Australian & New Zealand Journal of Statistics

Aust. N. Z. J. Stat. **56**(1), 2014, 73–87 doi: 10.1111/anzs.12063

ON VARIATIONAL BAYES ESTIMATION AND VARIATIONAL INFORMATION CRITERIA FOR LINEAR REGRESSION MODELS

CHONG YOU*, JOHN T. ORMEROD AND SAMUEL MÜLLER

University of Sydney

Summary

Variational Bayes (VB) estimation is a fast alternative to Markov Chain Monte Carlo for performing approximate Baesian inference. This procedure can be an efficient and effective means of analyzing large datasets. However, VB estimation is often criticised, typically on empirical grounds, for being unable to produce valid statistical inferences. In this article we refute this criticism for one of the simplest models where Bayesian inference is not analytically tractable, that is, the Bayesian linear model (for a particular choice of priors). We prove that under mild regularity conditions, VB based estimators enjoy some desirable frequentist properties such as consistency and can be used to obtain asymptotically valid standard errors. In addition to these results we introduce two VB information criteria: the variational Akaike information criterion and the variational Bayesian information criterion. We show that variational Akaike information criterion is asymptotically equivalent to the frequentist Akaike information criterion and that the variational Bayesian information criterion is first order equivalent to the Bayesian information criterion in linear regression. These results motivate the potential use of the variational information criteria for more complex models. We support our theoretical results with numerical examples.

Key words: Akaike information criterion; Bayesian information criterion; consistency; deviance information criterion; Markov Chain Monte Carlo.

1. Introduction

There is an ever increasing demand by society on the statistical community 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 (Rue, Martino & Chopin 2009; Volant, Magniette & Robin 2012). Variational approximations are a newly emerging class of alternatives to MCMC which may provide fast approximate Bayesian inference in such contexts.

Variational approximations are often criticised, typically on empirical grounds, for being unable to produce valid statistical inferences in several modelling contexts (again see, for example, Rue *et al.* 2009, section 1.6). Few theoretical developments about variational approximations have been made to prove or disprove such claims in general and the theory

*Author to whom correspondence should be addressed.
School of Mathematics and Statistics Carslaw Building (F07), University of Sydney, Sydney, NSW 2006, Australia.
e-mail: you@maths.usyd.edu.au

Acknowledgements. We are very thankful to the comments from the associate editor and the two referees on the original draft. These led to an improved presentation of the article. 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.

that does exist is context specific (Humphreys & Titterington 2000; Wang & Titterington 2006; Hall, Ormerod & Wand 2011; Hall *et al.* 2011; Ormerod & Wand 2012).

The main purpose of for this paper is to show that variational approximations can provide valid statistical inferences for a particular Bayesian linear model. In this article we focus on VB estimation (MacKay 1995; Bishop 2006), which is a special type of variational approximation, and build upon the growing body of theory in this area. For this Bayesian linear model we choose the coefficient prior to be independent of the response variance. We make this choice because this model is one of the simplest models where the marginal likelihood is *not analytically available*. Hence, while exact Bayesian inference is not possible for the Bayesian linear model we consider, it is still of theoretical interest to determine whether VB estimation provides valid statistical inferences.

In contrast, Ren *et al.* (2011, section 3) and Murphy (2012, chapter 21) consider the case where the coefficient prior depends, in a particular way, on the response variance. This choice facilitates analytic integration so that the marginal posterior distributions are available for the regression coefficients and response variance (see also, for example, Robert & Marin 2007). For this alternative Bayesian linear model the VB posterior approximations of these quantities can be shown to approach the true posterior distributions as the sample size increases. However, because for our model the marginal likelihood is not analytically available, we require different techniques to analyse our model.

