manuscript No.
(will be inserted by the editor)

On Variational Bayes Estimation and Variational Bayes Information Criteria for Linear Regression Models

Chong You · John T. Ormerod · Samuel Müller

Received: date / Accepted: date

Abstract Variational Bayes is a fast alternative to Markov chain Monte Carlo for performing approximate Bayesian inference. It can be an efficient and effective means of analyzing large datasets. However, variational approximations are often criticized, typically based on empirical grounds, for being unable to produce valid statistical inferences in several modeling contexts. In this article, we briefly summarize variational Bayes and describe how the method can be applied to a Bayesian linear model. We prove that under mild regularity conditions, the estimators based on variational Bayes enjoy some desirable frequentist properties such as consistency and can be used to obtain asymptotically valid standard errors for Bayesian linear regression models. This result partially contradicts the criticism that variational Bayes is not useful for inference. Furthermore, we introduce two variational Bayes information criteria: the variational Akaike information criterion (VAIC) and the variational Bayesian information criterion (VBIC). The variational Akaike information criterion is a variational Bayes approximation to the deviance information criterion, we show that it shares the first order asymptotic properties of the Akaike information criterion under mild regularity conditions. We also show that the proposed variational Bayesian information criterion shares the same first order asymptotic properties as the Bayesian information criterion. We support our theoretical results by numerical examples.

Chong You
School of Mathematics and Statistics, University of Sydney, Sydney 2006, AUSTRALIA
Tel.: +61(2)91141266
E-mail: chong.y@maths.usyd.com

John T. Ormerod
School of Mathematics and Statistics, University of Sydney, Sydney 2006, AUSTRALIA

Samuel Müller
School of Mathematics and Statistics, University of Sydney, Sydney 2006, AUSTRALIA

Keywords Deviance Information Criterion · Akaike Information Criterion · Bayesian Information Criterion · Markov Chain Monte Carlo · Consistency

1 Introduction

There has been an ever increasing demand by society on Statistics to develop efficient and effective means of analyzing large datasets. In contexts where decisions need to be made quickly Markov chain Monte Carlo (MCMC) methods for the analysis of Bayesian models can be deemed to be too slow in practice (Volant et al, 2012). Variational approximations are a newly emerging class alternative to MCMC for fast approximate Bayesian inference for such contexts.

Variational approximations are often criticized, typically based on empirical grounds, for being unable to produce valid statistical inferences in several modeling contexts Rue et al (2009, Section 1.6). Few theoretical developments for variational approximations have been made to prove or disprove such claims in general and the theory that does exist is context specific (Humphreys and Titterington, 2000; Wang and Titterington, 2006; Hall et al, 2011a,b; Ormerod and Wand, 2012). In this article we focus on a variational Bayes (VB), a special type of variational approximation, for a specific Bayesian linear model. We prove that the VB estimators in this context enjoy desirable frequentist properties such as consistency and can be used to obtain asymptotically valid standard errors. Furthermore, we show that a VB approximation of the deviance information criterion (DIC) of Spiegelhalter et al (2002), which we call the variational Akaike information criterion (VAIC), chooses models that share the same optimality properties as models selected by the Akaike information criterion (AIC) of Akaike (1973) for this class of models. We also propose a variational Bayes version of the Bayesian information criterion (BIC) (Schwarz, 1978), which we call the variational Bayesian information criterion (VBIC). In this article, we do not want to compare the model selection performance of criteria, but provide variational Bayes based analogs with frequentist AIC and BIC.

Closely related to our work is that of Ren et al (2011, Section 3). The model considered here is slightly different from the one considered in (Ren et al, 2011, Section 3), in that the prior for the coefficients is selected to depend on the response variance. This assumption facilitates analytic integration so that the marginal posterior distributions are available for the regression coefficients and response variance. The VB posterior approximations of these quantities can then be shown to approach the true posterior distributions. The priors considered here for the Bayesian linear model are different and so require different techniques to analyze the model.

In Section 2 we briefly summarize VB and describe how the method can be applied to a Bayesian linear model. Theory for VB estimators for increasingly diffuse priors and as the number of samples increases is presented in Section 3. Properties of the variational information criteria are derived in Section 4.

The numerical examples are shown in Section 5. We conclude in Section 6. The proof of the propositions are postponed to the Appendix.

2 Variational Bayes for Linear Regression

Let $\mathbf{y}$ denote a vector of observed data modeled by $p(\mathbf{y}|\boldsymbol{\theta})$ where $\boldsymbol{\theta} \in \Theta$ is a vector of parameters with prior $p(\boldsymbol{\theta})$. Let $q(\boldsymbol{\theta}) = \prod_{i=1}^{K} q_i(\boldsymbol{\theta}_i)$ where $\{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K\}$ is a partition of the parameter vector $\boldsymbol{\theta}$. It can be shown that the $q_i(\boldsymbol{\theta}_i)$, also called q-densities, which minimize the Kullback-Leibler (KL) distance between $p(\boldsymbol{\theta}|\mathbf{y})$ and $q(\boldsymbol{\theta})$ satisfy

$$q_i(\boldsymbol{\theta}_i) \propto \exp\left[\mathbb{E}_{-q(\boldsymbol{\theta}_i)}\left\{\log p(\mathbf{y}, \boldsymbol{\theta})\right\}\right], \ 1 \leq i \leq K,$$

where $\mathbb{E}_{-q(\boldsymbol{\theta}_i)}$ denotes expectation with respect to $\prod_{j \neq i} q_j(\boldsymbol{\theta}_j)$. Furthermore, a lower bound for the marginal log-likelihood is given by

$$\log p(\mathbf{y}) \geq \mathbb{E}_q \left[\log \left\{\frac{p(\mathbf{y}, \boldsymbol{\theta})}{q(\boldsymbol{\theta})}\right\}\right] \equiv \log \underline{p}_q(\mathbf{y}).$$

Finally, by iteratively calculating $q_i(\boldsymbol{\theta}_i)$ for fixed $\{q(\boldsymbol{\theta}_j)\}_{j \neq i}$ the lower bound is increased monotonically over each iteration so that convergence to a local maximizer of $\log p(\mathbf{y})$ occurs under mild regularity conditions. For more details and examples see Bishop (2006) or Ormerod and Wand (2010).

