AMSI
AUSTRALIAN MATHEMATICAL SCIENCES INSTITUTE

National Collaboration in the Mathematical Sciences

Theory for Gaussian Variational Approximation of Bayesian Generalised Linear Models

Gemma Moran

February 28, 2013

1 Introduction

Variational approximation methods are an emerging class of deterministic techniques for analytically approximating high dimensional intractable integrals. In this regard, they provide alternatives to the predominant Markov chain Monte Carlo (MCMC) methods in the fitting of and inference for complex statistical and probabilistic models. Whilst variational approximations sacrifice some of the accuracy of MCMC, they are significantly faster to compute, particularly when applied to large datasets or complex models (Ormerod and Wand 2012).

Variational approximations have become a key component of inference in Computer Science, having applications in such diverse areas as speech recognition, document retrieval and genetic linkage analysis (Jordan 2004). Despite this, variational approximations have yet to attain widespread attention in statistical settings.

This project develops theory for the particular technique of Gaussian variational approximation in the inference of Bayesian generalised linear models, proving that estimators in this context possess useful frequentist properties such as consistency and can be used to calculate asymptotically valid standard errors.

Postal Address: 111 Barry Street
c/- The University of Melbourne
Victoria 3010 Australia

Email: enquiries@amsi.org.au
Phone: +61 3 8344 1777
Fax: +61 3 9349 4106
Web: www.amsi.org.au

2 Bayesian Generalised Linear Model

The observed data are $(y_i, \mathbf{x}_i), 1 \leq i \leq n$. For each $1 \leq i \leq n$, define the $p \times 1$ vectors $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})^T$ and $\mathbf{1}_n$ as a vector of length n consisting of ones. The entries of the explanatory vectors $\mathbf{x}_i$ are unrestricted real numbers, while the response variables y_i are subject to restrictions such as being binary or non-negative integers.

We consider one-parameter exponential family models of the form

$$p(\mathbf{y}|\boldsymbol{\beta}) = \exp\left\{\mathbf{y}^T\mathbf{X}\boldsymbol{\beta} - \mathbf{1}_n^T b(\mathbf{X}\boldsymbol{\beta}) + \mathbf{1}_n^T c(\mathbf{y})\right\}$$

where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)$ is a vector of parameters with prior $\boldsymbol{\beta} \sim \mathcal{N}(\mathbf{0}, \sigma_\beta^2\mathbf{I})$.

Common examples of these exponential family models include the Poisson case, for which $b(x) = e^x$ and $c(x) = \log(x!)$, and the logistic case, for which $b(x) = \log(1 + e^x)$ and $c(x) = 0$.

The marginal likelihood is given by

$$p(\mathbf{y}) = \frac{1}{(2\pi\sigma_\beta^2)^{p/2}} \int_{\mathbb{R}^p} \exp\left\{\mathbf{y}^T\mathbf{X}\boldsymbol{\beta} - \mathbf{1}_n^T b(\mathbf{X}\boldsymbol{\beta}) + \mathbf{1}_n^T c(\mathbf{y}) - \frac{\|\boldsymbol{\beta}\|^2}{2\sigma_\beta^2}\right\} d\boldsymbol{\beta}. \qquad (1)$$

We note that $p(\mathbf{y})$, and hence $p(\boldsymbol{\beta}|\mathbf{y})$, involves a potentially intractable integral over $\mathbb{R}^p$. We use the method of *Gaussian variational approximation* to approximate $p(\mathbf{y})$.

3 Gaussian Variational Approximation of Bayesian Generalised Linear Models

Gaussian variational approximation involves the derivation of a lower bound for the intractable marginal likelihood using a Gaussian density. This is achieved by employing the notion of *Kullbalck-Leibler divergence*, the details of which follow.

Let $q(\boldsymbol{\beta})$ be the p-variate Gaussian density function with mean $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_p)$ and

covariance matrix $\boldsymbol{\Sigma}$. Then we have

$$\begin{aligned}\log p(\mathbf{y}) &= \int q(\boldsymbol{\beta}) \log p(\mathbf{y}) d\boldsymbol{\beta} \\ &= \int q(\boldsymbol{\beta}) \log \left[\frac{p(\mathbf{y}, \boldsymbol{\beta})}{q(\boldsymbol{\beta})}\right] d\boldsymbol{\beta} + \int q(\boldsymbol{\beta}) \log \left[\frac{q(\boldsymbol{\beta})}{p(\boldsymbol{\beta}|\mathbf{y})}\right] d\boldsymbol{\beta} \\ &\geq \int q(\boldsymbol{\beta}) \log \left[\frac{p(\mathbf{y}, \boldsymbol{\beta})}{q(\boldsymbol{\beta})}\right] d\boldsymbol{\beta}.\end{aligned}$$

The inequality arises from the fact that

$$\int q(\boldsymbol{\beta}) \log \left[\frac{q(\boldsymbol{\beta})}{p(\boldsymbol{\beta}|\mathbf{y})}\right] d\boldsymbol{\beta} \geq 0$$

for all densities q, with equality if and only if $q(\boldsymbol{\beta}) = p(\boldsymbol{\beta}|\mathbf{y})$ almost everywhere (Kullback and Leibler 1951).

Hence, a variational lower bound for the marginal log-likelihood is given by

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

where

$$\begin{aligned}\log \underline{p}(\mathbf{y}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \mathbf{y}^T \mathbf{X}\boldsymbol{\mu} - \mathbf{1}_n^T B(\mathbf{X}\boldsymbol{\mu}, \mathrm{dg}(\mathbf{X}\boldsymbol{\Sigma}\mathbf{X}^T)) + \mathbf{1}_n^T c(\mathbf{y}) - \tfrac{p}{2}\log(\sigma_\beta^2) \\ - \frac{1}{2\sigma_\beta^2}\left[\|\boldsymbol{\mu}\|^2 + \mathrm{tr}(\boldsymbol{\Sigma})\right] + \tfrac{1}{2}\log|e\boldsymbol{\Sigma}| \quad (2)\end{aligned}$$

is the Gaussian variational approximation to $p(\mathbf{y})$, $B^{(r)}(\mu, \sigma^2) \equiv \int_{-\infty}^{\infty} b^{(r)}(\sigma x + \mu)\phi(x)dx$, ϕ is the $\mathcal{N}(0, 1)$ density function and $\mathrm{dg}(\mathbf{A})$ is the vector consisting of the diagonal entries of $\mathbf{A}$. For the Poisson case, $B(\mu, \sigma^2) = \exp\{\mu + \frac{1}{2}\sigma^2\}$. For the logistic case, numerical integration is required.

