

Bounded Approximations for Marginal Likelihoods

Chunlin Ji^a, Haige Shen^b, Mike West^c

^a*Department of Statistics, Harvard University, Science Center, 1 Oxford Street,
Cambridge, MA 02138-2901, USA*

^b*Novartis Oncology, Biometrics, 180 Park Avenue,
Florham Park, NJ 07932, USA*

^c*Department of Statistical Science, Duke University, Durham, NC 27708-0251, USA*

Abstract

We discuss novel approaches to evaluation of both upper and lower bounds on log marginal likelihoods for model comparison in Bayesian analysis. From posterior Monte Carlo samples, we show how existing variational approximation methods defining lower bounds on marginal likelihoods can be extended to also define upper bounds, and develop optimization methods to minimize such upper bounds. Further, using this new approach to upper bound evaluation, we suggest and exemplify a new quasi-optimized lower bound that can often be obtained with trivial computations compared to current methods. We further discuss the use of partial analytic marginalization of some model parameters as a way of significantly reducing the differences between upper and lower bounds to improve marginal likelihood approximation. To implement this, however, traditional variational methods are intractable, and we provide solution in terms of a novel Monte Carlo Stochastic Approximation (MCSA). We provide theoretical results on convergence of the resulting approximations to true bounds, and several simulation examples in regression and mixture models to demonstrate the accuracy and efficacy of the new methods.

Keywords: Bayesian Model Comparison, Marginal Likelihood, Kullback-Leibler Divergence, Variational Methods

*Corresponding author: Chunlin Ji, *tel/fax:* 1(617)495-5496/496-8057
Email addresses: chunlin.ji@gmail.com (Chunlin Ji), haigeshen@yahoo.com
(Haige Shen), mw@stat.duke.edu (Mike West)

1. Introduction

The marginal likelihood is the essential quantity in Bayesian model selection, representing the evidence of a model. However, evaluating marginal likelihoods often involves intractable integration and relies on numerical integration and approximation. Mean-field variational methods, initially developed in statistical physics and extensively studied by machine learning and Bayesian learning communities for deterministic approximation of marginal distributions (MacKay, 1995; Jordan et al., 1999; Jaakkola and Jordan, 2000; Humphreys and Titterton, 2000; Ueda and Ghahramani, 2002; Jordan, 2004; Wang and Titterton, 2004), have been implemented in the model selection context (Corduneanu and Bishop, 2001; Beal, 2003).

For a specified model with parameters $\boldsymbol{\theta} = \{\theta_1, \dots, \theta_K\} \in \Theta$ and prior density $p(\boldsymbol{\theta})$, the marginal likelihood based on observed data D is the quantity

$$p(D) = \int_{\Theta} p(D, \boldsymbol{\theta}) d\boldsymbol{\theta} = \int_{\Theta} p(D|\boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta} \quad (1)$$

In many practical contexts, the required integration must be approximated numerically. In certain cases, some of the parameters can be analytically integrated out, reducing the dimension of the integral while leaving a numerical problem of the same form.

For any density function $q(\boldsymbol{\theta}|\boldsymbol{\gamma})$ parameterized by $\boldsymbol{\gamma} = \{\gamma_1, \dots, \gamma_J\} \in \Gamma$ and with the same support as the posterior $p(\boldsymbol{\theta}|D)$, Jensen's inequality gives

$$\log p(D) \geq L(\boldsymbol{\gamma}) = \int_{\Theta} q(\boldsymbol{\theta}|\boldsymbol{\gamma}) \log \frac{p(D, \boldsymbol{\theta})}{q(\boldsymbol{\theta}|\boldsymbol{\gamma})} d\boldsymbol{\theta}. \quad (2)$$

Maximizing $L(\boldsymbol{\gamma})$ with respect to $\boldsymbol{\gamma}$ provides an optimized lower bound on log marginal likelihood. This optimization is equivalent to minimization of the Kullback-Leibler (KL) divergence of the true posterior $p(\boldsymbol{\theta}|D)$ from $q(\boldsymbol{\theta}|\boldsymbol{\gamma})$.

Current mean-field methods typically use a variational density form factorized over hidden variables and model parameters (or construct such settings by treating certain model parameters as hidden variables), and rely on EM style iterative algorithms to provide solutions to the lower bound optimization (Beal, 2003). Similar to application in variational MLE with missing data (Celeux and Diebolt, 1992; Delyon et al., 1999), stochastic approximation (SA) algorithms based on an iterative Monte Carlo procedures can be used in cases where the expectation step in the EM algorithm can-

not be performed in closed form. Wang and Titterton (2004) have shown that, for the mean-field variational densities of exponential family form, this optimization converges to the true local maximized lower bound.

The first question addressed here is that of adding an upper bound to properly bracket the exact log marginal likelihood. Again using a class of variational densities, we define a theoretical upper bound and develop a computational approach to its minimization. This optimization is equivalent to minimization of the KL divergence of the variational density $q(\boldsymbol{\theta}|\boldsymbol{\gamma})$ from the true posterior $p(\boldsymbol{\theta}|D)$, complementing the lower bound approximation that minimizes the (directional) KL divergence in the other direction. Further, for variational densities of exponential family form, we show convergence to the true global minimum upper bound.

We also discuss a quasi-optimized lower bound that can be obtained with trivial computation – compared to current approaches – based on the result of the optimized upper bound. In addition, we demonstrate that, if we can marginalize with respect to a subset of parameters analytically, then we can often significantly reduce the range between the upper and lower bounds and hence improve the estimation. Further, we present a method that directly uses a method of Monte Carlo stochastic approximation (MCSA) to maximize the lower bound, and prove the convergence to the true local maximum lower bound when $q(\boldsymbol{\theta}|\boldsymbol{\gamma})$ takes an exponential family form.

Accuracy and performance of these new methods is demonstrated with two examples. The first example is a Bayesian linear regression model, in which the analytical form of marginal likelihood is available for assessment of numerical approximations. The second example concerns finite mixture models. We also refer to additional examples in applications.