Our prior is one of the simplest priors for the Bayesian linear model. Other more complicated priors, such as sparsity inducing priors, are also used in a Bayesian framework. Sparsity inducing priors ‘shrink’ some regression coefficients towards zero and hence lead to variable selection. The most common shrinkage priors, spike and slab priors (Mitchell & Beauchamp 1988), are discrete point mass mixture priors. They are natural priors to incorporate sparsity in a Bayesian framework and possess attractive theoretical properties. However they usually induce computational problems in high dimensions when marginal likelihoods are not analytically available. You, Ormerod & Müller (2013) incorporate spike and slab priors with a VB algorithm in the Bayesian linear regression model and show some desirable properties of the VB based estimators. Continuous shrinkage priors are alternatives to discrete-mixture priors. They have substantial computational advantages over discrete-mixture priors and can potentially deal with high dimensional datasets ($p \gg n$).

Moreover, a number of frequentist regularisation procedures are equivalent to calculating posterior modes under certain continuous shrinkage priors. For example, the Lasso (Tibshirani 1996) is equivalent to a double exponential prior. Continuous shrinkage priors have received little attention with regard to VB methodology. Notable exceptions are Armagan, Dunson & Clyde (2011) and Neville, Ormerod & Wand (2013). Neville, Ormerod & Wand (2013) develop algorithms to improve the performance of VB methods with regard to several continuous shrinkage priors. In this paper we do not compare the performance of other priors to the one we choose, but show that VB estimators in this simple Bayesian model can exhibit some desirable properties.

We prove that the VB estimators for our model enjoy desirable frequentist properties and can be used to obtain asymptotically valid standard errors. We consider the case where the design matrix is random (the fixed design case can be viewed as a special case of our results). Furthermore, our results can be considered quite general with regard to the distribution of the design matrix since they depend only on very mild regularity conditions on the distribution of the covariates (such as existence of the first and second moments).

After demonstrating these properties we develop two different variational information criteria for model selection. The first criterion is a VB approximation of the deviance information criterion (DIC) of Spiegelhalter *et al.* (2002), which we call the variational Akaike information criterion (VAIC). We show that the 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. The second variational information criterion we propose is a VB version of the Bayesian information criterion (BIC) (Schwarz 1978), which we call the variational Bayesian information criterion (VBIC). We show that the VBIC is first order equivalent to the BIC. In this article, rather than comparing these criteria to AIC, BIC or even to model selection using shrinkage priors, we show that the new VB based analogs of the frequentist AIC and BIC are sensible asymptotically in a linear regression context. These results give some motivation for the use of VB based information criteria for more complex models.

In Section 2 we briefly summarise VB estimation and describe how the method can be applied to a Bayesian linear model. Theory for VB estimators for increasingly diffuse priors and increasing sample size is presented in Section 3. Properties of the variational information criteria are derived in Section 4. A numerical example is given in Section 5. Section 6 presents the conclusion. The proofs of the propositions are postponed to the Appendix.

2. Variational Bayes estimation 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 minimise the Kullback–Leibler (KL) distance between $p(\boldsymbol{\theta}|\mathbf{y})$ and $q(\boldsymbol{\theta})$, satisfy

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

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

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

Finally, by iteratively calculating $q_i(\boldsymbol{\theta}_i)$ for fixed $\{q(\boldsymbol{\theta}_j)\}_{j \neq i}$ the lower bound increases after each iteration so that convergence to a local maximiser of $\log p(\mathbf{y})$ occurs under mild regularity conditions. For more details and examples see Beal (2003), Bishop (2006), or Ormerod & Wand (2010).

