DigitalCommons @WayneState

Journal of Modern Applied Statistical Methods

Volume 1 | Issue 2 Article 42

11-1-2002

Null Distribution Of The Likelihood Ratio Statistic For Feed-Forward Neural Networks

Douglas Landsittel
University of Pittsburgh, dpl12@pitt.edu

Harshinder Singh
West Virginia University

Vincent C. Arena
University of Pittsburgh

Stewart J. Anderson
University of Pittsburgh

Recommended Citation

Landsittel, Douglas; Singh, Harshinder; Arena, Vincent C.; and Anderson, Stewart J. (2002) "Null Distribution Of The Likelihood Ratio Statistic For Feed-Forward Neural Networks," *Journal of Modern Applied Statistical Methods*: Vol. 1: Iss. 2, Article 42.
DOI: 10.22237/jmasm/1036110120
Available at: http://digitalcommons.wayne.edu/jmasm/vol1/iss2/42

Journal of Modern Applied Statistical Methods
Fall 2002, Vol. 1, No 2, 333-342

1538 – 9472/02/$30.00

Null Distribution Of The Likelihood Ratio Statistic For Feed-Forward Neural Networks

Douglas Landsittel
Dept. of Biostatistics
University of Pittsburgh

Harshinder Singh
Department of Statistics
West Virginia University

Vincent C. Arena
Dept. of Biostatistics
University of Pittsburgh

Stewart J. Anderson
Dept. of Biostatistics
University of Pittsburgh

Despite recent publications exploring model complexity with modern regression methods, their dimensionality is rarely quantified in practice and the distributions of related test statistics are not well characterized. Through a simulation study, we describe the null distribution of the likelihood ratio statistic for several different feed-forward neural network models.

Key words: degrees of freedom, model complexity, chi-square distribution.

Introduction

Neural networks have become a popular regression method for classification and prediction of high-dimensional and/or highly non-linear data (Ripley, 1994). Their appeal in such circumstances is due to their implicitly non-linear model structure, which does not require the user to explicitly define the presence, or degree, of interactions and non-linear terms, and subsequent ability to universally approximate any function (Ripley, 1996). In cases where complex models are needed to fit the underlying associations, but the nature of those associations is not well understood, neural networks are hypothesized to offer a more effective approach to classification. Other consequences of this implicit non-linearity, however, are 1) the propensity of neural networks to over-fit the training data, and 2) the inability to equate the number of model parameters with the effective model dimension.

Other studies have rigorously investigated the issue of model complexity, both specifically for neural networks, and more generally for non-parametric and non-linear regression models. Hastie and Tibshirani (1990), and Loader (1999) calculated degrees of freedom for scatterplot smoothers, local regression, and other nonparametric models using the trace of the hat matrix. For more complex models or model selection procedures, where the hat matrix cannot be explicitly specified, Ye (1998) proposes the generalized degrees of freedom, which estimates the hat matrix diagonal based on the sensitivity of fitted values to changes in observed response values. Hodges and Sargent (2001) extended degrees of freedom to random effects, hierarchical models, and other regression methods (and show a connection to Hastie & Tibshirani, 1990; and Ye, 1998) using a re-parameterization of the trace of the hat matrix.

More specific to neural networks, Moody (1992) and others (Ripley, 1995; Liu, 1995; Amari & Murata, 1993; Murata, Yoshizawa, & Amari, 1991) calculated the effective number of model parameters based on approximating the test set error as a function of the training set error plus model complexity. Other methods (as summarized by Ripley, 1996; and Tetko, Villa, & Livingstone, 1996) include cross-validation, and eliminating variables based on small (absolute) parameter values, or variables with a small effect on predicted values (i.e. sensitivity methods). Bayesian approaches have also been proposed (Ripley, 1995; Ripley, 1996; Paige & Butler,

Douglas Landsittel (landsittel@upci.pitt.edu) is Research Assistant Professor, Biostatistics Dept., University of Pittsburgh, and Statistician, Pittsburgh Cancer Institute. Harshinder Singh (his6@cdc.gov) is Research Professor, Statistics Department, West Virginia University, and Senior Researcher, Biostatistics Branch, NIOSH/HELD. Vincent C. Arena (arena@pitt.edu) and Stewart Anderson (andersons@nsapb.pitt.edu) are Associate Professor, Biostatistics Department, University of Pittsburgh.