2. Upper Bound on Marginal Likelihood

When $q(\boldsymbol{\theta}|\boldsymbol{\gamma}) = p(\boldsymbol{\theta}|D)$, the inequality in (2) turns into equality

$$\log p(D) = \int_{\Theta} p(\boldsymbol{\theta}|D) \log p(D, \boldsymbol{\theta}) d\boldsymbol{\theta} - \int_{\Theta} p(\boldsymbol{\theta}|D) \log p(\boldsymbol{\theta}|D) d\boldsymbol{\theta}.$$

The second term here is the entropy of $p(\boldsymbol{\theta}|D)$ which, using Gibbs' inequality, satisfies

$$- \int_{\Theta} p(\boldsymbol{\theta}|D) \log p(\boldsymbol{\theta}|D) d\boldsymbol{\theta} \leq - \int_{\Theta} p(\boldsymbol{\theta}|D) \log q(\boldsymbol{\theta}|\boldsymbol{\gamma}) d\boldsymbol{\theta}$$

for any $q(\boldsymbol{\theta}|\boldsymbol{\gamma})$. We deduce the upper bound on log marginal likelihood

$$U(\boldsymbol{\gamma}) = \int_{\boldsymbol{\Theta}} p(\boldsymbol{\theta}|D) \log \frac{p(\boldsymbol{\theta}, D)}{q(\boldsymbol{\theta}|\boldsymbol{\gamma})} d\boldsymbol{\theta} \geq \log p(D). \quad (3)$$

Minimizing $U(\boldsymbol{\gamma})$ with respect to $\boldsymbol{\gamma}$ provides the optimum. It is clear that the optimal parameter, $\boldsymbol{\gamma}_U$, also minimizes the KL divergence of $q(\boldsymbol{\theta}|\boldsymbol{\gamma})$ from $p(\boldsymbol{\theta}|D)$, namely

$$\mathcal{D}(p||q) = \int_{\boldsymbol{\Theta}} p(\boldsymbol{\theta}|D) \log \frac{p(\boldsymbol{\theta}|D)}{q(\boldsymbol{\theta}|\boldsymbol{\gamma})} d\boldsymbol{\theta}. \quad (4)$$

The minimized KL value is then just the discrepancy between the implied estimate $U(\boldsymbol{\gamma}_U)$ and the true $\log p(D)$.

Assuming $q(\boldsymbol{\theta}|\boldsymbol{\gamma})$ comes from the exponential family, $\mathcal{D}(p||q)$ as a linear functional of $\log q(\boldsymbol{\theta}|\boldsymbol{\gamma})$ is convex with respect to $\boldsymbol{\gamma}$. Hence, the global minimum can be found by solving the $j = 1, \dots, J$ equations

$$\frac{\partial}{\partial \gamma_j} \mathcal{D}(p||q) = - \int_{\boldsymbol{\Theta}} p(\boldsymbol{\theta}|D) \left[\frac{\partial}{\partial \gamma_j} \log q(\boldsymbol{\theta}|\boldsymbol{\gamma}) \right] d\boldsymbol{\theta} = 0. \quad (5)$$

For $q(\boldsymbol{\theta}|\boldsymbol{\gamma})$ of exponential family form, these equations can be solved analytically in terms of posterior moments. Typically, the posterior is represented by a simulation sample such as from MCMC, and moments are approximated via Monte Carlo. Given a Monte Carlo posterior sample $\{\boldsymbol{\theta}^{(i)} : i = 1, \dots, N\}$, it is trivial to prove that the resulting Monte Carlo estimated solution $\hat{\boldsymbol{\gamma}}_U$ converges (almost surely) with N to $\boldsymbol{\gamma}_U$. The resulting global minimum upper bound of log marginal likelihood U_o is then estimated by

$$\hat{U}_o = \frac{1}{N} \sum_{i=1}^N \log \frac{p(\boldsymbol{\theta}^{(i)}, D)}{q(\boldsymbol{\theta}^{(i)}|\hat{\boldsymbol{\gamma}}_U)} \quad (6)$$

which converges almost surely to U_o .

3. Optimal Lower Bound Approximation

We comment on the standard variational lower bound approximation and then introduce two new methods: quasi-optimized lower bounds, for fast and efficient computations relative to the standard method, and Monte Carlo stochastic approximation, for problems with non-factored variational densities.

3.1. Standard Variational Methods

The standard lower bound is based on equation (2); maximization of $L(\boldsymbol{\gamma})$ over $\boldsymbol{\gamma}$ provides the optimized lower bound on the log marginal likelihood, equivalent to minimizing the KL divergence $\mathcal{D}(p||q)$ (Jordan et al., 1999; Ghahramani and Beal, 2001; Xing et al., 2003; Blei and Jordan, 2004). Computational tractability is achieved with factorized variational densities $q(\boldsymbol{\theta}|\boldsymbol{\gamma}) = \prod_{j=1}^J q(\boldsymbol{\theta}_j|\boldsymbol{\gamma}_j)$, where $\boldsymbol{\gamma} = \{\boldsymbol{\gamma}_1, \dots, \boldsymbol{\gamma}_J\}$, and when the component densities have exponential family forms, i.e. $q(\boldsymbol{\theta}_j|\boldsymbol{\gamma}_j) = h(\boldsymbol{\theta}_j) \exp\{\boldsymbol{\gamma}'_j \boldsymbol{\theta}_j - a(\boldsymbol{\gamma}_j)\}$. Most practical examples assume this.

Substantial further simplification is achieved if each complete conditional posterior $p(\boldsymbol{\theta}_j|D, \boldsymbol{\theta}_{-j})$ is also of exponential family form. This structure arises in many contexts, and conditional exponential family forms of complete conditional posteriors can sometimes be induced by use of latent variables that augment initial parameters. Under such an assumption, that for each $j = 1, \dots, J$,

$$p(\boldsymbol{\theta}_j|D, \boldsymbol{\theta}_{-j}) = h(\boldsymbol{\theta}_j) \exp\{\mathbf{g}(\boldsymbol{\theta}_{-j}, D)' \boldsymbol{\theta}_j - b(\mathbf{g}(\boldsymbol{\theta}_{-j}, D))\},$$

where $\mathbf{g}(\boldsymbol{\theta}_{-j}, D)$ is the posterior complete conditional natural parameter for $\boldsymbol{\theta}_j$, then the conditional optimization steps run through iterative updates of variational parameter subsets using

$$\boldsymbol{\gamma}_j = \int_{\boldsymbol{\theta}_{-j}} \mathbf{g}(\boldsymbol{\theta}_{-j}, D) q(\boldsymbol{\theta}_{-j}|\boldsymbol{\gamma}_{-j}) d\boldsymbol{\theta}_{-j}, \quad j = 1, \dots, J, \quad (7)$$

i.e., matching the natural parameter of the variational density with the expected value of that of the complete conditional posterior. In such cases, computation using coordinate ascent algorithms can be achieved, in which we iteratively maximize the bound with respect to each $\boldsymbol{\gamma}_j$ holding all other variational parameters $\boldsymbol{\gamma}_{-j}$ fixed at “current” values; see Ghahramani and Beal (2001), Blei and Jordan (2004).

