A note on mean-field variational approximate Bayesian inference for latent variable models*

Guido Consonni[†] and Jean-Michel Marin[‡]

Abstract

The ill-posed nature of missing variable models offers a challenging testing ground for new computational techniques. This is the case for the mean-field variational Bayesian inference (Jaakkola, 2001; Beal and Ghahramani, 2002; Beal, 2003). In this note, we illustrate the behavior of this approach in the setting of the Bayesian probit model. We show that the mean-field variational method always underestimates the posterior variance and, that, for small sample sizes, the mean-field variational approximation to the posterior location could be poor.

Keywords: Bayesian inference; Bayesian probit model; Gibbs sampling; Latent variable models; Marginal distribution; Mean-field variational methods

1 Introduction

Bayesian inference typically requires the computation of the posterior distribution for a collection of random variables (parameters or unknown observables). Save for stylized settings, this task leads often to intractable calculations. To solve this problem, numerous simulation-based methods have been developed and implemented within the Bayesian paradigm, e.g. importance sampling (Ripley, 1987), Markov Chains Monte Carlo (MCMC) algorithms (Casella and Robert, 1999), particle filtering (Doucet et al., 2001).

*The authors are grateful to Bo Wang and Mike Titterton for comments and suggestions which helped to improve this work. Feedback from the audience of IACS - CSDA 2005 conference, Cyprus, is also gratefully acknowledged.

[†]Department of Economics and Quantitative Methods, University of Pavia, Pavia, Italy guido.consonni@unipv.it

[‡]Corresponding author: INRIA FUTURS, Team SELECT, Department of Mathematics, University Paris-Sud, Orsay, France marin@ceremade.dauphine.fr Address: Université d'Orsay, Laboratoire de Mathématiques, Équipe de Probabilités, Statistique et Modélisation, 91425 Orsay, France

However, for large and complex models, simulation methods are often quite expensive, in terms of time and storage; in particular MCMC algorithms still exhibit technical difficulties, especially with regard to the assessment of the convergence of the chain to its stationary distribution. An important and difficult problem in Bayesian inference, which cannot be addressed directly even resorting to MCMC methods, is represented by the computation of the marginal likelihood of a model. This problem arises in several contexts: e.g. when evaluating the normalizing constant of a posterior distribution, or when computing the Bayes factor (Kass and Raftery, 1995) within the setting of Bayesian model determination (which includes model selection and model averaging).

Over the years, the machine learning community has been developing techniques for the deterministic approximations of marginal distributions, in particular the mean-field variational method, (Jordan et al., 1999; Jaakkola, 2001; Wierginck, 2000; Lauritzen, 2003; Jordan, 2004), which have strong connections to the mean-field approximations of statistical physics. Notable implementations of variational methods in Bayesian statistics may be found in Ghahramani and Beal (2001); Humphreys and Titterton (2000, 2001). The validity of the mean-field variational approach are still to be properly assessed: for example Humphreys and Titterton (2001) have noted that in many instances the mean-field variational approximation may be quite inferior to other simple approximations, besides being computationally quite intensive. On the other hand, Beal and Ghahramani (2002) and Beal (2003) promote the mean-field variational method in order to discover the structure of discrete Bayesian networks (Cowell et al., 1999) when latent variables are present.

In this note, we first present the mean-field variational approximation method in the specific context of Bayesian inference for latent structure models, (Everitt, 1984; Little and Rubin, 1987; McLachlan and Krishnan, 1997; Casella and Robert, 1999). Secondly, we apply the methodology to the analysis of binary data using a Bayesian probit model (Albert and Chib, 1993) and evaluate its performances on some simulated data sets.