Suppose that we have observed the pairs $(y_i, \mathbf{x}_i)$, $1 \leq i \leq n$, where $\mathbf{x}_i \in \mathbb{R}^p$ and suspect $y_i|\mathbf{x}_i \overset{\text{ind.}}{\sim} N(\mathbf{x}_i^T\boldsymbol{\beta}, \sigma^2)$, $1 \leq i \leq n$, where $\boldsymbol{\beta}$ is a $p \times 1$ vector of coefficients and σ^2 is the noise variance. Using conjugate priors for $\boldsymbol{\beta}$ and σ^2 a Bayesian version of the linear regression model may be written as

$$\mathbf{y}|\boldsymbol{\beta}, \sigma^2 \sim N(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}), \quad \boldsymbol{\beta} \sim N(\mathbf{0}, \sigma^2_{\boldsymbol{\beta}}\mathbf{I}) \quad \text{and} \quad \sigma^2 \sim \text{IG}(A, B), \tag{1}$$

where $\mathbf{X}$ is a $n \times p$ design matrix whose ith row is $\mathbf{x}_i^T$. If $x \sim \text{IG}(A, B)$ then $p(x) = B^A x^{-A-1}\exp(-B/x)/\Gamma(A)$. The parameters $\sigma^2_{\boldsymbol{\beta}}$, A and B are fixed prior hyperparameters. Let $\boldsymbol{\theta} = [\boldsymbol{\beta}^T, \sigma^2]^T$ then the optimal VB q-densities corresponding to the restriction $q(\boldsymbol{\theta}) = q_{\boldsymbol{\beta}}(\boldsymbol{\beta})q_{\sigma^2}(\sigma^2)$ have the form

$$q^*_{\boldsymbol{\beta}}(\boldsymbol{\beta}) \sim N(\boldsymbol{\mu}_{q(\boldsymbol{\beta})}, \boldsymbol{\Sigma}_{q(\boldsymbol{\beta})}) \quad \text{and} \quad q^*_{\sigma^2}(\sigma^2) \sim \text{IG}(A + \tfrac{n}{2}, B_{q(\sigma^2)}),$$

where

$$\boldsymbol{\Sigma}_{q(\boldsymbol{\beta})} = \left[\left(\frac{A + n/2}{B_{q(\sigma^2)}}\right)\mathbf{X}^T\mathbf{X} + \sigma_\beta^{-2}\mathbf{I}\right]^{-1}, \tag{2}$$

$$\boldsymbol{\mu}_{q(\boldsymbol{\beta})} = \left(\frac{A + n/2}{B_{q(\sigma^2)}}\right)\boldsymbol{\Sigma}_{q(\boldsymbol{\beta})}\mathbf{X}^T\mathbf{y}, \tag{3}$$

$$B_{q(\sigma^2)} = B + \frac{1}{2}||\mathbf{y} - \mathbf{X}\boldsymbol{\mu}_{q(\boldsymbol{\beta})}||^2 + \frac{1}{2}\text{tr}\left(\mathbf{X}^T\mathbf{X}\boldsymbol{\Sigma}_{q(\boldsymbol{\beta})}\right). \tag{4}$$

Note that equation (2)–(4) must hold simultaneously for the q-densities to be optimal.

Algorithm 1 *Iterative scheme for obtaining $q^*_{\boldsymbol{\beta}}(\boldsymbol{\beta})$ and $q^*_{\sigma^2}(\sigma^2)$ for model (1).*

Initialize: $B_{q(\sigma^2)} > 0$.
Cycle:

$$\boldsymbol{\Sigma}_{q(\boldsymbol{\beta})} \leftarrow \left[\left(\frac{A+n/2}{B_{q(\sigma^2)}}\right)\mathbf{X}^T\mathbf{X} + \sigma_{\boldsymbol{\beta}}^{-2}\mathbf{I}\right]^{-1} \quad ; \quad \boldsymbol{\mu}_{q(\boldsymbol{\beta})} \leftarrow \left(\frac{A+n/2}{B_{q(\sigma^2)}}\right)\boldsymbol{\Sigma}_{q(\boldsymbol{\beta})}\mathbf{X}^T\mathbf{y}$$
$$B_{q(\sigma^2)} \leftarrow B + \tfrac{1}{2}\|\mathbf{y} - \mathbf{X}\boldsymbol{\mu}_{q(\boldsymbol{\beta})}\|^2 + \tfrac{1}{2}\mathrm{tr}(\mathbf{X}^T\mathbf{X}\boldsymbol{\Sigma}_{q(\boldsymbol{\beta})})$$

until the increase of $\underline{p}_q(\mathbf{y})$ is negligible.

Algorithm 1 describes a process for finding these values. At the bottom of the main loop of Algorithm 1 below the lower bound $\underline{p}_q(\mathbf{y})$ simplifies to:

$$\log \underline{p}_q(\mathbf{y}) = \tfrac{p}{2} - \tfrac{n}{2}\log(2\pi) - \tfrac{p}{2}\log(\sigma_\beta^2) + \tfrac{1}{2}\log|\boldsymbol{\Sigma}_{q(\boldsymbol{\beta})}| - \frac{\|\boldsymbol{\mu}_{q(\boldsymbol{\beta})}\|^2 + \mathrm{tr}\left(\boldsymbol{\Sigma}_{q(\boldsymbol{\beta})}\right)}{2\sigma_\beta^2}$$
$$+A\log(B) - \log\Gamma(A) - \left(A + \tfrac{n}{2}\right)\log(B_{q(\sigma^2)}) + \log\Gamma\left(A + \tfrac{n}{2}\right).$$

3 Main Results

Henceforth we assume $\boldsymbol{\mu}_{q(\boldsymbol{\beta})}$, $\boldsymbol{\Sigma}_{q(\boldsymbol{\beta})}$ and $B_{q(\sigma^2)}$ are the optimal parameters of the q-densities. The following result describes the asymptotic behavior of the quantities defined in equation (2)–(4).

Result 1: *As $\sigma_\beta^2 \to \infty$ (for fixed n and p), provided $2A + n > p$,*

$$\boldsymbol{\Sigma}_{q(\boldsymbol{\beta})} = \left(\frac{2B + \|\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}_{\mathrm{LS}}\|^2}{2A + n - p}\right)(\mathbf{X}^T\mathbf{X})^{-1} + O(\sigma_\beta^{-2}),$$
$$\boldsymbol{\mu}_{q(\boldsymbol{\beta})} = \widehat{\boldsymbol{\beta}}_{\mathrm{LS}} + O(\sigma_\beta^{-2}) \quad \text{and} \quad B_{q(\sigma^2)} = \frac{B + \frac{1}{2}\|\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}_{\mathrm{LS}}\|^2}{1 - p/(2A+n)} + O(\sigma_\beta^{-2}).$$

where $\widehat{\boldsymbol{\beta}}_{LS} \equiv (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$ is the least squares estimate of $\boldsymbol{\beta}$.