3.2. Quasi-Optimized Lower Bound

As mentioned, the lower bound optimization via the standard variational method can be computationally challenging, so raising interest in modifications and alternatives. A first path is suggested by the relative ease of computation of the new optimized upper bound of Section 2. This suggests the use of $\boldsymbol{\gamma}_U$, or realistically the MC estimate $\hat{\boldsymbol{\gamma}}_U$, in *lower* bound evaluation

too. While γ_L and γ_U generally differ – since they are minimizers of the generally different $\mathcal{D}(p||q)$ and $\mathcal{D}(q||p)$, respectively, they tend to be close in problems with unimodal and increasingly concentrated posteriors, in many model contexts.

Thus we define the *quasi-optimized lower bound* as follows. Draw a Monte Carlo sample $\{\boldsymbol{\theta}^{(i)} : i = 1, \dots, N\}$ from $q(\boldsymbol{\theta}|\hat{\gamma}_U)$ and use it to define and evaluate the MC estimated quasi-optimized lower bound

$$\hat{L}_o = \frac{1}{N} \sum_{i=1}^N \log \frac{p(\boldsymbol{\theta}^{(i)}, D)}{q(\boldsymbol{\theta}^{(i)}|\hat{\gamma}_U)}. \quad (8)$$

This is a consistent estimate of $L(\gamma_U)$, a lower bound that is not theoretically optimized but offers a computationally attractive, and often effective, proxy for the optimal $L(\gamma_L)$. Examples below explore this.

3.3. Lower Bound Optimization by MCSA

If the variational distribution $q_\gamma(\boldsymbol{\theta})$ is not a fully factorized distribution, and/or the posterior complete conditionals intractable, then the analytical iterative update equation for variational parameters derived in the variational algorithm is inapplicable. This is common. We will show later, for example, that analytic marginalization of some model parameters – when feasible – can significantly reduce the discrepancy between the bounds on log marginal likelihood; doing so, however, moves us out of the context in which the lower bound optimization is tractable. For these cases, we need an alternative computational strategy, and here introduce a Monte Carlo stochastic approximation (MCSA) to numerically maximize the lower bound directly. We also show that the algorithm converges to the true local maximum lower bound when $q(\boldsymbol{\theta}|\gamma)$ takes an exponential family form.

First note that

$$\dot{L}(\gamma) \equiv \frac{dL(\gamma)}{d\gamma} = - \int_{\Theta} h(\boldsymbol{\theta}|\gamma) q(\boldsymbol{\theta}|\gamma) d\boldsymbol{\theta} = \mathbf{0} \quad (9)$$

where

$$h(\boldsymbol{\theta}|\gamma) = \left[1 + \log \frac{q(\boldsymbol{\theta}|\gamma)}{p(\boldsymbol{\theta}, D)} \right] \frac{d}{d\gamma} \log q(\boldsymbol{\theta}|\gamma).$$

Our novel MCSA approach aims to solve equations (9) numerically. This involves a modification of the standard stochastic approximation (SA) algo-

rithm (Robbins and Monro, 1951; Kushner and Yin, 1997) to define recursive approximation of the solution of $\dot{L}(\boldsymbol{\gamma}) = \mathbf{0}$ based on noisy, approximate evaluation of the defining integral in equation (9).

For a current value of $\boldsymbol{\gamma}$, assume we generate a Monte Carlo sample $\{\boldsymbol{\theta}^{(i)} : i = 1, \dots, N\}$ from $q(\boldsymbol{\theta}|\boldsymbol{\gamma})$, defining the consistent (in N) MC approximation to equation (9) as

$$\boldsymbol{\lambda}(\boldsymbol{\gamma}) \equiv -\frac{1}{N} \sum_{i=1}^N h(\boldsymbol{\theta}^{(i)}|\boldsymbol{\gamma}). \quad (10)$$

A Robbins-Monroe SA algorithm then iteratively updates a sequence of values $\boldsymbol{\gamma}^{(t)}$ over $t = 1, 2, \dots$ via the recursion

$$\boldsymbol{\gamma}^{(t+1)} = \boldsymbol{\gamma}^{(t)} + s^{(t+1)} \boldsymbol{\lambda}(\boldsymbol{\gamma}^{(t)})$$

for some initial $\boldsymbol{\gamma}^{(0)}$ and sequence of scalar step sizes $s^{(t)}$. The latter sequence must satisfy standard conditions, $\sum_{t=1}^{\infty} s^{(t)} = \infty$ and $\sum_{t=1}^{\infty} [s^{(t)}]^2 < \infty$. Under these conditions, it can be shown that the SA sequence $\boldsymbol{\gamma}^{(t)}$ converges to the lower bound optimizing value $\boldsymbol{\gamma}_L$ as $N, t \rightarrow \infty$; the proof relies on standard theory of convergence of SA algorithms from (Kushner and Yin, 1997); full mathematical details can be found in chapter 5 of the PhD thesis of the first author here (Ji, 2009).

Terminating the MCSA after a series of steps yielding a terminal value $\hat{\boldsymbol{\gamma}}_L$, a further MC sample $\{\boldsymbol{\theta}^{(i)} : i = 1, \dots, N\}$ from $q(\boldsymbol{\theta}|\hat{\boldsymbol{\gamma}}_L)$ defines the final estimate of the optimal lower bound

$$\hat{L}_o = \frac{1}{N} \sum_{i=1}^N \log \frac{p(\boldsymbol{\theta}^{(i)}, D)}{q(\boldsymbol{\theta}^{(i)}|\hat{\boldsymbol{\gamma}}_L)}. \quad (11)$$

4. Evaluation Examples

Two model classes provide evaluation and examples. The first involves linear regression model comparison, a context where the exact values of marginal likelihood are available so that bound approximations can be assessed against true values. The second is Gaussian mixture models where analytic marginal likelihoods are unavailable, and where comparisons are made with various methods to approximate bounds as well as the “gold standard” using Candidate’s formula (Chib, 1995).