2 The mean-field variational approximation method

Let $(y, z) = (y_1, \dots, y_n, z_1, \dots, z_d)$ be a continuous random vector taking values in $\mathbb{R}^{n+d}$. For given $\theta \in \mathbb{R}^p$, denote the joint density of (y, z) by $f(y, z|\theta)$. We suppose that y is observed while z is latent. From the Bayesian viewpoint the uncertainty on the parameter θ is modeled through a probability distribution having density $\pi(\theta)$. Let $\pi(\theta|y)$ denote the posterior density of θ conditional on y , so that

$$\pi(\theta|y) = \frac{\int_{\mathbb{R}^d} f(y, z|\theta)\pi(\theta)dz}{\int_{\mathbb{R}^p} \int_{\mathbb{R}^d} f(y, z|\theta)\pi(\theta)dzd\theta} = \frac{\int_{\mathbb{R}^d} g(y, z, \theta)dz}{m(y)},$$

where $g(y, z, \theta)$ is the joint density of (y, z, θ) and $m(y)$ the marginal density of y . For many latent structure models, an explicit calculation of $\pi(\theta|y)$ is not possible. We present here the mean-field variational approximation method developed by Beal and Ghahramani (2002).

Note first that the posterior distribution of θ may be obtained by marginalizing out z in the joint density of (z, θ) given y , $p(z, \theta|y)$, say. We now describe how the latter can be approximated using a variational argument.

- Let $q(z, \theta|y)$ denote the approximate variational density of (z, θ) given y , having the same support as the true distribution. Assume that under $q(z, \theta|y)$ z and θ are conditionally independent, so that

$$q(z, \theta|y) = q_1(z|y)q_2(\theta|y),$$

where $q_1(z|y)$ and $q_2(\theta|y)$ are, respectively, the approximate variational densities of z and θ , given y .

- We propose to choose $q_1(z|y)$ and $q_2(\theta|y)$ so that the Kullback-Leibler Divergence (KLD) between $p(z, \theta|y)$ and $q_1(z|y)q_2(\theta|y)$ is minimum.

Since $p(z, \theta|y)$ is unknown, it is not apparent how to carry out the minimization of the KLD.

Let $S(p(\cdot))$ denote the support of p and recall that $S(p(\cdot)) = S(q(\cdot))$. Then the Kullback-Leibler divergence between the variational and the true probability density is equal to:

$$\begin{aligned}
& K(q(z, \theta|y), p(z, \theta|y)) \\
&= \int_{S(p(\cdot))} q_1(z|y) q_2(\theta|y) \log \left(\frac{q_1(z|y) q_2(\theta|y)}{p(z, \theta|y)} \right) dz d\theta \\
&= \int_{S(p(\cdot))} q_1(z|y) q_2(\theta|y) \log \left(\frac{q_1(z|y) q_2(\theta|y) m(y)}{p(z, \theta|y) m(y)} \right) dz d\theta \\
&= \log(m(y)) + \int_{S(p(\cdot))} q_1(z|y) q_2(\theta|y) \log \left(\frac{q_1(z|y) q_2(\theta|y)}{g(y, z, \theta)} \right) dz d\theta
\end{aligned}$$

Since $g(y, z, \theta)$ is known, the minimization of $\int_{S(p(\cdot))} q_1(z|y) q_2(\theta|y) \log \left(\frac{q_1(z|y) q_2(\theta|y)}{g(y, z, \theta)} \right) dz d\theta$ with respect to $q_1(\cdot|y)$ and $q_2(\cdot|y)$ is well defined. Unfortunately, such minimization step is very hard to implement, since the functional K is neither convex in $q_1(\cdot|y)$ nor in $q_2(\cdot|y)$. In addition the minimization step must satisfy the following constraints:

- $q_1(\cdot|y)$ and $q_2(\cdot|y)$ are strictly positive functions,
- $\int_{\mathbb{R}^d} q_1(z|y) dz = 1$ and $\int_{\mathbb{R}^p} q_2(\theta|y) d\theta = 1$.