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

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

where $\mathbf{X}$ is a $n \times p$ design matrix whose ith row is $\mathbf{x}_i^\top$ and $\text{IG}(A, B)$ is the inverse Gamma distribution with shape parameter A and scale parameter B. The parameters $\sigma_{\boldsymbol{\beta}}^2$, A and B

are fixed prior hyperparameters. In contrast Ren *et al.* (2011) consider the case where $\boldsymbol{\beta}|\sigma^2 \sim N(\mathbf{0}, \sigma^2\mathbf{V}_{\boldsymbol{\beta}})$ for some constant positive definite matrix $\mathbf{V}_{\boldsymbol{\beta}}$ (for example $\mathbf{V}_{\boldsymbol{\beta}} = \sigma^2_{\boldsymbol{\beta}}\mathbf{I}$). There are several examples in the literature which implement the VB algorithm in slightly different Bayesian linear models. For example, the model in Murphy (2012, chapter 21) is a special case of the model in Ren *et al.* (2011), where only the intercept is included, i.e., $\beta_0|\sigma^2 \sim N(0, \kappa\sigma^2)$ and κ is fixed. Drugowitsch (2008) uses the same priors as in Murphy (2012, chapter 21) but the hyperparameter κ is assigned an inverse-gamma hyper-prior. MacKay (2003, chapter 33) uses improper priors.

If $\boldsymbol{\theta} = [\boldsymbol{\beta}^\top, \sigma^2]^\top$ 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

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

where

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

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

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

Note that (2)–(4) must hold simultaneously for the q-densities to be optimal. Algorithm 1 describes a process for finding these values. The lower bound p_{-q} $(\mathbf{y})$ which appears at the end of the main loop of Algorithm 1 below can be expressed as:

$$\begin{aligned}\log \underline{p}_q(\mathbf{y}) = {} & \frac{p}{2} - \frac{n}{2}\log(2\pi) - \frac{p}{2}\log(\sigma^2_\beta) + \frac{1}{2}\log|\boldsymbol{\Sigma}_{q(\boldsymbol{\beta})}| - \frac{\|\boldsymbol{\mu}_{q(\boldsymbol{\beta})}\|^2 + \text{tr}\big(\boldsymbol{\Sigma}_{q(\boldsymbol{\beta})}\big)}{2\sigma^2_\beta} \\ & + A\log(B) - \log\Gamma(A) - \Big(A + \frac{n}{2}\Big)\log(B_{q(\sigma^2)}) + \log\Gamma\Big(A + \frac{n}{2}\Big).\end{aligned}$$

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

```
Initialise: $B_{q(\sigma^2)} > 0$.
Cycle:
    $\boldsymbol{\Sigma}_{q(\boldsymbol{\beta})} \leftarrow \left[\left(\frac{A+n/2}{B_{q(\sigma^2)}}\right)\mathbf{X}^\top\mathbf{X} + \sigma_{\boldsymbol{\beta}}^{-2}\mathbf{I}\right]^{-1}$; $\boldsymbol{\mu}_{q(\boldsymbol{\beta})} \leftarrow \left(\frac{A+n/2}{B_{q(\sigma^2)}}\right)\boldsymbol{\Sigma}_{q(\boldsymbol{\beta})}\mathbf{X}^\top\mathbf{y}$
    $B_{q(\sigma^2)} \leftarrow B + \frac{1}{2}\|\mathbf{y} - \mathbf{X}\boldsymbol{\mu}_{q(\boldsymbol{\beta})}\|^2 + \frac{1}{2}\text{tr}(\mathbf{X}^\top\mathbf{X}\boldsymbol{\Sigma}_{q(\boldsymbol{\beta})})$
until the increase of $\underline{p}_q(\mathbf{y})$ is negligible.
```

3. Main results

Henceforth we assume that $\boldsymbol{\mu}_{q(\boldsymbol{\beta})}$, $\boldsymbol{\Sigma}_{q(\boldsymbol{\beta})}$ 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 (2)–(4) as σ_β^2 increases.

