Priors for Neural Networks

Herbert K. H. Lee

Department of Applied Mathematics and Statistics

University of California, Santa Cruz

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. This paper reviews several priors in the literature and introduces Jeffreys priors for neural network models. The effect on the posterior is demonstrated through an example.

Key Words: nonparametric classification; nonparametric regression; Bayesian statistics; prior sensitivity

1 Introduction

Neural networks are a popular tool for nonparametric classification and regression. They offer a computationally tractable model that is fully flexible, in the sense of being able to approximate a wide range of functions (such as all continuous functions). Many references on neural networks are available (Bishop, 1995; Fine, 1999; Ripley, 1996). The Bayesian approach is appealing as it allows full accounting for uncertainty in the model and the choice of model (Lee, 2001; Neal, 1996). An important decision in any Bayesian analysis is the choice of prior. The idea is that your prior should reflect your current beliefs (either from previous data or from purely subjective sources) about the parameters before you have observed the data. This task turns out to be rather difficult for a neural network, because in most cases the parameters have no interpretable meaning, merely being coefficients in a basis expansion (a neural network can be viewed as using an infinite set of location-scale logistic functions to span the space of continuous functions). The model for neural network regression is:

$$y_i = \beta_0 + \sum_{j=1}^{k} \beta_j \frac{1}{1 + \exp\left(-\gamma_{j0} - \sum_{h=1}^{r} \gamma_{jh} x_{ih}\right)} + \varepsilon_i \, ,$$

where k is the number of logistic basis functions (hidden nodes), r is the number of explanatory variables (inputs), and ε_i are *iid* Gaussian error. The parameters of this model are k (the number of hidden nodes), β_j for $j \in \{0, \ldots, k\}$ (linear coefficients for the basis functions), and γ_{jh} for $j \in \{1, \ldots, k\}$, $h \in \{0, \ldots, r\}$ (location-scale parameters to create the basis functions). For a classification problem, the fitted values from the neural network are transformed into a multinomial likelihood via the softmax function, i.e., $p_{ig} = \frac{\exp(\hat{y}_{ig})}{\sum_{h=1}^{q} \exp(\hat{y}_{ih})}$, where g indexes the q possible categories, $g \in \{1, \ldots, q\}$.

In general, the parameters of a neural network have no intuitive interpretations, as they are merely coefficients of a basis expansion. Lee (2003) provides a discussion of the few specific cases when the parameters have physical meaning, as well as graphic example of how the parameters become uninterpretable in even the simplest situations. Another example of interpretation difficulties can be found in Robinson (2001a). He provides an example (pp. 19–20) of two fitted three-node neural networks which give very similar fitted curves, yet have completely different parameter values. We again see that there is no clear link between parameter values and their interpretations.

This paper will review a variety of proposals for neural network priors, introduce Jeffreys priors, and provide an example comparing these priors.

2 Proper Priors

A common class of priors in the literature for neural networks are hierarchical proper priors. A *proper* prior is one that is a valid probability distribution, putting probability one on the whole domain of the distribution. The alternative is an *improper* prior as discussed in the next section. Hierarchical priors are useful for neural networks because of the lack of interpretability of the parameters. Adding levels to the hierarchy reduces the influence of the choice made at the top level, so the resulting prior at the bottom level (the original parameters) will be more diffuse, more closely matching the lack of available information about the parameters themselves. This approach can let the data have more influence on the posterior.

Müller and Rios Insua (1998) proposed a three-stage hierarchical model with a relatively simple structure. Prior distributions are chosen to be conditionally conjugate. A tool for visualizing a hierarchical prior is a directed acyclic graph (DAG), where the arrows show the flow of dependency. For this model, the DAG is shown in Figure 1. The priors are Gaussian for β_j and μ_β, multivariate Gaussian for $\boldsymbol{\gamma}_j$ and $\boldsymbol{\mu}_\gamma$, inverse-gamma for σ^2 and σ^2_β, and inverse-Wishart for $\mathbf{S}_\gamma$.