To overcome the above difficulties, Ghahramani and Beal (2001) propose to use a mean-field algorithm, which minimizes $K(q(z, \theta|y), g(y, z, \theta))$ with respect to the unconstrained distributions $q_1(\cdot|y)$ and $q_2(\cdot|y)$. Using arguments based on elementary calculus of variations, they propose to take functional derivatives of $K(q(z, \theta|y), g(y, z, \theta))$ separately with respect $q_1(\cdot|y)$ and $q_2(\cdot|y)$, while holding the other fixed. This results in the following update equations at iteration $(t+1)$:

$$q_1^{(t+1)}(z|y) \propto \exp \left[\int_{S(q_2(\cdot))} \log(f(y, z|\theta)) q_2^{(t)}(\theta|y) d\theta \right] \quad (1)$$

$$q_2^{(t+1)}(\theta|y) \propto \pi(\theta) \exp \left[\int_{S(q_1(\cdot))} \log(f(y, z|\theta)) q_1^{(t+1)}(z|y) dz \right]. \quad (2)$$

The above iterative process should be continued until convergence is reached. Since the objective function is not convex, convergence could occur to a *local* minimum. However, recently, Wang

and Titterington (2004) have shown that, for the exponential family models with missing values, the previous iterative algorithm for obtaining the mean-field variational Bayes estimator, mean of the variational posterior distribution, converges locally to the true value with probability 1 as the sample size becomes indefinitely large. In the same way, Wang and Titterington (2006) have shown that, for a normal mixture model, the variational Bayes estimator converges locally to the maximum likelihood estimator at the rate $O(1/n)$.

3 Bayesian probit model

In this section we apply the variational procedure to a probit model with latent variables. Note that Jaakkola and Jordan (2000) have applied another type of variational approximation to the Bayesian logit model. Suppose that n independent continuous latent random variables $z_1, \dots, z_n$ are given, with $z_i|\theta \sim \mathcal{N}(x_i'\theta, 1)$, where $x_i \in \mathbb{R}^p$ are known covariates and x_i' is the transpose of x_i . We actually observe only the binary random variables $y_1, \dots, y_n$ defined by

$$y_i = \begin{cases} 1 & \text{if } z_i > 0 \\ 0 & \text{otherwise} \end{cases}. \quad (3)$$

It can be easily shown that, given θ , the y_i 's are independent Bernoulli random variables with $\mathbb{P}(y_i = 1|\theta) = \Phi(x_i'\theta)$, where Φ is the standard normal cumulative distribution function, which corresponds to the standard probit model. Our objective is to evaluate the performances of the mean-field variational approximation method).

To this end, let us suppose that $\theta \sim \mathcal{N}(0_p, I_p)$ (we use the probit model here Since the y_i 's are independent of θ given z_i , the probability density of y_i given z_i and θ is equal to

$$1_{y_i=0}1_{z_i \leq 0} + 1_{y_i=1}1_{z_i > 0}, \quad (4)$$

where 1_A is the indicator function of the event A .

As a consequence, if $f_i(y_i, z_i|\theta)$ denotes the density of y_i, z_i given θ , one obtains

$$f_i(y_i, z_i|\theta) = 1/\sqrt{2\pi} \exp\left(- (z_i - x_i'\theta)^2/2\right) (1_{y_i=0}1_{z_i\leq 0} + 1_{y_i=1}1_{z_i>0}).$$

Therefore,

$$z_i|y_i, \theta \sim \begin{cases} \mathcal{N}_+(x_i'\theta, 1, 0) & \text{if } y_i = 1 \\ \mathcal{N}_-(x_i'\theta, 1, 0) & \text{if } y_i = 0 \end{cases} \quad (5)$$

where $\mathcal{N}_+(x_i'\theta, 1, 0)$ is the normal distribution with mean $x_i'\theta$ and variance 1 left-truncated at 0, while $\mathcal{N}_-(x_i'\theta, 1, 0)$ denotes the same distribution right-truncated at 0. Let $y = (y_1, \dots, y_n)$ and $z = (z_1, \dots, z_n)$. One can establish that

$$\theta|y, z \sim \mathcal{N}_p((I_p + X'X)^{-1}X'z, (I_p + X'X)^{-1}) \quad (6)$$

where $X = [x_1 | x_2 | \dots | x_p]$.

