Default Priors for Neural Network Classification

Herbert K. H. Lee
University of California, Santa Cruz

August 4, 2005

Abstract

Feedforward neural networks are a popular tool for classification, offering a method for fully flexible modeling. This paper looks at the underlying probability model, so as to understand statistically what is going on in order to facilitate an intelligent choice of prior for a fully Bayesian analysis. The parameters turn out to be difficult or impossible to interpret, and yet a coherent prior requires a quantification of this inherent uncertainty. Several approaches are discussed, including flat priors, Jeffreys priors and reference priors.

Key Words: Bayesian neural network; nonparametric classification; noninformative prior

1 Introduction

Neural networks offer a flexible model for nonparametric classification. Their popularity has spread as they have been found to work well in practice. Operating within the Bayesian paradigm also allows statements about predictive uncertainty. Titterington (2004) provides a recent review of the Bayesian approach for neural networks. As demonstrated by the references in that paper, there is a general tendency to treat the procedure as a "black box", with little or no thought going into the actual probability model and its parameters. This treatment can lead to problems in the Bayesian approach, where one must choose a prior for the parameters. Without careful thought about the choice of prior, one can inadvertently negatively impact the posterior, which may also decrease the quality of predictions from the

model. Priors that have been proposed in the literature include hierarchical priors that use a conjugate style structure for computational convenience (Neal, 1996; Müller and Rios Insua, 1998), priors for parsimony based on deviations from orthogonality or additivity (Robinson, 2001a; Robinson, 2001b), and an empirical Bayes approach (MacKay, 1992).

This paper will begin with a review of the probability model underlying a neural network, discussing issues in the difficulty of interpreting the parameters. Next will be a presentation of several default priors, a more appropriate approach than choosing an arbitrary prior without full consideration of its impact on the posterior. In particular, we discuss flat priors, Jeffreys priors, and reference priors. The idea of default priors has a long history in Bayesian statistics (see, for example, the review paper by Kass and Wasserman, 1996). Finally some examples are given.

2 Neural Networks

A neural network, despite frequent misconceptions, is a probability model for the data, like other statistical models. It falls into the general class of statistical methods for nonparametric regression and classification, in the sense of not assuming a particular parametric form for the relationship between the explanatory and response variables (either a regression response or the probabilities for a multinomial likelihood), but letting the functional form be virtually arbitrary, such as any continuous function. Thus neural networks are closely related to methods such as CART (classification and regression trees), wavelets, splines, and mixture models. In particular, neural networks are a member of the family of methods that use an infinite basis representation to span the space of continuous functions. Analogous to using an infinite series of polynomials or using a Fourier series, a neural network uses location-scale

logistic functions to approximate any continuous function arbitrarily closely. In practice, a finite number of bases are used to get a close enough approximation.

To be specific, first the model is defined for regression, and then for classification. In the regression case, denote the explanatory variables by $\mathbf{x}$ (including a column for the intercept) and the response by $\mathbf{y}$. The particular model for a (single hidden layer feed-forward) neural network for univariate regression is:

$$y_i = \beta_0 + \sum_{j=1}^{k} \beta_j \Psi(\boldsymbol{\gamma}_j^t \mathbf{x}_i) + \varepsilon_i \,, \tag{1}$$

where Ψ is the logistic function

$$\Psi(z) = \frac{1}{1+\exp(-z)} \,,$$

k is the number of logistic basis functions, the γ's are location and scale parameters defining the basis functions, and the β's are the coefficients determining the linear combination of the bases. The error terms are *iid* Gaussian: $\varepsilon_i \overset{iid}{\sim} N(0, \sigma^2)$. It has been shown that location-scale logistic functions do span the space of continuous functions, square-integrable functions, and other cases of interest (Cybenko, 1989; Funahashi, 1989; Hornik et al., 1989). From Equation (1), it is easy to see that a neural network is simply a basis expansion model. It is also a special case of projection pursuit regression (Friedman and Stuetzle, 1981).

To expand this formulation for a multivariate response $\mathbf{y}$, let y_{ig} be the gth component of the ith case, $g \in \{1, \ldots, q\}$, $i \in \{1, \ldots, n\}$. Each dimension g is now fit with a different linear combination of the same logistic basis functions:

$$\begin{aligned} y_{ig} &= \beta_{0g} + \sum_{j=1}^{k} \beta_{jg} \Psi(\boldsymbol{\gamma}_j^t \mathbf{x}_i) + \varepsilon_{ig} \\ \varepsilon_{ig} &\overset{iid}{\sim} N(0, \sigma^2) \,. \end{aligned}$$

This model can be adapted for classification by converting to a multinomial likelihood. The probabilities of class membership are now given by a transformation of the neural network outputs. For each class observation y_i, define a vector of indicator variables as to whether the ith observation is in the gth class, i.e., $y_{ig} = 1$ if and only if y_i is a member of the gth category. Let n be the total number of observations and q be the number of possible classes. Then

$$f\left(\mathbf{y}|\mathbf{p}\right) = \prod_{i=1}^{n}\prod_{g=1}^{q} p_{ig}^{y_{ig}} \tag{2}$$

where the class membership probabilities are

$$p_{ig} = \frac{\exp\left(w_{ig}\right)}{\sum_{h=1}^{q}\exp\left(w_{ih}\right)}\ , \tag{3}$$

and the w's are the neural network outputs:

$$w_{ig} = \beta_{0g} + \sum_{j=1}^{k}\beta_{jg}\Psi_j(\boldsymbol{\gamma}_j^t\mathbf{x}_i)\ .$$

For identifiability, β_{0q} is defined to be zero. In computer science, the transformation of Equation (3) is called the *softmax* model (Bridle, 1989). In statistics, this transformation appears in areas such as generalized linear regression (e.g., McCullagh and Nelder, 1989, p. 159).

