Bayesian Inference in Neural Networks

Robert L. Paige\(^1\) and Ronald W. Butler  
Department of Statistics  
Colorado State University  
Fort Collins, CO 80523 USA

**Summary.** Approximate marginal Bayesian computation and inference are developed for neural network models. The marginal considerations include: (i) determination of approximate Bayes factors for model choice about the number of nonlinear sigmoidal terms; (ii) approximate predictive density computation for a future observable; (iii) determination of approximate Bayes estimates for the nonlinear regression function; and (iv) marginal density computation. Important use is made of the inherent partial linearity of the model which leads to an explicit marginal posterior on the nonlinear parameter when used with appropriate conjugate priors. Such exact marginalization simplifies marginal Bayes calculations so that further marginalization in the nonlinear parameter can be performed using a Laplace approximation. The choice of prior and the use of an alternative sigmoidal lead to posterior invariance in the nonlinear parameter which is discussed in connection with the lack of sigmoidal identifiability. The proposed methods are illustrated in the context of two nonlinear data sets: a nonlinear regression model and a nonlinear autoregressive time series.

*Keywords:* Bayesian computation; Laplace approximation; Model choice; Neural Network; Prediction

1 **Introduction**

Neural network (NN) models have enjoyed considerable popularity in recent years. They have been applied to many nonlinear regression problems and have been used extensively for time series prediction, see Faraway and Chatfield (1998) and Weigend and Gershenfeld (1993). They also have been successfully applied in pattern recognition as reviewed in Bishop (1995) and Ripley (1996).

In the NN model, the expectation of observation \(y_i\) is a "quasi-linear" function of independent variable \(x_i = (1, x_{i1}, \ldots, x_{im})^T\) for \(i = 1, \ldots, n\). With \(y = \)

\(^1\)Address for Correspondence: Department of Statistics, Colorado State University, Ft. Collins, CO 80523, USA  
E-mail: paige@stat.colostate.edu
\((y_1, y_2, \ldots, y_n)^T\), the model is

\[ y = X_\alpha \beta + \tau \varepsilon \]

where \(\varepsilon = (\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n)^T \sim N(0, I_n), \tau > 0\) is a scale parameter, \(X_\alpha\) is an \(n \times p\) design matrix of the form

\[
X_\alpha = \begin{bmatrix}
  x_1^T & \sigma(x_1^T \alpha_1) & \ldots & \sigma(x_1^T \alpha_q) \\
  x_2^T & \sigma(x_2^T \alpha_1) & \ldots & \sigma(x_2^T \alpha_q) \\
  \vdots & \vdots & \ddots & \vdots \\
  x_n^T & \sigma(x_n^T \alpha_1) & \ldots & \sigma(x_n^T \alpha_q)
\end{bmatrix},
\]

and \(\beta\) is a \((p \times 1)\) linear regression vector with \(p = m + 1 + q\). The elements of \(X_\alpha\) are functions of the nonlinear parameter matrix \(\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_q)^T\) where \(\alpha_i = (\alpha_{i0}, \alpha_{i1}, \ldots, \alpha_{im})^T\) and \(\sigma(x)\) is a sigmoidal usually taken to be the logistic function \((1 + e^{-x})^{-1}\). The convention has been to work with both \(y\) and \(x\)-component values translated and scaled to be in the range \([0, 1]\).

In neural network parlance this particular type of model is known as a "feed-forward neural network with one hidden layer, one skip layer, and a linear output unit", as described in Ripley (1996). "Feed-forward" refers to the vector \(x_i\) being mapped or "fed" forward to give \(E[y_i]\). This expectation is a linear combination of the nonlinear sigmoidal terms, \(\sigma(x_i^T \alpha_1), \ldots, \sigma(x_i^T \alpha_q)\), or "hidden layer" and the linear term, \(x_i^T\), or "skip layer" which "skips" past the sigmoidal transformation. In addition, this model has a "linear output unit" since the expectation of \(y\) is linear in the regression function \(X_\alpha \beta\).

The NN model is a highly flexible model. With sufficiently many sigmoidal terms, the regression function \(X_\alpha \beta\) can approximate any continuous function uniformly over a compact set to an arbitrary degree of precision as first shown by Cybenko (1988). Usually relatively few sigmoidal terms are required in statistical problems. Nonetheless there is a need to avoid overfitting the data with either abruptly changing or "saturated" sigmoidal terms or too many sigmoidals. We use a prior distribution on \((\alpha, \beta, \tau^2)\) that a prior favours smooth and non-abruptly changing sigmoidal terms and furthermore penalizes overfitting with too many sigmoidals. A similar approach
has been used in Ripley’s S-Plus routine "nnet" as described in Venables and Ripley (1997).

This paper addresses several aspects of marginal Bayesian inference for such NN models. First the issue of model choice is considered. If the model in (1) with \( q \) sigmoidals is referred to as \( M_q \), then marginal posterior probabilities for the various models \( \{M_q : q = 0, 1, \ldots \} \) are approximated using the method of Laplace. These computations determine approximate Bayes factors for use in model choice (determination of \( q \)) as well as in posterior mixing over models for prediction and estimation. We show in the numerical examples that the Bayes factors alone do not convey the information needed to select a parsimonious model with small \( q \). Models with extra sigmoidal terms that nest this parsimonious choice are weighted as heavily as the parsimonious model. However, with the additional information about the modal estimates used in Laplace approximations for Bayes factors, this parsimonious choice can be made.

Approximate predictive densities for future observable \( Y_f \) given its associated independent variable \( X_f \) are computed using Laplace approximations. Also computed are approximate Bayes estimates for the nonlinear regression relationship and approximate marginal posteriors on model parameters.