4.1. Linear Regression

Assume a linear model with $n \times p$ design matrix \mathbf{X} and $n \times 1$ response vector \mathbf{y} , viz.

$$\begin{aligned}\mathbf{y} &\sim N(\mathbf{X}\boldsymbol{\beta}, \sigma^2 \mathbf{I}_n) \\ \boldsymbol{\beta} &\sim N(\mathbf{0}_p, \tau \sigma^2 \mathbf{I}_p) \\ \sigma^2 &\sim IG(h_0, k_0)\end{aligned}$$

where hyperparameters τ , h_0 , and k_0 are assumed to be fixed and known. Here $D = \{\mathbf{y}\}$. Due to the conjugate setting, the exact marginal likelihood has a closed form

$$p(D) = (2\pi)^{-n/2} |\mathbf{C}|^{-1/2} k_0^{h_0} \{\Gamma(h_0 + n/2)/\Gamma(h_0)\} (k_0 + \mathbf{y}'\mathbf{C}^{-1}\mathbf{y}/2)^{-(h_0+n/2)}$$

where $\mathbf{C} = \mathbf{I}_n + \tau \mathbf{X}\mathbf{X}'$.

To assess marginal likelihood bound approximations we consider and compare the two choices:

- the *full parameter* context, setting $\boldsymbol{\theta} = \{\boldsymbol{\beta}, \sigma^2\}$;
- the *reduced parameter* context, in this case marginalizing $\boldsymbol{\beta}$ away analytically and performing numerical approximations for bounds using only $\boldsymbol{\theta} = \{\sigma^2\}$.

In the reduced parameter context we make use of

$$p(D|\sigma^2) = (2\pi)^{-n/2} |\sigma^2 \mathbf{C}|^{-1/2} \exp(-\mathbf{y}'\mathbf{B}^{-1}\mathbf{y}/(2\sigma^2)).$$

In each case, the prior conjugacy means we can trivially sample posteriors to provide MC samples for bound approximations.

We focus on simulated polynomial regression models and the problem of comparing polynomial order; marginal likelihoods for each model order are key ingredients. The design matrix has rows of the $\mathbf{x}'_i = (1, x_i, x_i^2, \dots, x_i^r)$ for some $r > 0$, with the n design points x_1, \dots, x_n drawn across a grid of values. Synthetic response data \mathbf{y} come from a model with $r = 3$, $n = 20$, $\boldsymbol{\beta} = [0.2, 2, -2, 0.5]^T$ and $\sigma^2 = 10$. The hyperparameters are set as $\tau = 0.1$, $h_0 = 1$, and $k_0 = 1$.

Seven log marginal likelihood values are reported for each model analysis, as follows.

- ML:* ML is the exact log marginal likelihood $p(D)$.
- $U_1:$ U_1 is the estimate of the upper bound of log marginal likelihood in the reduced parameter context. Here the variational distribution is $q(\boldsymbol{\theta}|\boldsymbol{\gamma}) = IG(\sigma^2|h, k)$ with $\boldsymbol{\gamma} = (h, k)$. Posterior samples of σ^2 yield optimal variational parameters (h_U, k_U) via equation (5) and resulting U_1 from equation (6).
- $U_2:$ U_2 is the estimate of the upper bound in the full parameter context. Here $q(\boldsymbol{\theta}|\boldsymbol{\gamma}) = N(\boldsymbol{\beta}|\boldsymbol{\mu}, \boldsymbol{\Omega})IG(\sigma^2|h, k)$ and optimal variational parameters $\boldsymbol{\gamma} = \{\boldsymbol{\mu}_U, \boldsymbol{\Omega}_U, h_U, k_U\}$ are estimated using equations (5), whereupon U_2 is approximated via equation (6).
- $L_1:$ L_1 is the estimate of the lower bound in the reduced parameter context as for U_1 , and using the MCSA method of Section 3.3 with estimate of equation (11).
- $L_2:$ L_2 is the estimate of the quasi-optimized lower bound in the reduced parameter context as for L_1 , but now with inverse gamma variational parameters set to the quasi-optimized values $(h, k) \leftarrow (h_U, k_U)$ of U_1 .
- $L_3:$ L_3 is the quasi-optimized lower bound estimate in the full parameter context. The normal and inverse gamma variational parameters are set to the quasi-optimized values $\boldsymbol{\gamma} = \{\boldsymbol{\mu}_U, \boldsymbol{\Omega}_U, h_U, k_U\}$ from the upper bound analysis defining U_2 , and used to estimate the quasi-optimized lower bound of equation (8).
- $L_4:$ L_4 is the lower bound using the standard variational method. The variational parameters $\boldsymbol{\gamma} = \{\boldsymbol{\mu}_U, \boldsymbol{\Omega}_U, h, k\}$ are estimated using the coordinate ascent algorithm noted in Section 3.1. At these parameters, samples of $\boldsymbol{\theta} = \{\boldsymbol{\beta}, \sigma^2\}$ are drawn from $q(\boldsymbol{\theta}|\boldsymbol{\gamma})$ and used to approximate the lower bound via equation (11).

We show one example of a synthetic data set generated from a polynomial regression with $r = 3$, together with fitted models of several orders, in Figure 1. Simulations were repeated 100 times and each of the 7 marginal likelihood bound approximations evaluated for each. Table 1 shows results. We see that U_1 is more accurate than than U_2 , and L_1, L_2 better than L_3, L_4 , empirically confirming the view that, by marginalizing out some parameters, we can significantly reduce the discrepancy between the marginal likelihood