The greatest lower bound for $p(\mathbf{y})$ is obtained by maximising $\log \underline{p}(\mathbf{y}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ over the variational parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$. We denote these maximum likelihood estimators by $\widehat{\boldsymbol{\mu}}$ and $\widehat{\boldsymbol{\Sigma}}$, respectively.

4 Consistency and Standard Error

4.1 Consistency

We now prove the consistency of the maximum likelihood estimators of the model parameters based on Gaussian variational approximation. In order to do so, we impose the following conditions:

(A1) for $1 \leq i \leq n$, $y_i|\mathbf{x}_i$ are independent;

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

(A3) the distribution of each y_i belongs to a natural exponential family with $\mathbb{E}[y_i] = b(\mathbf{x}_i^T\boldsymbol{\beta}_0)$, where $\boldsymbol{\beta}_0$ are the true values of $\boldsymbol{\beta}$;

(A4) for each $1 \leq i \leq n$, $\mathbb{E}[\mathbf{x}_i\mathbf{x}_i^T]$ is element-wise finite and positive definite;

(A5) $b(\mathbf{x}_i^T\boldsymbol{\xi})$ is infinitely differentiable in $\boldsymbol{\xi}$;

(A6) $b^{(1)}(\mathbf{x}_i^T\boldsymbol{\xi})$ is continuously differentiable in $\boldsymbol{\xi}$;

(A7) $\varphi^{(r)}(\boldsymbol{\beta}_0) = \mathbb{E}_X[x_{ij}b^{(r)}(\mathbf{x}_i^T\boldsymbol{\beta}_0)]$ exists for $r = 0, 1, 2$;

(A8) $\mathbf{x}_i$, $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are element-wise bounded;

Note that

$$\frac{\partial \log \underline{p}}{\partial \boldsymbol{\mu}} = \mathbf{X}^T\left(\mathbf{y} - B^{(1)}(\mathbf{X}\boldsymbol{\mu}, \mathrm{dg}(\mathbf{X}\boldsymbol{\Sigma}\mathbf{X}^T))\right) - \frac{\boldsymbol{\mu}}{\sigma_\beta^2} \tag{3}$$

$$\text{and} \quad \frac{\partial \log \underline{p}}{\partial \boldsymbol{\Sigma}} = \boldsymbol{\Sigma}^{-1} - \boldsymbol{X}^T\mathrm{diag}(B^{(2)}(\mathbf{X}\boldsymbol{\mu}, \mathrm{dg}(\mathbf{X}\boldsymbol{\Sigma}\mathbf{X}^T)))\mathbf{X} - \sigma_\beta^{-2}\mathbf{I} \tag{4}$$

(Opper and Archambeau, 2009)

The following theoretical results are used in the course of proving the consistency of the maximum likelihood estimators.

Lemma 1. (Horn and Johnson 1985, 7.7.5) *The inverse of an Hermitian matrix*

$$\begin{bmatrix}\mathbf{A} & \mathbf{B}\\ \mathbf{B}^T & \mathbf{C}\end{bmatrix}^{-1} = \begin{bmatrix}\mathbf{I} & \mathbf{0}\\ -\mathbf{C}^{-1}\mathbf{B}^T & \mathbf{I}\end{bmatrix}\begin{bmatrix}\left(\mathbf{A} - \mathbf{B}\mathbf{C}^{-1}\mathbf{B}^T\right)^{-1} & \mathbf{0}\\ \mathbf{0} & \mathbf{C}^{-1}\end{bmatrix}\begin{bmatrix}\mathbf{I} & -\mathbf{B}\mathbf{C}^{-1}\\ \mathbf{0} & \mathbf{I}\end{bmatrix}$$

Lemma 2. *Let* $\mathbf{M}$ *be a symmetric* $p\times p$ *matrix and* $\mathbf{a} = (a_1, \ldots, a_p)^T$. *Then* $\mathbf{a}^T \left[\mathbf{M} + \text{diag}(\mathbf{d})\right]^{-1} \mathbf{a}$ *is a strictly decreasing funtion of* d_i, *where* $\mathbf{d} = (d_1, \ldots, d_p)^T$ *and* $\text{diag}(\mathbf{d})$ *is the* $p \times p$ *diagonal matrix containing the entries of* $\mathbf{d}$ *along the main diagonal.*

Proof.

$$\mathbf{M} + \text{diag}(\mathbf{d}) = \begin{bmatrix} m_{11} + d_1 & \mathbf{m}_{12} \\ \mathbf{m}_{12}^T & \mathbf{M}_{22} + \mathbf{D}_2 \end{bmatrix}$$

where $\mathbf{m}_{12} = (m_{12}, \ldots, m_{1p})$, $\mathbf{D}_2 = \text{diag}(d_2, \ldots, d_p)$ and $\mathbf{M}_{22} = \begin{bmatrix} m_{22} & \cdots & m_{2p} \\ \vdots & \ddots & \vdots \\ m_{p2} & \cdots & m_{pp} \end{bmatrix}$

Then, by Lemma 1,

$$\begin{aligned} \mathbf{a}^T \left[\mathbf{M} + \text{diag}(\mathbf{d})\right]^{-1} \mathbf{a} = \mathbf{a}^T & \begin{bmatrix} 1 & \mathbf{0} \\ -(\mathbf{M}_{22} + \mathbf{D}_2)^{-1}\mathbf{m}_{12}^T & \mathbf{I}_{p-1} \end{bmatrix} \\ & \times \begin{bmatrix} \left(m_{11} + d_1 - \mathbf{m}_{12}(\mathbf{M}_{22} + \mathbf{D}_2)^{-1}\mathbf{m}_{12}^T\right)^{-1} & \mathbf{0} \\ \mathbf{0} & (\mathbf{M}_{22} + \mathbf{D}_2)^{-1} \end{bmatrix} \\ & \times \begin{bmatrix} 1 & -\mathbf{m}_{12}(\mathbf{M}_{22} + \mathbf{D}_2)^{-1} \\ \mathbf{0} & \mathbf{I}_{p-1} \end{bmatrix} \mathbf{a} \end{aligned}$$

Let $\mathbf{b} = \begin{bmatrix} 1 & -\mathbf{m}_{12}(\mathbf{M}_{22} + \mathbf{D}_2)^{-1} \\ \mathbf{0} & \mathbf{I}_{p-1} \end{bmatrix} \mathbf{a}$

Then

$$\begin{aligned} \mathbf{a}^T \left[\mathbf{M} + \text{diag}(\mathbf{d})\right]^{-1} \mathbf{a} = \frac{b_1^2}{d_1 + m_{11} - \mathbf{m}_{12}(\mathbf{M}_{22} + \mathbf{D}_2)^{-1}\mathbf{m}_{12}^T} \\ + \mathbf{b}_2^T(\mathbf{M}_{22} + \mathbf{D}_2)^{-1}\mathbf{b}_2 \quad (5) \end{aligned}$$

where $\mathbf{b}_2 = (b_2, \ldots, b_p)$.