2.1 Parameter Difficulties

Consider first a single basis function for a regression neural network,

$$\hat{y}_i = \beta_0 + \beta_1/\left(1 + \exp(-\gamma_0 - \gamma_1 x_i)\right)\ .$$

In this case the parameters are easily interpretable. The γ's control the location and scale along the x-axis, and the β's control the location and scale along the y-axis. The quantity

$-\frac{\gamma_0}{\gamma_1}$ specifies the center of the logistic and γ_1 controls how quickly the logistic rises from its lower value to its upper value. β_1 is the range of $\hat{y}$ and β_0 is the lower bound of $\hat{y}$ (which is the y-intercept if $\beta_1 > 0$ and $-\frac{\gamma_0}{\gamma_1}$ is sufficiently above 0).

Problems in interpretation arise when more than one basis function is considered. Figure 1 shows the fitted values for a two node neural network on the motorcycle accident data of Silverman (1985). This dataset relates the acceleration force on the head of a motorcycle rider in the first moments after an impact, with time after impact as the explanatory variable. The solid line in the plot is the maximum likelihood fit, which demonstrates intriguing behavior when examined closely. In particular, note that there are three inflection points in the fit, even though only two basis functions are used. This occurs because the two basis functions are centered at nearly the same point, and so their active areas interact and the interpretations given above no longer apply. The individual basis functions are shown in Figure 2. Note that the scale of the y-axis changes by two orders of magnitude. This example shows that even in the simplest cases, the parameters can be completely uninterpretable.

Even on the predictive scale, in terms of the observables, the parameters are extremely difficult to interpret. Robinson (2001a, pp. 19–20) demonstrates this by presenting two fitted three-node networks that give very similar predicted curves, despite having quite different parameter values.

Because the parameter values and predictions are not well understood, it is important to realize that the choice of prior can have unpredictable effects on the posterior. Choosing a prior out of convenience or heuristics is not only theoretically incoherent, because the prior is specifying beliefs about the parameters that the user cannot explain, but also potentially harmful to predictive ability because the prior may pull parameters toward a suboptimal

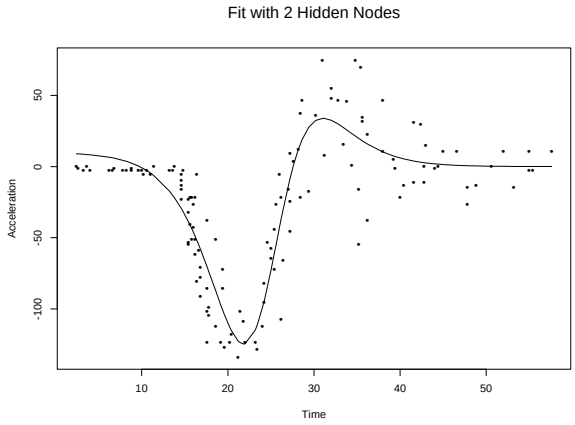

Figure 1: Maximum Likelihood Fit for a Two-Hidden Node Network

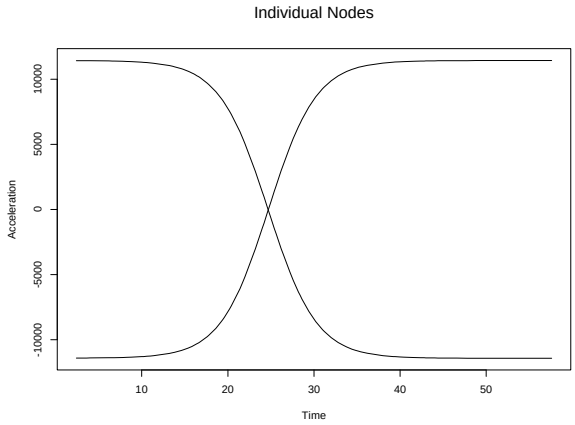

Figure 2: Logistic Basis Functions of the Fit in Figure 1

part of the parameter space.

3 Default Priors

The previous section demonstrated how difficult it can be to interpret the parameters even in basic cases. Yet under the Bayesian approach, one is required to produce a prior distribution for the parameters. Even when one does not have solid prior knowledge to incorporate, one may still take the Bayesian approach as it provides a mechanism for updating beliefs after observing data. An initial state of ignorance can be updated to produce posterior beliefs.

Such a process also provides a formal method for creating uncertainty estimates or intervals, which can be appealing regardless of philosphical arguments.

Rather than imposing a prior purely out of convenience, it makes more sense to choose a prior that in some way represents our ignorance about the parameters. Such a default prior would be derived from a formal statement of lack of information, which can be done in a variety of ways. Jeffreys (1961) was one of the first to develop a formal procedure for finding a default prior. Kass and Wasserman (1996) provide a thorough review of this now extensive literature, with additional arguments in favor of the use of default priors. Many of these priors have appealing invariance properties (Hartigan, 1964). Such priors can lead to confidence intervals with good (frequentist) coverage probabilities (Bayarri and Berger, 2004). This property sometimes appeals to non-Bayesians, who may use a default prior analysis because it can be a more convenient method for producing uncertainty intervals in complex problems. This approach is also helpful for fully accounting for uncertainty in multi-stage models, where a physical process (for example, fish spawning) may need to be fit first, and then the results of that model used in a second physical model (such as fish predation at sea). The Bayesian approach allows uncertainty from the first model to be propogated to the second model in a systematic manner, yet does require a choice of prior for the models. A default prior approach allows such propogation of uncertainty without requiring substantial prior knowledge.

One caveat is that in some cases, including that of neural networks, procedures for creating default priors can produce an improper prior, one with infinite probability mass. This is not a worry if the posterior is proper. For example, in linear regression, a flat prior can be used on the regression coefficients, and Gelman et al. (1995) present some theoretical

advantages of this family of priors. However, for neural networks, improper priors can result in an improper posterior, so one needs to take appropriate measures to ensure a valid posterior, as discussed in the next section. Typically truncation will be sufficient, and this can be done without practical effect in a double-precision computing environment.