2001) for model selection with neural networks. Implementation of such methods, however, has been limited by either computational issues, dependence on the specified test set, or lack of distributional theory.

To our knowledge, no previous studies have directly investigated the distribution of the likelihood ratio statistic with neural networks. In this study, simulations are conducted to empirically describe the distribution of the likelihood ratio statistic under the null assumption of the intercept model (versus the alternative of at least one non-zero covariate parameter). All simulations are conducted with a single binary response; in contrast, the previously cited literature primarily focuses on continuous outcomes. In cases where the likelihood ratio can be adequately approximated by a chi-square distribution, the degrees of freedom can be used to quantify neural network model complexity under the null. Derivation of the test statistic null distribution is pursued through simulation approaches, rather than theoretical derivations, because of the complexity of the network response function and the lack of maximum likelihood or other globally optimal estimation.

The two main objectives of this simulation study are to 1) verify that the chi-square distribution provides an adequate approximation to the empirical test statistic distribution in a limited number of simulated cases, both for the test of independence and tests of nested models, and 2) quantify how the distribution and number of covariates, and the number of hidden units affects model degrees of freedom. Adequacy of the chi-square approximation will be judged by how close the α-level based on the simulation distribution (i.e. the percent of the test statistic distribution greater than the corresponding chi-square quantile) is to various percentiles of the chi-square distribution. The variance, which should be approximately twice the mean under a chi-square distribution, is also displayed for each simulation condition.

Methodology

A Feed-Forward Neural Network Model

This study is restricted to feed-forward models, which are the most common type of neural networks implemented in classification of single dichotomous outcomes. We assume that y follows a Bernoulli distribution; x-values can follow any distribution, but are scaled to the interval [0,1] before fitting the model. Without doing so, the initial weights of the network would have to account for differences in magnitude, as would the process of weight decay (described later).

The predicted value, $\hat{y}$, for the k^{th} observation, with covariate values (or inputs) $\boldsymbol{x}_k = (x_{1k}, x_{2k}, ..., x_{pk})$, is given by

$$\hat{y}_k = f(v_0 + \sum_{j=1}^{H} v_j f\{w_{jo} + \sum_{i=1}^{p} w_{ji} x_{ik}\}) \qquad (1)$$

(Ripley, 1996), where $f(x)$ is the logistic function, $1/(1+e^{-x})$. Each logistic function of the weight sum of the data, $f\{w_{jo} + \sum_{i=1}^{p} w_{ji} x_{ik}\}$, is referred to as the j^{th} hidden unit. The predicted response of the neural network is calculated as a linear combination of these hidden unit values; the parameters $v_0, v_1, ..., v_H$ are referred to as the connections between the hidden and output layer. Each set of parameters $w_{j1}, w_{j2}, ..., w_{jp}$ then represents the weights of the p covariate values specific to the j^{th} hidden unit, or the connections between the input and hidden layer. One implication of this non-linear model structure is that none of the parameter values directly corresponds to any specific main effect or interaction.

Model fitting is typically accomplished through the procedure of back-propagation (Rumelhart, et al., 1995), where model parameters are iteratively updated using a gradient descent-based algorithm. We used the nnet function by Ripley in S-Plus (Venables & Ripley, 1997) to fit all neural network models in this study. The error criteria for dichotomous outcomes, namely minimization of

$$E = \sum_{k=1}^{n} [y_k \log \frac{y_k}{\hat{y}_k} + (1-y_k) \log \frac{1-y_k}{1-\hat{y}_k}], \qquad (2)$$

with respect to the parameters of interest is equivalent to finding global maxima of the corresponding likelihood function.