Clearly, (5) is strictly decreasing in d_1. The result follows for d_i after relabelling. □

Lemma 3.

$$\mathbf{x}_i^T \boldsymbol{\Sigma} \mathbf{x}_i = O_p(n^{-1})$$

Proof. Setting (4) to zero, we obtain

$$\boldsymbol{\Sigma} = \left[\sigma_\beta^{-2}\mathbf{I} + \boldsymbol{X}^T\mathbf{W}\mathbf{X}\right]^{-1}$$

where $\mathbf{W} = \text{diag}(B^{(2)}(\mathbf{X}\boldsymbol{\mu}, \text{dg}(\mathbf{X}\boldsymbol{\Sigma}\mathbf{X}^T)))$. Using the Sherman-Morrison-Woodbury matrix identity, we have

$$\begin{aligned}\mathbf{x}_i^T\boldsymbol{\Sigma}\mathbf{x}_i &= \mathbf{x}_i^T\left[\sigma_\beta^{-2}\mathbf{I} + \boldsymbol{X}^T\mathbf{W}\mathbf{X}\right]^{-1}\mathbf{x}_i \\ &= \sigma_\beta^2\|\mathbf{x}_i\|^2 - \sigma_\beta^4\mathbf{x}_i^T\mathbf{X}^T\left[\sigma_\beta^2\mathbf{X}\mathbf{X}^T + \mathbf{W}^{-1}\right]^{-1}\mathbf{X}\mathbf{x}_i\end{aligned}$$

Now, by Lemma 2, $\mathbf{x}_i^T\mathbf{X}^T\left[\sigma_\beta^2\mathbf{X}\mathbf{X}^T + \mathbf{W}^{-1}\right]^{-1}\mathbf{X}\mathbf{x}_i$ is a strictly decreasing function of w_i^{-1}. Let

$$\gamma = \inf_{\mathbf{x}_i,\boldsymbol{\mu},\boldsymbol{\Sigma}} B^{(2)}(\mathbf{x}_i^T\boldsymbol{\mu}, \mathbf{x}_i^T\boldsymbol{\Sigma}\mathbf{x}_i).$$

Then γ is bounded away from zero by assumption (A8) and the fact that $B^{(2)}(\cdot,\cdot)$ is bounded away from zero for finite arguments. Then,

$$\begin{aligned}\mathbf{x}_i^T\boldsymbol{\Sigma}\mathbf{x}_i &\leq \sigma_\beta^2\|\mathbf{x}_i\|^2 - \sigma_\beta^4\mathbf{x}_i^T\mathbf{X}^T\left[\sigma_\beta^2\mathbf{X}\mathbf{X}^T + \gamma^{-1}\mathbf{I}\right]^{-1}\mathbf{X}\mathbf{x}_i \\ &= \mathbf{x}_i^T\left[\gamma\mathbf{X}^T\mathbf{X} + \sigma_\beta^{-2}\mathbf{I}\right]^{-1}\mathbf{x}_i \\ &= \frac{1}{n}\mathbf{x}_i^T\left[\frac{\gamma}{n}\sum_{i=1}^n\mathbf{x}_i\mathbf{x}_i^T + \frac{\sigma_\beta^{-2}}{n}\mathbf{I}\right]^{-1}\mathbf{x}_i.\end{aligned}$$

Let

$$\bar{\mathbf{A}}_n = \frac{1}{n}\sum_{i=1}^n\mathbf{x}_i\mathbf{x}_i^T.$$

By the strong law of large numbers, and under the assumptions (A2)-(A3), we have

$$\bar{\mathbf{A}}_n \xrightarrow{a.s.} \mathbf{A}$$

where $\mathbf{A}_i = \mathbf{x}_i\mathbf{x}_i^T$. Now, using the Taylor series expansion about $\frac{1}{n}\sigma_\beta^{-2} = 0$, we obtain

$$\begin{aligned}\mathbf{x}_i^T\boldsymbol{\Sigma}\mathbf{x}_i &\leq \frac{1}{n}\gamma^{-1}\mathbf{x}_i^T\bar{\mathbf{A}}_n^{-1}\mathbf{x}_i + \frac{1}{n^2}\gamma^{-2}\sigma_\beta^{-1}\mathbf{x}_i^T\bar{\mathbf{A}}_n^{-2}\mathbf{x}_i + O_p(n^{-3}) \\ &\xrightarrow{p} \frac{1}{n}\gamma^{-1}\mathbf{x}_i^T\mathbf{A}^{-1}\mathbf{x}_i + O_p(n^{-2}) \\ &= O_p(n^{-1}).\end{aligned}$$

□

We note that the proof of Lemma 3 is identical for $\mathbf{e}_i^T\boldsymbol{\Sigma}\mathbf{e}_j$, where $\mathbf{e}_i$ is the $p \times 1$ vector with 1 in the i-th row and zero elsewhere. That is, $\Sigma_{ij} = O_p(n^{-1})$ for all $1 \leq i, j \leq p$.