The calculation of the posterior distribution of θ is intractable. Albert and Chib (1993) suggest to implement a Gibbs sampler. Given a previous value of θ , one cycle of the Gibbs algorithm would produce a sample value for z from the distribution (5), which, when substituted into (6) would produce a new value for θ . The starting value of θ may be taken to be the maximum likelihood estimate.

We plan instead to apply the mean-field variational approximation method described in the previous section, in order to evaluate its performances. The preliminary computations of (1) and (2) are required. Since $f(y, z|\theta) = \prod_{i=1}^n f_i(y_i, z_i|\theta)$, it can be easily shown that, at the iteration $(t+1)$ of the variational optimization algorithm, we have:

- $q_1^{(t+1)}(z|y)$ is such that the random variables $z_1, \dots, z_n$ are conditionally independent with:

$$z_i|y_i \sim \begin{cases} \mathcal{N}_+(x_i'\mathbb{E}_{q_2^{(t)}}(\theta), 1, 0) & \text{if } y_i = 1, \\ \mathcal{N}_-(x_i'\mathbb{E}_{q_2^{(t)}}(\theta), 1, 0) & \text{if } y_i = 0, \end{cases}$$

where $\mathbb{E}_{q_2^{(t)}}(\theta)$ is the expectation of θ under the distribution $q_2^{(t)}(\theta|y)$;

- $q_2^{(t+1)}(\theta|y)$ is distributed according to a p -dimensional Gaussian distribution with expectation

$$(I_p + X'X)^{-1} X' \mathbb{E}_{q_1^{(t+1)}}(z)$$

and variance-covariance matrix

$$(I_p + X'X)^{-1}$$

where $\mathbb{E}_{q_1^{(t+1)}}(z)$ is the expectation of z under the distribution $q_1^{(t+1)}(z|y)$.

The covariance matrix of $q_2^{(t+1)}(\theta|y)$ does not depend on t and therefore represents the mean-field variational approximation to $\mathbb{V}(\theta|y)$. Also,

$$\begin{aligned} \mathbb{V}(\theta|y) &= \mathbb{E}(\mathbb{V}(\theta|y, z)) + \mathbb{V}(\mathbb{E}(\theta|y, z)) \\ &= (I_p + X'X)^{-1} + \mathbb{V}\left((I_p + X'X)^{-1} X' z\right) \end{aligned}$$

Then, the mean-field variational covariance matrix of θ is smaller than the posterior covariance matrix (the difference between these two matrices is definite negative). Moreover, since the probit model belongs to the exponential family models with missing values, the mean-field variational mean of θ converges locally to the true value of θ with probability 1 as the sample size becomes indefinitely large, according to the results of Wang and Titterton (2004).

4 Simulation studies

First, we apply the mean-field variational algorithm on two simulated data sets. We consider the most simple probit model wherein $p = 2$, $x'_i = (1, q_i)$ and $\theta' = (\alpha, \beta)$. In such a case, there is only one predictor variable in the model and $\mathbb{P}(y_i = 1|\alpha, \beta) = \Phi(\alpha + \beta q_i)$.

Dataset 1, presented in Figure 1, contains $n = 100$ observations, generated using $\alpha = 1$ and $\beta = 2$. The values of the predictor variable have been fixed at random from the standard normal distribution. Figure 2 presents the sample paths of the Gibbs Markov chains for α and β over

Figure 1: Dataset 1: plot of simulated data and of $\Phi(1 + 2q)$ (curve)