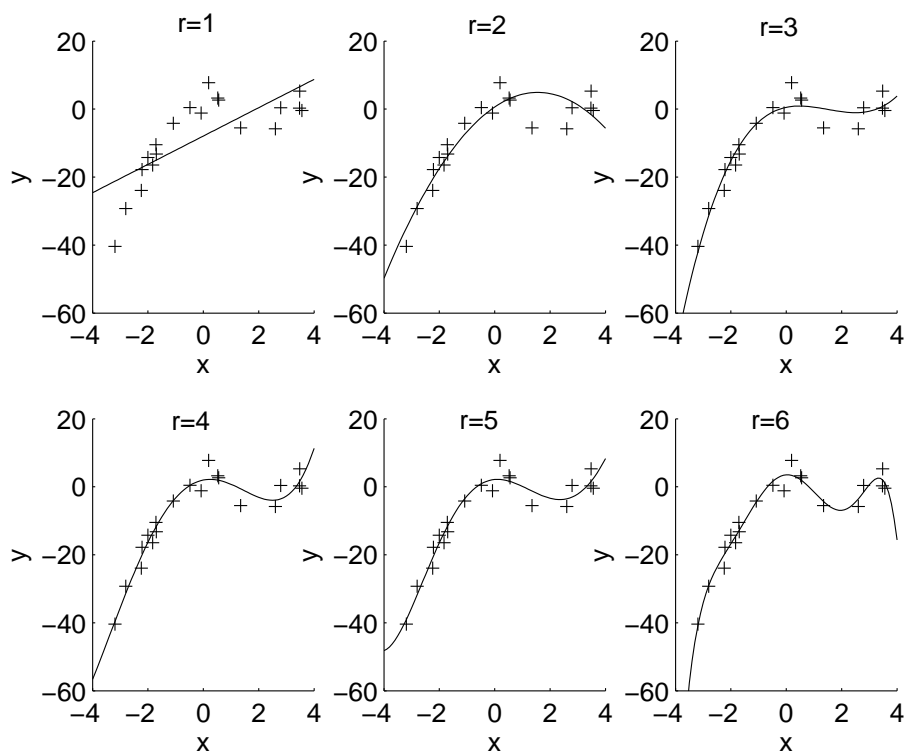


Figure 1: Synthetic data approximated by polynomials of varying orders.

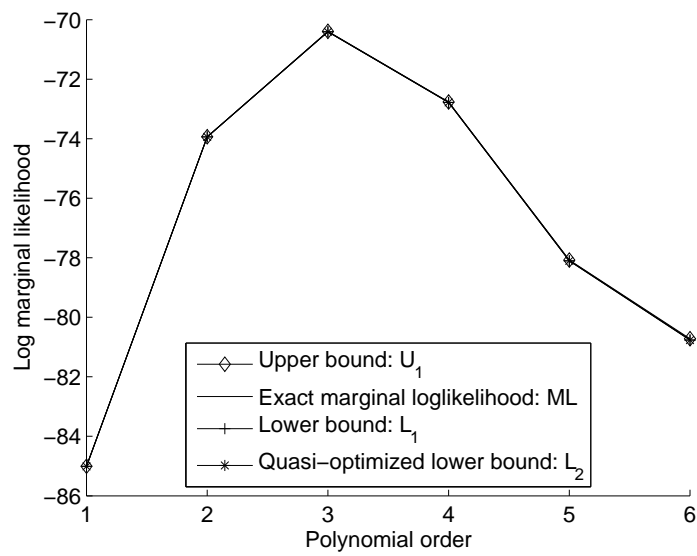


Figure 2: Plot of the analytic value of the log marginal likelihood of the Bayesian linear model with varying number of order q , and means of upper bound U_1 , lower bound L_1 and quasi-optimized lower bound L_2 of the log marginal likelihood for 100 Monte Carlo runs.

and its upper/lower bounds. Moreover, averages of U_1, L_1 and L_2 over the 100 simulations are plotted in Figure 2, illustrating the irrelevance of the approximation errors in estimating bounds in this example.

4.2. Mixture Model

A second study concerns evaluation of marginal likelihoods in multivariate normal, k -component mixture models, especially for comparison of models with respect to the number k of mixture components. Denoting inherent, latent mixture component indicators by z , the model for a random sample of p -variate observations $\mathbf{x}_i, i = 1, \dots, n$, is

$$(\mathbf{x}_i | z_i = j) \sim N(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j), \quad z_i \sim Mn(1, \boldsymbol{\pi}),$$

where $\boldsymbol{\pi} = (\pi_1, \dots, \pi_k)'$ is the vector of mixing probabilities and $Mn(1, \boldsymbol{\pi})$ denotes the multinomial with sample size 1 on cells $1, \dots, k$. We use traditional, conditionally conjugate priors: normal, inverse-Wishart priors for the p -vector means and $p \times p$ covariances matrices, $(\boldsymbol{\mu}_j | \boldsymbol{\Sigma}_j) \sim N(\mathbf{0}, \tau^{-1} \boldsymbol{\Sigma}_j)$ and $\boldsymbol{\Sigma}_j \sim IW(d, \mathbf{S})$ independently over $j = 1, \dots, k$, and with specified hyperparameters τ, d, \mathbf{S} ; a uniform Dirichlet prior for component probabilities $\boldsymbol{\pi} \sim Dir(\mathbf{1}/k)$ where $\mathbf{1}$ is the k -vector of ones. Under this specification, posterior simulation using Gibbs sampling is easy and widely used in routine applications (e.g. Lavine and West, 1992; Chan et al., 2008). This applies to generate posterior samples for the full set of uncertain parameters and latent variables

$$\{z_i; i = 1, \dots, n\}, \quad \{\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j; j = 1, \dots, k\} \quad \text{and} \quad \boldsymbol{\pi} \quad (13)$$

given data $D \equiv \{\mathbf{x}_i, i = 1, \dots, n\}$. To apply the methods for marginal likelihood bounds, we can consider several possibilities for partial analytic posterior marginalization in order to improve approximations. We consider the cases below, where in each case we include the latent z_i as part of the $\boldsymbol{\theta}$ parameter.

- the *standard* or *reduced- $\boldsymbol{\pi}$ context*, when $\boldsymbol{\theta}$ represents the quantities in (13) above with the exception of $\boldsymbol{\pi}$ that can always be analytically integrated away conditional on the indicators z_i ;
- the *reduced- $(\boldsymbol{\pi}, \boldsymbol{\mu})$ context*, when $\boldsymbol{\pi}$ and the $\boldsymbol{\mu}_j, (j = 1, \dots, k)$ are

Table 1: Exact value and approximate bounds on the log marginal likelihood in the linear regression model example. The Monte Carlo estimates of bounds are given in terms of Monte Carlo mean and standard deviations over simulations.