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

$$\boldsymbol{\Sigma}_{q(\boldsymbol{\beta})} = \left(\frac{2B + \|\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}_{\text{LS}}\|^2}{2A + n - p}\right)(\mathbf{X}^\top\mathbf{X})^{-1} + O(\sigma_\beta^{-2}),$$

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

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

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

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

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

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

Solving for $B_{q(\sigma^2)}$ we obtain the stated convergence result for $B_{q(\sigma^2)}$ provided that $2A + n > p$ (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 may 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 $y_i|\mathbf{x}_i \sim N(\mathbf{x}_i^\top\boldsymbol{\beta}_0, \sigma_0^2)$ for some 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

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

Note that $\text{Cov}(\boldsymbol{\beta}_{\text{LS}}|\mathbf{X}) = \sigma_0^2(\mathbf{X}^\top\mathbf{X})^{-1}$ can be used to calculate standard errors. Result 1 suggests that 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, that as $\sigma_\beta^2 \to \infty$ we have $\boldsymbol{\beta}_{\rm VB} = \boldsymbol{\beta}_{\rm LS} + O(\sigma_\beta^{-2})$ and

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

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

A1. For $1 \le i \le n$, $y_i = \mathbf{x}_i^\top \boldsymbol{\beta}_0 + \varepsilon_i$ where ε_i are independent $N(0, \sigma_0^2)$ and $\boldsymbol{\beta}_0$ 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 (and $p < n$);

A3. The $p \times p$ matrix $\mathbb{E}(\mathbf{x}_i\mathbf{x}_i^\top)$ is element-wise finite and $\mathbf{X} = [\mathbf{X}_1, \ldots, \mathbf{X}_p]$ where

$$P(\mathbf{a}^T\mathbf{X} = 0) < 1 \ \forall \mathbf{a} \ne \mathbf{0} \Rightarrow P(\text{rank}(\mathbf{X}) = p) \to 1 \text{ for } n \to \infty;$$

and

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

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

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

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

Proposition 1. *Assuming A1–A4 we have*

(a) $\mathbf{A}_n \overset{\text{a.s.}}{\to} \mathbb{E}(\mathbf{x}_i\mathbf{x}_i^\top)$ and $\mathbf{b}_n \overset{\text{a.s.}}{\to} \mathbb{E}(\mathbf{x}_i\mathbf{x}_i^\top)\boldsymbol{\beta}_0$,
(b) the estimator $\boldsymbol{\beta}_{\rm LS} \overset{\text{a.s.}}{\to} \boldsymbol{\beta}_0$,
(c) the sequence of random variables c_n satisfies $0 \le c_n \le p$ 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 $\boldsymbol{\beta}_{\rm VB}$ is a consistent estimator of $\boldsymbol{\beta}_{\rm o}$ and $\sigma_{\rm VB}^2$ is a consistent estimator of σ_0^2.

Proof. Firstly, $\boldsymbol{\beta}_{\rm VB}$ may be rewritten as $\boldsymbol{\beta}_{\rm VB} = (\mathbf{A}_n + \sigma_\beta^{-2}n^{-1}d_n\mathbf{I})^{-1}\mathbf{b}_n$. Using Proposition 1(d) the term d_n is $O_p(1)$ 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, $\boldsymbol{\beta}_{\mathrm{VB}} \xrightarrow{\mathrm{P}} [\mathbb{E}(\mathbf{x}_i\mathbf{x}_i^\top)]^{-1}\mathbb{E}(\mathbf{x}_i\mathbf{x}_i^\top)\boldsymbol{\beta}_0 = \boldsymbol{\beta}_0$. Consequently, $\boldsymbol{\beta}_{\mathrm{VB}}$ is a consistent estimator of $\boldsymbol{\beta}_{\mathrm{o}}$. Secondly, we may rewrite σ^2_{VB} as

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

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, $\|\mathbf{y}-\mathbf{X}\boldsymbol{\mu}_{q(\boldsymbol{\beta})}\|^2/(n-p) \xrightarrow{\mathrm{P}} \sigma_0^2$ 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 VB information criteria, VAIC and VBIC, and establish the first order asymptotic properties of these two criteria. The VAIC is a VB approximation to the DIC and we show that it shares the asymptotic properties of the AIC under mild regularity conditions. We also show that the proposed VBIC is a VB analogue of the BIC. 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 in a Bayesian context is the DIC introduced by Spiegelhalter *et al.* (2002). This criterion can be viewed as a hierarchical modelling generalisation of the AIC. It is defined as

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

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

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 knowledge of the posterior distribution, which is often difficult to obtain exactly.