This study also incorporated weight decay, which is almost universally used to improve optimization and generalization. Rather than minimizing E in Equation 2, the fitting algorithm is applied to minimize

$$E + \lambda \sum_{j=1}^{H} \sum_{i=1}^{p} [v_j^2 + w_{ji}^2], \qquad (3)$$

and thus penalize the network for large parameter values. To determine the magnitude of λ for dichotomous outcomes, Ripley (1996) recommended exploration in the range of [0.001,0.1], which is based on Bayesian arguments and the range of the logistic function. For this study, we utilized $\lambda = 0.01$ for most simulations; additional simulations were also conducted with $\lambda = 0.10$.

Likelihood Ratio Test of Independence

The likelihood ratio statistic for testing model independence with neural networks corresponds to the usual expression from logistic regression,

$$D = 2\{\sum_{k=1}^{n} [y_k \log \hat{y}_k + (1-y_k)\log(1-\hat{y}_k)] \\ -[n_1 \log n_1 + n_0 \log n_0 - n \log n]\}, \qquad (4)$$

where $n_1 = \sum_{k=1}^{n} y_k$, $n_0 = n - n_1$, and $\hat{y}_k$ is calculated from Equation 1 (Cox & Snell, 1989). As opposed to the logistic model, however, the $\hat{y}_k$ do not typically represent the maximum likelihood estimates, rather they represent only locally optimal parameter values. A primary aim of this study will therefore be to assess the adequacy of the chi-square distribution for approximating the null distribution of likelihood ratio test (of model independence) with neural networks.

This study will also investigate the null test statistic distribution for differences between nested models. Denoting D_R and D_F as the likelihood ratio statistics for model independence of the reduced and full models, respectively, $D_F - D_R$ gives the usual likelihood ratio test for significance of the covariates in the full but not the reduced model.

A Simulation Study

To investigate the null distribution (i.e. under the intercept model) of the likelihood ratio statistic (Equation 3), we simulated random data with the following characteristics. Covariate values $\{x_{ik}\}$ were simulated with n = 2,000 observations and between two and five covariates. Covariates and a single binary outcome were first randomly generated from a Bernoulli distribution with $\Pr[x_{ik}=1] = 0.5$ and $\Pr[y_k=1] = 0.5$. The first two covariates, $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$, were simulated with 75 percent concordance, i.e. $\Pr[x_{2k}=1| \ x_{1k}=1] = 0.75$ and $\Pr[x_{2k}=0| \ x_{1k}=0] = 0.75$; all other Bernoulli covariates were independently generated. Covariates were then generated from a standard normal distribution with a correlation of 0.50 between $\boldsymbol{x}_{11}$ and $\boldsymbol{x}_{12}$; all other normal covariates were independently generated. All simulations included the two correlated (either Bernoulli or standard normal) variables and 0 to 3 independent covariates. Neural network models with 2, 5, and 10 hidden units were fit to the simulated data. Model fitting incorporated weight decay (λ = 0.01 or 0.10) (as previously-described).

Means and variances of the simulated likelihood ratio statistics, D_s, are displayed for each simulation condition. Each simulated distribution (for a given number of inputs and hidden units) was then associated with the chi-square distribution having degrees of freedom equal to the mean (simulated) likelihood ratio ($\bar{D}$). Simulated α-levels ($\alpha_q^{(S)}$) were then defined as the percentage of simulated values greater than q^{th} percentile of the corresponding chi-square distribution. For instance, the nominal α-level for the simulated distribution is given by

$$\alpha_{0.05}^{(S)} = P[D \geq \chi_{0.05}^2(\bar{D})]. \qquad (5)$$

Simulated α-levels will then be compared to the chi-square percentiles at significance levels of 0.75, 0.50, 0.25, 0.10, and 0.05. Q-Q plots will also be presented to quantify agreement with the appropriate chi-square distribution.

Results

Simulations were first conducted to investigate the null distribution of the likelihood ratio for testing model independence with strictly binary input variables (Table 1, following page). Results indicate reasonable agreement between the simulated α-levels and the corresponding percentiles of the chi-square distribution. The average simulated α-levels, across the 12 conditions, were all within 0.02 of the expected values. Individually, none of the simulated α-levels varied more than 0.04 from the corresponding chi-square percentile. Based on this correspondence between the simulated results and the chi-square distribution, the mean likelihood ratio can be interpreted as model degrees of freedom.