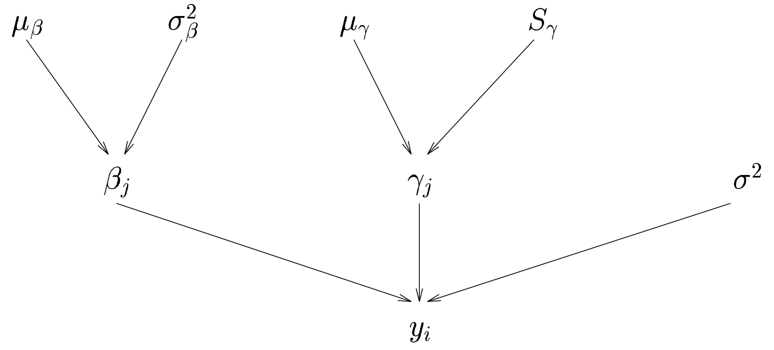

Figure 1: DAG for the Müller and Rios Insua prior

Neal (1996) suggests a more complex model. The DAG diagram of his prior is shown in Figure 2. Each

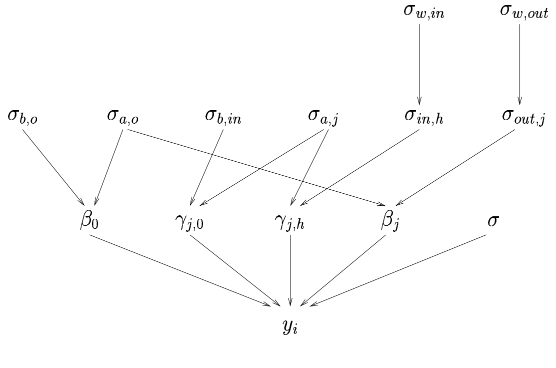

Figure 2: DAG for the Neal prior

of the network parameters (β_j and γ_{jh}) is treated as a Gaussian with mean zero and its own standard deviation which is the product of two hyperparameters, one for the originating node of the link in the graph, and one for the destination node. For example, the weight for the first input to the first hidden node, γ_{11}, has distribution $N\left(0, \sigma_{in,1} * \sigma_{a,1}\right)$, where $\sigma_{in,h}$ is the term for the links from the hth input and $\sigma_{a,j}$ is the term for the links into the jth hidden node; the weight from the first hidden node to the output (i.e. the regression coefficient), β_1, has distribution $N\left(0, \sigma_{out,1} * \sigma_o\right)$, where $\sigma_{out,j}$ is the term for the links from the

jth hidden node, and σ_o is the term for links to the output node. For all of the new σ parameters and for the original σ of the error term, there is an inverse-gamma distribution. There is another set of hyperparameters that must be specified for the inverse-gamma priors on these σ parameters.

3 Noninformative Priors

As was demonstrated in Lee (2003), the parameters of a neural network are typically not interpretable. This makes choosing an informative prior difficult. The previous section introduced various hierarchical priors defined to be proper, yet an attempt was made to let them be diffuse enough that they do not contain too much information, since one would not want to use a highly informative prior that may not come close to one's actual beliefs, or lack thereof. An alternative approach is to use a noninformative prior, one that attempts to quantify ignorance about the parameters in some manner. Early work on noninformative priors was done by Jeffreys (1961). Kass and Wasserman (1996) provide a thorough review of this extensive literature. Many noninformative priors, including those in this section, have appealing invariance properties (Hartigan, 1964).

In many cases, such as for a neural network, procedures for creating a noninformative prior result in a prior that is improper, in the sense that the integral of its density is infinite. This is not an issue as long as 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 a neural network, careless use of improper priors will result in an improper posterior, so measures will need to be taken to prevent this problem.

3.1 Flat Priors

Just as with linear regression, a flat prior can be applied to the parameters of a neural network. As the β parameters of a neural network are analogous to regression coefficients (for fixed values of the γ's, fitting the β's is exactly a linear regression problem), it is reasonable to consider a flat prior for them. A flat prior on the log of the variance is also natural by the same reasoning. The prior for the γ parameters is a more tender question, as they are the ones that are lacking in interpretation. One obvious approach is to also use a flat prior for them. The resulting flat prior for all parameters would be $P(\boldsymbol{\gamma}, \boldsymbol{\beta}, \sigma^2) \propto \frac{1}{\sigma^2}$. Since the prior is improper, the normalizing constant is arbitrary. A major problem with this prior is that it leads to an improper posterior. There are two ways in which things can go wrong: linear independence and tail behavior (Lee, 2003). Some additional notation is necessary. Denote the basis functions evaluated at the data points (i.e., the outputs of the hidden layer) as