3.1 Flat Priors

A simple quantification of ignorance is to claim that all values of the parameter are equally likely. This claim translates to a flat prior:

$$P(\boldsymbol{\gamma}, \boldsymbol{\beta}) \propto 1\,. \tag{4}$$

Since the prior is improper, it is not affected by multiplication by a constant, so the constant 1 is used here to keep things simple. Unfortunately, this impropriety also results in an improper posterior. In order to ensure a proper posterior, it is necessary to truncate the prior to be positive over a finite region. There are two problems that occur with the unrestricted prior. First, it is necessary for the logistic basis functions to be linearly independent (analogous to requiring a full-rank design matrix in linear regression). The second issue is that unlike in most problems, the likelihood does not necessary go to zero in the tails. In certain infinite regions, the limit is a non-zero value. For example, consider the case of a single explanatory variable, and then let $\gamma_0, \gamma_1 \to \infty$ such that $\frac{\gamma_0}{\gamma_1} \to c$ where c is any constant. In this case, the logistic basis function converges to an indicator function, and while this may not be the optimal basis function, the likelihood converges to a non-zero value for a substantial range of coefficients $\boldsymbol{\beta}$. Further details of these issues in the context of regression are in Lee (2003; 2004). It can also be shown that the truncated prior is asymptotically equivalent to the truncated one in both global and local senses (Wasserman, 2000).

In practice, truncation done correctly does not make any noticeable change in the fitted values. The logistic function reaches its limits rather quickly, so that in double precision only a fairly small range is necessary. In particular, for the logistic function $\Psi(z) = 1/(1+\exp(z))$, if the argument z is larger than about 40, $\Psi(z)$ is exactly one in double precision, and if $z < -750$, $\Psi(z) = 0$. So beyond certain values, large γs are redundant, not changing the fitted values at all. Unlike some problems where the choice of truncation point can greatly affect the results, as long as the truncation point is reasonably large, nothing is lost because of the truncation here.

For classification, this flat prior has the potentially appealing property of treating all class predictions equivalently, leading to equal mean prior predictive class probabilities. Thus the statement of prior ignorance also translates to the observables.

3.2 Jeffreys Priors

Flat priors are only one possible approach to specifying ignorance. One major issue with flat priors is that if the model is re-parameterized using a non-linear transformation of the parameters, then the same transformation applied to the prior will not result a flat prior. Jeffreys (1961) introduced a rule for generating a prior that is invariant to differentiable one-to-one transformations of the parameters. The Jeffreys prior is the square root of the determinant of the Fisher information matrix:

$$P_J(\boldsymbol{\theta}) = \sqrt{|I(\boldsymbol{\theta})|} \tag{5}$$

where the Fisher information matrix, $I(\boldsymbol{\theta})$, has elements

$$I_{ij}(\boldsymbol{\theta}) = \mathrm{Cov}_{\boldsymbol{\theta}}\left[\left(\frac{\partial}{\partial\theta_i}\log f(\mathbf{y}|\boldsymbol{\theta})\right)\left(\frac{\partial}{\partial\theta_j}\log f(\mathbf{y}|\boldsymbol{\theta})\right)\right] \tag{6}$$

where $f(\mathbf{y}|\boldsymbol{\theta})$ is the likelihood and the expectation is over $\mathbf{y}$ for fixed $\boldsymbol{\theta}$. The Jeffreys prior is frequently intuitively reasonable and leads to a proper posterior. However, the prior can sometimes fail to produce a proper posterior (e.g., Berger et al. 2001; Jeffreys 1961). Indeed for neural networks, the Jeffreys prior does lead to an improper posterior, so truncation will be necessary as it was with the flat prior.

In some cases, Jeffreys (1961) argued that treating the classes of parameters as independent, and computing the priors independently (treating parameters from other classes as fixed) will produce more reasonable priors. This does seem to be the case for linear regression and neural network regression (Lee, 2004). To distinguish this approach from the joint approach described above, the collective prior (Equation 5) is sometimes called the *Jeffreys-rule prior*. In contrast, the *independence Jeffreys prior* is the product of the Jeffreys-rule priors for each class of parameters independently, while treating the other parameters as fixed. However, for neural network classification, the independence Jeffreys prior is quite similar to the Jeffreys-rule prior because the complex multinomial likelihood prevents any separation of the parameters. The only difference is that the determinant is over a block-diagonal matrix, without any of the $Cov_{\boldsymbol{\theta}}\left(\frac{\partial}{\partial\beta_{ab}}\log f(\mathbf{y}|\boldsymbol{\theta}),\ \frac{\partial}{\partial\gamma_{cd}}\log f(\mathbf{y}|\boldsymbol{\theta})\right)$ terms from the full Fisher information matrix. The quantities in the diagonal blocks are identical. For the rest of this paper, we will focus on the independence Jeffreys prior, as it seems to be generally better behaved in multivariate settings.

In order to proceed, we require the Fisher information matrix. First, define $\Gamma_{ij} = (1 + \exp(-\sum_{h=0}^{r}\gamma_{jh}x_{ih}))^{-1}$ to be the jth basis function evaluated for the i case, with r being the dimension of x and $x_{i0} = 1$ providing an intercept term. Define $\Gamma_{i0} = 1$ for all i. The full

likelihood is

$$f(\mathbf{y}|\boldsymbol{\theta}) = f(\mathbf{y}|\mathbf{x}, \boldsymbol{\gamma}, \boldsymbol{\beta}) = \prod_{i=1}^{n} \prod_{g=1}^{q} {p_{ig}}^{y_{ig}} \ ,$$

where

$$p_{ig} = \frac{\exp\left(\sum_{i=0}^{k} \beta_{jg}\Gamma_{ij}\right)}{\sum_{h=1}^{q} \exp\left(\sum_{i=0}^{k} \beta_{jh}\Gamma_{ij}\right)}$$

for $i = 1, \ldots, n$ and $g = 1, \ldots, q$, as was defined in Equation (3). The loglikelihood is

$$\begin{aligned} \log f(\mathbf{y}|\boldsymbol{\theta}) &= \sum_{i=1}^{n}\sum_{g=1}^{q} y_{ig} \log p_{ig} \\ &= \sum_{i=1}^{n}\left[\sum_{g=1}^{q}\sum_{j=0}^{k} y_{ig}\beta_{jg}\Gamma_{ij} - \log\left(\sum_{g=1}^{q}\exp\left(\sum_{j=0}^{k}\beta_{jg}\Gamma_{ij}\right)\right)\right] \ . \end{aligned}$$