Our marginal inference is facilitated by making important use of the "partial linearity" in the regression parameters. This term refers to the fact that if \( \alpha \) were known, then the NN model would be a linear regression model in \( \beta \). As such, any marginal inference may be simplified considerably by marginalizing first with respect to the linear and scale terms. Using standard conjugate priors on \( (\beta, \tau^2) \), the exact marginal posterior density on nonlinear parameter \( \alpha \) is easily computed analytically. Then Laplace's approximation need only be applied for marginalizing in \( \alpha \). Previous use of Laplace's method has not taken advantage of partial linearity to allow exact marginalization in the parameters \( (\beta, \tau^2) \). The obvious benefits are simplicity as well as the greater expected numerical accuracy provided when Laplace’s approximation is used in lower dimensional integration. Partial linearity has been noted in Ripley (1996) but has only been used for parameter estimation in frequentist models by

The marginal posterior for $\alpha$ can have many local maxima. Therefore, when marginalizing in $\alpha$, a sum of Laplace approximations at the various local maxima is required. This procedure was proposed for marginalizing over $\alpha$ and $\beta$ in Buntine and Weigand (1991) and further discussed in Mackay (1992) and Ripley (1994a).

The methodology is illustrated on a small nonlinear regression data set from Bates and Watts (1988,§3.13). The form of the regression is not specified by the underlying theory, but the fit of the NN model may be interpreted as a "soft" switching regimes linear regression. A second larger financial time series data set from Lee, White, & Granger (1993) is also modelled using a lag one nonlinear autoregressive model. The NN fit may be interpreted as a "soft" threshold autoregressive model.

2 An Alternate Invariant Sigmoidal

There are several troubling aspects of these NN models related to our Bayesian computations. First, the parameter $(\alpha, \beta)$ lacks identifiability. To see this, consider a NN model with one independent variable, no linear term, and two sigmoidal terms. The expected value of $y$ is given by

$$E[y] = \beta_0 + \beta_1 \sigma (\alpha_{10} + \alpha_{11} x) + \beta_2 \sigma (\alpha_{20} + \alpha_{21} x).$$

(2)

This expectation is unchanged if one interchanges the parameter sets $(\beta_1, \alpha_{10}, \alpha_{11})$ and $(\beta_2, \alpha_{20}, \alpha_{21})$. Identifiability however is not an issue in prediction or Bayes regression estimation where the parameter $(\alpha, \beta)$ is integrated out. We shall return to this issue later as it concerns marginal inference for the components of $\alpha$.

Secondly, certain sets of parameter values are not equivalent but which should be. Consider a change in the sign of $(\beta_1, \alpha_{10}, \alpha_{11})$ while also adding $\beta_1$ to $\beta_0$. This parameter change leaves the expectation in (2) unchanged, or

$$\beta_0 + \beta_1 - \beta_1 \sigma (-\alpha_{10} - \alpha_{11} x) = \beta_0 + \beta_2 \sigma (\alpha_{10} + \alpha_{11} x).$$

These two sets of parameter values are said to be "expectation equivalent". These values receive equal weight from the data through the likelihood function but may
not have equivalent posterior weight because they may receive different prior weights. This lack of "posterior equivalence" is troublesome because in effect one arbitrary parametrization is assigned a higher posterior probability than another.

This posterior inequivalence is corrected by working instead with the translated sigmoidal,

\[ \sigma^*(x) = \sigma(x) - \frac{1}{2} = \frac{1}{2} \tanh \left( \frac{x}{2} \right) \]

which is an odd function about 0. The choice of prior for \((\alpha, \beta)\) also bears upon this issue and in this regard we suggest an exchangeable multivariate T prior with common mean 0 in the next subsection. Such a prior is an even function in each component and guarantees prior equivalence for parameter values that are expectation equivalent; thus posterior and expectation equivalence are one in the same. This corrects the conceptual difficulty above.

For ease of discussion, we shall refer to \(\sigma^*\) as a sigmoidal. Strictly speaking it is a translated sigmoidal since sigmoidals must be distribution functions.

The use of \(\sigma^*\) together with our choice of prior greatly reduce the number of local maxima that must be averaged when using Laplace's approximation to marginalize in \(\alpha\). This occurs because they lead to a marginal posterior in \(\alpha\) that is exchangeable and which therefore generates equivalence classes of local maxima over certain permutations and sign changes for \(\alpha\). This phenomenon has previously been pointed out by Bishop (1995) when using the tanh sigmoidal.

The use of sigmoidal \(\sigma^*\) also necessitates that we translate the \(y\)- and \(x\)-component values by subtracting 1/2 so they now fall in the range \([-1/2, 1/2]\). The first row entries of \(\{x_i\}\), which are all 1, are also replaced with the value 1/2.

3 Choice of Prior

To motivate the choice of prior, we must first understand the nature of the sigmoidal data transformations. Consider the graph of \(\sigma^*(ax)\) versus \((a, x)\) for \(-15 \leq a \leq 15\) and \(-1/2 \leq x \leq 1/2\) as given in Figure 1 below. Values of \(a\) centered about 0 result in a functions which are flat or gradual whereas larger values of \(|a|\) yield functions
which are steep. As \(|a| \to \infty\) the sigmoidal approaches a unit step function and is said to approach "saturation". More generally, \(\beta \sigma^*(\alpha_0 + \alpha_1 x)\) will avoid saturation and be gradual and smooth if the values of \(\alpha_0, \alpha_1,\) and \(\beta\) fall within the range \((-6, 6)\). A summation of such terms yields a regression function which is also smooth and which tends not to overfit the data.