$$\Gamma_{ij} = \left[1 + \exp\left(-\gamma_{j0} - \sum_{h=1}^{r} \gamma_{jh} x_{ih}\right)\right]^{-1} \tag{1}$$

with $\Gamma_{i0} = 1$ and let $\mathbf{\Gamma}$ be the matrix with elements (z_{ij}). Thus the fitting of the vector $\boldsymbol{\beta}$ is a least-squares regression on the design matrix $\mathbf{\Gamma}$, $\hat{\mathbf{y}} = \mathbf{\Gamma}^t \boldsymbol{\beta}$.

To understand the linear independence problem, consider the linear regression parallel. When using the standard noninformative prior for linear regression, the posterior will be proper as long as the design matrix is full rank (its columns are linearly independent). For a neural network, we need the k logistic basis functions to be linearly independent, i.e., we need $\mathbf{\Gamma}$ to be full rank. A straightforward way to ensure linear independence is to require that the determinant of $\mathbf{\Gamma}^t\mathbf{\Gamma}$ is positive.

The second possible problem is that unlike in most problems, the likelihood does not necessarily go to zero in the tails, converging to non-zero constants in some infinite regions. If the tails of the prior also do not go to zero, the posterior will not have a finite integral. An obvious way to avoid this problem is to bound the parameter space for $\boldsymbol{\gamma}$. It is worth noting that truncating the parameter space tends to not have much impact on the posterior, as the fitted functions being eliminated are numerically indistinguishable (in double precision) from those in the valid range of the parameter space.

Thus instead of using the flat prior introduced above, a restricted flat prior should be used:

$$P(\boldsymbol{\gamma}, \boldsymbol{\beta}, \sigma^2) \propto \frac{1}{\sigma^2} \mathrm{I}_{\{(\boldsymbol{\gamma}, \boldsymbol{\beta}, \sigma^2) \in \mathbf{\Omega}\}}$$

where $\mathrm{I}_{\{\}}$ is an indicator function, and $\mathbf{\Omega}$ is the parameter space restricted such that $\left|\mathbf{\Gamma}^T\mathbf{\Gamma}\right| > C$ and $|\gamma_{jh}| < D$ for all j, h, where $C > 0$ and $D > 0$ are constants with C small and D large. Lee (2003) shows that this restricted flat prior guarantees a proper posterior, as well as showing that the "interesting" parts of the parameter space are the same under the restricted and unrestricted priors, and thus the posteriors are essentially the same, in the sense that they are both asymptotically globally and locally equivalent.

3.2 Jeffreys Priors

Flat priors are not without drawbacks. In particular, if the problem is re-parameterized using a non-linear transformation of the parameters, then the same transformation applied to the prior will result in something other than 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:

$$P_J(\boldsymbol{\theta}) = \sqrt{|I(\boldsymbol{\theta})|} \tag{2}$$

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

$$I_{ij}(\boldsymbol{\theta}) = \mathrm{Cov}_{\boldsymbol{\theta}}\left[\left(\frac{\partial}{\partial\theta_i}\log f(\mathbf{y}|\boldsymbol{\theta})\right)\left(\frac{\partial}{\partial\theta_j}\log f(\mathbf{y}|\boldsymbol{\theta})\right)\right] \tag{3}$$

where $f(\mathbf{Y}|\boldsymbol{\theta})$ is the likelihood and the expectation is over $\mathbf{Y}$ for fixed $\boldsymbol{\theta}$. Often the Jeffreys prior is intuitively reasonable and leads to a proper posterior. Occasionally the prior may fail to produce a reasonable or even proper posterior (e.g., Berger et al. 2001, Jeffreys 1961), which also turns out to be the case for a neural network.

Jeffreys (1961) argued that it is often better to treat classes of parameters as independent, and compute the priors independently (treating parameters from other classes as fixed). To distinguish this approach from the previous one which treated all parameters collectively, the collective prior (Equation 2) is referred to as the *Jeffreys-rule prior*. In contrast, the *independence Jeffreys prior* (denoted P_{IJ}) 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 a neural network, separate Jeffreys-rule priors would be computed for each of $\boldsymbol{\gamma}$, $\boldsymbol{\beta}$, and σ^2, and the independence Jeffreys prior is the product of these separate priors.