	U_2	U_1	ML	L_1	L_2	L_3	L_4
1	-84.9591 ± 0.0045	-85.0046 ± 0.0012	-85.0068	-85.0074 ± 0.0006	-85.0089 ± 0.0008	-85.0521 ± 0.0049	-85.0475 ± 0.0028
2	-73.8644 ± 0.0067	-73.9329 ± 0.0015	-73.9374	-73.9395 ± 0.0010	-73.9414 ± 0.0011	-74.0038 ± 0.0052	-73.9974 ± 0.0032
3	-70.3021 ± 0.0079	-70.3933 ± 0.0021	-70.4009	-70.4054 ± 0.0012	-70.4077 ± 0.0013	-70.4871 ± 0.0060	-70.4791 ± 0.0039
4	-72.6467 ± 0.0093	-72.7603 ± 0.0027	-72.7720	-72.7792 ± 0.0016	-72.7825 ± 0.0020	-72.8781 ± 0.0065	-72.8676 ± 0.0040
5	-77.9364 ± 0.00105	-78.0741 ± 0.0029	-78.0906	-78.1013 ± 0.0016	-78.1045 ± 0.0018	-78.2152 ± 0.0078	-78.2040 ± 0.0042
6	-80.5663 ± 0.00127	-80.7266 ± 0.0039	-80.7477	-80.7622 ± 0.0020	-80.7656 ± 0.0020	-80.8920 ± 0.0085	-80.8781 ± 0.0048

integrated out analytically, so that

$$\boldsymbol{\theta} = \{(z_i, i = 1, \dots, n), (\boldsymbol{\Sigma}_j, j = 1, \dots, k)\};$$

- the reduced $-(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma})$ context, when $\boldsymbol{\pi}$ and the $(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)$, $(j = 1, \dots, k)$ are integrated out analytically, so that

$$\boldsymbol{\theta} = \{(z_i, i = 1, \dots, n)\}.$$

Given the prior conjugacy, in each case, the required densities $p(D|\boldsymbol{\theta})$, $p(\boldsymbol{\theta})$ and $p(\boldsymbol{\theta}|D)$ are analytically available.

We consider evaluation of seven log marginal likelihood estimates, as follows.

U_1, L_1 : U_1 is the upper bound estimate in the reduced $-(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma})$ context, when $\boldsymbol{\theta}$ represents just the configuration indicators z_i . The variational distribution is $q(\boldsymbol{\theta}|\boldsymbol{\gamma}) = \prod_{i=1}^n Mn(z_i|1, \boldsymbol{w}_i)$ with variational parameters $\boldsymbol{\gamma} = \{\boldsymbol{w}_i, i = 1, \dots, n\}$ where $\boldsymbol{w}_i = (w_{i1}, \dots, w_{ik})'$ is a separate multinomial probability vector for each i . The posterior simulation samples immediately provide the optimizing \boldsymbol{w} using equation (5), and the upper bound then follows from equation (6).

L_1 is the quasi-optimized lower bound estimate in this case. Using the above optimized variational parameters, the quasi-optimized lower bound is estimated by equation (8).

U_2, L_2 : U_2 is the upper bound estimate in the reduced $-(\boldsymbol{\pi}, \boldsymbol{\mu})$ context where $\boldsymbol{\theta}$ comprises the $\boldsymbol{\Sigma}_j$ and z_i . We use variational distribution

$$q(\boldsymbol{\theta}|\boldsymbol{\gamma}) = \left\{ \prod_{j=1}^k IW(\boldsymbol{\Sigma}_j|\kappa_j, \boldsymbol{\Psi}_j) \right\} \prod_{i=1}^n Mn(z_i|1, \boldsymbol{w}_i)$$

with variational parameter $\boldsymbol{\gamma}$ comprised of the set of κ_j , $\boldsymbol{\Psi}_j$ and \boldsymbol{w}_i quantities. From the posterior MCMC summaries, the optimizing $\boldsymbol{\gamma}$ is evaluated using equation (5) and the upper bound follows from equation (6).

L_2 is the quasi-optimized lower bound estimate in this case. Using the above optimized variational parameters, the quasi-optimized lower bound is estimated by equation (8).

U_3, L_3 : U_3 is upper bound estimate in the reduced- π context, using variational distribution

$$q(\boldsymbol{\theta}|\boldsymbol{\gamma}) = \left\{ \prod_{j=1}^k N(\boldsymbol{\mu}_j|\boldsymbol{\nu}_j, \boldsymbol{\Omega}_j) IW(\boldsymbol{\Sigma}_j|\kappa_j, \boldsymbol{\Psi}_j) \right\} \prod_{i=1}^n Mn(z_i|1, \boldsymbol{w}_i) \quad (14)$$

with variational parameter $\boldsymbol{\gamma}$ comprising all $\boldsymbol{\nu}_j, \boldsymbol{\Omega}_j, \kappa_j, \boldsymbol{\Psi}_j$ and \boldsymbol{w}_i quantities. From the posterior MCMC summaries, the optimizing $\boldsymbol{\gamma}$ is evaluated using equation (5) and the upper bound follows from equation (6).

L_3 is the quasi-optimized lower bound estimate in this case. Using the above optimized variational parameters, the quasi-optimized lower bound is estimated by equation (8).

L_4 : L_4 is the estimate of lower bound of log marginal likelihood using the traditional variational method (Corduneanu and Bishop, 2001; Wang and Titterington, 2004). With variational distribution of equation (14), optimal parameters $\boldsymbol{\gamma}$ are estimated using the coordinate ascent algorithm. Samples are then generated from $q(\boldsymbol{\theta}|\boldsymbol{\gamma})$ at the optimized values and used to estimate the lower bound via equation (11).

These bounds are evaluated on synthetic data analogous to an example in Corduneanu and Bishop (2001): $n = 600$ data points generated from a mixture of five bivariate normals with means $[0, 0]'$, $[3, -3]'$, $[3, 3]'$, $[-3, 3]'$, $[-3, -3]'$ and covariance matrices $[1, 0; 0, 1]$, $[1, 0.5; 0.5, 1]$, $[1, -0.5; -0.5, 1]$, $[1, 0.5; 0.5, 1]$, $[1, -0.5; -0.5, 1]$; see Figure 3. We ran 20 repeat simulations and report the mean and standard deviations across replicates for each of the bounds in Table 2. Evidently, U_1, L_1 are more accurate than U_2, L_2 , which themselves dominate U_3, L_3 . This again shows that marginalizing out some parameters can significantly reduce the spread between the upper/lower bounds, and that the resulting bounds can define accurate estimates of the exact but non-computable value – each pair of upper and lower bounds is guaranteed to bracket the true value so that, when the spread is small on the log likelihood scale, they can define practically acceptable values even though they may not be quite the “optimal” bounds. The traditional variational method for lower bounds, which can only be applied to the standard parameter context, fails to achieve performance as good as either L_1 or L_2 ; as shown in Figure 4 this standard method clearly fails to give anything close to a practically useful lower bound in many cases.