In this regard, we assume a marginal prior structure for \((\alpha, \beta)\) that is an exchangeable multivariate \(T\) and which gives each component the common mean 0 and variance 4. The support of each component is concentrated in the range \((-6, 6)\) so that prior weight is given to smooth regression functions. As previously mentioned, this prior is an even function in each component and assures that expectation and posterior equivalence are the same. Such a marginal prior is specified hierarchically using a conditional prior for \((\alpha, \beta)\) given \(\tau^2\) that is \(\text{N}(0, (\tau^2/\lambda) I_{p+q(m+1)})\) with the fixed value of \(\lambda = 10^{-3}/4\) explained below. A conjugate prior is used for \(\tau^2\) which is the inverse gamma distribution \(\Gamma^{-1} (\nu/2, \gamma/2)\) (see Lee, 1989, §A.5) with density

\[
\pi (\tau^2) \propto (\tau^2)^{-\nu/2-1} \exp \left\{ -\frac{\gamma}{2\tau^2} \right\} \tag{3}
\]

and parameters \(\nu = 3\) and \(\gamma = 4\lambda\). The parameter choices represent vague prior knowledge in the following sense. The first parameter measures the prior information for \(\tau^2\) and \(\nu = 3\) is its smallest integer value for which (3) has finite mean. The second parameter assures that \(E [\tau^2] = 4\lambda\) so that the marginal prior mean for components of \((\alpha, \beta)\) is 0 with variance 4. These parameter choices assure a prior in which

\[
\tau^2/\lambda \sim \Gamma^{-1} (3/2, 2)
\]

for any choice of \(\lambda > 0\).

Our choice of \(\lambda = 10^{-3}/4\) has been motivated by the desire to seek prior choices using sigmoidal \(\sigma^*\) that are compatible with the penalty choices of Ripley (1996) when working with sigmoidal \(\sigma\). Using \(\lambda = 10^{-3}/4\), then our choice of prior combines with the normal likelihood to produce a posterior that depends on the penalized least squares error defined as
\[ SSE^*(\alpha, \beta) = \| y - X_\alpha \beta \|^2 + \lambda \| (\alpha, \beta) \|^2. \] 

The latter penalty term in (4) is called a "weight decay". Expression (4) may be minimized over \((\alpha, \beta)\) by using Ripley's "nnet" routine. For \(x\)- and \(y\)-values in \([0, 1]\), and a moderate number of parameters, Ripley (1996) has suggested the penalty weight \(\lambda \in (10^{-4}, 10^{-2})\). Our selected value \(\lambda = 10^{-3}/4\) is consistent with Ripley's choice of \(\lambda = 10^{-3}\) for the following reason. Our \(x\)- and \(y\)-values fall in the range \([-1/2, 1/2]\) instead of \([0, 1]\) with half the magnitude; this allows us to double the value of \((\alpha, \beta)\) but this must also be compensated for in the weight decay factor \(\| (\alpha, \beta) \|^2\) by using a quarter of \(\lambda \in \left(10^{-4}, 10^{-2}\right)\).

To summarize, the choice of prior has been motivated in several ways. An exchangeable multivariate \(T\) prior on \((\alpha, \beta)\) with mean 0 and variance 4 that incorporates the penalty weight \(\lambda = 10^{-3}/4\) makes saturation of sigmoidals difficult and leads to a smooth well-behaved marginal posterior surface in \(\alpha\) with a decreased number of local maxima and few false maxima (saddlepoints). This smooth surface provides improved performance for quasi-Newton maximization routines when implementing Laplace's method in \(\alpha\). From a Bayesian perspective, the prior choice expresses an a priori preference for smooth regression functions, perhaps as an expression of Occam's razor for simple smooth relationships.

4 Posterior Calculation

The priors on \((\alpha, \beta, \tau^2)\) combine with normal likelihood to yield posterior

\[ \pi(\alpha, \beta, \tau^2 | y) \propto (\tau^2)^{-(n+p+q(m+1)+v)/2} \exp \left\{ -\frac{SSE^*(\alpha, \beta) + \gamma}{2\tau^2} \right\}. \]

This expression admits explicit marginalization in \(\beta\) followed by \(\tau^2\). Completing the square in \(\beta\) and integrating out yields

\[ \pi(\alpha, \tau^2 | y) \propto (\tau^2)^{-(n+q(m+1)+v)/2} |X_\alpha^T X_\alpha + \lambda I_p|^{-1/2} \exp \left\{ -\frac{E(\alpha)}{2\tau^2} \right\} \]

with

\[ E(\alpha) = s_\alpha^2 + B_\alpha + \gamma + \lambda \| \alpha \|^2. \] 

(5)
In (5), $s^2_\alpha$ is the error sum of squares of linear regression treating $\alpha$ as fixed, or

$$s^2_\alpha = y^T (I - X_\alpha X_\alpha^+ ) y$$

with $X_\alpha^+$ as the Moore-Penrose inverse of $X_\alpha$, and

$$B_\alpha = \hat{\beta}_\alpha X_\alpha^T \left[ I_n - X_\alpha (X_\alpha^T X_\alpha + \lambda I_p)^{-1} X_\alpha^T \right] X_\alpha \hat{\beta}_\alpha$$  

(6)

with $\hat{\beta}_\alpha$ as the least squares estimate of $\beta$ with $\alpha$ fixed or $\hat{\beta}_\alpha = X_\alpha^+ y$. Marginalization in $\tau^2$ yields

$$\pi (\alpha | y) \propto \bar{\pi}(\alpha | y) = c |X_\alpha^T X_\alpha + \lambda I_p|^{-1/2} E(\alpha)^{-(n+q(m+1)+\nu)/2}$$  

(7)

where $\pi$ is the true posterior, $\bar{\pi}$ is the unnormalized posterior from exact marginalization in $\beta$ and $\tau^2$, and

$$c = \Gamma \left( \frac{n+q+\nu}{2} \right) \lambda^{(p+q)/2} \gamma^{\nu/2} / \Gamma \left( \frac{\nu}{2} \right) \pi^{(n+q)/2}.$$  