100,000 iterations. Their behavior appear satisfactory and convergence to the stationary distribution is very quick. The empirical means, based on the last 20,000 iterations, are equal to 0.94 for α and 1.85 for β . They represent MCMC estimates of the posterior expectations of α and β , and appear satisfactory. Moreover, using the empirical variances based on the last 20,000 iterations, MCMC estimates of the posterior variance of α and β are respectively given by 0.048 and 0.12. The maximum likelihood estimates (MLEs) of α and β are equal to 0.92 and 1.79. The MLEs have been calculated using the EM algorithm: convergence was reached after 254 iterations. The optimization algorithm of the variational approximation method has converged after 179 iterations and the variational distributions of α and β are respectively $\mathcal{N}(0.83, 0.01)$ and $\mathcal{N}(1.59, 0.01)$. Figure 3 shows that these approximate distributions are significantly different from the true posterior distributions estimated through their accurate Gibbs approximations. Nevertheless, for this simulated dataset, the expected values of α and β calculated through the variational approximation are acceptable. However, the variance of the approximate variational posterior distribution of α and β

Figure 2: Dataset 1: evolution of the Gibbs Markov chains over 100,000 iterations

are too small.

Dataset 2, presented in Figure 4, has been generated for $\alpha = 1$ and $\beta = 5$. The values of the predictor variable have been fixed at random from the standard normal distribution. Figure 5 presents the behavior of the Gibbs chains over 100,000 iterations. We observe that the convergence of these chains to their stationary distributions is as fast as in the case of dataset 1. The empirical means over the last 20,000 iterations are equal to 1.25 for α and 4.30 for β . This is the Gibbs estimates of the posterior means of α and β . Moreover, using the empirical variances relative to the last 20,000 iterations, MCMC estimates of the posterior variance of α and β are respectively given by 0.10 and 0.84. The maximum likelihood estimations of α and β are respectively equal to 1.19 and 4.07. These estimates have been calculated using the EM algorithm which has converged after 1859 iterations. The optimization algorithm of the variational approximation method has converged after 438 iterations and the variational distributions of α and β are respectively $\mathcal{N}(0.81, 0.01)$ and $\mathcal{N}(2.58, 0.01)$. Figure 6 shows that these distributions are very different from the

Figure 3: Dataset 1: histograms of the last 20,000 Gibbs simulations, variational densities (curve) and maximum likelihood estimations (straight line)

true posterior distributions estimated by their accurate Gibbs approximations. For this simulated dataset, variational approximations are very poor.

As we have just verified on a specific example, the mean-field variational approach can fail to identify the posterior location. In order to provide a further check on the performance of the method, we consider now the probit model with $p = 2$ and $x_i = (u_i, v_i)$. In this case, there two predictor variables, chosen at random from the normal distribution with mean 0 and variance 2, and no intercept. For different sample sizes with $\theta' = (\theta_1, \theta_2) = (1, 2)$ and 100 Monte Carlo replications, Table 1 gives a Monte Carlo approximation to the square root of the mean square difference between the Gibbs sampling approximation to the posterior mean and the mean-field variational location. For each dataset, we used 100,000 Gibbs iterations and checked the convergence. In this case, the mean-field variational approximation to the posterior mean is reasonable even though, for small sample sizes, this approximation could be poor.

Figure 4: Dataset 2: plot of simulated data and of $\Phi(1 + 5q)$ (curve)

5 Conclusion

Variational methods have been recently popularized in order to carry out complex Bayesian inferences and predictions, especially in the presence of latent variables. In this note, we have described the mean-field variational method and evaluated its performance with regard to the Bayesian probit model. The results we obtain are not unambiguous, and reinforce the conclusion that further investigation and testing is required before the mean-field variational method can be accepted as a reliable computational tool for Bayesian inference in a variety of applied settings.

References

- Albert, J. and Chib, S. (1993). Bayesian analysis of binary and polychotomous response data. *J. American Statist. Assoc.*, 88:669–679.
- Beal, M. (2003). Variational Algorithms for Approximate Bayesian Inference. Technical report, Gatsby Computational Neuroscience Unit. Ph.D. Thesis.
- Beal, M. and Ghahramani, Z. (2002). The variational Bayesian EM algorithm for incomplete data: with application to scoring graphical model structures. In *Bayesian Statistics, 7*, Oxford. Oxford University Press.