Proof: After a Taylor series expansion around σ_β^{-2} in equation (2) we have as $\sigma_\beta^2 \to \infty$ that

$$\boldsymbol{\Sigma}_{q(\boldsymbol{\beta})} = \left(\frac{B_{q(\sigma^2)}}{A + n/2}\right)[\mathbf{X}^T\mathbf{X}]^{-1} + O(\sigma_\beta^{-2}). \tag{5}$$

Substituting (5) into the expression for $\boldsymbol{\mu}_{q(\boldsymbol{\beta})}$ in (3) and simplifying establishes $\boldsymbol{\mu}_{q(\boldsymbol{\beta})} = \widehat{\boldsymbol{\beta}}_{\mathrm{LS}} + O(\sigma_\beta^{-2})$. Similarly, substituting (5) into the expression for $B_{q(\sigma^2)}$ in (4) we obtain

$$B_{q(\sigma^2)} = B + \tfrac{1}{2}\|\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}_{\mathrm{LS}}\|^2 + B_{q(\sigma^2)}p/(2A+n) + O(\sigma_\beta^{-2}).$$

Solving for $B_{q(\sigma^2)}$ we obtain the stated convergence for $B_{q(\sigma^2)}$ provided $2A + n > p$ (to insure positivity of $B_{q(\sigma^2)}$). The stated limit for $\boldsymbol{\Sigma}_{q(\boldsymbol{\beta})}$ is then obtained by substituting the limit for $B_{q(\sigma^2)}$ into (5).

□

The above result is useful as a caution against using diffuse priors for $\boldsymbol{\beta}$ in situations where $2A + n < p$. From Result 1, we see that when σ_β^2 is large it is essential that we require that $2A + n > p$. Otherwise Algorithm 1 will not converge. This is consistent with our empirical experience.

3.1 Theory

Henceforth we will treat y_i and $\mathbf{x}_i$ as random quantities, where $y_i|\mathbf{x}_i \sim N(\mathbf{x}_i^T\boldsymbol{\beta}_0, \sigma_0^2)$ for some true vector of coefficients $\boldsymbol{\beta}_0$ and variance σ_0^2. The commonly used unbiased estimators for $\boldsymbol{\beta}_0$ and σ_0^2 are

$$\boldsymbol{\beta}_{\text{LS}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} \quad \text{and} \quad \sigma^2_{\text{unbiased}} = \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}_{\text{LS}}\|^2/(n-p).$$

Note, $\text{Cov}(\boldsymbol{\beta}_{\text{LS}}) = \sigma_0^2(\mathbf{X}^T\mathbf{X})^{-1}$. Result 1 suggests the VB based estimators $\boldsymbol{\beta}_{\text{VB}} = \boldsymbol{\mu}_{q(\boldsymbol{\beta})}$ and $\sigma^2_{\text{VB}} = \mathbb{E}_q(\sigma^2) = B_{q(\sigma^2)}/(A + n/2 - 1)$ may have reasonable properties. Note, using Result 1, as $\sigma_\beta^2 \to \infty$ we have $\boldsymbol{\beta}_{\text{VB}} = \boldsymbol{\beta}_{\text{LS}} + O(\sigma_\beta^{-2})$ and

$$\sigma^2_{\text{VB}} = \frac{2B + (n-p)\sigma^2_{\text{unbiased}}}{2A + n - p - 2\left(1 - \frac{p}{2A+n}\right)} + O(\sigma_\beta^{-2}).$$

We notice that as $\sigma_\beta^2 \to \infty$ and $n \to \infty$ that σ^2_{VB} approaches $\sigma^2_{\text{unbiased}}$ in probability. Also, as $\sigma_\beta^2 \to \infty$, $A \to 0$ and $B \to 0$ we have $\boldsymbol{\Sigma}_{q(\boldsymbol{\beta})}$ approaching $\sigma^2_{\text{unbiased}}(\mathbf{X}^T\mathbf{X})^{-1}$, which can be used for estimating standard errors for $\boldsymbol{\beta}$. Thus, Result 1 suggests that the estimators $\boldsymbol{\beta}_{\text{VB}}$ and σ^2_{VB} may have good frequentist properties. In order to establish such properties we use the following assumptions:

(A1) For $1 \le i \le n$ the $y_i = \mathbf{x}_i^T\boldsymbol{\beta}_0 + \varepsilon_i$ where ε_i are independent $N(0, \sigma_0^2)$ where $\boldsymbol{\beta}_0$ and $0 < \sigma_0^2 < \infty$ are the true values of $\boldsymbol{\beta}$ and σ^2 respectively with $\boldsymbol{\beta}_0$ being element-wise finite;

(A2) For $1 \le i \le n$ the random vectors $\mathbf{x}_i \in \mathbb{R}^p$ are independent and identically distributed with p fixed;

(A3) The $p \times p$ matrix $\mathbb{E}(\mathbf{x}_i\mathbf{x}_i^T)$ is element-wise finite and positive definite and $\mathbf{X}^T\mathbf{X}$ is positive definite for all finite n; and

(A4) For $1 \le i \le n$ the random vectors $\mathbf{x}_i$ and random variables ε_i are independent.

Let $\mathbf{U}\text{diag}(\boldsymbol{\lambda})\mathbf{U}^T$ be the eigenvalue decomposition of $\mathbf{X}^T\mathbf{X}$ where $\mathbf{U}$ is an orthonormal matrix and $\boldsymbol{\lambda} = [\lambda_1, \ldots, \lambda_p]^T$ is the vector of eigenvalues with $\lambda_i > 0$ for $i = 1, \ldots, p$. Also, let

$$\mathbf{A}_n = n^{-1}\mathbf{X}^T\mathbf{X}, \quad \mathbf{b}_n = n^{-1}\mathbf{X}^T\mathbf{y},$$
$$c_n = \text{tr}(\mathbf{A}_n(\mathbf{A}_n + \sigma_\beta^{-2}n^{-1}d_n\mathbf{I})^{-1}) \quad \text{and} \quad d_n = B_{q(\sigma^2)}/(A + n/2).$$

Properties of A_n, b_n, c_n and d_n are described in Proposition 1.