Result 1. *$\widehat{\boldsymbol{\mu}}$ is a consistent estimator of $\boldsymbol{\beta}_0$.*

Proof. The Taylor series expansion of $B^{(r)}(\mathbf{x}_i^T\boldsymbol{\mu}, \mathbf{x}_i^T\boldsymbol{\Sigma}\mathbf{x}_i)$ about $\mathbf{x}_i^T\boldsymbol{\Sigma}\mathbf{x}_i = 0$ gives

$$\begin{aligned} B^{(r)}(\mathbf{x}_i^T\boldsymbol{\mu}, \mathbf{x}_i^T\boldsymbol{\Sigma}\mathbf{x}_i) &\approx b^{(r)}(\mathbf{x}_i^T\boldsymbol{\mu}) + \tfrac{1}{2!}b^{(r+2)}(\mathbf{x}_i^T\boldsymbol{\mu})\mathbf{x}_i^T\boldsymbol{\Sigma}\mathbf{x}_i + \tfrac{1}{3!}b^{(r+4)}(\mathbf{x}_i^T\boldsymbol{\mu})\left[\mathbf{x}_i^T\boldsymbol{\Sigma}\mathbf{x}_i\right]^2 + \dots \\ &= b^{(r)}(\mathbf{x}_i^T\boldsymbol{\mu}) + O_p(n^{-1}) \end{aligned}$$

Hence, for each $j = 1, \dots, p$, (3) gives

$$\begin{aligned} \frac{1}{n}\sum_{i=1}^{n} y_i x_{ij} &= \frac{1}{n}\sum_{i=1}^{n} x_{ij} b^{(1)}(\mathbf{x}_i^T\boldsymbol{\mu}) - \frac{\mu_j}{n\sigma_\beta^2} + O_p(n^{-2}) \\ &= \frac{1}{n}\sum_{i=1}^{n} x_{ij} b^{(1)}(\mathbf{x}_i^T\boldsymbol{\mu}) + O_p(n^{-1}). \end{aligned} \tag{6}$$

Now, by the strong law of large numbers and under assumptions (A1)-(A2),

$$\frac{1}{n}\sum_{i=1}^{n} y_i x_{ij} \xrightarrow{a.s.} \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} y_i x_{ij}\right] = \mathbb{E}_X\left[x_{ij} b^{(1)}(\mathbf{x}_i^T\boldsymbol{\beta}_0)\right] \equiv \varphi(\boldsymbol{\beta}_0)$$

and

$$\frac{1}{n}\sum_{i=1}^{n} x_{ij} b^{(1)}(\mathbf{x}_i^T\boldsymbol{\mu}) \xrightarrow{a.s.} \mathbb{E}_X\left[x_{ij} b^{(1)}(\mathbf{x}_i^T\boldsymbol{\mu})\right] \equiv \varphi(\boldsymbol{\mu}).$$

Hence, we have

$$\varphi(\boldsymbol{\beta}_0) = \varphi(\boldsymbol{\mu}) + o_p(1).$$

As $\varphi(\boldsymbol{\xi})$ is continuous in $\boldsymbol{\xi}$, the multivariate mean value theorem applies and so there exists a $\mathbf{c}^* \in (\boldsymbol{\mu}, \boldsymbol{\beta}_0)$ such that $\varphi(\boldsymbol{\beta}_0) - \varphi(\boldsymbol{\mu}) = \nabla\varphi(\mathbf{c}^*)^T(\boldsymbol{\beta}_0 - \boldsymbol{\mu})$. Thus

$$\boldsymbol{\beta}_0 = \boldsymbol{\mu} + o_p(1).$$

Hence, $\widehat{\boldsymbol{\mu}}$ is a consistent estimator of $\boldsymbol{\beta}_0$. □

Result 2. *The rate of convergence for $\widehat{\boldsymbol{\mu}}$ is $O_p(n^{-1/2})$.*

Proof. Consider the maximum likelihood estimator $\widehat{\boldsymbol{\beta}}$ of $\log p(\mathbf{y}|\boldsymbol{\beta})$. Note that

$$\frac{\partial \log p(\mathbf{y}|\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \mathbf{X}^T\mathbf{y} - \mathbf{X}^T b(\mathbf{X}\boldsymbol{\beta}) \tag{7}$$

Hence for each $j = 1, \ldots, p$, we have

$$\sum_{i=1}^{n} y_i x_{ij} = \sum_{i=1}^{n} x_{ij} b^{(1)}(\mathbf{x}_i^T \widehat{\boldsymbol{\beta}}). \tag{8}$$

Solving (6) and (8) gives

$$\sum_{i=1}^{n} x_{ij} b^{(1)}(\mathbf{x}_i^T \widehat{\boldsymbol{\beta}}) = \sum_{i=1}^{n} x_{ij} b^{(1)}(\mathbf{x}_i^T \widehat{\boldsymbol{\mu}}) + O_p(n^{-1}). \tag{9}$$

As $b^{(1)}(\mathbf{x}_i^T\boldsymbol{\xi})$ is continuous in $\boldsymbol{\xi}$, the multivariate mean value theorem applies and so there exists a $\mathbf{c}^* \in (\widehat{\boldsymbol{\mu}}, \widehat{\boldsymbol{\beta}})$ such that $b^{(1)}(\mathbf{x}_i^T \widehat{\boldsymbol{\beta}}) - b^{(1)}(\mathbf{x}_i^T \widehat{\boldsymbol{\mu}}) = \nabla b^{(1)}(\mathbf{c}^*)^T(\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\mu}})$. Hence, we have

$$\widehat{\boldsymbol{\beta}} = \widehat{\boldsymbol{\mu}} + O_p(n^{-1}). \tag{10}$$

From the standard asymptotic properties of maximum likelihood estimators, we have $\widehat{\boldsymbol{\beta}} = \boldsymbol{\beta}_0 + O_p(n^{-1/2})$ (Bishop et al. 1975, Theorem 14.4-1). Thus, we obtain

$$\widehat{\boldsymbol{\mu}} = \boldsymbol{\beta}_0 + O_p(n^{-1/2}).$$

□

4.2 Standard Error

Result 3.

$$\widehat{\boldsymbol{\Sigma}} \xrightarrow{p} \mathcal{I}_n(\boldsymbol{\beta}_0)^{-1}$$

Proof. The Fisher information matrix is given by:

$$\begin{aligned}
\mathcal{I}_n(\boldsymbol{\beta}) &= -\mathbb{E}\left[\frac{\partial^2}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T}\log p(\mathbf{y}|\boldsymbol{\beta})\right] \\
&= -\mathbb{E}\left[-\mathbf{X}^T \mathrm{diag}(b^{(2)}(\mathbf{X}\boldsymbol{\beta}))\mathbf{X}\right] \\
&= \mathbf{X}^T \mathrm{diag}(b^{(2)}(\mathbf{X}\boldsymbol{\beta}))\mathbf{X}
\end{aligned}$$