The Q-Q plot of the likelihood ratio statistic (for testing model independence) with 5 binary inputs and 10 hidden units is displayed in Figure 1, which is generally representative of the other Q-Q plots. The diagonal line through $x = y$ represents perfect agreement between the two distributions. The somewhat greater than expected test statistic variance (66.8 as opposed to twice the mean, which is 57.6) is evidenced by larger values of the statistic at the upper end of the distribution; slightly lower test statistic values were observed at the lower end of the distribution. This deviation in the variance, however, led to only slightly liberal α-levels.

The degrees of freedom varied between approximately 3 for 2 binary inputs, to almost 30 for five binary inputs (with 10 hidden units). The number of hidden units seemed to have a greater effect on the resulting degrees of freedom with 5 inputs than with 2-4 inputs. The model with 5 inputs and 10 hidden units had nearly twice the degrees of freedom as the model with 5 inputs and 2 hidden units.

Table 2 (next page) displays simulation results for comparing the reduced model with between 2 and 4 binary covariates to the full model with all 5 binary covariates. The reduced models were specified by removing x_5 to x_3 in reverse order. For instance, a model reduced to 3 covariates, $\{x_1, x_2, x_3\}$, would be compared to the full model with all 5 covariates.

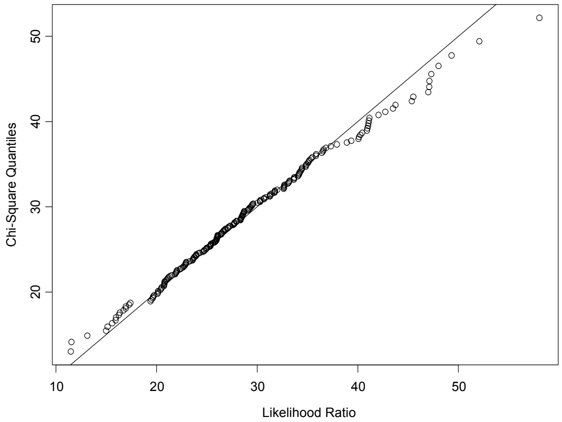

Figure 1. Q-Q Plot of the Likelihood Ratio with 5 Binary Covariates and 10 Hidden Units

The average simulated α-levels, across the 12 conditions, were all within 0.02 of the expected values. With one exception (2 hidden units and 4 inputs in the reduced model), none of the simulated α-levels individually varied more than 0.04 from the corresponding chi-square percentile, and most simulated results were within 0.02 of the chi-square percentile.

The degrees of freedom varied between approximately 5 when adding 1 binary input to the reduced model with 4 inputs (and 2 hidden units), to 26 when adding 3 binary inputs to the reduced model with 2 inputs (and 10 hidden units). The number of hidden units seemed to have a greater effect on the resulting degrees of freedom using the reduced model with 4 inputs. Testing the addition of a single binary input to the reduced model with 4 inputs equated to 15 degrees of freedom with 10 hidden units, as opposed to 5 degrees of freedom with 2 hidden units.

Table 3 (following page) presents simulation results for the case of standard normal covariates. Results again indicated reasonable agreement between the simulated α-levels and the corresponding percentiles of the chi-square distribution. The average simulated α-levels, across the 12 conditions, were all within 0.02 of the expected values. Individually, all of the simulated α-levels were within approximately 0.05 of the corresponding chi-square percentile plot in Figure 1 was also generally representative of the Q-Q plots for testing nested models.

Table 1. Likelihood Ratio Statistic for Model Independence with Binary Inputs

	Hidden	Likelihood Ratio		Simulated α-levels				
Inputs	Units	Mean	Variance	0.75	0.50	0.25	0.10	0.05
2	2	2.8	6.2	0.715	0.535	0.245	0.090	0.055
	5	2.8	6.1	0.720	0.530	0.240	0.085	0.050
	10	2.8	6.1	0.720	0.530	0.240	0.085	0.050
3	2	5.9	13.5	0.700	0.480	0.285	0.120	0.060
	5	6.2	14.3	0.710	0.485	0.270	0.095	0.060
	10	6.3	14.3	0.710	0.480	0.270	0.100	0.060
4	2	10.5	22.6	0.730	0.495	0.265	0.105	0.040
	5	13.7	34.4	0.735	0.490	0.245	0.105	0.070
	10	13.8	34.5	0.740	0.490	0.245	0.105	0.070
5	2	15.6	33.3	0.750	0.520	0.235	0.125	0.080
	5	27.4	61.7	0.755	0.475	0.240	0.115	0.065
	10	28.8	66.8	0.740	0.490	0.265	0.125	0.065
	Mean Simulated α -levels			0.727	0.500	0.254	0.105	0.060