For large $n$, the term $B_\alpha$ is close to 0 and negligible so that it may be left out of posterior computation. This occurs with informative designs in which the likelihood contribution $X_\alpha^T X_\alpha$ "washes out" prior term $\lambda I_p$ in the centre matrix of (6). In this setting the centre matrix approximates the residual projection matrix that is orthogonal to the columns of $X_\alpha$ so  that $B_\alpha \approx 0$. The expansion

$$(X_\alpha^T X_\alpha + \lambda I_p)^{-1} = (X_\alpha^T X_\alpha)^{-1/2} \left[ I_n + \sum_{k=1}^{\infty} \left\{ -\lambda (X_\alpha^T X_\alpha)^{-1} \right\}^k \right] (X_\alpha^T X_\alpha)^{-1/2}$$

$$= (X_\alpha^T X_\alpha)^{-1} - \lambda (X_\alpha^T X_\alpha)^{-2} + \lambda^2 (X_\alpha^T X_\alpha)^{-3} + \cdots$$

when substituted into (6) gives

$$B_\alpha = 0 - \lambda \hat{\beta}_\alpha \hat{\beta}_\alpha + \lambda^2 \hat{\beta}_\alpha \hat{\beta}_\alpha \left( X_\alpha^T X_\alpha \right)^{-1} \hat{\beta}_\alpha + \cdots$$

$$= O(\lambda/n) + O \left( (\lambda/n)^2 \right) + \cdots.$$

Further simplification occurs with large $n$. The factor $|X_\alpha^T X_\alpha + \lambda I_{p+1}|$ is dominated by the $E(\alpha)$ term whose important components now include

$$\bar{\pi}(\alpha | y) \simeq c \left( s^2_\alpha + \lambda \| \alpha \|^2 \right)^{-(n+q(m+1)+\nu)/2}.$$
More generally, for any nonlinear normal regression model with a partially linear component, the marginal posterior on nonlinearity parameter \( \alpha \) is given in (7) when computed with conjugate priors. Letting \( \lambda \to 0 \) and setting \( \nu = 0 = \gamma \) corresponds to using Jeffreys' priors for which the posterior on \( \alpha \) is

\[
\pi (\alpha | y) \propto C |X_\alpha^T X_\alpha|^{-1/2} |s_\alpha|^{-(n+q(m+1))}.
\]

Modal estimates of \( \alpha \) essentially minimize the empirical error \( s_\alpha^2 \) for large \( n \).

5 Posterior Symmetry in \( \alpha \)

The posterior on \( \alpha \) is invariant under two groups of transformations. First, because the sigmoidals are not identified, it is invariant to the \( q! \) permutations of the \( q \) parameter sets within common sigmoidals. Secondly, because of its symmetry about 0, it is invariant to the \( 2^q \) possible sign changes applied to these same parameter sets. Together the posterior is invariant under a single group of \( 2^q q! \) transformations. The symmetry of the marginal posterior on \( \alpha \) can be easily seen by expressing it as the integral of likelihood \( \times \) prior and noting that this integral is symmetric because of the choice of symmetric priors and the oddness property for \( \sigma^* \).

The group under which \( \pi (\alpha | y) \) is invariant forms a subgroup of the orthogonal group of transformations on \( \alpha \). This subgroup is the largest such subgroup which has relevance to the NN model. To show this, consider the nonlinear part of the design matrix

\[
\begin{bmatrix}
\sigma (x_1^T \alpha_1) & \cdots & \sigma (x_1^T \alpha_q) \\
\sigma (x_2^T \alpha_1) & \cdots & \sigma (x_2^T \alpha_q) \\
\vdots & \ddots & \vdots \\
\sigma (x_n^T \alpha_1) & \cdots & \sigma (x_n^T \alpha_q)
\end{bmatrix}
:= \sigma
\left(
\begin{bmatrix}
x_1^T \\
x_2^T \\
\vdots \\
x_n^T
\end{bmatrix}
\begin{bmatrix}
\alpha_1 \\
\alpha_2 \\
\vdots \\
\alpha_q
\end{bmatrix}
\right).
\]

The group maps \( \alpha \to O\alpha \) and an orthogonal matrix \( O \) with more than one non-zero entry in the \( i \)th row will result in a transformed parameter matrix \( O\alpha \) whose \((i, j)\)th term is a linear combination of \( \alpha_{1, j-1}, \alpha_{2, j-1}, \ldots, \alpha_{q, j-1} \) which is not permissible in the NN model. Matrix \( O \) must therefore have only one non-zero entry in each row.
which necessarily must be either "1" or " - 1" if it is to be orthogonal; thus \( O \) is a member of the subgroup of size \( 2^q q! \).

The group invariance leads to equivalence classes of parameter values sharing common orbits under the invariant group of size \( 2^q q! \). Members of the same orbit are expectation equivalent parameter values. In each orbit, one member has location parameters for the \( q \) sigmoids that satisfy

\[
0 \leq \alpha_{10} \leq \ldots \leq \alpha_{q0}
\]  

(8)

and which can be used to identify the orbit. Starting with arbitrary matrix parameter \( \alpha \), it is transformed into its identified form by ordering the vectors \( \{ \alpha_{i \text{sgn} (\alpha_{i0})} : i = 1, \ldots, q \} \) according to their first components and letting these form the columns of the matrix \( \alpha_{MI}^T \). This identified version is the "maximal invariant" parameter under the invariant group. We shall denote the set of maximal invariant parameter values as \( O = \{ \alpha_{MI} \} \subset \mathbb{R}^{q(m+1)} \). For parameter \( \alpha_{MI} \in O \), the sigmoids essentially end up numbered from left to right according to how the sigmoids are centred in the regression.

6 Approximate Posterior Expectation in \( \alpha \)

6.1 Laplace's Approximation with Multiple Modes

Laplace's method was introduced for Bayesian marginalization by Leonard (1982), Davison (1986), and Tierney and Kadane (1986). More recently it has found application as a means for computing approximate Bayes factors as discussed in Berger and Pericchi (1996), Kass and Raferty (1995), and DiCiccio et al. (1997). All of these authors, however, consider the unimodal situation and not the multimodal setting commonly arising with NN models. This latter setting has been discussed by Buntine and Weigand (1991), Mackay (1992), Bishop (1995), and Ripley (1994a). Multimodal marginalization has also been considered with MCMC by Neal (1993, 1995).