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

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

where $\boldsymbol{\theta}^* = \mathbb{E}_q(\boldsymbol{\theta})$ and $P_D^* = 2\log p(\mathbf{y}|\boldsymbol{\theta}^*) - 2\mathbb{E}_{\mathbb{q}}[\log \mathbb{p}(\mathbf{y}|\boldsymbol{\theta})]$. Note that McGrory & Titterington (2007) do not call their approximation the DIC.

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

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

where $P = p+1$ and the maximum likelihood estimates are $\hat{\boldsymbol{\beta}}_{\mathrm{ML}} \equiv \hat{\boldsymbol{\beta}}_{\mathrm{LS}}$ and $\hat{\sigma}^2_{\mathrm{ML}} \equiv n^{-1}\|\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}_{\mathrm{LS}}\|^2$ with $\hat{\boldsymbol{\theta}}_{\mathrm{ML}} = [\hat{\boldsymbol{\beta}}_{\mathrm{ML}}^\top, \hat{\sigma}^2_{\mathrm{ML}}]^\top$.

Proposition 2. Assuming *A1–A4* and $n \to \infty$ we have

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

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, as B approaches 0 we have* $P_D^* \xrightarrow{\mathrm{P}} P$ *and* $\mathrm{VAIC} \xrightarrow{\mathrm{P}} \mathrm{AIC}$.

Proof. Note that VAIC − AIC simplifies to

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

where P_D^* simplifies to

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

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

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

while the second term in P_D^* approaches -1 since $\psi(x) = \log(x) - 1/(2x) + O(x^{-2})$ (see Abramowitz & 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 $\lim_{n\to\infty}[n\log(1 + (2A-2)/n)] \to 2A - 2$. Applying Proposition 2(a), we are able to show that

$$(n+2A-2)\|\mathbf{y}-\mathbf{X}\boldsymbol{\mu}_{q(\boldsymbol{\beta})}\|^2/(2B_{q(\sigma^2)}) - n - n\log(1+(2A-2)/n) \xrightarrow{\mathrm{P}} 0.$$

Proposition 2(b) implies that the first term in (8) approaches 0 in probability. Hence, $\mathrm{VAIC} - \mathrm{AIC} \xrightarrow{\mathrm{P}} 0$.

Note that the VAIC is not uniquely defined in (7) since it depends on the type of variation approximation used and also depends on the choice of marginalisation. That is we can marginalise out part of $\boldsymbol{\theta}$ analytically and use variational approximation to approximate the remaining elements of $\boldsymbol{\theta}$. A very recent paper Gelman, Hwang & Vehtari (2013) shows that the exact DIC and AIC are the same when $y_i \sim N(\mu, 1), i = 1, \ldots, n$, where μ has a noninformative prior, i.e., $p(\mu) \propto 1$.

4.2. Variational Bayesian information criterion

A Bayesian model selection procedure chooses the model which is a posteriori most likely, where the marginal likelihood $p(\mathbf{y})$ can be used to construct a selection criterion. As $p(\mathbf{y})$ is typically analytically unavailable, the BIC is derived by approximating $-2\log p(\mathbf{y})$. The 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

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

where $P, \hat{\boldsymbol{\beta}}_{\mathrm{ML}}$ and $\hat{\sigma}^2_{\mathrm{ML}}$ are defined as in Section 4. Using the Laplace approximation (Claeskens & Hjort 2008), we can obtain

$$-2\log p(\mathbf{y}) + 2\log p(\hat{\boldsymbol{\theta}}) = \mathrm{BIC} - p\log(2\pi) + \log\| - n^{-1}\mathfrak{I}(\hat{\boldsymbol{\theta}})\| + O(n^{-1}), \qquad (9)$$

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

Motivated by (9), we define the VBIC as

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

An advantage of using this definition instead of $-2\mathbb{E}_q \log p(\mathbf{y}, \boldsymbol{\theta}) + P\log(n)$ is that even when P and n are not clearly defined, for example when data are missing, 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* $\mathrm{VBIC} = \mathrm{BIC} + O_p(1)$.