The next step is to compute the Fisher information matrix. We shall consider only univariate regression predictions here, but these results are readily extended to a multivariate regression or classification scenario. For notational and conceptual simplicity, it is easier to work with the precision, $\tau = \frac{1}{\sigma^2}$, the reciprocal of the variance. Thus our parameter vector is $\boldsymbol{\theta} = (\boldsymbol{\gamma}, \boldsymbol{\beta}, \tau)$ and the full likelihood is

$$f(\mathbf{y}|\boldsymbol{\theta}) = f(\mathbf{y}|\mathbf{x}, \boldsymbol{\gamma}, \boldsymbol{\beta}, \tau) = (2\pi)^{-n/2}\tau^{n/2}\exp\left\{-\frac{\tau}{2}(\mathbf{y} - \boldsymbol{\Gamma}^t\boldsymbol{\beta})^t(\mathbf{y} - \boldsymbol{\Gamma}^t\boldsymbol{\beta})\right\} ,$$

where $\boldsymbol{\Gamma}$ is as defined in Equation (1). The loglikelihood, without the normalizing constant, is

$$\log f = \frac{n}{2}\log\tau - \frac{\tau}{2}(\mathbf{y} - \boldsymbol{\Gamma}^t\boldsymbol{\beta})^t(\mathbf{y} - \boldsymbol{\Gamma}^t\boldsymbol{\beta}) .$$

The individual elements of the information matrix are given by Equation (3), and it is straightforward to show that:

$$\begin{aligned}
Cov_{\boldsymbol{\theta}}\left(\frac{\partial}{\partial\beta_j}\log f(\mathbf{y}|\boldsymbol{\theta}),\ \frac{\partial}{\partial\beta_g}\log f(\mathbf{y}|\boldsymbol{\theta})\right) &= \tau\sum_{i=1}^{n}\Gamma_{ij}\Gamma_{ig} \\
Cov_{\boldsymbol{\theta}}\left(\frac{\partial}{\partial\beta_j}\log f(\mathbf{y}|\boldsymbol{\theta}),\ \frac{\partial}{\partial\gamma_{gh}}\log f(\mathbf{y}|\boldsymbol{\theta})\right) &= \tau\beta_g\sum_{i=1}^{n}x_{ih}\Gamma_{ij}\Gamma_{ig}(1-\Gamma_{ig}) \\
Cov_{\boldsymbol{\theta}}\left(\frac{\partial}{\partial\gamma_{jh}}\log f(\mathbf{y}|\boldsymbol{\theta}),\ \frac{\partial}{\partial\gamma_{gl}}\log f(\mathbf{y}|\boldsymbol{\theta})\right) &= \tau\beta_j\beta_g\sum_{i=1}^{n}x_{ih}x_{il}\Gamma_{ij}(1-\Gamma_{ij})\Gamma_{ig}(1-\Gamma_{ig}) \\
Cov_{\boldsymbol{\theta}}\left(\frac{\partial}{\partial\beta_j}\log f(\mathbf{y}|\boldsymbol{\theta}),\ \frac{\partial}{\partial\tau}\log f(\mathbf{y}|\boldsymbol{\theta})\right) &= 0 \\
Cov_{\boldsymbol{\theta}}\left(\frac{\partial}{\partial\gamma_{jh}}\log f(\mathbf{y}|\boldsymbol{\theta}),\ \frac{\partial}{\partial\tau}\log f(\mathbf{y}|\boldsymbol{\theta})\right) &= 0 \\
Var_{\boldsymbol{\theta}}\left(\frac{\partial}{\partial\tau}\log f(\mathbf{y}|\boldsymbol{\theta})\right) &= \frac{n}{2\tau^2}\ .
\end{aligned}$$