The posterior expectation of an arbitrary smooth positively-valued function \( g(\alpha) \) is computed using Laplace's approximation adapted to deal with the multiple local
maxima found in this context. Our context requires marginalization in $\alpha$ and the approximation is

$$
\int g(\alpha) \pi(\alpha|y) d\alpha \simeq \sum_{\hat{\alpha}} g(\hat{\alpha}) L(\hat{\alpha})
$$

(9)

where

$$
L(\hat{\alpha}) = \frac{c(E(\hat{\alpha}))^{-(n+q(m+1)+\nu) / 2}}{\sqrt{|X^T_{\hat{\alpha}}X_{\hat{\alpha}} + \lambda I_p| \times \|H(\hat{\alpha})\|}}
$$

The values $\{\hat{\alpha}\}$ comprise the set of local maxima for the dominant portion of $\ln \pi(\alpha|y)$ taken to be

$$
-\frac{1}{2} (n + q(m + 1) + \nu) \ln E(\alpha).
$$

(10)

Matrix $H(\alpha)$ is the Hessian of (10). Determination of the collection of local maxima is the most computationally intensive part of this computation. For this we used Gauss-Newton searches for critical values starting the iterations at randomly selected values of $\alpha$.

When $g(\cdot)$ is invariant under the permutation group of Section 5, the summation in (9) may be restricted to summing over only the identified local maxima parameters in $O$ when the multiplicity factor $2^q q!$ is also used. Then (9) is

$$
\int g(\alpha) \pi(\alpha|y) d\alpha \simeq \sum_{\hat{\alpha}} g(\hat{\alpha}) L(\hat{\alpha}) = 2^q q! \sum_{\hat{\alpha} \in O} g(\hat{\alpha}) L(\hat{\alpha}).
$$

(11)

Thus the search for local maxima $\hat{\alpha}$ is accordingly restricted to starting Gauss-Newton interaction within $O$.

6.2 Further Approximation Issues

Critical values of $\ln E(\cdot)$ require the computation of $\partial \ln E(\alpha)/\partial \alpha$ which is straightforward except for the dependence of $E(\cdot)$ on $X^+_\alpha$. Golub and Pereyra (1973) determine the $(i,j)^{th}$ component of this derivative as

$$
\partial X^+_\alpha / \partial \alpha_{ij} = -X^+_\alpha (\partial X_\alpha / \partial \alpha_{ij}) X^+_\alpha + X^+_\alpha (X^+_\alpha)^T (\partial X_\alpha / \partial \alpha_{ij})^T (I - X_\alpha X^+_\alpha)
$$

$$
+ (I - X^+_\alpha X_\alpha) (\partial X_\alpha / \partial \alpha_{ij})^T (X^+_\alpha)^T X^+_\alpha
$$
when $X_\alpha^T X_\alpha$ has locally constant rank which might not be its full rank. From this we are able to derive an exact expression for the gradient of $\ln E (\alpha)$. For Hessian computation, sufficient accuracy was obtained by using finite differences of the gradient.

Special care must be taken when working with the numerical Moore-Penrose inverse of $X_\alpha$. This requires singular value decomposition and also a determination about which of the small singular values should be considered negligible and taken as 0. We used the IMSL routine DLSGRR and found that a tolerance for negligibility set at $10^{-8}$, approximately the square root of machine precision, worked well. This setting was required due to the amount of relative error caused by the use of other numerical routines.

The stable estimate $\hat{\alpha}_s$ is defined as the dominant Bayesian mode with $\lambda = 0$ and is nearly the dominant Bayesian mode, $\hat{\alpha}$ with $\lambda = 10^{-3}/4$ since

$$\hat{\alpha}_s = \hat{\alpha} + O \left( n^{-1} \right).$$

Despite this, it is still preferable to use a weight decay term for large $n$. The weight decay term makes the maximization of $-\ln E (\alpha)$ easier, gives fewer maxima, results in Hessians that are usually nonsingular and yields a surface which is more quadratic in appearance. This last point is particularly relevant for achieving accuracy with Laplace's approximation.

7 Marginal Inference

Model Choice. The NN model with $q$ sigmoidal terms is denoted $M_q$ and has Bayes factor

$$\Pr \{ M_q | y \} \propto \pi (y | M_q) \pi (M_q),$$

where

$$\pi (y | M_q) = \int \pi (\alpha | y, M_q) d\alpha \quad (12)$$

is approximated using Laplace's method with $g (\alpha) \equiv 1$. Variable selection from among the $m$ independent variables in $\{ x_i \}$ could also be implemented with or without sigmoidal selection but this has not been attempted.
In this context, the "best" model is not always that which maximizes \( \Pr \{ M_q | y \} \) in \( q \). Suppose, for example, the true model has a single sigmoidal so \( q = 1 \). The NN models are nested with \( M_0 \subset M_1 \subset M_2 \subset \cdots \) so that, for example, a model with 2 sigmoidals is a 1 sigmoidal model if one of the sigmoidals has all its non-intercept \( \alpha \)-values set to zero. Thus there is no unique correct choice for \( q \) when \( M_1 \) is the true parsimonious model; the choice of any value of \( q \in \{ 1, 2, \ldots \} \) is as good as any other. We shall see the occurrence of this phenomenon in the examples where the Bayes factors for 2 and 3 sigmoidals are greater than that for 1 but the parsimonious choice of model is clearly a single sigmoidal. The Bayes factors in combination with their fitted parameter values need to be examined in order to make a parsimonious model choice for value \( q \).