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 obtain

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

Note that $\boldsymbol{\Sigma}_{q(\boldsymbol{\beta})} = n^{-1}d_n(A_n + \sigma_\beta^2 n^{-1} d_n \mathbf{I})^{-1} = O_p(n^{-1})$ since $d_n = O_p(1)$ by 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

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

The difference between BIC and VBIC is thus equal to

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

From Proposition 2(b) we see that the first term in (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 standardised pseudo random vector with entries from the standard normal distribution. We consider six simulation settings: the sample size n varies over 10, 100 and 1000 and the hyperparameter B varies over 0.1 and 10^{-8}. We keep the hyperparameters $A = 0.01$, $\sigma_\beta^2 = 10^8$ and the coefficient vector $\boldsymbol{\beta} = [1,1,1,1,1]^\top$ constant. Each scenario is repeated 100 times. The quantities $\|\boldsymbol{\mu}_{q(\boldsymbol{\beta})} - \hat{\boldsymbol{\beta}}_{\text{LS}}\|^2$ and $\|\boldsymbol{\Sigma}_{q(\beta)} - \hat{\sigma}^2_{\text{unbiased}}(\mathbf{X}^\top\mathbf{X})^{-1}\|_\infty$ are used to show empirically that the VB estimators are

TABLE 1
$|VAIC - AIC|$

n	Mean	Standard Error
(a) $B = 0.1$		
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) $B = 10^{-8}$		
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 2
$\|\boldsymbol{\mu}_{q(\boldsymbol{\beta})} - \hat{\boldsymbol{\beta}}_{\mathrm{LS}}\|^2$

n	Mean	Standard Error
(a) $B = 0.1$		
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) $B = 10^{-8}$		
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}

consistent and able to produce valid standard errors. The quantity |VBIC – BIC| is used to measure the difference between VBIC and BIC. The means and standard errors of the above four measures are shown in Tables 1–4.

In Table 1 the values get closer to 0 as n increases and when B is changed from 0.1 to 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 given Table 4 are almost constant but have smaller standard errors with increasing n. This means that 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 substantially.

6. Conclusions

This article shows that for the Bayesian linear model presented here, the corresponding VB estimators $\boldsymbol{\beta}_{VB}$ and σ^2_{VB} are consistent estimators of $\boldsymbol{\beta}_0$ and σ^2_0 under mild regularity conditions. This finding partially contradicts the criticism that VB are not valid for statistical inferences. The works of Teh, Newman & Welling (2006); Teh, Kurihara & Welling (2007) and Welling, Teh & Kappen (2008) on collapsed VB estimation represents another approach to improving VB estimation which may also mitigate these criticisms. Furthermore we proved that the variational Akaike information criterion shares the same first order asymptotic properties as the Akaike information criterion and that the

TABLE 3

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

n	Mean	Standard Error
(a) $B = 0.1$		
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) $B = 10^{-8}$		
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 4

$|VBIC - BIC|$

n	Mean	Standard Error
(a) $B = 0.1$		
10	8.883	4.218×10^{-1}
100	8.833	1.110×10^{-1}
1000	8.890	3.222×10^{-2}
(b) $B = 10^{-8}$		
10	8.819	4.213×10^{-1}
100	8.828	1.111×10^{-1}
1000	8.890	3.222×10^{-2}

variational Bayesian information criterion shares the same first order asymptotic properties as the Bayesian information criterion in a linear regression model. Computationally, in the context of linear regression models, VAIC and VBIC have no advantages over AIC and BIC as they are not as easy to calculate. However they are naturally derived in Bayesian contexts with VB estimates and the asymptotic properties in linear regression modelling motivate the potential use of VB based information criteria for more complex models. In particular, VAIC and VBIC can be used when n and P are not appropriate values for sample size and model size, while AIC and BIC are not applicable in such settings. Examples of such settings include the missing data context (for n) and mixed models (for P).