Table 2. Likelihood Ratio Statistic for Nested Models with Binary Inputs

Reduced	Hidden	Likelihood Ratio		Simulated α-levels				
Model	Units	Mean	Variance	0.75	0.50	0.25	0.10	0.05
2 inputs	2	12.8	26.1	0.760	0.515	0.240	0.105	0.065
	5	24.6	51.7	0.755	0.490	0.240	0.105	0.070
	10	26.0	56.4	0.750	0.455	0.265	0.110	0.085
3 inputs	2	9.7	23.2	0.750	0.500	0.285	0.110	0.060
	5	21.2	43.6	0.755	0.475	0.255	0.105	0.070
	10	22.6	47.5	0.745	0.490	0.265	0.100	0.075
4 inputs	2	5.1	18.1	0.695	0.535	0.305	0.145	0.090
	5	13.7	26.3	0.750	0.490	0.240	0.095	0.055
	10	15.1	28.2	0.750	0.495	0.250	0.090	0.050
	Mean Simulated α -levels			0.746	0.494	0.261	0.107	0.069

Table 3. Likelihood Ratio Statistic for Model Independence with Standard Normal Inputs

	Hidden	Likelihood Ratio				Simulated α-levels		
Inputs	Units	Mean	Variance	0.75	0.50	0.25	0.10	0.05
2	2	9.1	19.3	0.750	0.540	0.290	0.105	0.045
	5	21.8	50.8	0.735	0.500	0.270	0.100	0.045
	10	39.4	101.3	0.725	0.540	0.280	0.135	0.040
3	2	13.8	24.2	0.765	0.555	0.250	0.085	0.030
	5	34.9	65.6	0.760	0.505	0.270	0.095	0.040
	10	69.4	133.7	0.755	0.540	0.250	0.075	0.025
4	2	19.1	31.0	0.795	0.520	0.250	0.085	0.040
	5	47.5	84.7	0.775	0.525	0.255	0.075	0.045
	10	100.4	158.1	0.800	0.530	0.220	0.075	0.030
5	2	23.5	49.6	0.765	0.495	0.240	0.110	0.045
	5	61.3	110.9	0.775	0.495	0.225	0.095	0.025
	10	128.5	206.4	0.780	0.520	0.205	0.085	0.025
	Mean Simulated α -levels			0.765	0.522	0.250	0.093	0.036

Table 4. Likelihood Ratio Statistic for Nested Models with Standard Normal Inputs

Reduced	Hidden	Likelihood Ratio				Simulated α-levels		
Model	Units	Mean	Variance	0.75	0.50	0.25	0.10	0.05
2 inputs	2	14.4	54.1	0.705	0.510	0.315	0.150	0.090
	5	39.5	150.3	0.705	0.540	0.320	0.155	0.085
	10	88.1	262.5	0.710	0.505	0.300	0.140	0.100
3 inputs	2	9.7	52.5	0.660	0.510	0.340	0.215	0.135
	5	26.4	135.8	0.685	0.515	0.340	0.210	0.145
	10	58.1	266.0	0.665	0.505	0.355	0.230	0.130
4 inputs	2	4.4	56.6	0.605	0.515	0.400	0.245	0.195
	5	13.8	152.8	0.615	0.535	0.350	0.260	0.205
	10	27.1	260.3	0.630	0.500	0.385	0.270	0.230
	Mean Simulated α -levels			0.664	0.515	0.345	0.208	0.146

Table 5. Likelihood Ratio Statistic for Nested Models with Standard Normal Inputs and Weight Decay of 0.10