Prediction. The posterior density for future observable \( Y_f \) at \( y_f \), given its associated variable \( x_f = (1, x_{f1}, \ldots, x_{fm}) \) and model \( M_q \) is easily computed by including \( y_f \) and \( x_f \) into the data set and marginalizing as in model choice. Then

\[
\pi (y_f | x_f, y, M_q) = \frac{\pi (y, y_f | M_q)}{\pi (y | M_q)} = \frac{\int \pi (\alpha | y, y_f, M_q) \, d\alpha}{\int \pi (\alpha | y, M_q) \, d\alpha}
\]

where both Laplace approximations take \( g(\alpha) \equiv 1 \). The denominator is the computation from model choice and the numerator is the same computation but including the prospective future observable \( y_f \) in the data. Further marginalization over the models \( \{ M_q \} \) with a priori weighting \( \pi (M_q) \propto 1 \) is determined as

\[
\pi (y_f | x_f, y) = \frac{\sum_q \pi (y_f | x_f, y, M_q) \, \pi (y | M_q)}{\sum_q \pi (y | M_q)} = \frac{\sum_q \pi (y, y_f | M_q)}{\sum_q \pi (y | M_q)}
\]

and may be based entirely upon Laplace approximations as used to compute (12) and (13). When a clear choice for a parsimonious model exists, as occurs in our examples, then the mixing of models in (14) is not much different from the use of (13) based upon the parsimonious choice. In such instances, use of the parsimonious model can greatly simplify the computations.

Bayes Estimation of the Regression Function. Taking \( \theta = (\alpha, \beta, \tau^2) \), then the posterior expectation of the regression function is
\[
E_y^\theta (X_\alpha \beta | y, M_q) = \int X_\alpha \beta \pi (\theta | y, M_q) \, d\theta \\
= \int X_\alpha \Sigma_1 X_\alpha \beta \hat{\alpha} \pi (\alpha | y, M_q) \, d\alpha
\]  \hspace{1cm} (15)

since
\[
\beta | \alpha, \tau^2, y \sim N \left( \Sigma_1 X_\alpha X_\alpha \hat{\beta} \alpha, \tau^2 \Sigma_1 \right)
\]

with \( \Sigma_1 = (X_\alpha^T X_\alpha + \lambda I_p)^{-1} \). In terms of \( \hat{\pi} \), (15) is
\[
E_y^\theta (X_\alpha \beta | y, M_q) = \frac{\int X_\alpha \Sigma_1 X_\alpha^T X_\alpha \beta \hat{\alpha} \hat{\pi} (\alpha | y, M_q) \, d\alpha}{\int \hat{\pi} (\alpha | y, M_q) \, d\alpha}.
\]  \hspace{1cm} (16)

Laplace’s approximation in the numerator of (16) takes
\[
g (\alpha) = X_\alpha \Sigma_1 X_\alpha^T X_\alpha \beta \alpha.
\]

Since the regression function is invariant under the permutation group, Laplace’s approximation as in (11) may be used. This invariance exists because expectation and posterior equivalence are the same in our framework. The denominator is \( \pi (y | M_q) \) in (12) and its Laplace approximation also determines the Bayes factor for model \( M_q \). Further marginalization over the models \( \{ M_q \} \) is possible and leads to a Bayes estimator that is a mixture of the Bayes estimates in (16) of the form (14). This is not particularly beneficial, however, when there is a clear cut choice for a parsimonious model.

Marginal Distributions. Inference about marginal densities in \( \alpha \) is only meaningful once the various sigmoidals have been identified as described in Section 5. Consider such an identified NN model in which \( \alpha \in O \) has been partitioned into \( \alpha(1) \) and \( \alpha(2) \) where \( \alpha(1) \) is one-dimensional and \( \alpha(2) \) contains all the other components. The marginal posterior of \( \alpha(1) \) is
\[
\pi (\alpha(1) | y, M_q) = \frac{\int_{O_1} \hat{\pi} (\alpha(1), \alpha(2) | y, M_q) \, d\alpha(2)}{\int \hat{\pi} (\alpha | y, M_q) \, d\alpha},
\]  \hspace{1cm} (17)

where \( O_1 = \{ \alpha(2) : (\alpha(1), \alpha(2)) \in O \} \). A Laplace approximation to the denominator is the Bayes factor computation above and is summed over the identified members in
An approximation to the numerator holds \( \alpha_{(1)} \) fixed and applies Laplace's method in parameter \( \alpha_{(2)} \) with \( g = 1 \). Only identified values of \( (\alpha_{(1)}, \hat{\alpha}_{(2)}) \in O \) are used where \( \hat{\alpha}_{(2)} \) is an identified maximum determined by holding \( \alpha_{(1)} \) fixed.

8 Examples

8.1 Nonlinear Regression

Our first data set was analyzed in Bates and Watts (1988) and concerns the utilization of nitrate (y-variable) in bush beans as a function of light intensity (x-variable). The primary leaves of three 16-day-old bean plants were subjected to eight levels (2.2, 5.5, 9.6, 17.5, 27.0, 46.0, 94.0, 170.0) of light intensity (\( \mu E/m^2s \)) and the nitrate utilizations (mmol/g hr) were measured. The experiment was performed on two different days. Even though no theoretical model has been proposed for this data, it has been hypothesized, by the researchers, that utilization should be zero for zero light intensity and should also approach an asymptote as light intensity increases. Incremental parameter effects for the two days were included in the analysis of Bates and Watts, for models having an asymptote, but were not found to be significant. Therefore, we combine the data from both days and fit a common model to the \( n = 48 \) cases of data.

We consider NN models with up to three sigmoidal terms using \( \lambda = 10^{-3}/4 \). The factors \( \{ \pi (y|M_q) \} \) which determine the posterior weights of the various models are approximated using Laplace’s method and displayed in Table 1. For each \( M_q \), the local maxima were determined by fitting the model 100 times starting the Gauss-Newton iterations at randomly selected points in \( \alpha \in [-9, 9]^{q(m+1)} \cap O \). The various maxima are given in maximal invariant form \( \hat{\alpha}_{MI} \) along with the associated contributions to the Laplace approximation from the orbits of maxima. Starred entries are used in place of parameter estimates whose values are less than \( 10^{-12} \) in magnitude.

