\text{Naive Mean Field Algorithm}}
\]

Consider the simplest case when all edges are removed from

being factorized as

Mean Field (NMF) setting, with the probability function

replaced by

\[x_i\text{ and its neighbouring node}\]

exact probability

field algorithms is as follows: instead of computing the

algorithms. Given a graph

ods which include, for example, the family of mean field

agraph G, we define a subgraph \(G' = (V, E')\), where \(E' \subseteq E\). 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 \(G\). Thus, one approximates \(q(X)\) by \(q'(X) = \prod_{h \in H} q_h(x_h)\), resulting in the variational E step:

\[
q'(X) = \arg \max_{q(X)} L(q(X), q(\theta)).
\]

Naive Mean Field Algorithm Consider the simplest case when all edges are removed from \(G\), forming a subgraph \(G' = (V, E')\) with \(E' = \emptyset\) 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 NMRF approximation, since \(E' = \emptyset\), for any node \(i\) and its neighbouring node \(j\), the correlation term \(x_i x_j\) is replaced by \(x_i \hat{\mu}_j\) where \(\hat{\mu}_j\) is the estimated mean value of \(x_j\). Thus we compute in the E step:

\[
q_{ik} = \omega_k \hat{s}_k \exp\left\{-\frac{1}{2} E_{q_k}[S_k](y_i - \hat{\mu}_k)^2 + \hat{s}_k^{-1}\right\} \tag{12}
\]

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

where \(\log \omega_k\), \(\hat{s}_k\) and \(E_{q_k}[S_k]\) denote the weight, precision and expected precision estimates, respectively.

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

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

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. However for the sake of simplicity, here we solely consider the naive mean field approximation. Define the log-partition function as:

\[
Z(\theta; Y) = \frac{\prod_{x} \exp \left\{ \sum_{e} \log w_{e,k} + \alpha \sum_{e} \phi(x_e) + \log p(Y | X, \beta) \right\}}{\prod_{X} \exp \left\{ \sum_{e} \log w_{e,k} + \alpha \sum_{e} \phi(x_e) + \log p(Y | X, \beta) \right\}}
\]

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,

\[
Z(\alpha, W) \approx Z^{mf}(\alpha, W) + \triangle Z^{mf}(\alpha, W) = Z^{GBF}(\alpha, W)
\]

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

<table>
<thead>
<tr>
<th>$K$</th>
<th>$\alpha = 1.5$</th>
<th>$K = 2$</th>
<th>$K = 3$</th>
<th>$K &gt; 3$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Estimated $BIC^{GBF}$</td>
<td>50</td>
<td>49</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>VB-HMRF</td>
<td>50</td>
<td>47</td>
<td>1</td>
<td></td>
</tr>
</tbody>
</table>

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 $K \in \{2 \ldots 6\}$. 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 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 $\mu$, the mean $\mu_0$ is set as the mean of the image data, and the precision $\gamma_0$ is set to a real value that is closely related to the sensor variance, here we use

\[ \gamma_0 = \frac{1}{\text{sensor variance}} \]

\[ \mu_0 = \text{mean of image data} \]

from a known model (class number) and parameters ($\theta$) 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.

The prediction posterior distributions of means ($\mu_1, \mu_2$) and precisions ($s_1, s_2$) of the hand image, by applying the VB-HMRF method, and the predicted expectation of $\alpha$ value is 0.8. The model dimension is $K = 2$ as in Fig.7.

\[ \mu \quad 32.3 \quad 66.4 \quad 120.5 \quad 159.2 \quad 174.0 \quad 178.9 \]

\[ s \quad 0.008 \quad 0.005 \quad 0.003 \quad 0.078 \quad 1.294 \quad 0.024 \]

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.

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.
Figure 6. Top-left shows the $BIC^{GBF}$ values as function of model space ($K \in \{2, \ldots, 6\}$), 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 ($\theta$) with $K = 6$; the green curve is the VB-HMRF predicted posterior distribution estimated from the parameters ($\theta$) with $K = 2$.

Figure 7. Top-left shows the $BIC^{GBF}$ values as function of model space ($K \in \{2, \ldots, 6\}$), 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 ($\theta$) with $K = 6$; the green curve is the VB-HMRF predicted posterior distribution estimated from the parameters ($\theta$) with $K = 2$, and the predicted expectation of $\alpha$ value is 1.5.
Variational Bayesian Image Modelling

\[ \gamma_0 = \frac{\left( \sup_y - \inf_y \right)^2}{4}. \] For the gamma prior of the precision \( s \), the scale \( s_0 \) is set to \( 10^4 \) and the shape parameter \( N_0 \) to \( 3 \times 10^{-3} \). For the normal prior of the Potts weight \( \alpha \), we set the mean \( \alpha_0 \) to zero and the precision \( \gamma_0 \) 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 \( X, \mu, s, \alpha \) and \( W \) by Eq.(12),(5),(6),(7) and Eq.(14), respectively.
   - (c) \( t \leftarrow t + 1 \), goto 2

2. Compute the lower bound \( L \) 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 \( m \in M \), 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