Now, by the multivariate mean value theorem and Result 2

$$\begin{aligned}\boldsymbol{\Sigma} &= \left[\sigma_\beta^{-2}\mathbf{I} + \boldsymbol{X}^T \mathrm{diag}(B^{(2)}(\mathbf{X}\boldsymbol{\mu}, \mathrm{dg}(\mathbf{X}\boldsymbol{\Sigma}\mathbf{X}^T)))\mathbf{X}\right]^{-1} \\ &= \frac{1}{n}\left[O_p(n^{-1}) + \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i\mathbf{x}_i^T b^{(2)}(\mathbf{x}_i^T\boldsymbol{\mu})\right]^{-1} \\ &= \mathcal{I}_n^{-1}(\boldsymbol{\beta}) + O_p(n^{-1}).\end{aligned}$$

Thus,

$$\widehat{\boldsymbol{\Sigma}} \xrightarrow{p} \mathcal{I}_n(\boldsymbol{\beta}_0)^{-1}$$

□

5 Invariance under Linear Transformation

Let $\boldsymbol{\beta} = \mathbf{A}\mathbf{u} + \mathbf{b}$ where $\mathbf{A}$ is an invertible $p \times p$ matrix and $\mathbf{b} = (b_1, \ldots, b_p)$ is a $p \times 1$ vector.

The transformation formula gives

$$p(\mathbf{y}, \mathbf{u}) \propto \exp\left\{\mathbf{y}^T\mathbf{X}\mathbf{A}\mathbf{u} - \mathbf{1}_n^T b(\mathbf{X}\mathbf{A}\mathbf{u} + \mathbf{X}\mathbf{b})\right\} \exp\left\{-\frac{1}{2\sigma_\beta^2}\left(\mathbf{u}^T\mathbf{A}^T\mathbf{A}\mathbf{u} + \mathbf{u}^T\mathbf{A}^T\mathbf{b} + \mathbf{b}^T\mathbf{A}\mathbf{u}\right)\right\}. \quad (11)$$

Then, the lower bound for the marginal likelihood is given by

$$\mathbb{E}_q\left[\log\left(\frac{p(\mathbf{y}, \mathbf{u})}{q(\mathbf{u})}\right)\right] \equiv \log \widetilde{p}(\mathbf{y}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) \quad (12)$$

We solve for the maximum likelihood estimators $\widetilde{\boldsymbol{\mu}}$ and $\widetilde{\boldsymbol{\Sigma}}$ to obtain:

$$\mathbf{A}^T\mathbf{X}^T(\mathbf{y} - \tilde{\mathbf{v}}) - \frac{1}{\sigma_\beta^2}\left(\mathbf{A}^T\mathbf{A}\widetilde{\boldsymbol{\mu}} + \mathbf{A}^T\mathbf{b}\right) = \mathbf{0} \quad \text{and} \quad (13)$$

$$\widetilde{\boldsymbol{\Sigma}} = \mathbf{A}^{-1}\left[\mathbf{X}^T\widetilde{\mathbf{W}}\mathbf{X} + \sigma_\beta^{-2}\mathbf{I}\right]^{-1}\mathbf{A}^{-T} \quad (14)$$

where

$$\begin{aligned}\tilde{\mathbf{v}} &= B^{(1)}(\mathbf{X}\mathbf{A}\widetilde{\boldsymbol{\mu}} + \mathbf{X}\mathbf{b}, \mathrm{dg}(\mathbf{X}\mathbf{A}\widetilde{\boldsymbol{\Sigma}}\mathbf{A}^T\mathbf{X}^T)) \quad \text{and} \\ \widetilde{\mathbf{W}} &= \mathrm{diag}(B^{(2)}(\mathbf{X}\mathbf{A}\widetilde{\boldsymbol{\mu}} + \mathbf{X}\mathbf{b}, \mathrm{dg}(\mathbf{X}\mathbf{A}\widetilde{\boldsymbol{\Sigma}}\mathbf{A}^T\mathbf{X}^T))).\end{aligned}$$

Result 4. *The optimal values of* $\boldsymbol{\mu}$ *and* $\boldsymbol{\Sigma}$ *are invariant under the transformation.*

Proof. Let $\boldsymbol{\xi} = \mathbf{A}\widetilde{\boldsymbol{\mu}} + \mathbf{b}$ and $\boldsymbol{\Lambda} = \mathbf{A}\widetilde{\boldsymbol{\Sigma}}\mathbf{A}^T$. Now,

$$\begin{aligned}
\tilde{\mathbf{v}} &= \frac{1}{|2\pi\tilde{\boldsymbol{\Sigma}}|^{1/2}} \int b^{(r)}(\mathbf{X}\mathbf{A}\mathbf{u} + \mathbf{X}\mathbf{b}) \exp\left\{-\frac{1}{2}(\mathbf{u} - \widetilde{\boldsymbol{\mu}})^T \widetilde{\boldsymbol{\Sigma}}^{-1}(\mathbf{u} - \widetilde{\boldsymbol{\mu}})\right\} d\mathbf{u} \\
&= \frac{1}{|2\pi\boldsymbol{\Lambda}|^{1/2}} \int b^{(r)}(\mathbf{X}\boldsymbol{\beta}) \exp\left\{-\frac{1}{2}(\boldsymbol{\beta} - \boldsymbol{\xi})^T \boldsymbol{\Lambda}^{-1}(\boldsymbol{\beta} - \boldsymbol{\xi})\right\} d\boldsymbol{\beta} \\
&= B^{(r)}(\mathbf{X}\boldsymbol{\xi}, \mathrm{dg}(\mathbf{X}\boldsymbol{\Lambda}\mathbf{X}^T))
\end{aligned}$$

Then, substitution into (13) gives

$$\mathbf{X}^T\left(\mathbf{y} - B^{(1)}(\mathbf{X}\boldsymbol{\xi}, \mathrm{dg}(\mathbf{X}\boldsymbol{\Lambda}\mathbf{X}^T))\right) - \sigma_\beta^{-2}\boldsymbol{\xi} = \mathbf{0}.$$

Now, substitution into (14) gives

$$\boldsymbol{\Lambda} = \left[\mathbf{X}^T \mathrm{diag}(B^{(2)}(\mathbf{X}\boldsymbol{\xi}, \mathrm{dg}(\mathbf{X}\boldsymbol{\Lambda}\mathbf{X}^T)))\mathbf{X} + \sigma_\beta^{-2}\mathbf{I}\right]^{-1}$$

Hence, $\boldsymbol{\xi}$ and $\boldsymbol{\Lambda}$ satisfy first order optimality conditions. □

6 Variational Information Criterion

6.1 Variational Bayesian Information Criterion (VBIC)

In this section we discuss the properties of the Bayesian information criterion in relation to Gaussian variational approximation.