The Bayes factor totals favour 1, 2, and 3 nodes but these alone do not fully explain the regression relationship. An examination of the modal parameter estimates from Laplace’s method is more informative about parsimonious model choice. There
are two maxima when fitting the single node models in $M_1$. When fitting two node models, these same two nodes are essentially the two dominant maxima as may be seen from their weights in the first sigmoidal which are all less than $10^{-12}$ in magnitude. The percentage of the Bayes factor contributed by these two dominant maxima is 96.0%. Thus the fit of two node models also suggests that one sigmoidal is a parsimonious fit. The fit of three node models explains nothing that hasn’t already been explained with two node models, since all three maxima are essentially those from the two node models. The two dominant maxima that represent a single node model combine for 97% of the Bayes factor when fitting three nodes.

To assess the accuracy of Laplace’s method in determining the Bayes factors of the models above, we evaluated these quantities using numerical integration. For 1, 2, and 3 node models these factors were $1.733 \times 10^{20}$, $1.577 \times 10^{21}$, and $6.865 \times 10^{22}$. They agree closely with the totals given in Table 1.

Figure 2 shows a scatterplot of the regression data (circles) along with several estimates of the regression function: (i) the true Bayes estimate (solid) as in (16) assuming a one node model as determined from numerical integration; (ii) Laplace’s approximation to this estimate (dashed); and (iii) Laplace’s approximation to the mixture of Bayes estimates (dot-dash) where the mixing is over models with up to three nodes and weights are proportional to Bayes factors. Numerical integration and Laplace approximation yield indistinguishable estimates. A comparison of the dashed and dot-dashed lines reveals that very little is lost by restricting attention to the parsimonious model with a single node.

Figure 3 shows a predictive density for future value $Y_f$ given $x_f = 187$ which provides an extrapolation for $Y_f$ to the right of the data in the regression plot. The true predictive density based upon a single node (solid) is determined with numerical integration and its normalized Laplace approximation (dashed) is reasonably accurate.

The researchers who gathered the data expected a "levelling off" of the regression as light intensity increases. This can be seen in all the regression fits in Figure 2. It also is reflected in the location of the predictive density plot in Figure 3; the
approximate mean of $1.8 \times 10^4$ is consistent with the asymptotic level determined from Figure 2. By contrast, the frequentist model eventually proposed by Bates and Watts did not suggest this "leveling off" behavior.

For the parsimonious fit of a single node, Figure 4 shows a surface plot for the marginal bivariate posterior of nonlinear parameters $\alpha_{10}$ (intercept) and $\alpha_{11}$ (slope). Figure 5 is the marginal posterior on $\alpha_{11}$ (slope) as determined from numerical integration (solid) and Laplace's method (dashed) when normalized. The Laplace approximation is extremely accurate.

8.2 Nonlinear Time Series

Our next data set consists of monthly U.S. personal income data from January, 1959 to July, 1990 (inclusive) for a total of 379 observations and is shown in Figure 6. It is taken from Lee, White, & Granger (1993) where tests for neglected nonlinearity are compared against their newly proposed NN test. These authors transformed the personal income data $\{z_t\}$ into a stationary sequence of seasonally adjusted $y_t = \ln(z_{t+1}/z_t)$ values. A time plot for $\{y_t\}$ is shown in Figure 7. The $y_t$-values were then fit to an $AR(k)$ model where $k$ was found to be one using the SIC criterion. All the tests for the presence of nonlinearity, including the Keenan, Tsay, Ramsay RESET (1 & 2), White (1, 2 & 3), Mcleod and Li, BDS, Bispectrum tests and their newly proposed NN test, gave p-values below 0.016 which strongly suggests the presence of neglected nonlinearity. We modelled this time series as a non-linear autoregressive process via a NN model with q sigmoidal terms where our distributions are conditional upon the observed value of $Y_t$. Once again we take $\lambda = 10^{-3}/4$. Since the data consist of a large number of observations, we might expect the prior to be dominated by the likelihood and so its choice is not so critical.

Our parameter estimation results are given in Table 2. A careful examination of the table suggests a parsimonious fit with one sigmoidal. As seen with the previous example, the best fits with one, two and three sigmoidals have increasing Bayes factors. Their associated parameter estimates, however, support the parsimonious fit of a single sigmoidal.
We may assess the accuracy of Laplace’s method for determining Bayes factors in the models above by using numerical integration. For 1, 2, and 3 nodes these factors are $3.756 \times 10^{14}$, $1.278 \times 10^{15}$, and $1.523 \times 10^{16}$ which show reasonable agreement with the totals in Table 2.

Figure 8 shows a scatterplot of the regression data (circles) along with several estimates of the regression function: (i) the true Bayes estimate (solid) as in (16) assuming a one node model as determined from numerical integration; (ii) Laplace’s approximation to this estimate (dashed); and (iii) Laplace’s approximation to the mixture of Bayes estimates (dot-dash) where the mixing is over models with up to three nodes and weights are proportional to Bayes factors. Numerical integration and Laplace approximation yield indistinguishable estimates. A comparison of the dashed and dot-dashed lines reveals that virtually nothing is lost in working with the NN model with a single node. This model also corresponds to a ”soft” threshold autoregressive model. Fuller (1976) considers such models in his analysis of the Canadian Lynx time series.

Figure 9 shows a predictive density for future value $Y_{378}$ given $y_{377} = 5.779 \times 10^{-3}$. The true predictive density based upon a single node (solid) is determined with numerical integration and its normalized Laplace approximation (dashed) is once again quite accurate. The predictive mean of approximately 0.007 appears to be a reasonable extrapolation of the data in Figure 7.