Appendix: Proofs

Proof of Proposition 1. 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, $\mathbf{b}_n \overset{\text{a.s}}{\to} \mathbb{E}(\mathbf{x}_i y_i) = \mathbb{E}_X[\mathbf{x}_i(\mathbf{x}_i^\top \boldsymbol{\beta}_0 + \epsilon_i)] = \mathbb{E}(\mathbf{x}_i\mathbf{x}_i^\top)\boldsymbol{\beta}_0$.

Note that $\boldsymbol{\beta}_{\text{LS}} = \mathbf{A}_n^{-1}\mathbf{b}_n$. Using Proposition 1(a) we obtain the stated result.

Let $\alpha = \sigma_\beta^{-2} d_n$. Note that $c_n = \text{tr}[\mathbf{X}^\top\mathbf{X}(\mathbf{X}^\top\mathbf{X} + \alpha\mathbf{I})^{-1}] = \sum_{i=1}^p \lambda_i/(\lambda_i + \alpha) = \sum_{i=1}^p (1 - \alpha/(\lambda_i + \alpha))$. The result follows since $\lambda_i > 0$ for $i = 1, \ldots, p$ and $\alpha > 0$.

Next, 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,

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

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

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

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

By definition $\|\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}\|^2 \leq \|\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}_{\mathrm{LS}}\|^2$ for any choice of $\hat{\boldsymbol{\beta}}$, and so

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

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

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

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

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

Note that if the first term in the expansion converges to zero as n increases then so will higher order terms. We may rewrite the first term as

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

by using (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 the right hand side of (12) approaching 0.

References

Abramowitz, M. & Stegun, I.A. (1964). *Handbook of Mathematical Functions.* New York: Dover.

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In *Proceedings of the 2nd International Symposium on Information Theory*, eds. B.N. Petrov & F. Csaki, pp. 267–281. Budapest: Akadémiai Kiadó.

Armagan, A., Dunson, D. & Clyde, M. (2011). Generalized beta mixtures of gaussians. *Adv. Neural Inf. Process. Syst.* **24**, 523–531.

Beal, M.J. (2003). *Variational algorithms for approximate bayesian inference.* Unpublished doctoral thesis, University College London, Gatsby Computational Neuroscience Unit.

Bishop, C.M. (2006). *Pattern Recognition and Machine Learning.* New York: Springer.

Claeskens, G. & Hjort, N.L. (2008). Model Selection and Model Averaging. *Cambridge Series in Statistical and Probabilistic Mathematics.* Cambridge: Cambridge University Press.

Drugowitsch, J. (2008). Bayesian linear regression. Technical report, University of Rochester, Rochester, NY.

Erdélyi, A., Magnus, W., Oberhettinger, F. & Tricomi, F.G. (1981). *Higher Transcendental Functions*, Vol. **III**. Melbourne, FL: Robert E. Krieger Publishing Co. Inc. Based on notes left by Harry Bateman, Reprint of the 1955 original.

Gelman, A., Hwang, J. & Vehtari, A. (2013). Understanding predictive information criteria for Bayesian models. *Statist. Comput.* Arxiv.

Hall, P., Ormerod, J.T. & Wand, M.P. (2011). Theory of Gaussian variational approximation for a Poisson mixed model. *Statist. Sinica* **21**, 369–389.