To combine these into the matrix $I(\boldsymbol{\theta})$, the exact ordering of parameters within $\boldsymbol{\theta} = (\boldsymbol{\gamma}, \boldsymbol{\beta}, \tau)$ must be specified. The $\boldsymbol{\beta}$ section of $k+1$ elements are $(\beta_0, \ldots, \beta_k)$ and the final element is τ, but $\boldsymbol{\gamma}$ is a matrix. In the presentation here, it appears as row-order, so that $\boldsymbol{\gamma} = (\gamma_{10}, \gamma_{11}, \gamma_{12}, \ldots, \gamma_{1r}, \gamma_{20}, \gamma_{21}, \ldots, \gamma_{2r}, \gamma_{30}, \gamma_{31}, \ldots)$. Now define the n x $(r{+}1)\,k$ matrix $\mathbf{G}$ to have 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., $h = j - (r+1) * g$. With this notation, the full Fisher information matrix is

$$I(\boldsymbol{\theta}) = \begin{bmatrix} \tau\mathbf{G}^t\mathbf{G} & \tau\mathbf{G}^t\boldsymbol{\Gamma} & 0 \\ \tau\boldsymbol{\Gamma}^t\mathbf{G} & \tau\boldsymbol{\Gamma}^t\boldsymbol{\Gamma} & 0 \\ 0 & 0 & \frac{n}{2\tau^2} \end{bmatrix}\ .$$

Thus the Jeffreys-rule prior is

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

The prior is stated as a proportionality because any constants are irrelevant since the prior is improper. The large power on τ seems rather odd, and so Jeffreys would probably recommend the independence prior instead, as this situation is similar to the linear regression setting where analogous problems occur with the prior for the precision. The independence Jeffreys prior is simpler in form, as the Jeffreys-rule prior for $\boldsymbol{\beta}$ with other parameters fixed is a flat prior.

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

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. It is unfortunate that both of these priors are improper, and both lead to improper posteriors. For any particular dataset, it is possible to construct an infinite region of the γ space such that all of the entries of $\mathbf{\Gamma}$ are nonnegative and its columns are linearly independent. Thus $\left|\mathbf{\Gamma}^t\mathbf{\Gamma}\right| > 0$ over this infinite region, so the integral of $\left|\mathbf{\Gamma}^t\mathbf{\Gamma}\right|$ over the whole parameter space will be infinite. This same region of the parameter space also leads to strictly positive $|\mathbf{F}^t\mathbf{F}|$ and $\begin{vmatrix} \mathbf{G}^t\mathbf{G} & \mathbf{G}^t\mathbf{\Gamma} \\ \mathbf{\Gamma}^t\mathbf{G} & \mathbf{\Gamma}^t\mathbf{\Gamma} \end{vmatrix}$. One can also find ranges of $\boldsymbol{\beta}$ so that the likelihood is larger than some positive constant over this same region of the $\boldsymbol{\gamma}$ parameter space. Thus the posterior will also be improper, for both the Jeffreys-rule prior and the independence Jeffreys prior. As with the flat prior, this can be worked around by suitably truncating the parameter space.

4 Hybrid Priors

Some of the priors proposed in the literature combine elements of proper priors and noninformative priors. A basic prior for a neural network would be to combine the noninformative priors for $\boldsymbol{\beta}$ and σ^2 with independent normal priors for each γ_{jh}, i.e., $P(\boldsymbol{\beta}) \propto 1,\ P(\sigma^2) \propto \frac{1}{\sigma^2},\ P(\gamma_{jh}) \sim N(0, \nu)$. This prior gives a proper posterior, and it is notable because it is equivalent to using *weight decay*, a popular (non-Bayesian) method in machine learning for reducing overfitting. The usual specification of weight decay is as penalized maximum likelihood, where there is a penalty of $\frac{||\boldsymbol{\gamma}_j||}{\nu}$ for each of the hidden nodes. Thus instead of maximizing only the likelihood $f(\mathbf{y}|\boldsymbol{\theta})$, one might maximize $f(\mathbf{y}|\boldsymbol{\theta}) - \sum_j \sum_h \frac{\gamma_{jh}^2}{\nu}$, which results in shrinkage of the parameters toward zero. The β_j parameters could also be shrunk if desired. Of course, the choice of ν is important, and various rules of thumb have been developed. Just as with ridge regression, this penalized maximum likelihood approach is equivalent to using a simple prior in a Bayesian context.

Robinson (2001a; 2001b) proposes priors for parsimony on an effective domain of interest. He starts with the basic weight decay prior above, adds one level of hierarchy, putting an inverse-gamma prior with parameters a and b on ν, and then notes that ν can be integrated out leaving the marginal prior distribution for $\boldsymbol{\gamma}_j$ as a multivariate t, i.e., $P(\boldsymbol{\gamma}_j) \propto \left(1 + \frac{1}{b}||\boldsymbol{\gamma}_j||^2\right)^{-(a+r)/2}$. Parsimony is then imposed either through orthogonality or additivity by adding appropriate penalty terms to the prior.