We define the BIC as follows:

$$\mathrm{BIC} = -2\log p(\mathbf{y}|\widehat{\boldsymbol{\beta}}) + p\log n.$$

(Claeskens and Hjort 2010, 3.1)

In the variational approximation context, we approximate the BIC by the following result, which we call the variational Bayesian information criterion (VBIC):

$$\mathrm{VBIC} = -2\log \underline{p}(\mathbf{y}) + 2\mathbb{E}_q[\log p(\boldsymbol{\beta})].$$

Result 5. *The* VBIC *is first order equivalent in probablity to the* BIC. *That is,*

$$\text{BIC} = \text{VBIC} + O_p(1).$$

Proof.

$$\begin{aligned}
\text{BIC} - \text{VBIC} &= 2\mathbf{y}^T\mathbf{X}(\boldsymbol{\mu} - \widehat{\boldsymbol{\beta}}) - \mathbf{1}_n^T[2B(\mathbf{X}\boldsymbol{\mu}, \text{dg}(\mathbf{X}\boldsymbol{\Sigma}\mathbf{X}^T)) - 2b(\mathbf{X}\widehat{\boldsymbol{\beta}})] \\
&\quad + p\log(2\pi) + \log|e\boldsymbol{\Sigma}| + p\log n \\
&= -\mathbf{1}_n^T[2B(\mathbf{X}\boldsymbol{\mu}, \text{dg}(\mathbf{X}\boldsymbol{\Sigma}\mathbf{X}^T)) - 2b(\mathbf{X}\widehat{\boldsymbol{\beta}})] + p\log(2\pi) \\
&\quad + p + \log|\boldsymbol{\Sigma}| + p\log n + O_p(n^{-1}) \qquad \text{by (10)}
\end{aligned}$$

Consider the term $\mathbf{1}_n^T[2B(\mathbf{X}\boldsymbol{\mu}, \text{dg}(\mathbf{X}\boldsymbol{\Sigma}\mathbf{X}^T)) - 2b(\mathbf{X}\widehat{\boldsymbol{\beta}})]$. Using the Taylor series expansion we obtain

$$\begin{aligned}
\mathbf{1}_n^T[2B(\mathbf{X}\boldsymbol{\mu}, \text{dg}(\mathbf{X}\boldsymbol{\Sigma}\mathbf{X}^T)) - 2b(\mathbf{X}\widehat{\boldsymbol{\beta}})] &= 2\sum_{i=1}^{n}\frac{1}{2!}b^{(2)}(\mathbf{x}_i^T\boldsymbol{\mu})\mathbf{x}_i^T\boldsymbol{\Sigma}\mathbf{x}_i + \frac{1}{3!}b^{(4)}(\mathbf{x}_i^T\boldsymbol{\mu})\left[\mathbf{x}_i^T\boldsymbol{\Sigma}\mathbf{x}_i\right]^2 \\
&\quad + O_p(\left[\mathbf{x}_i^T\boldsymbol{\Sigma}\mathbf{x}_i\right]^3) \\
&= \left[\sum_{i=1}^{n} b^{(2)}(\mathbf{x}_i^T\boldsymbol{\mu})\mathbf{x}_i^T\boldsymbol{\Sigma}\mathbf{x}_i\right] + O_p(n^{-2}). \qquad (15)
\end{aligned}$$

Now,

$$\begin{aligned}
\sum_{i=1}^{n} b^{(2)}(\mathbf{x}_i^T\boldsymbol{\mu})\mathbf{x}_i^T\boldsymbol{\Sigma}\mathbf{x}_i &= \sum_{i=1}^{n} b^{(2)}(\mathbf{x}_i^T\boldsymbol{\mu})\text{tr}\left(\mathbf{x}_i\mathbf{x}_i^T\boldsymbol{\Sigma}\right) \\
&= \text{tr}\left(\sum_{i=1}^{n} b^{(2)}(\mathbf{x}_i^T\boldsymbol{\mu})\mathbf{x}_i\mathbf{x}_i^T\boldsymbol{\Sigma}\right) \\
&= \text{tr}\left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i^T b^{(2)}(\mathbf{x}_i^T\boldsymbol{\mu})\left[\frac{\sigma_\beta^{-2}}{n}\mathbf{I} + \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i^T b^{(2)}(\mathbf{x}_i^T\boldsymbol{\mu})\right]^{-1}\right) \\
&= \text{tr}\left(\left[\frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i^T b^{(2)}(\mathbf{x}_i^T\boldsymbol{\mu})\right]\left[\frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i^T b^{(2)}(\mathbf{x}_i^T\boldsymbol{\mu})\right]^{-1} + O_p(n^{-1})\right) \\
&= \text{tr}(\mathbf{I}_p) + O_p(n^{-1}) \\
&= p + O_p(n^{-1}).
\end{aligned}$$

Postal Address: 111 Barry Street
c/- The University of Melbourne
Victoria 3010 Australia

Email: enquiries@amsi.org.au
Phone: +61 3 8344 1777
Fax: +61 3 9349 4106
Web: www.amsi.org.au

Hence,

$$\mathbf{1}_n^T[2B(\mathbf{X}\boldsymbol{\mu}, \mathrm{dg}(\mathbf{X}\boldsymbol{\Sigma}\mathbf{X}^T)) - 2b(\mathbf{X}\widehat{\boldsymbol{\beta}})] = p + O_p(n^{-1}). \tag{16}$$