Evaluating Equation (6) gives the individual elements of the information matrix:

$$\begin{aligned} Cov_{\boldsymbol{\theta}}\left(\tfrac{\partial}{\partial\gamma_{ab}}\log f(\mathbf{y}|\boldsymbol{\theta}),\ \tfrac{\partial}{\partial\gamma_{cd}}\log f(\mathbf{y}|\boldsymbol{\theta})\right) &= \\ \sum_{i=1}^{n} x_{ib}x_{id}\Gamma_{ia}(1-\Gamma_{ia})\Gamma_{ic}(1-\Gamma_{ic}) &\left[\sum_{h=1}^{q}\beta_{ah}\beta_{ch}p_{ih} - \sum_{g=1}^{q}\sum_{h=1}^{q}\beta_{ag}\beta_{ch}p_{ig}p_{ih}\right] \\ Cov_{\boldsymbol{\theta}}\left(\tfrac{\partial}{\partial\beta_{ab}}\log f(\mathbf{y}|\boldsymbol{\theta}),\ \tfrac{\partial}{\partial\gamma_{cd}}\log f(\mathbf{y}|\boldsymbol{\theta})\right) &= \sum_{i=1}^{n} x_{id}\Gamma_{ia}\Gamma_{ic}(1-\Gamma_{ic})p_{ib}\left[\beta_{cb} - \sum_{h=1}^{q}\beta_{ch}p_{ih}\right] \\ Cov_{\boldsymbol{\theta}}\left(\tfrac{\partial}{\partial\beta_{ab}}\log f(\mathbf{y}|\boldsymbol{\theta}),\ \tfrac{\partial}{\partial\beta_{cd}}\log f(\mathbf{y}|\boldsymbol{\theta})\right) &= \begin{cases} \sum_{i=1}^{n}\Gamma_{ia}\Gamma_{ic}p_{ib}(1-p_{ib}) & \text{if } b = d \\ \sum_{i=1}^{n}\Gamma_{ia}\Gamma_{ic}p_{ib}p_{id} & \text{if } b \neq d \end{cases} \ . \end{aligned}$$

The Jeffreys-rule prior is now the determinant of the complete Fisher information matrix of all $k(r+1)$ γ's and $(k+1)q$ β's, while the independence Jeffreys prior is the product of the determinant of the $k(r+1)$ by $k(r+1)$ matrix of γ entries of the Fisher information matrix, and the determinant of the $(k+1)q$ by $(k+1)q$ matrix of β entries.

3.3 Reference Priors

An information-theoretic approach is to create a prior that will minimize its effect on the posterior. Bernardo (1979) introduced a class of *reference priors* that are based on maximizing the change in information provided by the data, as measured by a variant of the Shannon information. A key idea is that parameters are separated into groups, with more important parameters listed first, nuisance parameters at the end. The goal is to maximize the effect of the data on the parameters of interest. Note that if all parameters are treated as a single group, this approach reduces to the Jeffreys-rule prior. A more recent discussion of this approach is given in Berger and Bernardo (1992), along with an in-depth description of algorithms for the construction of these priors. Because of the frequent collaboration of those authors on this topic, these priors are sometimes called "Berger-Bernardo priors".

The full derivation of a reference prior is given in the Appendix. The parameter space is partitioned into $\boldsymbol{\theta} = (\boldsymbol{\gamma}, \boldsymbol{\beta})$ and the Fisher information matrix is partitioned into four corresponding parts

$$I(\boldsymbol{\theta}) = \begin{bmatrix} A_{11} & A_{21}^t \\ A_{21} & A_{22} \end{bmatrix} \tag{7}$$

with A_{11} corresponding to the $Cov_{\boldsymbol{\theta}}\left(\frac{\partial}{\partial\gamma_{ab}}\log f(\mathbf{y}|\boldsymbol{\theta}),\ \frac{\partial}{\partial\gamma_{cd}}\log f(\mathbf{y}|\boldsymbol{\theta})\right)$ entries, A_{22} to the $Cov_{\boldsymbol{\theta}}\left(\frac{\partial}{\partial\beta_{ab}}\log f(\mathbf{y}|\boldsymbol{\theta}),\ \frac{\partial}{\partial\beta_{cd}}\log f(\mathbf{y}|\boldsymbol{\theta})\right)$ entries, and A_{21} the cross terms. A resulting reference prior is

$$\pi_R(\boldsymbol{\theta}) \propto \lim_{l\to\infty} |A_{22}|^{1/2} \exp\left\{\frac{1}{2}E^l(\boldsymbol{\gamma})\right\}$$

where

$$E^l(\boldsymbol{\gamma}) = \int_{\{\boldsymbol{\beta}:\boldsymbol{\theta}\in\boldsymbol{\Theta}^l\}} \left(\log|A_{11} - A_{21}^t A_{22}^{-1} A_{21}|\right) \pi_2^l(\boldsymbol{\beta}|\boldsymbol{\gamma}) d\boldsymbol{\beta}$$

with

$$\boldsymbol{\Theta}^l = \left(-\frac{l}{2}, \frac{l}{2}\right)^{k(r+1)+(k+1)q} \qquad \text{and} \qquad \pi_2^l(\boldsymbol{\beta}|\boldsymbol{\gamma}) = \frac{|A_{22}|^{1/2} 1_{\left\{\boldsymbol{\theta}\in\boldsymbol{\Theta}^l\right\}}}{\int_{\{\boldsymbol{\beta}:\boldsymbol{\theta}\in\boldsymbol{\Theta}^l\}} |A_{22}|^{1/2} d\boldsymbol{\beta}} \,.$$

Note that the integral $E^l(\boldsymbol{\gamma})$ is analytically intractable, and thus would be quite difficult to use in practice, since each MCMC iteration would require a numerical integration to evaluate the prior. Switching the order of the parameters does not improve the situation. As the $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ parameters cannot be untangled, as they can in a number of other problems where the reference prior works out nicely, this may be a sign that these two groups of parameters are not clearly distinct, and so perhaps they should not be separated. Leaving them together as a single clump reduces the reference prior to be the same as the Jeffreys-rule prior.