Proposition 1: *Assuming (A1)–(A4) we have*

(a) $\mathbf{A}_n \xrightarrow{\text{a.s.}} \mathbb{E}(\mathbf{x}_i\mathbf{x}_i^T)$ *and* $\mathbf{b}_n \xrightarrow{\text{a.s.}} \mathbb{E}(\mathbf{x}_i\mathbf{x}_i^T)\boldsymbol{\beta}_0$,
(b) the estimator $\boldsymbol{\beta}_{\text{LS}} \xrightarrow{\text{a.s.}} \boldsymbol{\beta}_0$,
(c) the sequence of random variables c_n *satisfies* $c_n \leq p$ *for all* n, *and*
(d) $d_n = O_p(1)$ and $d_n^{-1} = O_p(1)$.

The proof of Proposition 1 is postponed to the Appendix. The next result establishes the consistency of the VB estimator.

Result 2: *Assuming (A1)–(A4) the estimator* $\boldsymbol{\beta}_{VB}$ *is a consistent estimator of* $\boldsymbol{\beta}$ *and* σ^2_{VB} *is a consistent estimator of* σ_0^2.

Proof: Firstly, $\boldsymbol{\beta}_{\text{VB}}$ may be rewritten as $\boldsymbol{\beta}_{\text{VB}} = (\mathbf{A}_n + \sigma_\beta^{-2} n^{-1} d_n \mathbf{I})^{-1}\mathbf{b}_n$. Using Proposition 1(d) the term d_n is $O_p(n^{-1})$ and so the term $\sigma_\beta^{-2} n^{-1} d_n$ is $O_p(n^{-1})$ and hence negligible. Since almost sure convergence implies convergence in probability we have, $\boldsymbol{\beta}_{\text{VB}} \xrightarrow{\text{P}} [\mathbb{E}(\mathbf{x}_i\mathbf{x}_i^T)]^{-1}\mathbb{E}(\mathbf{x}_i\mathbf{x}_i^T)\boldsymbol{\beta}_0 = \boldsymbol{\beta}_0$. Consequently, $\boldsymbol{\beta}_{\text{VB}}$ is a consistent estimator of $\boldsymbol{\beta}$. Secondly, we may rewrite σ^2_{VB} as

$$\sigma^2_{\text{VB}} = \frac{2B}{2A+n-2} + \left(\frac{n-p}{2A+n-2}\right)\frac{\|\mathbf{y} - \mathbf{X}\boldsymbol{\mu}_{q(\boldsymbol{\beta})}\|^2}{n-p} + \frac{c_n d_n}{2A+n-2}. \tag{6}$$

Using Proposition 1(c–d) the first and last terms on the right hand side of (6) are $O(n^{-1})$ and $O_p(n^{-1})$ respectively. Finally, $\|\mathbf{y} - \mathbf{X}\boldsymbol{\mu}_{q(\boldsymbol{\beta})}\|^2/(n-p) \xrightarrow{\text{P}} \sigma_0^2$ since $\boldsymbol{\mu}_{q(\boldsymbol{\beta})}$ is a consistent estimator for $\boldsymbol{\beta}_0$. Hence, the second term on the right hand side of (6) approaches σ_0^2 in probability and the result follows.

□

4 Variational Information Criteria

In this section, we introduce two variational Bayes information criteria: VAIC and VBIC, and establish the first order asymptotic properties of these two information criteria. The VAIC is a variational Bayes approximation to the DIC, we show that it shares the asymptotic properties of the AIC under mild regularity conditions. We also show that the proposed VBIC is a variational Bayes based analogue of Bayesian information criterion. As a consequence the model selection criterion VAIC selects, as the AIC does a model which is minimax rate optimal for selecting the regression function and VBIC tends to select the same linear regression model as the BIC (Yang, 2005).

4.1 Variational Akaike Information Criterion

A popular criterion for scoring individual models within a Bayesian context is the DIC introduced by Spiegelhalter et al (2002) and can be viewed as a hierarchical modeling generalization of AIC. It is defined as

$$\text{DIC} \equiv -2\log p(\mathbf{y}|\widetilde{\boldsymbol{\theta}}) + 2P_D,$$

where $\boldsymbol{\theta}$ is a vector of parameters, $\widetilde{\boldsymbol{\theta}}$ is a Bayesian estimator for $\boldsymbol{\theta}$, e.g., $\widetilde{\boldsymbol{\theta}} = \mathbb{E}(\boldsymbol{\theta}|\mathbf{y})$ and $P_D = 2\log p(\mathbf{y}|\widetilde{\boldsymbol{\theta}}) - 2\mathbb{E}_{\boldsymbol{\theta}|\mathbf{y}}[\log p(\mathbf{y}|\boldsymbol{\theta})]$.

Smaller values of DIC are preferable. The first term in DIC represents a measure of goodness of fit for the model, whereas the second term is a penalty for model complexity whose purpose is to prevent overfitting. The DIC can be useful for comparing models when improper priors are employed. Explicit calculation of the DIC requires the knowledge of the posterior distribution, which is often difficult to obtain exactly.

Instead, following McGrory and Titterington (2007), we approximate the DIC by replacing $p(\boldsymbol{\theta}|\mathbf{y})$ with $q(\boldsymbol{\theta})$ and call the result the variational Akaike information criterion (VAIC), i.e.,

$$\text{VAIC} \equiv -2\log p(\mathbf{y}|\boldsymbol{\theta}^*) + 2P_D^*, \tag{7}$$

where $\boldsymbol{\theta}^* = \mathbb{E}_q(\boldsymbol{\theta})$ and $P_D^* = 2\log p(\mathbf{y}|\boldsymbol{\theta}^*) - 2\mathbb{E}_q[\log p(\mathbf{y}|\boldsymbol{\theta})]$.

As VAIC is an approximation to DIC, smaller values of VAIC are preferable. Note, for comparative purposes, that for the classical linear model the AIC is given by

$$\text{AIC} \equiv -2\log p(\mathbf{y}|\widehat{\boldsymbol{\theta}}_{\text{ML}}) + 2P,$$

where $P = p + 1$ and the maximum likelihood estimates are $\widehat{\boldsymbol{\beta}}_{\text{ML}} \equiv \widehat{\boldsymbol{\beta}}_{\text{LS}}$ and $\widehat{\sigma}^2_{\text{ML}} \equiv n^{-1}\|\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}_{\text{LS}}\|^2$.

Proposition 2: *Assuming (A1)–(A4) and $n \to \infty$ we have*

(a) $\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}_{\text{LS}}\|^2 / B_{q(\sigma^2)} \xrightarrow{P} 2$ *and*

(b) $n\log\left(\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}_{\text{LS}}\|^2 / 2B_{q(\sigma^2)}\right) \xrightarrow{P} 0$.