Table 2: Monte Carlo estimates of bounds in the mixture model example. Table displays the mean and standard deviation over 20 repeat simulations.

	U_3	U_2	U_1	L_1	L_2	L_3	L_4
k=2	-2895.4 ± 22.6	-2895.7 ± 21.6	-2901.9 ± 25.5	-2908.9 ± 27.5	-2909.4 ± 27.7	-2911.2 ± 27.5	-2910.2 ± 27.3
k=3	-2836.7 ± 80.6	-2848.3 ± 38.8	-2859.8 ± 21.1	-2863.4 ± 14.3	-2877.0 ± 15.4	-2884.8 ± 26.9	-2884.2 ± 24.8
k=4	-2781.6 ± 21.8	-2782.3 ± 21.9	-2784.1 ± 21.1	-2785.1 ± 21.4	-2786.3 ± 22.5	-2786.9 ± 22.6	-2787.2 ± 22.6
k=5	-2683.0 ± 0.15	-2683.8 ± 0.23	-2686.0 ± 0.09	-2687.2 ± 0.06	-2688.6 ± 0.08	-2689.2 ± 0.07	-2689.2 ± 0.08
k=6	-2777.1 ± 1.73	-2779.5 ± 1.93	-2782.5 ± 1.57	-2785.4 ± 1.20	-2788.2 ± 1.12	-2821.6 ± 4.05	-2822.2 ± 4.58
k=7	-2857.4 ± 2.44	-2860.5 ± 2.39	-2863.7 ± 2.31	-2869.4 ± 1.77	-2873.4 ± 1.60	-2946.0 ± 9.48	-2945.7 ± 8.68
k=8	-2925.7 ± 2.88	-2931.3 ± 3.42	-2934.4 ± 3.31	-2941.2 ± 2.12	-2946.5 ± 1.95	-3059.7 ± 8.18	-3058.5 ± 8.41
k=9	-2982.6 ± 8.01	-2990.9 ± 8.92	-2993.7 ± 7.23	-3002.7 ± 4.23	-3010.9 ± 3.98	-3153.5 ± 17.8	-3151.9 ± 17.1
k=10	-3038.9 ± 3.10	-3049.0 ± 4.14	-3052.3 ± 3.98	-3060.5 ± 2.19	-3067.6 ± 1.89	-3246.1 ± 10.5	-3243.8 ± 10.4

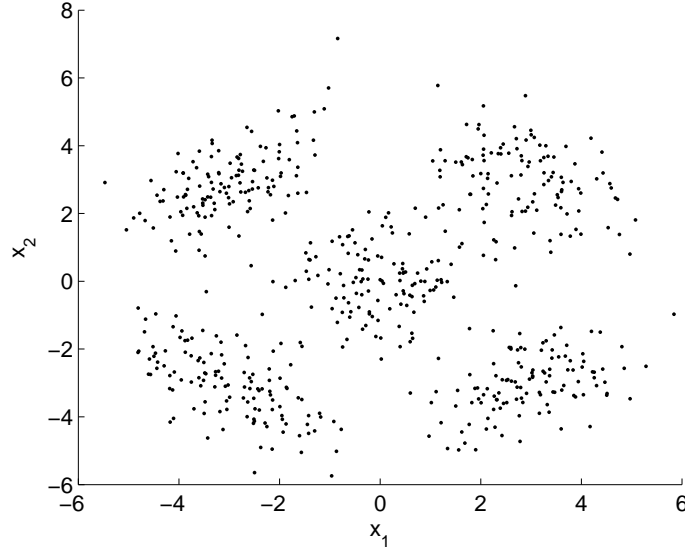


Figure 3: One sample of 600 data points sampled from the mixture of 5 bivariate Gaussians.

5. Discussion

The new approach to defining both upper and lower bounds for log marginal likelihoods extends the prior, standard approach using variational methods in several ways, and the examples show the major benefit and practical utility. First, the utility of our approach to evaluation of optimized upper bounds to couple with lower bounds is obvious. Having bounds that “bracket” the true value, even though the bounds may not be absolutely optimized, provides opportunity to clearly assess practical adequacy of the bounds based on the spread between upper and lower values. If the spread is tight on the log likelihood scale, we can be comfortable with the bound values as defining practically useful estimates. On the other hand, a large spread will indicate that the chosen class of variational densities does not provide a satisfactory approximation to the posterior. Second, the novel quasi-optimized lower bound, that uses the same variational parameters as our optimized upper bound, can in many cases define practically satisfactory lower bounds at minimal computational costs relative to the formally optimized value. Third, it is evident that our methods apply to models in which

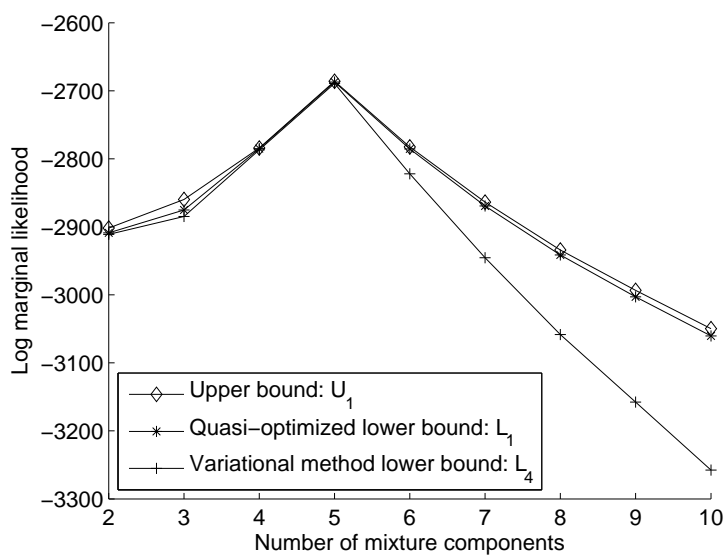


Figure 4: Plot of log marginal likelihood bounds U_1, L_1 , in the mixture model example; the model was fitted at each different value of k indicated, and the bound computations repeated across 20 replicate samples; the average bound values across these samples are plotted. For comparison, the lower bound L_4 estimated by the variational method is also shown.

we can marginalize with respect to a subset of the model parameters and latent variables, and this can be expected to reduce the bound spread and hence improve accuracy, often very substantially.