4 Examples

Here we demonstrate on several examples that the proposed methods give reasonable results in both frequentist and Bayesian contexts. We start with a simulated example, where the fitted results can be compared to the known truth. We follow with analyses on two real datasets.

4.1 Simulated Test

A good check on proposed methodology is that it can recover a known truth in a simulated test. Here we consider a binary response whose true underlying probability function is a three-node network with a single input:

$$P(Y=1|X=x) = \frac{6}{1+\exp(12-16x)} - \frac{5}{1+\exp(10-20x)} + \frac{2}{1+\exp(6-20x)} \qquad (8)$$

A dataset was generated by drawing 500 X values uniformly from the unit interval $[0, 1]$, then drawing the respective responses from a Bernoulli distribution with probability given by Equation (8). Three-node networks were fit, finding the maximum likelihood fit, and the posterior means with each of a flat prior, the Jeffreys-rule prior, and the independence Jeffreys prior. The MLE was found using the R code of Venables and Ripley (1999), while the posterior means were found with code programmed in C. This whole process (starting by generating a new binary dataset) was repeated a total of 50 times (in a partially automated process, with some runs being repeated when there were obvious convergence problems). The mean probability fits are shown in Figure 3 as the grey lines, with the truth as the solid black line.

Note that in practice we are interested in recovering the predictive surface, not the particular values of the parameters. Robinson (2001a, pp. 19–20) gives an example of equivalent predictions from two networks with different parameter values. Thus we see that from a predictive standpoint, these methods all perform reasonably.

4.2 Iris Data

The first example on real data is the well-studied iris data from Fisher (1936). In order to be able to create pictures to help with the intuition, we first consider only a single explanatory variable, sepal length. From this we attempt to predict which of three species of iris each of the 150 samples belongs to, with the possible species being Setosa, Versicolor, and Virginica. The 150 samples are comprised of 50 of each type. Neural networks are fit using just two hidden nodes, to keep the pictures simple. The results are summarized in Figure 4. Each row shows the data and fitted probabilities for one of the three species

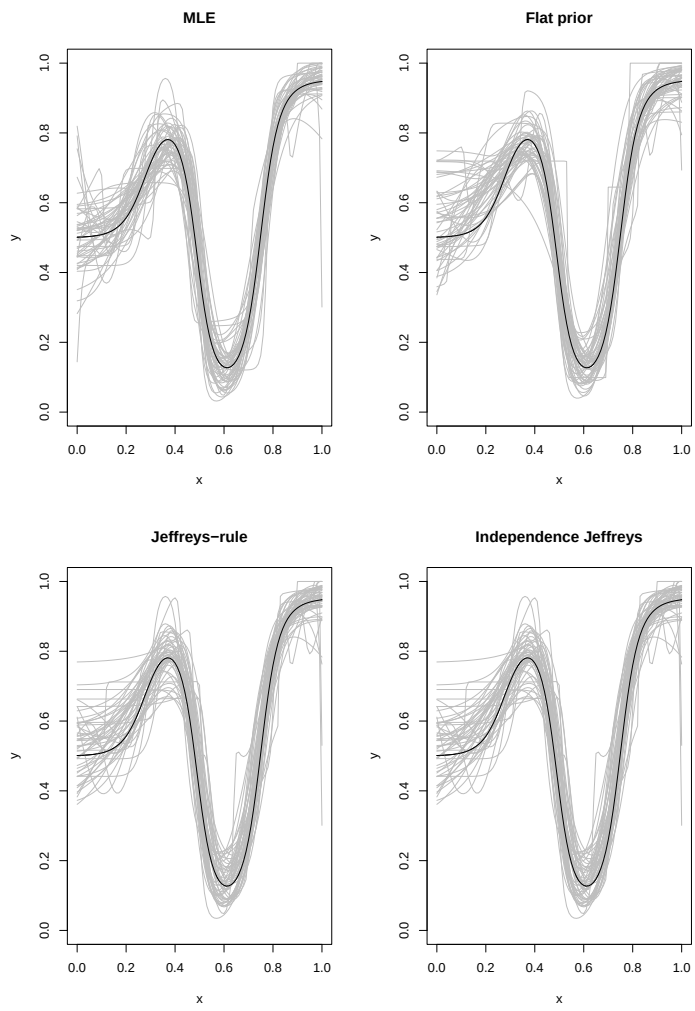

Figure 3: Fitted probabilities of class membership from simulated trials, MLE and posterior means of individual trials shown in grey, truth in black.

of iris. The left column shows the conditional probabilities of the data (for a given sepal length) as a probability histogram, and the probabilities of class membership as estimated by maximum likelihood. The middle column shows the posterior mean (solid line) and 95% pointwise credible intervals (dashed lines) for the fitted probabilities using the flat prior from Section 3.1. The right column shows the corresponding posterior mean and 95% pointwise credible intervals for the Jeffreys-rule prior from Section 3.2. Notice that the MLE and the

posterior mean from the flat prior are very similar, as one would expect them to be. The Jeffreys prior leads to posterior means that are a little less smooth in this case, with the