Reduced	Hidden	Likelihood Ratio		Simulated α-levels				
Model	Units	Mean	Variance	0.75	0.50	0.25	0.10	0.05
2 inputs	2	10.8	21.6	0.780	0.550	0.235	0.120	0.060
	5	35.2	95.1	0.710	0.495	0.255	0.160	0.090
	10	73.0	158.5	0.725	0.525	0.255	0.125	0.060
3 inputs	2	7.5	20.1	0.745	0.520	0.280	0.105	0.075
	5	24.1	94.5	0.695	0.500	0.315	0.185	0.120
	10	51.9	181.0	0.695	0.520	0.305	0.165	0.090
4 inputs	2	4.1	21.3	0.585	0.450	0.365	0.210	0.130
	5	12.3	72.9	0.655	0.515	0.380	0.210	0.135
	10	25.6	134.6	0.675	0.520	0.350	0.240	0.140
	Mean Simulated α -levels			0.696	0.511	0.304	0.169	0.100

The degrees of freedom varied between approximately 9 for 2 binary inputs (with 2 hidden units), to 128 for five binary inputs (with 10 hidden units). The number of hidden units greatly affected the resulting degrees of freedom for all simulated cases. The model with 5 hidden units corresponded to approximately twice the degrees of freedom as the model with 2 hidden units, and half the degrees of freedom as the model with 10 hidden units.

The Q-Q plot of the likelihood ratio statistic (for testing model independence) with 5 standard normal inputs and 10 hidden units is displayed in Figure 2. It is generally representative of the other Q-Q plots. The somewhat lesser than expected test statistic variance (206.4 as opposed to twice the mean, which is 257.0) is evidenced by smaller values of the statistic at the upper end of the distribution. The nominal α -level were subsequently somewhat conservative.

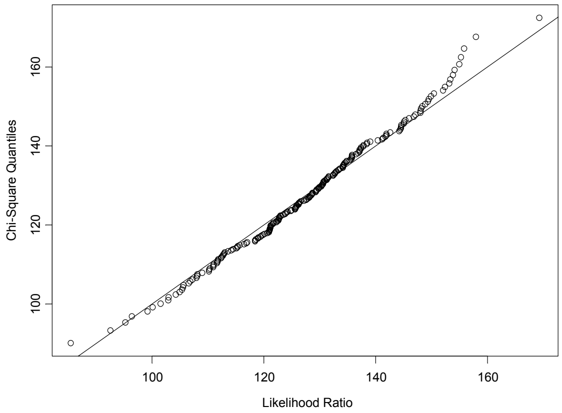

Figure 2. Q-Q Plot of the Likelihood Ratio with 5 Standard Normal Covariates and 10 Hidden Units

Table 4 (previous page) displays simulation results for comparing the reduced model with between 2 and 4 standard normal covariates to the full model with all 5 standard normal covariates. These results, as opposed to previous simulations, do not reflect correspondence to a chi-square distribution. The simulated distributions for testing nested models with continuous covariates are far more skewed; the variance was often 4 or more times greater than the mean (in contrast to the expected 1:2 mean-variance ratio). On average, across the 12 conditions, the difference between simulated α -levels and chi-square percentiles was approximately 10 percent.

To address the substantial discrepancies in Table 4, simulations were rerun using a weight decay of 0.10. Results in Table 5 show a slightly better correspondence to the chi-square distribution under some conditions, but still reflect far greater variability in the test statistic, and subsequently large differences from the chi-square percentiles. The nominal 0.05 α-level, for instance, was between 0.06 and 0.09 for testing the reduced model with 2 standard normal covariates, but was at least 13 percent for testing the reduced model with 4 covariates.

Conclusion

The chi-square distribution appears to provide an adequate approximation to the null distribution (assuming no association between covariates and response) for likelihood ratio tests of independence with feed-forward neural networks. Tests between nested models are approximately chi-square for strictly binary inputs, but not for standard normal covariates. Apart from significance testing, one contribution of these simulations is to quantify the model complexity (under the null) for various neural network models. Although the implicitly non-linear nature of neural networks is commonly known, specifically quantifying the effective number of model parameters remains a difficult task.