For the parsimonious fit of a single node, Figure 10 shows a surface plot for the marginal bivariate posterior of nonlinear parameters $\alpha_{10}$ (intercept) and $\alpha_{11}$ (slope). Figure 11 is the marginal posterior on $\alpha_{11}$ as determined from numerical integration (solid) and Laplace’s method (dashed) when normalized. The Laplace approximation is extremely accurate.

9 Acknowledgments

The authors thank Dr. Kelly McArthur for her advice on singular value decomposition and Dr. Tae-Hwy Lee for his assistance in procuring the time series data set. The research was partially supported by the National Science Foundation under contract DMS-9625396.
References


Table 1. Contributions to the Bayes factors from Laplace's approximation associated with the various orbits of local maxima designated in the first six columns. Stars indicate that entries are smaller than $10^{-12}$ in magnitude and $\emptyset$ indicates the entry is not applicable.

<table>
<thead>
<tr>
<th>Local maxima vec ($\hat{\alpha}<em>{MF}$) = ($\hat{\alpha}</em>{11}^1$, $\hat{\alpha}<em>{12}^2$, $\hat{\alpha}</em>{13}^3$)</th>
<th>Laplace Approx.</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\hat{\alpha}_{10}$</td>
<td>$\hat{\alpha}_{11}$</td>
</tr>
<tr>
<td>Zero nodes ($q = 0$)</td>
<td></td>
</tr>
<tr>
<td>$\emptyset$</td>
<td>$\emptyset$</td>
</tr>
<tr>
<td>One node ($q = 1$)</td>
<td></td>
</tr>
<tr>
<td>$\emptyset$</td>
<td>$\emptyset$</td>
</tr>
<tr>
<td>$\emptyset$</td>
<td>$\emptyset$</td>
</tr>
<tr>
<td>$\emptyset$</td>
<td>$\emptyset$</td>
</tr>
<tr>
<td>Total</td>
<td>$1.6666 \times 10^{20}$</td>
</tr>
<tr>
<td>Two nodes ($q = 2$)</td>
<td></td>
</tr>
<tr>
<td>$\emptyset$</td>
<td>$\emptyset$</td>
</tr>
<tr>
<td>$\emptyset$</td>
<td>$\emptyset$</td>
</tr>
<tr>
<td>$\emptyset$</td>
<td>$\emptyset$</td>
</tr>
<tr>
<td>$\emptyset$</td>
<td>$\emptyset$</td>
</tr>
<tr>
<td>Total</td>
<td>$1.5408 \times 10^{21}$</td>
</tr>
<tr>
<td>Three nodes ($q = 3$)</td>
<td></td>
</tr>
<tr>
<td>$\emptyset$</td>
<td>$\emptyset$</td>
</tr>
<tr>
<td>$\emptyset$</td>
<td>$\emptyset$</td>
</tr>
<tr>
<td>$\emptyset$</td>
<td>$\emptyset$</td>
</tr>
<tr>
<td>$\emptyset$</td>
<td>$\emptyset$</td>
</tr>
<tr>
<td>Total</td>
<td>$2.0785 \times 10^{22}$</td>
</tr>
</tbody>
</table>
Table 2. Contributions to the Bayes factors from Laplace’s approximation associated with the various orbits of local maxima designated in the first six columns.

<table>
<thead>
<tr>
<th>Local maxima $\vec{\alpha}_{MI} = (\hat{\alpha}_1^0, \hat{\alpha}_2^0, \hat{\alpha}_3^0)$</th>
<th>Laplace Approx.</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\hat{\alpha}_{10}$</td>
<td>$\hat{\alpha}_{11}$</td>
</tr>
<tr>
<td>Zero nodes ($q = 0$)</td>
<td>(0)</td>
</tr>
<tr>
<td>One node ($q = 1$)</td>
<td>(0)</td>
</tr>
<tr>
<td></td>
<td>(0)</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>Two nodes ($q = 2$)</td>
<td>(0)</td>
</tr>
<tr>
<td></td>
<td>(0)</td>
</tr>
<tr>
<td></td>
<td>(0)</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>Three nodes ($q = 3$)</td>
<td>*</td>
</tr>
<tr>
<td></td>
<td>*</td>
</tr>
<tr>
<td></td>
<td>*</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Fig. 1. Plot of $\sigma^*(ax)$ versus $(a,x)$. 

Fig. 2. Baye's estimates of the regression function: exact (solid) fitting a single node, its Laplace approximation (dashed), and a Laplace approximation to the estimate from mixing over models $M_0 - M_3$ (dot-dashed).
Fig. 3. True predictive density of $Y_f$ for $x_f = 187$ (solid), assuming a single node, and its normalized Laplace approximation (dashed).

Fig. 4. Joint marginal posterior of the nonlinear parameters $\alpha_{10}$ (intercept) and $\alpha_{11}$ (slope) for a one node model.
Fig. 5. The true marginal posterior on slope parameter \( \alpha_{11} \) from Figure 4 (solid) and its Laplace’s approximation (dashed).

Fig. 6. Personal income data from Jan. 1959 to July 1990 (inclusive).
Fig. 7. The stationary time series.

Fig. 8. Baye's estimates of the regression function: exact (solid) fitting a single node, its Laplace approximation (dashed), and a Laplace approximation to the estimate from mixing over models $M_0 - M_3$ (dot-dashed).
Fig. 9. True predictive density of $Y_{377}$ for $y_{377} = 5.779 \times 10^{-3}$ (solid), assuming a single node, and its renormalized Laplace approximation (dashed).

Fig. 10. Joint marginal posterior of the nonlinear parameters $\alpha_{10}$ (intercept) and $\alpha_{11}$ (slope) for a one node model.
Fig. 11. The true marginal posterior on slope parameter $\alpha_{11}$ from Figure 10 (solid) and its Laplace's approximation (dashed).