Hall, P., Pham, T., Wand, M.P.& Wang, S.S.J. (2011). Asymptotic normality and valid inference for Gaussian variational approximation. *Ann. Statist.* **39**, 2502–2532.

Humphreys, K. & Titterington, D.M. (2000). Approximate bayesian inference for simple mixtures. In *Proceedings of Computational Statistics*, eds. J.G. Bethlehem & P.G.M. van der Heijden, pp. 2502–2532. Heidelberg: Physica.

MacKay, D.J.C. (1995). Ensemble learning and evidence maximization. Technical report. Cavendish Laboratory University of Cambridge, Cambridge.

MacKay, D.J.C. (2003). *Information Theory, Inference and Learning Algorithms.* New York: Cambridge University Press.

McGrory, C.A. & Titterington, D.M. (2007). Variational approximations in Bayesian model selection for finite mixture distributions. *Comput. Statist. Data Anal.* **51**, 5352–5367.

Mitchell, T.J. & Beauchamp, J.J. (1988). Bayesian variable selection in linear regression. *J. Amer. Statist. Assoc.* **83**, 1023–1036.

Murphy, K.P. (2012). *Machine Learning: A Probabilistic Perspective.* London: The MIT Press.

Neville, S.E., Ormerod, J.T. & Wand, M.P. (2013). Mean field variational bayes for continuous sparse signal shrinkage: pitfalls and remedies. Preprint.

Ormerod, J.T. & Wand, M.P. (2010). Explaining variational approximations. *Amer. Statist.* **64**, 140–153.

Ormerod, J.T. & Wand, M.P. (2012). Gaussian variational approximate inference for generalized linear mixed models. *J. Comput. Graph. Statist.* **21**, 2–17.

Pauler, D.K. (1998). The schwarz criterion and related methods for normal linear models. *Biometrika* **85**, 13–27.

Ren, Q., Banerjee, S., Finley, A.O. & Hodges, J.S. (2011). Variational Bayesian methods for spatial data analysis. *Comput. Statist. Data Anal.* **55**, 3197–3217.

ROBERT, C.P. & MARIN, J.M. (2007). *Bayesian Core, A Practical Approach to Computational Bayesian Statistics.* New York: Springer.

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**, 461–464.

SPIEGELHALTER, D.J., BEST, N.G., CARLIN, B.P. & VAN DER LINDE, A. (2002). Bayesian measures of model complexity and fit (with Discussion). *J. R. Stat. Soc. Ser. B Stat. Methodol.* **64**, 583–639.

TEH, Y.W., KURIHARA, K. & WELLING, M. (2007). Collapsed variational inference for HDP. *Adv. Neural Inf. Process. Syst.* **20.**

TEH, Y.W., NEWMAN, D. & WELLING, M. (2006). A collapsed variational bayesian inference algorithm for latent dirichlet allocation. *Adv. Neural Inf. Process. Syst.* **19**, 1353–1360.

TIBSHIRANI, R. (1996). Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **58**, 267–288.

VOLANT, S., MAGNIETTE, M.L. & ROBIN, S. (2012). Variational Bayes approach for model aggregation in unsupervised classification withMarkovian dependency. *Comput. Statist. Data Anal.* **56**, 2375–2387.

WANG, B. & TITTERINGTON, D.M. (2006). Convergence properties of a general algorithm for calculating variational Bayesian estimates for a normal mixture model. *Bayesian Anal.* **1**, 625–650.

WELLING, M., TEH, Y.W. & KAPPEN, H.J. (2008). Hybrid variational/gibbs collapsed inference in topic models. *Proc. Int. Conf. Uncertainty Artif. Intell.* **24**, 587–591.

YANG, Y. (2005). Can the strengths of AIC and BIC be shared? A conflict between model indentification and regression estimation. *Biometrika* **92**, 937–950.

YOU, C., ORMEROD, J.T. & MÜLLER, S. (2013). A variational Bayes approach to variable selection. Preprint.