Now,

$$\begin{aligned}\mathrm{BIC} - \mathrm{VBIC} &= p\log(2\pi) + \log|\boldsymbol{\Sigma}| + p\log n + O_p(n^{-1}) \\ &= p\log(2\pi) + p\log(O_p(n^{-1})) + p\log n + O_p(n^{-1}) \\ &= O_p(1)\end{aligned}$$

as required. □

6.2 Variational Akaike Information Criterion (VAIC)

Following McGrory and Titterington (2007), we define the variational Akaike information criterion (VAIC) as follows:

$$\mathrm{VAIC} \equiv -2\log p(\mathbf{y}|\boldsymbol{\mu}) + 2P$$

where $P = 2\log p(\mathbf{y}|\boldsymbol{\mu}) - 2\mathbb{E}_{q(\boldsymbol{\beta})}[\log p(\mathbf{y}|\boldsymbol{\beta})]$. We have:

$$\log p(\mathbf{y}|\boldsymbol{\mu}) = \mathbf{y}^T\mathbf{X}\boldsymbol{\mu} - \mathbf{1}_n^T b(\mathbf{X}\boldsymbol{\mu}) + \mathbf{1}_n^T c(\mathbf{y})$$

$$\mathbb{E}_{q(\boldsymbol{\beta})}[\log p(\mathbf{y}|\boldsymbol{\beta})] = \mathbf{y}^T\mathbf{X}\boldsymbol{\mu} - \mathbf{1}_n^T B(\mathbf{X}\boldsymbol{\mu}, \mathrm{dg}(\mathbf{X}\boldsymbol{\Sigma}\mathbf{X}^T)) + \mathbf{1}_n^T c(\mathbf{y})$$

Note that the AIC is given by:

$$\mathrm{AIC} = -2\log p(\mathbf{y}|\widehat{\boldsymbol{\beta}}) + 2p$$

where $\widehat{\boldsymbol{\beta}} = \mathrm{argmax}_{\boldsymbol{\beta}}\, p(\mathbf{y}|\boldsymbol{\beta})$.

(Claeskens and Hjort 2010, 2.3)

Result 6. *Let the* AIC *and* VAIC *be defined as above. Then* VAIC $\xrightarrow{p}$ AIC.

Proof. Using similar expansions to (15) and by (9) and (10) we have

$$\begin{aligned}\mathrm{VAIC} - \mathrm{AIC} &= 2\mathbf{y}^T\mathbf{X}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\mu}) - \mathbf{1}_n^T(2b(\mathbf{X}\widehat{\boldsymbol{\beta}}) - 2b(\mathbf{X}\boldsymbol{\mu})) + O_p(n^{-1}) \\ &= O_p(n^{-1})\end{aligned}$$

□

7 Kullback-Leibler Dominance

In this section, we prove that the Gaussian variational approximation of the marginal likelihood $p(\mathbf{y})$ always yields a tighter lower bound than a number of other variational approximations.

7.1 Auxiliary Variables

Consider the auxiliary variable representation $p(\mathbf{y}, \boldsymbol{\beta}, \mathbf{a})$ where $p(\mathbf{y}, \boldsymbol{\beta}) = \int p(\mathbf{y}, \boldsymbol{\beta}, \mathbf{a})\, d\mathbf{a}$. Now, using Jensen's inequality we have

$$\log p(\mathbf{y}, \boldsymbol{\beta}) \geq \int q(\mathbf{a}) \log \left[\frac{p(\mathbf{y}, \boldsymbol{\beta}, \mathbf{a})}{q(\mathbf{a})} \right] d\mathbf{a}$$

for all $q(\mathbf{a})$. The Gaussian variational approximation gives:

$$\log p(\mathbf{y}) \geq \int q_G(\boldsymbol{\beta}) \log \left[\frac{p(\mathbf{y}, \boldsymbol{\beta})}{q_G(\boldsymbol{\beta})} \right] d\boldsymbol{\beta} \equiv \log \underline{p}_G(\mathbf{y})$$

where $q_G(\boldsymbol{\beta})$ is the optimal Gaussian density function. Hence

$$\log \underline{p}_G(\mathbf{y}) \geq \int q(\mathbf{a}) q_G(\boldsymbol{\beta}) \log \left[\frac{p(\mathbf{y}, \boldsymbol{\beta}, \mathbf{a})}{q(\mathbf{a}) q_G(\boldsymbol{\beta})} \right] d\mathbf{a} d\boldsymbol{\beta} \equiv \log \underline{p}_A(\mathbf{y}).$$

Thus, the method of Gaussian variational approximation always gives a better approximation to $\log p(\mathbf{y})$ than that of auxiliary variable representation.

7.2 Local Variational Approximation

In order to approximate $p(\mathbf{y})$, the local variational method involves the bounding of $p(\mathbf{y}|\boldsymbol{\beta})$ by a suitable funtion for which the integral $\int p(\mathbf{y}|\boldsymbol{\beta}) p(\boldsymbol{\beta}) d\boldsymbol{\beta}$ can be computed. We consider the local variational approximation for the logistic case, where

$$p(\mathbf{y}|\boldsymbol{\beta}) = \exp\left\{ \mathbf{y}^T \mathbf{X}\boldsymbol{\beta} - \mathbf{1}_n^T \log[\mathbf{1}_n^T + \exp(\mathbf{X}\boldsymbol{\beta})] \right\}.$$

We note the following representation of $-\log(1 + e^x)$ as the maxima of a family of parabolas:

$$-\log(1 + e^x) = \max_{\xi \in \mathbb{R}} \left\{ A(\xi) x^2 - \frac{1}{2} x + C(\xi) \right\} \quad \text{for all } x \in \mathbb{R}$$

where

$$A(\xi) \equiv -\tanh(\xi/2)/(4\xi)$$
$$\text{and} \quad C(\xi) \equiv \xi/2 - \log(1+e^{\xi}) + \xi\tanh(\xi/2)/4.$$

(Jaakkola and Jordan 2000).

This leads to the lower bound:

$$p(\mathbf{y}|\boldsymbol{\beta}) \geq c(\boldsymbol{\xi}) \exp\left\{\boldsymbol{\beta}^T \mathbf{F}(\boldsymbol{\xi})\boldsymbol{\beta} + \mathbf{f}(\boldsymbol{\xi})^T\boldsymbol{\beta}\right\} \tag{17}$$

where

$$\boldsymbol{\xi} = (\xi_1, \ldots, \xi_n) \text{ is a vector of variational parameters;}$$
$$\mathbf{F}(\boldsymbol{\xi}) = \mathbf{X}^T \text{diag}(A(\boldsymbol{\xi}))\mathbf{X};$$
$$\text{and} \quad \mathbf{f}(\boldsymbol{\xi}) = \left(\mathbf{y}^T\mathbf{X} - \tfrac{1}{2}\mathbf{1}_n^T\mathbf{X}\right)^T.$$

The following proof is adapted from Barber (2012, 28.5.2).