In all cases the tightness of the spread between upper and lower bounds is essentially determined by how good an approximation $q(\boldsymbol{\theta}|\boldsymbol{\gamma})$ is to $p(\boldsymbol{\theta}|D)$. We have an effective and efficient method for computing and optimizing the upper bound; though we can compute and optimize lower bounds similarly using our new MCSA approach, the computations are intensive compared to the upper bound analysis, and hence the quasi-optimized lower bound idea becomes attractive. More research is needed to theoretically understand this use of upper and lower bounds based on the variational parameter computed on the upper bound. We have performed simulation studies in several model classes that show that the quasi-optimized lower bound variational parameters are often very close to the optimal values, so expect this to be productive. Finally, we note additional examples in more complex settings are developed in Shen and West (2010) and Merl et al. (2010), based on original studies in Shen (2007). Those studies use our methods in problems where $\boldsymbol{\theta}$ includes thousands of binary variables, so the inherent dimension is very high. Shen and West (2010) also explore comparisons with other methods in restricted examples where other methods can be applied, and bear out the results of our examples here in demonstrating the utility of our strategy and specific algorithms for evaluation of upper and lower bounds on marginal likelihoods.

Acknowledgements

This work was performed while the first two authors were PhD students in the Department of Statistical Science at Duke University. We acknowledge support of the NSF (grant DMS-0342172) and NIH (grant U54-CA-112952). Any opinions, findings and conclusions or recommendations expressed in this work are those of the authors and do not necessarily reflect the views of the NSF or NIH.

References

- Beal, M., 2003. Variational algorithms for approximate Bayesian inference. Ph.D. thesis. Gatsby Computational Neuroscience Unit, University College London.
- Blei, D., Jordan, M., 2004. Variational inference for Dirichlet process mixtures. *Bayesian Analysis* 1, 121–144.
- Celeux, G., Diebolt, J., 1992. A stochastic approximation type EM algorithm for the mixture problem. *Stochastics and Stochastics Reports* 41, 127–146.
- Chan, C., Feng, F., West, M., Kepler, T., 2008. Statistical mixture modelling for cell subtype identification in flow cytometry. *Cytometry, A* 73, 693–701.
- Chib, S., 1995. Marginal likelihood from the Gibbs output. *Journal of the American Statistical Association* 90, 1313–1321.
- Corduneanu, A., Bishop, C., 2001. Variational Bayesian model selection for mixture distributions, in: Richardson, T., Jaakkola, T. (Eds.), *Proceedings Eighth International Conference on Artificial Intelligence and Statistics*, Morgan Kaufmann. pp. 27–34.
- Delyon, B., Lavielle, M., Moulines, E., 1999. Convergence of a stochastic approximation version of the EM algorithm. *The Annals of Statistics* 27, 94–128.
- Ghahramani, Z., Beal, M., 2001. Propagation algorithms for variational Bayesian learning, in: Leen, T., Dietterich, T., Tresp, V. (Eds.), *Advances in Neural Information Processing Systems*. MIT Press. volume 13, pp. 507–513.
- Humphreys, K., Titterton, D., 2000. Approximate Bayesian inference for simple mixtures, in: *Proceedings in Computational Statistics, COMP-STAT'2000*. Springer-Verlag.
- Jaakkola, T., Jordan, M., 2000. Bayesian parameter estimation via variational methods. *Statistics and Computing* 10, 25–37.
- Ji, C., 2009. *Advances in Bayesian modelling and computation: Spatio-temporal processes, model assessment and adaptive MCMC*. <http://>

- stat.duke.edu/people/theses/ChunlinJi.html. Ph.D. thesis, Department of Statistical Science, Duke University.
- Jordan, M., 2004. Graphical models. *Statistical Science* 19, 140–15.
- Jordan, M., Ghahramani, Z., Jaakkola, T., Saul, K., 1999. An introduction to variational methods for graphical models. *Machine Learning* 37, 183–233.
- Kushner, H.J., Yin, G.G., 1997. *Stochastic Approximation Algorithms and Applications*. Springer-Verlag, New York.
- Lavine, M., West, M., 1992. A Bayesian method for classification and discrimination. *Canadian Journal of Statistics* 20, 451–461.
- MacKay, D., 1995. Developments in probabilistic modelling with neural networks ensemble learning, in: *Neural Networks: Artificial Intelligence and Industrial Applications*. Proceedings of the 3rd Annual Symposium on Neural Networks, Nijmegen, Netherlands. pp. 191–198.
- Merl, D., Lucas, J., Nevins, J., Shen, H., West, M., 2010. Trans-study projection of genomic biomarkers using sparse factor regression models, in: O’Hagan, A., West, M. (Eds.), *The Handbook of Applied Bayesian Analysis*. Oxford University Press, pp. 118–154.
- Robbins, H., Monro, S., 1951. A stochastic approximation method. *Annals of Mathematical Statistics* 22, 400–407.
- Shen, H., 2007. Bayesian analysis in cancer pathway studies and probabilistic pathway annotation. <http://stat.duke.edu/people/theses/ShenH.html>. Ph.D. thesis, Department of Statistical Science, Duke University.
- Shen, H., West, M., 2010. Bayesian modelling for biological pathway annotation of gene expression pathway signatures, in: Chen, M.H., Dey, D., Müller, P., Sun, D., Ye, K. (Eds.), *Frontiers of Statistical Decision Making and Bayesian Analysis*. New York: Springer-Verlag.
- Ueda, N., Ghahramani, Z., 2002. Bayesian model search for mixture models based on optimizing variational bounds. *Neural Networks* 15, 1223–1241.
- Wang, B., Titterton, D., 2004. Convergence and asymptotic normality of variational Bayesian approximations for exponential family models with

missing values, in: Chickering, M., Halpern, J. (Eds.), Proceedings of the 20th Conference in Uncertainty in Artificial Intelligence, pp. 577–584.

Xing, E., Jordan, M., Russell, S., 2003. A generalized mean field algorithm for variational inference in exponential families, in: Meek, C., Kjerulff, U. (Eds.), Proceedings of the 19th Annual Conference on Uncertainty in AI, Morgan Kaufmann Publishers. pp. 583–591.