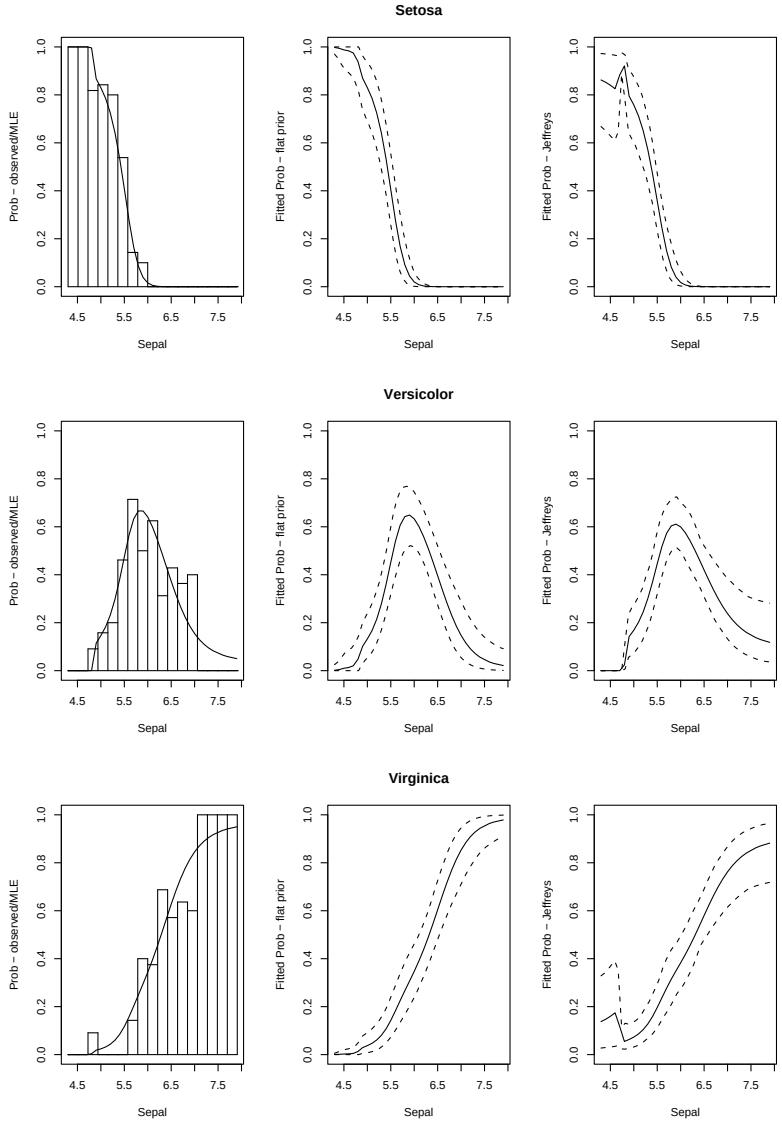

Figure 4: Fitted probabilities for iris species using only sepal length. Species are shown in the rows, the left column shows conditional probabilities of the observed data (histogram) and the MLE fit (solid line), the middle column shows the posterior mean fit using the flat prior and its pointwise 95% credible intervals, and the right column shows posterior mean fits with the Jeffreys prior (solid line) and its pointwise 95% credible intervals

interesting feature that it is attempting to fit some probability to the third class (Virginica) for small sepal lengths because of one observation with sepal length 4.9, whereas the MLE and flat prior models basically ignore this one observation. However the posterior intervals are noticeably wider for low values with the Jeffreys prior, indicating that it is less certain about its fitted probabilities there than in the rest of the space, where it does match the other two models more closely. In terms of selecting a fitted class by choosing the class whose fitted probability is the highest of the three, the three different formulations agree on all observations except for a sepal length of 6.3, which the MLE assigns to Virginica while both Bayesian models assign it to Versicolor. As there are six Virginicas and three Versicolors in the sample with sepal length 6.3, this gives a slight advantage to the MLE in overall misclassification rate. Across the whole sample, the overall misclassification rates are 25% and 27% respectively.

Realistically, one is not usually dealing with just a single explanatory variable. The basic iris dataset contains four (sepal length and width, and petal length and width). Using all four variables and a neural network with two hidden nodes leads to all three approaches (MLE, flat prior, Jeffreys prior) fitting quite well, misclassifying only one observation out of the 150. Note that throughout this example, all of the data were used for training, so the misclassification rates given here are probably lower than they would be for predicting on new data.

4.3 Liver Data

The final example is the liver disorder dataset available from the UCI Machine Learning Repository (`http://www.ics.uci.edu/~mlearn/MLRepository.html`). This dataset con-

sists of 345 observations of males, sorted into two groups based on their sensitivity to liver disorders. There are six explanatory variables: mean corpuscular volume, alkaline phosphotase, alamine aminotransferase, aspartate aminotransferase, gamma-glutamyl transpeptidase, and the number of half-pint equivalents of alcoholic beverages drunk per day. The first five of these variables are blood test measurements, with the first one being a general blood characteristic and the other four being associated primarily with liver function. The sixth covariate is a measure of alcohol consumption, which is important because alcohol is processed in the liver.

The dataset was randomly divided into a training set of 145 observations, which was used to fit the model, and a test set of the remaining 200 observations, which were used to validate the model. Table 1 shows the accuracy on the training set for maximum likelihood estimation (MLE), the posterior predictions with the flat prior, and the posterior predictions with the independence Jeffreys prior. For example, the upper left block shows that fitting a two hidden node network with maximum likelihood resulted in 48 observations being corrected classified as being in group 1, 91 correctly classified into group 2, 19 incorrectly classified into group 1, and 42 incorrectly classified into group 2, for an overall 30.5% misclassification rate.

For all methods, the three hidden node model was optimal. For a fixed network size, the three different methods give rather similar answers. This is to be expected, since the idea of default priors is to try to put a minimal amount of information in the prior, and to let the data have as much influence as possible. Note that the maximum a posteriori predictions of the flat prior are identical to those of the MLE. The differences arise from the averaging over the whole posterior, which more fully accounts for uncertainty.