The proof of Proposition 2 is postponed to the Appendix. The theorem below establishes the asymptotic behavior of the VAIC.

Theorem 1: *Let AIC and VAIC be defined as above. Then assuming (A1)–(A4) and as B approaches 0 we have $P_D^* \xrightarrow{\text{P}} P$ and* $\text{VAIC} \xrightarrow{\text{P}} \text{AIC}$.

Proof: Note that VAIC − AIC simplifies to

$$\begin{aligned}\text{VAIC} - \text{AIC} = -n\log\left(\frac{\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}_{\text{LS}}\|^2}{2B_{q(\sigma^2)}}\right) + (n + 2A - 2)\left(\frac{\|\mathbf{y} - \mathbf{X}\boldsymbol{\mu}_{q(\boldsymbol{\beta})}\|^2}{2B_{q(\sigma^2)}}\right) \\ -n - n\log\left(1 + \tfrac{2A-2}{n}\right) + 2(P_D^* - P),\end{aligned} \tag{8}$$

where P_D^* simplifies to,

$$P_D^* = c_n + n\left[\log\left(A + \tfrac{n}{2} - 1\right) - \psi\left(A + \tfrac{n}{2}\right)\right] + \frac{\|\mathbf{y} - \mathbf{X}\boldsymbol{\mu}_{q(\boldsymbol{\beta})}\|^2}{B_{q(\sigma^2)}}$$

and $\psi(x) = d\log\Gamma(x)/dx$ is the digamma function. From Proposition 1(a) and Proposition 1(c), the first term

$$c_n = \text{tr}\left(\frac{\mathbf{X}^T\mathbf{X}}{n}\left(\frac{\mathbf{X}^T\mathbf{X}}{n} + (\sigma_\beta^{-2}d_n/n)\mathbf{I}\right)^{-1}\right) \xrightarrow{\text{P}} \text{tr}(\mathbb{E}(\mathbf{x}_i\mathbf{x}_i^T)(\mathbb{E}(\mathbf{x}_i\mathbf{x}_i^T))^{-1}) = p,$$

while the second term in P_D^* approaches -1 since $\psi(x) = \log(x) - 1/(2x) + O(x^{-2})$ (see Abramowitz and Stegun, 1964, Formula 6.3.18). From proposition 2(a), the third term in P_D^* approaches 2 in probability. Hence, P_D^* approaches $P = p + 1$ in probability.

Next, using L'Hopital's rule, we have $\lim_{n\to\infty}[n\log(1+(2A-2)/n)] \to 2A-2$. Together with Proposition 2(a), we are able to show that $(n+2A-2)\|\mathbf{y} - \mathbf{X}\boldsymbol{\mu}_{q(\boldsymbol{\beta})}\|^2/\big(2B_{q(\sigma^2)}\big) - n - n\log(1+(2A-2)/n) \xrightarrow{\text{P}} 0$. Proposition 2(b) shows the first term in (8) approaching 0 in probability. Hence, VAIC $-$ AIC $\xrightarrow{\text{P}} 0$.

□

Note that the VAIC is not uniquely defined as in equation (7), it depends on the type of variation approximation used and also depends on the way of marginalization, i.e., we can marginalize out part of $\boldsymbol{\theta}$ analytically and use variational approximation to approximate the remaining elements of $\boldsymbol{\theta}$.

4.2 Variational Bayesian Information Criterion

A Bayesian model selection procedure chooses the model which is posteriorly most likely, hence the marginal likelihood $p(\mathbf{y})$ can be used to construct a selection criterion. As it is computational intractable most of the time, the Bayesian information criterion (BIC) is derived to approximate $-2\log p(\mathbf{y})$. BIC is one of the most popular choices for the consistent selection of an optimal model among a set of potential models. It is defined as

$$\text{BIC} \equiv -2\log p(\mathbf{y}|\widehat{\boldsymbol{\theta}}_{\text{ML}}) + P\log(n),$$

where $P, \widehat{\boldsymbol{\beta}}_{\text{ML}}$ and $\widehat{\sigma}^2_{\text{ML}}$ are defined as in Section 4. By using the Laplace approximation method, we can obtain

$$-2\log p(\mathbf{y}) + 2\log p(\widehat{\boldsymbol{\theta}}) = \text{BIC} - p\log(2\pi) + \log\| - n^{-1}\mathbf{I}(\widehat{\boldsymbol{\theta}})\| + O(n^{-1}), \quad (9)$$

where $\mathbf{I}(\boldsymbol{\theta}) = \partial^2 \log p(\mathbf{y}|\boldsymbol{\theta})/\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T$. Note that under mild condition, the second and third terms on the right-hand-side of equation (9) are $O_p(1)$ while the first terms on both sides of equation (9) are $O_p(n)$. See Claeskens and Hjort (2008) and Pauler (1998) for more details of the derivation of BIC.

Motivated from equation (9), we define the variational Bayesian information criterion (VBIC) as

$$\text{VBIC} \equiv -2\mathbb{E}_q \log \underline{p}_q(\mathbf{y}) + 2\mathbb{E}_q \log p(\boldsymbol{\theta}).$$

The advantage of having this definition instead of $-2\mathbb{E}_q \log p(\mathbf{y}, \boldsymbol{\theta}) + P\log(n)$ is even when P and n are not clearly defined, for example when missing data is present, the VBIC can still be used. Next, we establish the first order asymptotic behavior of VBIC.

Theorem 2: *Let BIC and VBIC be defined as above, assuming (A1)–(A4) we have VBIC* $=$ *BIC* $+ O_p(1)$.

