Neural Networks and Default Priors

Herbert K. H. Lee
Department of Applied Math and Statistics, University of California, Santa Cruz
1156 High St, Santa Cruz, CA 95064, herbie@ams.ucsc.edu

Abstract:

Neural networks are commonly used for classification and regression. The Bayesian approach may be employed, but choosing a prior for the parameters presents challenges. As the parameters are not easily interpretable, it can make sense to try to perform a default Bayesian analysis. Examples discussed here include Jeffreys priors and reference priors.

Key Words: Bayesian Statistics; Noninformative Prior; Jeffreys Prior; Reference Prior

1 Introduction

Neural networks have gained popularity as a method for nonparametric classification and regression, as they often work well in practice. Operating within the Bayesian paradigm also allows statements about predictive uncertainty. Titterington (2004) gives a recent review of the Bayesian approach for neural networks. As seen in the references of that paper, there is a general tendency to treat the procedure as a "black box", with little or no thought going into the underlying probability model and its parameters. This approach can lead to problems in the Bayesian paradigm, 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).

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: $\varepsilon_i \stackrel{iid}{\sim} N(0, \sigma^2)$. 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).

This work was partially supported by National Science Foundation grants DMS-9873275 and DMS-0233710.

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

It is important to note that the parameters are difficult or impossible to interpret in any meaningful manner. Lee (2004, pp. 32–34) provides an example where for fitting a two hidden node network to real one-dimensional data results in the maximum likelihood estimates of the γ parameters that are two orders of magnitude larger than the scale of the original data. Robinson (2001a, pp. 19–20) gives an example on the predictive scale where even in terms of the observables the parameters are extremely difficult to interpret.

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 part of the parameter space.

3 Default Priors

Because it can be difficult to interpret the parameters in even basic cases, 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 could 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. 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).

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, $P(\boldsymbol{\gamma}, \boldsymbol{\beta}) \propto 1$ in the case of classification, or $P(\boldsymbol{\gamma}, \boldsymbol{\beta}, \sigma^2) \propto 1/\sigma^2$ for regression (which is flat with respect to the log of the variance; this sort of prior is well-established in least-squares regression, see for example, Gelman et al. (1995)). As the prior is improper, the choice of constant is unimportant, so 1 is used here for simplicity. Note that this prior is improper, and it results in an improper posterior. To ensure posterior propriety, 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 untruncated 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 41, $\Psi(z)$ 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

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. Denote the parameter vector by $\boldsymbol{\theta}$ (which consists of $\boldsymbol{\gamma}$, $\boldsymbol{\beta}$, and σ^2 in the regression case, and just $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$ for classification). The Jeffreys prior is the square root of the determinant of the Fisher information matrix:

$$P_J(\boldsymbol{\theta}) = \sqrt{|I(\boldsymbol{\theta})|} \qquad (4)$$

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

$$I_{ij}(\boldsymbol{\theta}) = \mathrm{Cov}_{\boldsymbol{\theta}} \left[\left(\frac{\partial}{\partial\theta_i} \log f(\mathbf{y}|\boldsymbol{\theta}) \right) \left(\frac{\partial}{\partial\theta_j} \log f(\mathbf{y}|\boldsymbol{\theta}) \right) \right] \qquad (5)$$

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 4) 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.

In the case of regression, working with the precision $\tau = 1/\sigma^2$, the Jeffreys-rule prior is (Lee, 2004):

$$P_J(\boldsymbol{\theta}) \propto \tau^{((r+2)k-1)/2} \begin{vmatrix} \mathbf{G}^t\mathbf{G} & \mathbf{G}^t\boldsymbol{\Gamma} \\ \boldsymbol{\Gamma}^t\mathbf{G} & \boldsymbol{\Gamma}^t\boldsymbol{\Gamma} \end{vmatrix}^{1/2} ,$$

where $\boldsymbol{\Gamma}$ has elements Γ_{ij} and the n x $(r+1)\,k$ matrix $\mathbf{G}$ has elements $G_{ij} = \beta_g x_{ih} \Gamma_{ig}(1 - \Gamma_{ig})$, where g is the integer part of $\frac{j}{r+1}$ and h is the remainder, i.e., $h = j - (r+1) * g$. The independence Jeffreys prior is:

$$P_{IJ}(\boldsymbol{\theta}) \propto \frac{1}{\tau} \left|\mathbf{F}^t\mathbf{F}\right|^{1/2} ,$$

where $\mathbf{F}$ is just $\mathbf{G}$ without any of the β_g terms, i.e., $F_{ij} = x_{ih}\Gamma_{ig}(1 - \Gamma_{ig})$ where g is the integer part of $\frac{j}{r+1}$ and h is the remainder. As with the flat prior, both of these priors are improper and also lead to improper posteriors, so the parameter space needs to be suitably truncated. Note that the large exponent on the τ term is eliminated in the independence Jeffreys prior, analogously to the linear regression case.