Net Size	True Group	MLE Fitted Group 1	MLE Fitted Group 2	Flat Prior Fitted Group 1	Flat Prior Fitted Group 2	Jeffreys Prior Fitted Group 1	Jeffreys Prior Fitted Group 2
	1	48	42	56	34	38	52
2	2	19	91	24	86	6	104
		30.5% error		29% error		29% error	
	1	44	46	40	50	43	47
3	2	8	102	5	105	8	102
		27% error		27.5% error		27.5% error	
	1	59	31	49	41	47	43
4	2	26	84	25	85	22	88
		28.5% error		33% error		32.5% error	

Table 1: Results of MLE, flat prior, and Jeffreys prior predictions for 2, 3, and 4 hidden-node models on the liver disorders data

5 Conclusions

When the parameters are difficult or impossible to interpret, one should admit ignorance and attempt to choose a prior consistent with this ignorance. This paper has introduced some examples of the quantification of ignorance for neural networks. These priors do not unduly restrict the posterior to a part of the space with low likelihood values. One can thus obtain good models in practice while still being a coherent Bayesian. Alternatively, one can be a "practical Bayesian", getting approximately the same fits as standard maximum likelihood while also gaining the ability to directly estimate uncertainty.

It is important to note that since little or no information is being specified in the prior, the issue of model selection becomes important. Left to its own devices, a neural network with too many basis functions will tend to overfit the data, such as the four-node network in the liver example, as shown by the worse error rates in Table 1. Thus choosing an appropriate number of basis functions is critical. The problem of model selection (or Bayesian model averaging) has a wide variety of proposed solutions in the literature, and many can easily be combined with the priors of this paper. Some examples of methodology that have been applied specifically to neural networks include Lee (2001), MacKay (1994), and Murata et al. (1994).

Finally, the focus of this paper is on the case when little or no prior information is available. Should the practitioner have some information on the relationship between covariates and class membership, or even marginal information about classes, it is probably better to use a different model where this information can be coherently incorporated into the prior. Neural networks are at their best when flexibility is desired, when interactions may occur in higher dimensions, and when little is known a priori.

Appendix — Reference Prior Derivation

The derivation and notation here follow that in Section 2 of Berger and Bernardo (1992). An ordering of groups of parameters is required, and because $\boldsymbol{\gamma}$ is more difficult to understand, it is placed first, with $\boldsymbol{\beta}$ second. As there is no reason to distinguish among the components of these, only two groups are considered here. Thus the parameter vector is $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) = (\boldsymbol{\gamma}, \boldsymbol{\beta}) \in \boldsymbol{\Theta} = \mathbb{R}^{k(r+1)+(k+1)q}$.

Computations require selecting a nested sequence of compact subsets of the parameter space, with their infinite union being the whole of the space. These sets are chosen here to be $\boldsymbol{\Theta}^l = \left(-\frac{l}{2}, \frac{l}{2}\right)^{k(r+1)+(k+1)q}$. Define $\boldsymbol{\Theta}_2^l(\boldsymbol{\gamma}) = \left\{\boldsymbol{\beta} : (\boldsymbol{\gamma}, \boldsymbol{\beta}) \in \boldsymbol{\Theta}^l\right\}$ and $\boldsymbol{\Theta}_1^l = \left\{\boldsymbol{\gamma} : (\boldsymbol{\gamma}, \boldsymbol{\beta}) \in \boldsymbol{\Theta}^l \text{ for some } \boldsymbol{\beta}\right\}$. Denote the indicator function $1_{\boldsymbol{\Theta}^l}(\boldsymbol{\theta}) = 1$ if $\boldsymbol{\theta} \in \boldsymbol{\Theta}^l$ and 0 otherwise.

Partition the Fisher information matrix into four pieces as in Equation (7). Denote its inverse by S:

$$S = S(\boldsymbol{\theta}) = I(\boldsymbol{\theta})^{-1} = \begin{bmatrix} B_{11} & B_{21}^t \\ B_{21} & B_{22} \end{bmatrix}.$$

We will require the lower right block of the inverses of successive upper left blocks of S, i.e., $h_1 = B_{11}^{-1} = \left[\left(A_{11} - A_{21}^t A_{22}^{-1} A_{21}\right)^{-1}\right]^{-1} = A_{11} - A_{21}^t A_{22}^{-1} A_{21}$ and h_2 is the lower right partition of the inverse of S (which is of course just $I(\boldsymbol{\theta})$), so $h_2 = A_{22}$.

The prior is computed iteratively by group, in the reverse order of $\boldsymbol{\theta}$.

$$\pi_2^l(\boldsymbol{\beta}|\boldsymbol{\gamma}) = \frac{|h_2(\boldsymbol{\theta})|^{1/2} 1_{\{\boldsymbol{\Theta}_2^l(\boldsymbol{\gamma})\}}(\boldsymbol{\beta})}{\int_{\boldsymbol{\Theta}_2^l(\boldsymbol{\gamma})} |h_2(\boldsymbol{\theta})|^{1/2} d\boldsymbol{\beta}} = \frac{|A_{22}|^{1/2} 1_{\{\boldsymbol{\Theta}_2^l(\boldsymbol{\gamma})\}}}{\int_{\{\boldsymbol{\beta}:\boldsymbol{\theta}\in\boldsymbol{\Theta}^l\}} |A_{22}|^{1/2} d\boldsymbol{\beta}}$$

$$\pi_1^l(\boldsymbol{\theta}) = \frac{\pi_2^l(\boldsymbol{\beta}|\boldsymbol{\gamma}) \exp\left\{\frac{1}{2} E^l(\boldsymbol{\gamma})\right\} 1_{\{\boldsymbol{\Theta}_1^l\}}(\boldsymbol{\gamma})}{\int_{\boldsymbol{\Theta}_1^l} \exp\left\{\frac{1}{2} E^l(\boldsymbol{\gamma})\right\} d\boldsymbol{\gamma}}$$

where

$$E^l(\boldsymbol{\gamma}) = \int_{\{\boldsymbol{\beta}:\boldsymbol{\theta}\in\boldsymbol{\Theta}^l\}} (\log|h_1(\boldsymbol{\theta})|)\,\pi_2^l(\boldsymbol{\beta}|\boldsymbol{\gamma})d\boldsymbol{\beta} = \int_{\{\boldsymbol{\beta}:\boldsymbol{\theta}\in\boldsymbol{\Theta}^l\}} \left(\log|A_{11} - A_{21}^t A_{22}^{-1} A_{21}|\right)\pi_2^l(\boldsymbol{\beta}|\boldsymbol{\gamma})d\boldsymbol{\beta}\,.$$