These simulations illustrate that even a neural network with only 5 strictly binary inputs (and ten hidden units) can implicitly fit nearly 29 degrees of freedom. Testing the significance of a single binary input, against the reduced model with 4 binary inputs, equates to approximately 15 degrees of freedom. Neural networks with continuous covariates resulted in even greater model complexity; the neural network with 5 standard normal covariates and 10 hidden units equated to approximately 129 degrees of freedom.

The degrees of freedom with strictly binary inputs can be conceptualized as the number of main effects and interaction terms fit by the neural network model; other non-linear functions of a binary term are still 0 or 1, and therefore not relevant. In a related technical report (Landsittel, et al., 2002a), we explored these same models (of strictly binary data) using globally optimal parameter estimates; numerous initial weights were implemented to conduct a grid search of the likelihood surface. In that study, the degrees of freedom was equal to the number of covariate patterns minus one for the intercept (i.e. 2^p-1, where p is the number of parameters) given a sufficient number of hidden units. For simulations where there was an insufficient number of model parameters to fit the saturated model (i.e. the number of parameters was less than 2^p-1), the degrees of freedom was greater than the number of model parameters, but less than the number of covariate patterns. In the current study, based on the usual algorithm which picks only one randomly chosen set of initial parameters, the degrees of freedom was always less than the number of covariate patterns. For instance, 2 binary inputs equates to 2 main effects and 1 interaction term yielding 3 degrees of freedom. The simulated degrees of freedom subsequently equaled 3.0 in the previously-described technical report (based on globally optimal models), and was slightly less, at 2.8, in this current study.

The neural network models with standard normal covariates implicitly fit not only main effects and interactions, but also an indeterminate number of non-linear terms (of an indeterminate nature). This is evidenced by the greater degrees of freedom associated with standard normal covariates (i.e. Table 3 versus Table 1). Consider, for instance, the Taylor series expansion (using the first q terms) of the neural network response function for the k^{th} observation with a single continuous covariate.

$$\begin{aligned}\operatorname{logit}(\hat{y}_k) = v_0 + \sum_{j=1}^{H} v_j f(w_{j0}) + x_k\left(\sum_{j=1}^{H} v_j w_{j1} f'(w_{j0})\right) \\ + \tfrac{1}{2} x_k^2\left(\sum_{j=1}^{H} v_j w_{j1}^2 f''(w_{j0})\right) \\ + \ldots + \tfrac{1}{q!} x_k^q\left(\sum_{j=1}^{H} v_j w_{j1}^q f^{(q)}(w_{j0})\right)\end{aligned} \tag{6}$$

No clear correspondence can be derived between the number of parameters and the number of implicitly fit non-linear terms. This approximation underscores both the implicitly nonlinear structure and the lack of interpretable coefficients. Each expansion term is a function of multiple network parameters and, with the exception of v_0 (the hidden layer intercept term), each network parameter is involved in calculating multiple expansion terms.

The results of this simulation reflect the unpredictable nature of model complexity with neural networks. The degrees of freedom varies both according to the number of input variables and the distribution of these covariates, as well as the number of hidden units. Furthermore, the degrees of freedom will also depend significantly on other issues not investigated here, such as the underlying association (all simulations here were under the null), use of additional training modifications (e.g. model averaging or early stopping of training based on a test set), and further variations in the covariate distributions. This would imply that, from these simulations, we can still only specify the appropriate degrees of freedom in very limited cases.

To address this limitation, we are currently investigating an explicit approach to calculate degrees of freedom with neural networks and dichotomous outcomes. The approach is based on a simple modification to Ye's (1998) procedure for generalized degrees of freedom in the continuous case. The resulting measure for a binary outcome corresponds to Fay's range of influence (ROI) statistic for logistic regression. In a recent commentary (Landsittel, et al., 2002), we empirically show that Fay's ROI statistic asymptotically corresponds to the hat matrix diagonal, and therefore (the sum of these ROI statistics) provides a potential measure of degrees of freedom. Additional simulations will focus on connecting this statistic to the mean likelihood ratio over simulated distributions with neural networks.