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. These priors do not have compact representations, as they do in the regression case. The full equations and other details are available in Lee (2005).

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 derivations of reference priors are available in Lee (2004) for regression and in Lee (2005) for classification. In both cases, the $\boldsymbol{\gamma}$ parameter is put first, with $\boldsymbol{\beta}$ next, and σ^2 last in the regression case. For regression, a reference prior is

$$P_R \propto \frac{1}{\tau}|\mathbf{F}^t\left(\mathbf{I}-\mathbf{\Gamma}(\mathbf{\Gamma}^t\mathbf{\Gamma})^{-1}\mathbf{\Gamma}^t\right)\mathbf{F}|^{1/2}\,.$$

Again, this prior requires truncation for propriety of the posterior. For classification, an intractable integral is reached, resulting in a less useful prior. Leaving both parameter groups together reduces to the Jeffreys-rule prior.

4 Examples

4.1 Regression

To provide an illustration of the posteriors resulting from the regression priors discussed herein, Figure 1 shows the posterior means from these priors for neural networks with six hidden nodes. The data come from Breiman and Friedman (Breiman and Friedman, 1985) where groundlevel ozone concentration (a pollutant) is being modeled as a function of various meteorological variables. To keep the example simple and easily visualized, the only covariate used here is day of the year. The priors shown are the flat prior, the independence Jeffreys prior, and the reference prior. They all pick up a fair amount of movement in the data. One could make the case that six hidden nodes is too many, and that these posterior means are overfitting the data. But this does demonstrate the flexibility of the model, and that these default priors result in little smoothing, allowing the data to have the largest influence.

4.2 Classification

As an example in classification, we turn to 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 2. Each row shows the data and fitted probabilities for one of the three species of iris. The left column shows the actual data as a probability density histogram, and the probabilities of class membership as estimated by maximum likelihood using the R code of Venables and Ripley (1999). The right column shows the posterior mean fitted probabilities using two of the priors from this paper, a flat prior and a Jeffreys-rule prior. The flat prior is shown with the solid lines and the Jeffreys prior with dashed lines. 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 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. 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

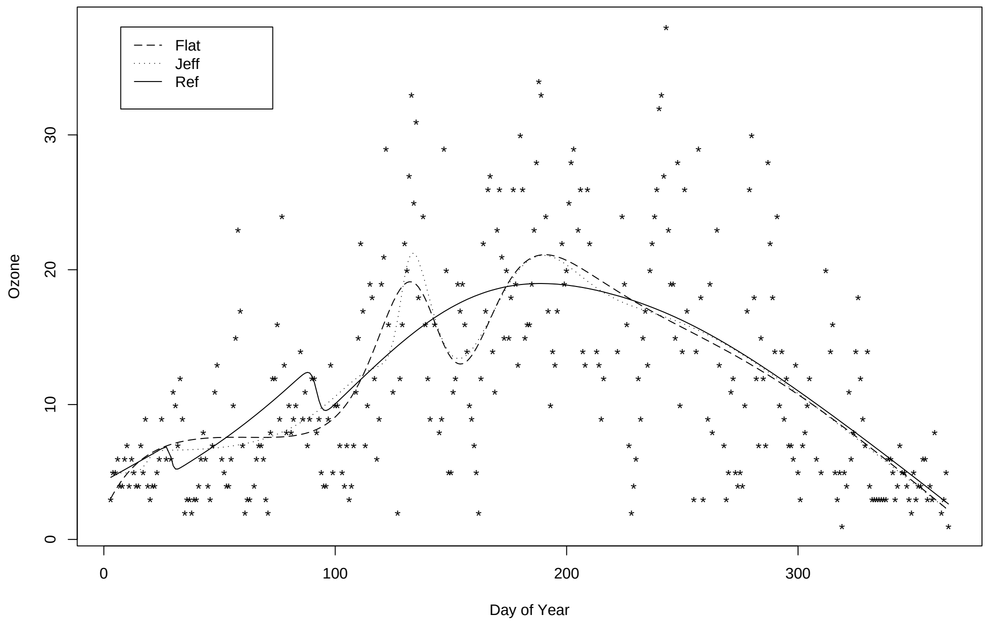

Figure 1: Comparison of posteriors from default priors

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.