Proof: First note that $\log\Gamma(x) = x\log(x) - x - (1/2)\log(x) + (1/2)\log(2\pi) + O(x^{-1})$ (Erdélyi et al, 1981). Also note that $\log(A + n/2) = \log(n) - \log(2) + O(n^{-1})$. Therefore, we can obtain

$$\begin{aligned}
\text{VBIC} &= -p + n\log(2\pi) - \log|\boldsymbol{\Sigma}_{q(\boldsymbol{\beta})}| + (n-2)\log(B_q) - 2\log\Gamma\left(A + \tfrac{n}{2}\right) \\
&\quad -p\log(2\pi) + (2A+2)\psi(A + \tfrac{n}{2}) - 2B\tfrac{A+n/2}{B_q} \\
&= -p + n\log(2\pi) + p\log(n) + \log|d_n(A_n + \sigma_\beta^{-2}n^{-1}d_n\mathbf{I})| \\
&\quad +(n-2)\log(B_q) - 2\big((A+\tfrac{n}{2})\log(A+\tfrac{n}{2}) - (A+\tfrac{n}{2}) - \tfrac{1}{2}\log(A+\tfrac{n}{2}) \\
&\quad +\tfrac{1}{2}\log(2\pi) + O(n^{-1})\big) - p\log(2\pi) + (2A+2)\big(\log(A+\tfrac{n}{2}) \\
&\quad +O(n^{-1})\big) - 2B\,d_n^{-1}.
\end{aligned}$$

Note that $\boldsymbol{\Sigma}_{q(\boldsymbol{\beta})} = n^{-1}d_n\big(A_n + \sigma_\beta^2 n^{-1}d_n\mathbf{I}\big)^{-1} = O_p(n^{-1})$ as $d_n = O_p(1)$ from Proposition 1(d). Hence $\log|\boldsymbol{\Sigma}_{q(\boldsymbol{\beta})}| = -p\log(n) + O_p(1)$ and $\text{tr}(\boldsymbol{\Sigma}_{q(\boldsymbol{\beta})}) = O_p(n^{-1})$. Next, we eliminate $O_p(n^{-1})$ terms to obtain

$$\begin{aligned}
\text{VBIC} &= n\log(2\pi) + p\log(n) + \log(A + \tfrac{n}{2}) + n + (n-2)\log(B_q) \\
&\quad -(n-2)\log(A+\tfrac{n}{2}) + O_p(1) \\
&= n\log(2\pi) + P\log(n) + n + (n-2)\log(B_q) - (n-2)\log(A+\tfrac{n}{2}) \\
&\quad +O_p(1).
\end{aligned}$$

Comparing the difference

$$\begin{aligned}
\text{BIC} - \text{VBIC} &= n\log(2\pi) + n\log(\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}_{\text{LS}}\|^2) - n\log(n) + n + P\log(n) \\
&\quad -n\log(2\pi) - P\log(n) - n - (n-1)\log(B_q) \\
&\quad +(n-1)\log(A+\tfrac{n}{2}) + O_p(1) \\
&= n\log\left(\tfrac{\|\mathbf{y}-\mathbf{X}\boldsymbol{\beta}_{\text{LS}}\|^2}{2B_q}\right) - n\log\left(\tfrac{n}{2A+n}\right) + \log(d_n) + O_p(1). \quad (10)
\end{aligned}$$

From Proposition 2(b) we see that the first term in equation (10) approaches to 0. The second term converges to $2A$. Using Proposition 1(d), the third term is $O_p(1)$ and the result follows.

□

5 Numerical Example

We illustrate our theoretical findings through simple numerical examples. Consider a linear model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

where $\boldsymbol{\beta} \in \mathbb{R}^5$. Each column of $\mathbf{X}$ is a standardized pseudo random vector from the standard normal distribution. We consider 6 simulation settings: the sample size n varies over 10, 100 and 1000 and the hyperparameter B varies over 0.1 and 1×10^{-8}. We keep the hyperparameters $A = 0.01$, $\sigma_\beta^2 = 10^8$ and the coefficient vector $\boldsymbol{\beta} = [1,1,1,1,1]^T$ constant. Each scenario is repeated 100

times. The absolute difference is used to measure the difference between VAIC and AIC. The quantities $\|\boldsymbol{\mu}_{q(\boldsymbol{\beta})} - \hat{\boldsymbol{\beta}}_{\text{LS}}\|^2$ and $\|\boldsymbol{\Sigma}_{q(\beta)} - \hat{\sigma}^2_{\text{unbiased}}(\mathbf{X}^T\mathbf{X})^{-1}\|_\infty$ are used to show empirically that the VB estimators are consistent and able to obtain valid standard errors. The term $|\text{VBIC} - \text{BIC}|$ is used to measure the difference between VBIC and BIC. The means and standard errors of the above 4 measures are shown in Table 1–4. In Table 1, the values are getting closer to 0 from top to bottom and when changing from $B = 0.1$ to 1×10^{-8}, which is consistent with Theorem 1. The values in Table 2 (a) and (b) both decrease with increasing sample size n. This supports Result 2 in Section 3. The values in Table 3 also decrease as n increases which is evidence that the VB estimates have valid standard errors. The average differences are almost constant but smaller standard error with increasing n in Table 4. This means the difference does not vary with n which is consistent with Theorem 2. Note in Table 2, 3 and 4 the hyperparameter B does not change the results.

(a) $B = 0.1$

n	Mean	Standard Error
10	9.76×10^{-1}	1.83×10^{-2}
100	2.17×10^{-2}	1.46×10^{-4}
1000	1.73×10^{-3}	4.81×10^{-6}

(b) $B = 10^{-8}$

n	Mean	Standard Error
10	7.51×10^{-1}	1.17×10^{-3}
100	1.48×10^{-2}	3.91×10^{-5}
1000	1.11×10^{-3}	2.69×10^{-6}

Table 1 $|\text{VAIC} - \text{AIC}|$

(a) $B = 0.1$

n	Mean	Standard Error
10	2.38×10^{-22}	8.01×10^{-23}
100	6.41×10^{-26}	2.34×10^{-27}
1000	5.54×10^{-28}	6.59×10^{-30}

(b) $B = 10^{-8}$

n	Mean	Standard Error
10	2.27×10^{-22}	7.78×10^{-23}
100	6.38×10^{-26}	2.33×10^{-27}
1000	5.54×10^{-28}	6.75×10^{-30}

Table 2 $||\boldsymbol{\mu}_{q(\boldsymbol{\beta})} - \hat{\boldsymbol{\beta}}_{\text{LS}}||^2$

(a) $B = 0.1$

n	Mean	Standard Error
10	5.38×10^{-2}	6.90×10^{-4}
100	3.23×10^{-5}	1.13×10^{-7}
1000	1.96×10^{-7}	3.48×10^{-10}

(b) $B = 10^{-8}$

n	Mean	Standard Error
10	7.41×10^{-3}	6.82×10^{-4}
100	3.50×10^{-6}	1.08×10^{-7}
1000	2.19×10^{-8}	3.47×10^{-10}

Table 3 $\|\boldsymbol{\Sigma}_{q(\beta)} - \sigma^2_{\text{unbiased}}(\mathbf{X}^T\mathbf{X})^{-1}\|_\infty$

(a) $B = 0.1$

n	Mean	Standard Error
10	8.883	4.218×10^{-1}
100	8.833	1.110×10^{-1}
1000	8.890	3.222×10^{-2}