In addition to the methods employed here, numerous other training modifications, such as committees of networks, or early stopping of training based a test set, are frequently used and do affect model complexity. Additional simulations (not shown here) indicated that neither network committees nor early stopping lead to correspondence with a chi-square distribution. Greater values of weight decay, or other modifications to model fitting, may lead to a better correspondence with chi-square percentiles in the case of testing nested models with standard normal covariates. In addition to slight improvement of the chi-square approximation, increasing the weight decay tends to reduce the mean likelihood ratio implicitly fit under the null. Further variations on neural network models, such as other covariate distributions, will likely effect the model complexity in an unpredictable manner. These issues can be better explored once an explicit measure is derived for calculating degrees of freedom with a binary outcome.

Although other methods exist for inference and quantifying model complexity with neural networks, these approaches are not widely implemented because of associated computational issues (see Introduction). Use of the likelihood ratio statistic provides a more widely utilized approach, which is easily calculated from the observed and predicted response values (using common statistical programs such as S-Plus). Results of this approach can also be easily interpreted by applied researchers.

References

Amari, S. and Murata, N. (1993). Statistical theory of learning curves under entropic loss criterion. *Neural Computation*, *5*, 140-153.

Cox, D.R. and Snell, E.J. (1989). *Analysis of binary data*. New York, NY: Chapman and Hall, 26-102.

Faraggi, D., & Simon, R. (1995). Maximum likelihood neural network prediction models. *Biometrical Journal*, *37*(6), 713-725.

Fay, M. (2002). Measuring a binary response's range of influence in logistic regression. *The American Statistician*, *56*(1), 5-9.

Hastie, T., & Tibshirani, R. (1990). *Generalized additive models*. New York: Chapman and Hall, 150-152.

Hodges, J.S., & Sargent, D.J. (2001). Counting degrees of freedom in hierarchical and other richly-parameterised models. *Biometrika*, *88*(2), 367-379.

Landsittel, D., Singh, H., Arena V.C. (2002a). Likelihood ratio test of independence for binary data with neural networks. *Technical Report Series – Methods #37*, Department of Biostatistics, University of Pittsburgh.

Landsittel, D., Singh, H., Arena, V.C., & Anderson, S. (2002). Comment on "Measuring a binary response's range of influence in logistic regression." *The American Statistician*.

Liu, Y. (1995). Unbiased estimate of generalization error and model selection in neural networks. *Neural Networks*, *8*(2), 215-219.

Moody, J.E. (1992). The effective number of parameters: an analysis of generalization and regularization in nonlinear learning systems. In J. E. Moody, S. J. Hanson, & R. P. Lippmann (Eds.) *Advances in neural information processing systems 4*. San Mateo, CA: Morgan Kaufmann, 847-854.

Murata, N., Yoshizawa, S., & Amari, S. (1991). A criterion for determining the number of parameters in an artificial neural network model. In T. Kohonen, K Makisara, O. Simula, & J. Kangas (Eds.) *Artificial neural networks*. North Holland: Elsevier Science Publishers, 9-14.

Paige, R. L. and Butler, R. W. (2001). Bayesian inference in neural networks. *Biometrika*, *88*(3), 623-641.

Ripley, B.D. (1996). *Pattern recognition and neural networks*. Cambridge: Cambridge University Press, 143-180.

Rumelhart, D. E., Durbin, R., Golden, R., & Chauvin, Y. (1995). Backpropagation: The basic theory. In Y. Chauvin & D.E. Rumelhart (Eds.) *Backpropagation: Theory, architectures, and applications*. Hillsdale, NJ: Lawrence Erlbaum Associates, 1-34.

Tetko, I.V., Villa, A.E., & Livingstone, D.J. (1996). Neural network studies 2: Variable selection. *Journal of Chemical Informatics and Computer Science*, *36*(4), 794-803.

Venables, W. N., & Ripley, B. D. (1997). *Modern applied statistics with S-Plus*. New York: Springer-Verlag, 337-341.

Ye, J. (1998). On measuring and correcting the effects of data mining and model selection. *Journal of the American Statistical Association*, *93*(441), 120-131.