MacKay (1992) takes an empirical Bayes approach. Starting with a simple two-stage hierarchical model, he attempts to use flat and improper priors at the top level. Since this would lead to an improper posterior, he uses the data to fix the values for these hyperparameters (α_m and ν) at their posterior modes. This approach is essentially using a data-dependent prior for a one-stage model, and represents a slightly different approach to get around putting too much information in the prior for parameters that do not have straightforward interpretations.

5 Example Comparing Priors

To demonstrate some of the differences between the priors discussed in this paper, Figure 3 shows the posterior means for several choices of prior. The data are from Breiman and Friedman (1985) and the goal is to model groundlevel ozone concentration (a pollutant) as a function of several meteorological variables. Only day of the year is used here so that the results can be plotted and visually interpreted. A regression example was chosen because it can be easily visualized, but the consequences translate to classification problems as well. Included are the proper priors of Müller and Rios Insua and of Neal, two noninformative priors (flat and independence Jeffreys), and the hybrid weight decay prior. The Neal and weight decay priors have user-specified hyperparameters that greatly affect the behavior of the resulting posteriors, and values were picked here just as examples. The suggested default levels of Müller and Rios Insua were used, and the noninformative priors do not need user specification.

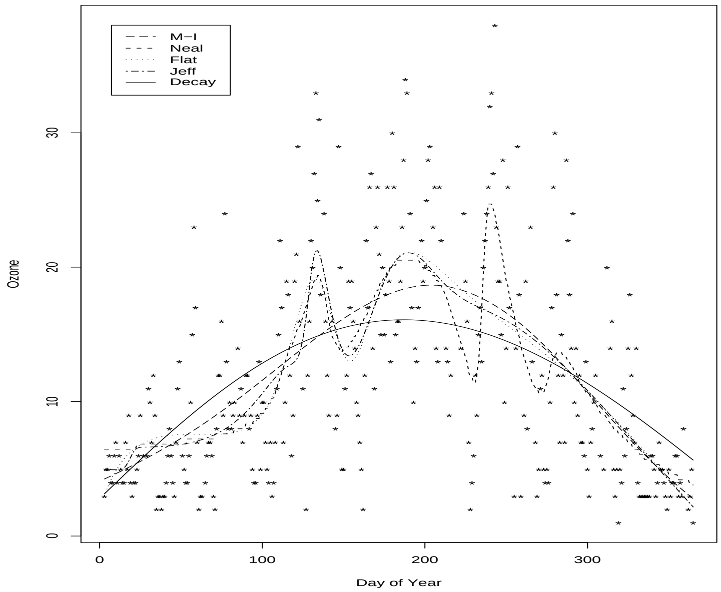

Figure 3: Comparison of priors

Figure 3 shows that the posterior mean fit can have a wide variety of behaviors depending on the choice of prior. This weight decay prior produces the most shrinkage, resulting in a very smooth posterior mean. The Müller and Rios Insua is also highly informative and results in a smooth fit. In contrast, the other priors result in posterior means with more features that try to capture more of the variability in the data.

It is important to note that the weight decay and Neal priors can be adjusted by using different choices of hyperparameters. The examples provided here are meant to show a large part of the range of their flexibility. The weight decay prior has limiting cases with full shrinkage (the posterior mean is just a constant at the mean of the data) and no shrinkage (equivalent to the flat prior). The pictured plot shows a large amount of shrinkage (a more informative prior). The Neal prior similarly has a wide range, with the pictured plot representing very little shrinkage, but it could also be tuned to produce a large amount of shrinkage.

Acknowledgments

This work was supported by National Science Foundation grant DMS 0233710.

References

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.

Bishop, C. M. (1995). *Neural Networks for Pattern Recognition*. Oxford: Clarendon.

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

Fine, T. L. (1999). *Feedforward Neural Network Methodology*. New York: Springer.

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.

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, 197–2152.

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

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

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

Ripley, B. D. (1996). *Pattern Recognition and Neural Networks*. Cambridge: Cambridge University Press.

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.