(b) $B = 10^{-8}$

n	Mean	Standard Error
10	8.819	4.213×10^{-1}
100	8.828	1.111×10^{-1}
1000	8.890	3.222×10^{-2}

Table 4 $|\text{VBIC} - \text{BIC}|$

6 Conclusion

This article shows that for the Bayesian linear model presented here the corresponding variational Bayes based estimators $\boldsymbol{\beta}_{\text{VB}}$ and σ^2_{VB} are consistent estimators of $\boldsymbol{\beta}_0$ and σ_0^2 under mild regularity conditions. This finding partially contradicts the criticism that variational Bayes is not used for statistical inferences. Furthermore, it is proved that the variational Akaike information criterion shares the same the first order asymptotic properties as the Akaike information criterion and variational Bayesian information criterion shares the same the first order asymptotic properties as the Bayesian information criterion. While the results concern the well understood linear regression model they do represent an advancement in the emerging area of variational Bayes. Our results also give some motivation for the justification of variational Bayes based information criteria for more complex models.

Acknowledgements This research was partially supported by Australian Research Council Discovery Project DP110100061 (JTO) and Australian Research Council Discovery Project DP110101998 (SM). We thank Michael Stewart for helpful discussions.

Appendix

Proof of Proposition 1(a): The stated convergence of $\mathbf{A}_n$ follows from (A2) and (A3) and the strong law of large numbers. Similarly, assuming also (A1) and (A4), $\mathbf{b}_n \xrightarrow{\text{a.s.}} \mathbb{E}(\mathbf{x}_i y_i) = \mathbb{E}_X[\mathbf{x}_i(x_i^T\boldsymbol{\beta}_0 + \epsilon_i)] = \mathbb{E}(\mathbf{x}_i\mathbf{x}_i^T)\boldsymbol{\beta}_0$.

□

Proof of Proposition 1(b): Note $\boldsymbol{\beta}_{\text{LS}} = \mathbf{A}_n^{-1}\mathbf{b}_n$. Using Proposition 1(a) obtains the stated result.

□

Proof of Proposition 1(c): Let $\alpha = \sigma_\beta^{-2} d_n$. Note that $c_n = \text{tr}(\mathbf{X}^T\mathbf{X}(\mathbf{X}^T\mathbf{X} + \alpha\mathbf{I})^{-1}) = \sum_{i=1}^p \lambda_i/(\lambda_i + \alpha) = \sum_{i=1}^p (1 - \alpha/(\lambda_i + \alpha)) \leq p$ since $\lambda_i > 0$ for $i = 1, \ldots, p$ and $\alpha > 0$.

□

Proof of Proposition 1(d): Firstly, note that $B_{q(\sigma^2)}$ satisfies $B_{q(\sigma^2)} = B + \|\mathbf{y} - \mathbf{X}\boldsymbol{\mu}_{q(\boldsymbol{\beta})}\|^2/2 + B_{q(\sigma^2)}c_n/(2A + n)$, hence

$$B_{q(\sigma^2)} = \frac{B + \frac{1}{2}\|\mathbf{y} - \mathbf{X}\boldsymbol{\mu}_{q(\boldsymbol{\beta})}\|^2}{1 - c_n/(2A + n)}. \tag{11}$$

Using Proposition 1(c) and the triangle inequality,

$$d_n = \frac{2B + \|\mathbf{y} - \mathbf{X}\boldsymbol{\mu}_{q(\boldsymbol{\beta})}\|^2}{2A + n - c_n} \leq \frac{2B + \|\mathbf{y}\|^2 + \|\mathbf{X}\boldsymbol{\mu}_{q(\boldsymbol{\beta})}\|^2}{2A + n - p}.$$

Next consider,

$$\|\mathbf{X}\boldsymbol{\mu}_{q(\boldsymbol{\beta})}\|^2 = \mathbf{y}^T\mathbf{X}(\mathbf{X}^T\mathbf{X}+\alpha\mathbf{I})^{-1}\mathbf{X}^T\mathbf{X}(\mathbf{X}^T\mathbf{X}+\alpha\mathbf{I})^{-1}\mathbf{X}^T\mathbf{y} = \sum_{i=1}^{p}\frac{v_i^2\lambda_i}{(\lambda_i+\alpha)^2}$$

where $[v_1,\ldots,v_p]^T = \mathbf{U}^T\mathbf{X}^T\mathbf{y}$. However, $\|\mathbf{X}\boldsymbol{\mu}_{q(\boldsymbol{\beta})}\|^2$ is clearly maximized with respect to α when $\alpha = 0$. Hence,

$$\begin{aligned} d_n &\le \frac{2B}{2A+n-p} + \frac{n}{2A+n-p}\left(\frac{\|\mathbf{y}\|^2}{n} + \boldsymbol{\beta}_{\text{LS}}^T\mathbf{A}_n\boldsymbol{\beta}_{\text{LS}}\right) \\ &= \frac{2B}{2A+n-p} + \frac{n}{2A+n-p}\left(\frac{\|\boldsymbol{\varepsilon}\|^2+2\boldsymbol{\varepsilon}^T\mathbf{X}\boldsymbol{\beta}_0+\boldsymbol{\beta}_0^T\mathbf{X}^T\mathbf{X}\boldsymbol{\beta}_0}{n} + \boldsymbol{\beta}_{\text{LS}}^T\mathbf{A}_n\boldsymbol{\beta}_{\text{LS}}\right), \end{aligned}$$

where $\boldsymbol{\varepsilon} = [\varepsilon_1,\ldots,\varepsilon_n]^T$. Now, using assumptions (A1)–(A4) and the strong law of large numbers we have $\|\boldsymbol{\varepsilon}\|^2/n \xrightarrow{\text{a.s.}} \sigma_0^2$, $\boldsymbol{\varepsilon}^T\mathbf{X}\boldsymbol{\beta}_0/n \xrightarrow{\text{a.s.}} 0$ (due to the independence of ε_i and $\mathbf{x}_i$), $\boldsymbol{\beta}_0^T\mathbf{X}^T\mathbf{X}\boldsymbol{\beta}_0/n \xrightarrow{\text{a.s.}} \boldsymbol{\beta}_0^T\mathbb{E}(\mathbf{x}_i\mathbf{x}_i^T)\boldsymbol{\beta}_0$, and using Proposition 1(a) and Proposition 1(b) we have $\boldsymbol{\beta}_{\text{LS}}^T\mathbf{A}_n\boldsymbol{\beta}_{\text{LS}} \xrightarrow{\text{a.s.}} \boldsymbol{\beta}_0^T\mathbb{E}(\mathbf{x}_i\mathbf{x}_i^T)\boldsymbol{\beta}_0 + p\sigma_0^2$. The result follows from almost sure convergence implying convergence in probability.