We have

$$\begin{aligned} p(\mathbf{y}) &\geq \frac{c(\boldsymbol{\xi})}{(2\pi\sigma_\beta^2)^{p/2}} \int \exp\left\{-\frac{\|\boldsymbol{\beta}\|^2}{2\sigma_\beta^2}\right\} \exp\left\{\boldsymbol{\beta}^T\mathbf{F}(\boldsymbol{\xi})\boldsymbol{\beta} + \mathbf{f}(\boldsymbol{\xi})^T\boldsymbol{\beta}\right\} d\boldsymbol{\beta} \\ &= \frac{c(\boldsymbol{\xi})}{(2\pi\sigma_\beta^2)^{p/2}} \int \exp\left\{-\frac{1}{2}\boldsymbol{\beta}^T\mathbf{A}\boldsymbol{\beta} + \mathbf{f}(\boldsymbol{\xi})^T\boldsymbol{\beta}\right\} d\boldsymbol{\beta} \end{aligned}$$

where $\mathbf{A} = \sigma_\beta^{-2}\mathbf{I} - 2\mathbf{F}(\boldsymbol{\xi})$. Completing the square and integrating, we have

$$\log p(\mathbf{y}) \geq \log c(\boldsymbol{\xi}) + \frac{1}{2}\mathbf{f}(\boldsymbol{\xi})^T\mathbf{A}^{-1}\mathbf{f}(\boldsymbol{\xi}) - \frac{p}{2}\log(\sigma_\beta^2) - \frac{1}{2}\log|\mathbf{A}| \equiv B(\boldsymbol{\xi}).$$

Now,

$$\begin{aligned} \log \underline{p}_G(\mathbf{y}) &= \int q_G(\boldsymbol{\beta}) \log\left[\frac{p(\mathbf{y},\boldsymbol{\beta})}{q_G(\boldsymbol{\beta})}\right] d\boldsymbol{\beta} \\ &\geq \log c(\boldsymbol{\xi}) + \mathbb{E}_{q_G}\left[-\tfrac{1}{2}\boldsymbol{\beta}^T\mathbf{A}\boldsymbol{\beta} + \mathbf{f}(\xi)^T\boldsymbol{\beta}\right] - \frac{p}{2}\log(2\pi\sigma_\beta^2) - \mathbb{E}_{q_G}[\log q_G(\boldsymbol{\beta})] \\ &= \log c(\boldsymbol{\xi}) + \mathbb{E}_{q_G}[\log \tilde{q}(\boldsymbol{\beta})] - \mathbb{E}_{q_G}[\log q_G(\boldsymbol{\beta})] - \frac{p}{2}\log(2\pi\sigma_\beta^2) \\ &\quad + \tfrac{1}{2}\log|2\pi\mathbf{A}^{-1}| + \tfrac{1}{2}\mathbf{f}(\boldsymbol{\xi})^T\mathbf{A}^{-1}\mathbf{f}(\boldsymbol{\xi}) \end{aligned} \tag{18}$$

where $\tilde{q}(\boldsymbol{\beta}) = \mathcal{N}(\boldsymbol{\beta}|\mathbf{A}^{-1}\mathbf{f}(\boldsymbol{\xi}, \mathbf{A}^{-1})$. Now, $\mathbb{E}_{q_G}[\log q_G(\boldsymbol{\beta})] - \mathbb{E}_{q_G}[\log \tilde{q}(\boldsymbol{\beta})]$ is the Kullback-Leibler distance between q_G and $\tilde{q}$. This is minimised when $q_G(\boldsymbol{\beta}) = \tilde{q}(\boldsymbol{\beta})$. Hence, maximising (18) gives

$$\log \underline{p}_G(\mathbf{y}) \geq \log c(\boldsymbol{\xi}) + \frac{1}{2}\mathbf{f}(\boldsymbol{\xi})^T\mathbf{A}^{-1}\mathbf{f}(\xi) - \frac{p}{2}\log(\sigma_\beta^2) - \frac{1}{2}\log|\mathbf{A}| \equiv B(\boldsymbol{\xi}).$$

Thus, the method of Gaussian variational approximation always gives a better approximation to $\log p(\mathbf{y})$ than that of local variational approximation.

8 Conclusion

Variational approximations have the potential to become an important tool in statistical inference, particularly in problems involving large datasets where the use of MCMC becomes untenable. This project shows that for Bayesian generalised linear models, the specific technique of Gaussian variational approximation yields estimators which can be used for valid statistical inferences.

9 Acknowledgements

I would like to thank my supervisor, Dr John Ormerod, for his support, without which this project would not have been possible, and in particular for his assistance in the theoretical results and for supplying Lemma 2.

Additionally, I would like to thank AMSI for their generous support in facilitating this valuable research experience and CSIRO for providing the opportunity to both present my findings and learn about the exciting research being undertaken by other vacation scholars across Australia at the Big Day In conference.

References

Barber, D. (2012), *Bayesian Reasoning and Machine Learning*, New York, United States of America: Cambridge University Press.

Bishop, Y. M. M., Fienberg, S. E., and Holland, P. W. (1975), *Discrete Multivariate Analysis: Theory and Practice*, Cambridge, Massachusetts, and London, England: The MIT Press.

Claeskens, G. and Hjort, N. L. (2010), *Model Selection and Model Averaging*, Cambridge, United Kingdom: Cambridge University Press.

Horn, R. A. and Johnson, C. R. (1985), *Matrix Analysis*, Cambridge, United Kingdom: Cambridge University Press.

Jaakkola, T. S. and Jordan, M. I. (2000), "Bayesian Parameter Estimation via Variational Methods," *Statistics and Computing*, 10, 25–37.

Jordan, M. (2004), "Graphical Models," *Statistical Science*, 19, 140–155.

Kullback, S. and Leibler, R. A. (1951), "On Information and Sufficiency," *The Annals of Mathematical Statistics*, 22, 79–86.

McGrory, C. A. and Titterington, D. M. (2007), "Variational approximations in Bayesian model selection for finite mixture distributions," *Computational Statistics and Data Analysis*, 51, 5352–5367.

Ormerod, J. T. and Wand, M. P. (2012), "Gaussian Variational Approximate Inference for Generalized Linear Mixed Models," *Journal of Computational and Graphical Statistics*, 21, 2–17.