Figure 5: Dataset 2: evolution of the Gibbs chains over 100,000 iterations

- Casella, G. and Robert, C. (1999). *Monte Carlo Statistical Methods*. Springer-Verlag.
- Cowell, R., Dawid, A., Lauritzen, S., and Spiegelhalter, D. (1999). *Probabilistic Networks and Expert Systems*. Springer-Verlag.
- Doucet, A., de Freitas, N., and Gordon, N. (2001). *Sequential Monte Carlo Methods in Practice*. Springer-Verlag.
- Everitt, B. (1984). *An Introduction to Latent Variable Models*. Chapman and Hall.
- Ghahramani, Z. and Beal, M. (2001). Propagation algorithms for variational Bayesian learning. In Leen, T., Dietterich, T., and Tresp, V., editors, *Advances in Neural Information Processing Systems*, 13. MIT Press.
- Humphreys, K. and Titterton, D. (2000). Approximate Bayesian inference for simple mixtures. In *COMPSTAT'2000, Proceedings in Computational Statistics*. Springer-Verlag.
- Humphreys, K. and Titterton, D. (2001). Some examples of recursive variational approximations for Bayesian inference. In Opper, M. and Saad, D., editors, *Advances Mean Field Methods: Theory and Practice*. MIT Press.
- Jaakkola, T. (2001). Tutorial on variational approximation methods. In Opper, M. and Saad, D., editors, *Advanced Mean Field Methods: Theory and Practice*. MIT Press.
- Jaakkola, T. and Jordan, M. (2000). Bayesian parameter estimation via variational methods. *Statistics and Computing*, 10:25–37.

Figure 6: Dataset 2: histograms of the last 20,000 Gibbs simulations, variational densities (curve) and maximum likelihood estimations (straight line)

Jordan, M. (2004). Graphical Models. *Statist. Science*, 19:140–155.

Jordan, M., Ghahramani, Z., Jaakkola, T., and Saul, K. (1999). An introduction to variational methods for graphical models. *Machine Learning*, 37(2):183–233.

Kass, R. and Raftery, A. (1995). Bayes factor and model uncertainty. *J. American Statist. Assoc.*, 90:773–795.

Lauritzen, S. (2003). Some modern applications of graphical models. In Green, P., Hjort, N., and Richardson, S., editors, *Highly Structured Stochastic Systems*. Oxford University Press, Oxford.

Little, R. and Rubin, D. (1987). *Statistical Analysis with Missing Data*. John Wiley, New York.

McLachlan, G. and Krishnan, T. (1997). *The EM Algorithm and Extensions*. John Wiley, New York.

Ripley, B. (1987). *Stochastic Simulation*. John Wiley, New York.

Wang, B. and Titterton, D. (2004). Convergence and asymptotic normality of variational Bayesian approximations for exponential family models with missing values. In Chickering, M. and Halpern, J., editors, *Proceedings of the 20th Conference in Uncertainty in Artificial Intelligence*. AUAI Press.

Wang, B. and Titterton, D. (2006). Convergence properties of a general algorithm for calculating variational Bayesian estimates for a normal mixture model. *Bayesian Analysis*. to appear.

	θ_1	θ_2
$n = 50$	0.0881	0.1151
$n = 100$	0.0466	0.0588
$n = 500$	0.0299	0.0561

Table 1: For different sample sizes and 100 Monte Carlo replications, comparison between Gibbs sampling approximation to the posterior mean and the mean-field variational location

Wiegerinck, W. (2000). Variational approximations between mean field theory and the junction tree algorithm. In Boutilier, C. and Goldszmidt, M., editors, *Proceedings of the 16th Conference in Uncertainty in Artificial Intelligence*. Morgan Kaufmann.