It is known that $||\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}}||^2 \le ||\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}}_{LS}||^2$ for any choice of $\hat{\boldsymbol{\beta}}$,

$$d_n^{-1} = \frac{2A+n-c_n}{2B+\|\mathbf{y}-\mathbf{X}\boldsymbol{\mu}_{q(\boldsymbol{\beta})}\|^2} \le \frac{2A+n}{2B+\|\mathbf{y}-\mathbf{X}\boldsymbol{\beta}_{LS}\|^2} = \frac{2A+n}{2B+||\epsilon||^2} \xrightarrow{\text{a.s.}} \sigma_0^{-2}.$$

So $d_n^{-1} = O_p(1)$.

□

Proof of Proposition 2(a): Note $\|\mathbf{y}-\mathbf{X}\boldsymbol{\mu}_{q(\boldsymbol{\beta})}\|^2/n \xrightarrow{\text{P}} \sigma_0^2$ since $\boldsymbol{\mu}_{q(\boldsymbol{\beta})} \xrightarrow{\text{P}} \boldsymbol{\beta}_0$ and $B_q = (A+n/2-1)\sigma_{\text{VB}}^2$. Combining this with Result 2 we have $(n/(A+n/2-1))(\|\mathbf{y}-\mathbf{X}\boldsymbol{\mu}_{q(\boldsymbol{\beta})}\|^2/n\sigma_{\text{VB}}^2) \xrightarrow{\text{P}} 2$.

□

Proof of Proposition 2(b): First note that $\log(t) = (t-1) - (t-1)^2/2 + O((t-1)^3)$. Then consider

$$n\log\left(\frac{\|\mathbf{y}-\mathbf{X}\boldsymbol{\beta}_{\text{LS}}\|^2}{2B_{q(\sigma^2)}}\right) = n\left(\frac{\|\mathbf{y}-\mathbf{X}\boldsymbol{\beta}_{\text{LS}}\|^2}{2B_{q(\sigma^2)}} - 1\right) + nO\left(\frac{\|\mathbf{y}-\mathbf{X}\boldsymbol{\beta}_{\text{LS}}\|^2}{2B_{q(\sigma^2)}} - 1\right)^2.$$

Note that if the first term in the expansion converges to zero as n diverges then so will higher order terms which we may rewrite as

$$n\frac{\|\mathbf{y}-\mathbf{X}\boldsymbol{\beta}_{\text{LS}}\|^2 - 2B_{q(\sigma^2)}}{2B_{q(\sigma^2)}} = \frac{n}{A+n/2}\left(\frac{\|\mathbf{y}-\mathbf{X}\boldsymbol{\beta}_{\text{LS}}\|^2 - \frac{2B+\|\mathbf{y}-\mathbf{X}\boldsymbol{\mu}_{q(\boldsymbol{\beta})}\|^2}{1-c_n/(2A+n)}}{2d_n}\right), \tag{12}$$

by using equation (11) on the numerator and $B_q = (A+n/2)d_n$ on the denominator. Applying Proposition 1(c) we have $c_n/(2A+n) \le p/(2A+n)$ approaching 0 as $n \to \infty$. Then using Proposition 1(b) and Result 2 and as $B \to 0$ we have (12) approaching 0.

□

References

Abramowitz M, Stegun IA (1964) Handbook of Mathematical Functions. Dover, New York

Akaike H (1973) Information theory and an extension of the maximum likelihood principle. In: Petrov BN, Csaki F (eds) In Proceedings of the 2nd International Symposium on Information Theory, Akademiai Kiad6, Budapest, pp 267–281

Bishop CM (2006) Pattern Recognition and Machine Learning. Springer, New York

Claeskens G, Hjort NL (2008) Model selection and model averaging. Cambridge Series in Statistical and Probabilistic Mathematics, Cambridge University Press, Cambridge

Erdélyi A, Magnus W, Oberhettinger F, Tricomi FG (1981) Higher transcendental functions. Vol. III. Robert E. Krieger Publishing Co. Inc., Melbourne, Fla., based on notes left by Harry Bateman, Reprint of the 1955 original

Hall P, Ormerod JT, Wand MP (2011a) Theory of Gaussian variational approximation for a Poisson mixed model. Statistica Sinica 21:369–389

Hall P, Pham T, Wand MP, Wang SSJ (2011b) Asymptotic normality and valid inference for Gaussian variational approximation. The Annals of Statistics 39:2502–2532

Humphreys K, Titterington DM (2000) Approximate Bayesian inference for simple mixtures. In: Bethlehem JG, van der Heijden PGM (eds) Proceedings of Computational Statistics, Physica, Heidelberg, pp 2502–2532

McGrory CA, Titterington DM (2007) Variational approximations in Bayesian model selection for finite mixture distributions. Computational Statistics & Data Analysis 51:5352–5367

Ormerod JT, Wand MP (2010) Explaining variational approximations. The American Statistician 64:140–153

Ormerod JT, Wand MP (2012) Gaussian variational approximate inference for generalized linear mixed models. Journal of Computational & Graphical Statistics 21:2–17

Pauler DK (1998) The schwarz criterion and related methods for normal linear models. Biometrika 85:13–27

Ren Q, Banerjee S, Finley AO, Hodges JS (2011) Variational Bayesian methods for spatial data analysis. Computational Statistics & Data Analysis 55:3197–3217

Rue H, Martino S, Chopin N (2009) Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. J R Stat Soc Ser B Stat Methodol 71:319–392

Schwarz G (1978) Estimating the dimension of a model. Ann Statist 6(2):461–464

Spiegelhalter DJ, Best NG, Carlin BP, Van der Linde A (2002) Bayesian measures of model complexity and fit (with Discussion). Journal of the Royal Statistical Society, Series B 64:583–639

Volant S, Magniette ML, Robin S (2012) Variational Bayes approach for model aggregation in unsupervised classification with Markovian dependency. Computational Statistics and Data Analysis 56:2375–2387

Wang B, Titterington DM (2006) Convergence properties of a general algorithm for calculating variational Bayesian estimates for a normal mixture model. Bayesian Analysis 1:625–650

Yang Y (2005) Can the strengths of AIC and BIC be shared? A conflict between model indentification and regression estimation. Biometrika 92:937–50