A reference prior π_R can now be found by choosing any fixed point $\boldsymbol{\theta}^* = (\boldsymbol{\gamma}^*, \boldsymbol{\beta}^*)$ with positive density for all π_1^l and evaluating the limit

$$\pi_R(\boldsymbol{\theta}) = \lim_{l\to\infty} \frac{\pi_1^l(\boldsymbol{\theta})}{\pi_1^l(\boldsymbol{\theta}^*)} = \lim_{l\to\infty} \frac{|A_{22}|^{1/2}\exp\{\frac{1}{2}E^l(\boldsymbol{\gamma})\}}{|A_{22}^*|^{1/2}\exp\{\frac{1}{2}E^l(\boldsymbol{\gamma}^*)\}} \propto \lim_{l\to\infty} |A_{22}|^{1/2}\exp\left\{\frac{1}{2}E^l(\boldsymbol{\gamma})\right\}$$

where A_{22}^* is A_{22} evaluated at $\boldsymbol{\theta}^*$.

References

Bayarri, M. J. and Berger, J. O. (2004). "The Interplay of Bayesian and Frequentist Analysis." *Statistical Science*, 19, 58–80.

Berger, J. O. and Bernardo, J. M. (1992). "On the Development of Reference Priors." In *Bayesian Statistics 4*, eds. J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith, 35–60. Oxford University Press.

Berger, J. O., De Oliveira, V., and Sansó, B. (2001). "Objective Bayesian analysis of spatially correlated data." *Journal of the American Statistical Association*, 96, 456, 1361–1374.

Bernardo, J. M. (1979). "Reference Posterior Distributions for Bayesian Inference (with discussion)." *Journal of the Royal Statistical Society Series B*, 41, 113–147.

Bridle, J. S. (1989). "Probabilistic Interpretation of Feedforward Classification Network Outputs, with Relationships to Statistical Pattern Recognition." In *Neuro-computing: Algorithms, Architectures and Applications*, eds. F. F. Soulié and J. Héault, 227–236. New York: Springer-Verlag.

Cybenko, G. (1989). "Approximation by Superpositions of a Sigmoidal Function." *Mathematics of Control, Signals and Systems*, 2, 303–314.

Fisher, R. A. (1936). "The Use of Multiple Measurements in Taxonomic Problems." *Annals of Eugenics*, 7, 179–188.

Friedman, J. H. and Stuetzle, W. (1981). "Projection Pursuit Regression." *Journal of the American Statistical Association*, 76, 817–823.

Funahashi, K. (1989). "On the Approximate Realization of Continuous Mappings by Neural Networks." *Neural Networks*, 2, 3, 183–192.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (1995). *Bayesian Data Analysis*. London: Chapman and Hall.

Hartigan, J. A. (1964). "Invariant Prior Distributions." *Annals of Mathematical Statistics*, 35, 2, 836–845.

Hornik, K., Stinchcombe, M., and White, H. (1989). "Multilayer Feedforward Networks are Universal Approximators." *Neural Networks*, 2, 5, 359–366.

Jeffreys, H. (1961). *Theory of Probability*. 3rd ed. New York: Oxford University Press.

Kass, R. E. and Wasserman, L. (1996). "The Selection of Prior Distributions by Formal Rules." *Journal of the American Statistical Association*, 91, 435, 1343–1370.

Lee, H. K. H. (2001). "Model Selection for Neural Network Classification." *Journal of Classification*, 18, 227–243.

— (2003). "A Noninformative Prior for Neural Networks." *Machine Learning*, 50, 197–212.

— (2004). *Bayesian Nonparametrics via Neural Networks*. ASA-SIAM Series on Statistics and Applied Probability. Philadelphia: Society for Industrial and Applied Mathematics.

MacKay, D. J. C. (1992). "Bayesian Methods for Adaptive Methods." Ph.D. thesis, California Institute of Technology, Program in Computation and Neural Systems.

— (1994). "Bayesian Non-Linear Modeling for the Energy Prediction Competition." *ASHRAE Transactions*, 100, pt. 2, 1053–1062.

McCullagh, P. and Nelder, J. A. (1983). *Generalized Linear Models*. London: Chapman and Hall.

Müller, P. and Rios Insua, D. (1998). "Issues in Bayesian Analysis of Neural Network Models." *Neural Computation*, 10, 571–592.

Murata, N., Yoshizawa, S., and Amari, S. (1994). "Network Information Criterion—Determining the Number of Hidden Units for an Artificial Neural Network Model." *IEEE Transactions on Neural Networks*, 5, 6, 865–871.

Neal, R. M. (1996). *Bayesian Learning for Neural Networks*. New York: Springer.

Robinson, M. (2001a). "Priors for Bayesian Neural Networks." Master's thesis, University of British Columbia, Department of Statistics.

— (2001b). "Priors for Bayesian Neural Networks." In *Computing Science and Statistics*, eds. E. J. Wegman, A. Braverman, A. Goodman, and P. Smyth, vol. 33, 122–127.

Silverman, B. W. (1985). "Some Aspects of the Spline Smoothing Approach to Non-Parametric Curve Fitting." *Journal of the Royal Statistical Society Series B*, 47, 1–52.

Titterington, D. M. (2004). "Bayesian Methods for Neural Networks and Related Methods." *Statistical Science*, 19, 128–139.

Venables, W. N. and Ripley, B. D. (1999). *Modern Applied Statistics with S-PLUS*. 3rd ed. New York: Springer-Verlag.

Wasserman, L. (2000). "Asymptotic Inference for Mixture Models by Using Data-Dependent Priors." *Journal of the Royal Statistical Society Series B*, 62, 159–180.