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. 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.

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 corre-

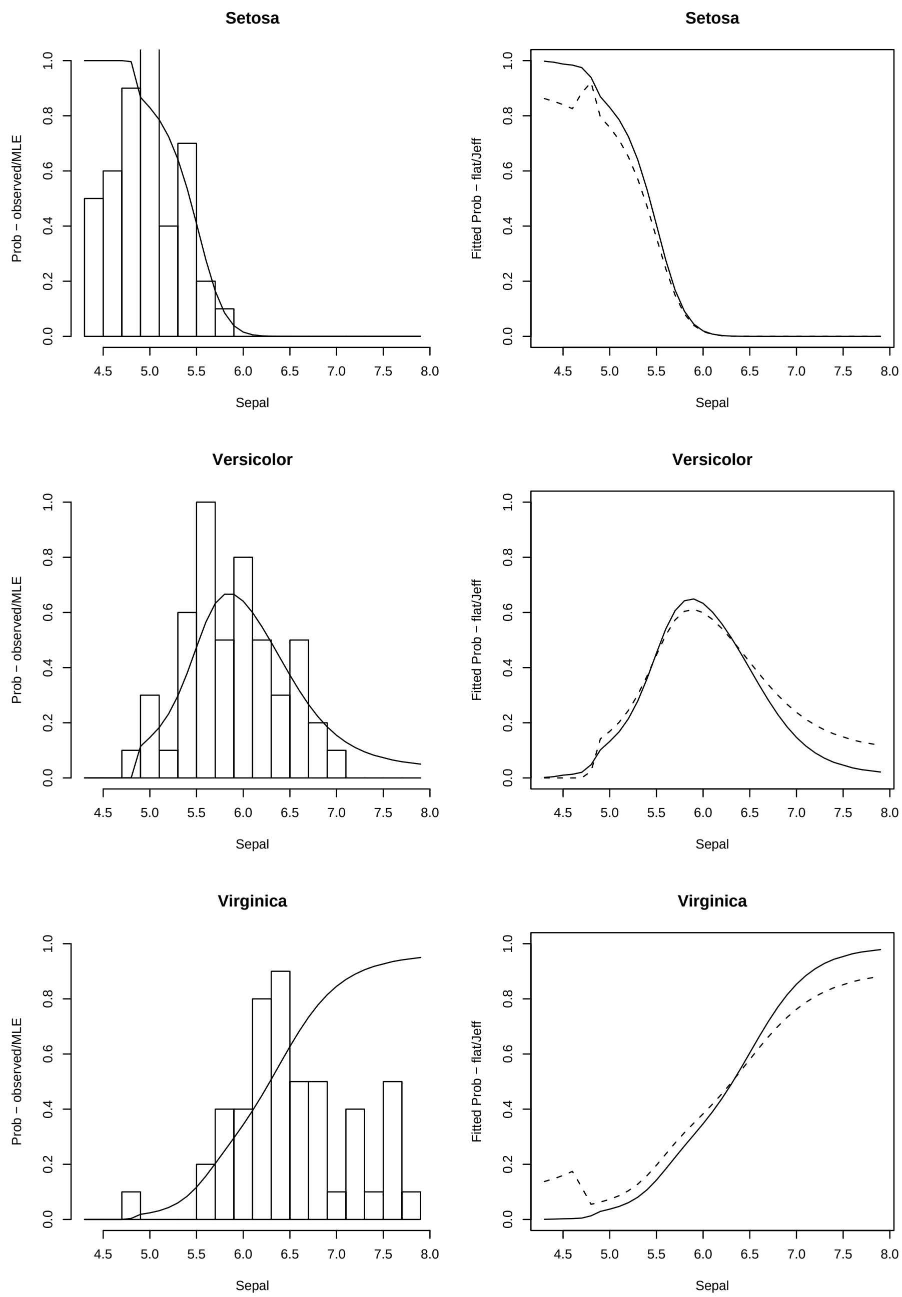

Figure 2: Fitted probabilities for iris species using only sepal length. Species are shown in the rows, the left column shows observed data (histogram) and MLE fit (solid line), and the right column shows posterior mean fits using the flat prior (solid line) and Jeffreys prior (dashed line)

lated 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.

Breiman, L. and Friedman, J. H. (1985). "Estimating Optimal Transformations for Multiple Regression and Correlation." *Journal of the American Statistical Association*, 80, 580–619.

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.

— (2005). "Default Priors for Neural Network Classification." Tech. Rep. 2005–16, University of California, Santa Cruz, Department of Applied Mathematics and Statistics.

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.

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.