Journal of Modern Applied Statistical Methods

Volume 8 | Issue 2 Article 13

11-1-2009

Estimating Model Complexity of Feed-Forward Neural Networks

Douglas Landsittel
University of Pittsburgh, dpl12@pitt.edu

DOI: 10.22237/jmasm/1257034320

Journal of Modern Applied Statistical Methods
November 2009, Vol. 8, No. 2, 488-504

1538 – 9472/09/$95.00

Estimating Model Complexity of Feed-Forward Neural Networks

Douglas Landsittel
University of Pittsburgh

In a previous simulation study, the complexity of neural networks for limited cases of binary and normally-distributed variables based the null distribution of the likelihood ratio statistic and the corresponding chi-square distribution was characterized. This study expands on those results and presents a more general formulation for calculating degrees of freedom.

Key words: Degrees of freedom, null distribution, chi-square distribution.

Introduction

Feed-forward neural networks are commonly utilized as a statistical tool for classification and prediction of high-dimensional and/or potentially highly non-linear data. Their popularity stems from an implicitly non-linear and flexible model structure, which does not require explicit specification of interactions or other non-linear terms, and can universally approximate any function (Ripley, 1996). In cases where epidemiologic data or the underlying theory of the specific problem suggest a complex association, but the exact nature of such associations is not well understood, neural networks represent a more flexible methodology for potentially modeling such associations. One significantly negative consequence of this implicit non-linearity and flexible model structure, however, is the resulting inability to quantify model complexity. The typical approach of counting model terms does not provide a rationale basis for quantifying the effective model dimension because the model parameters are inherently correlated to varying degrees.

Douglas Landsittel is an Associate Professor of Medicine and Clinical and Translational Science in the Division of General Internal Medicine. Email: dpl12@pitt.edu.

Previous work has sought to quantify model degrees of freedom for other nonlinear or nonparametric models through use of the hat matrix. For scatterplot smoothers, local regression, and other nonparametric models, Hastie and Tibshirani (1990) and others directly calculated the trace of the hat matrix to estimate degrees of freedom. In cases where the hat matrix cannot be explicitly specified, such as more complex models or model selection procedures, Ye (1998) proposes the generalized degrees of freedom, which estimates the diagonal terms based on the sensitivity of fitted values to changes in observed response values. To address random effects, hierarchical models, and other regression methods, Hodges and Sargent (2001) extended degrees of freedom using a re-parameterization of the trace of the hat matrix and subsequent linear model theory.

Other publications have specifically addressed the issue of model complexity for neural networks. For instance, Moody (1992) 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. A number of other articles (Liu, 1995; Amari & Murata, 1993; Murata, Yoshizawa, & Amari, 1991) have presented theorems to quantify model complexity, but, without a framework for practically applying such methods, none have been utilized in practice. Others have taken a more computational approach (as summarized by Ripley, 1996; and Tetko, Villa, & Livingstone, 1996) using methods such as cross-validation, eliminating variables based on small (absolute)

parameter values, or eliminating variables with a small effect on predicted values (i.e. sensitivity methods). Bayesian approaches have also been proposed (Ripley, 1996; Paige & Butler, 2001) for model selection with neural networks. Despite the noted advances, implementation of such methods has been limited by either computational issues, dependence on the specified test set, or lack of distributional theory. As a result, there are no established procedures for variable selection or determination of the optimal network structure (e.g. the number of hidden units) with neural networks.

Previously, a simulation study was conducted to investigate the distribution of the likelihood ratio statistic with neural networks. In the present 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, and (2) quantify how the distribution, 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

The Feed-Forward Neural Network Model

This study focuses strictly on a single Bernoulli outcome, such as presence or absence of disease. All neural network models utilized a feed-forward structure (Ripley, 1996) with a single hidden layer so that

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

where $\hat{y}$ is the predicted value for the k^{th} observation with covariate values $\boldsymbol{x}_k = (x_{1k}, x_{2k}, ..., x_{pk})$. The function $f(x)$ is the logistic function, $1/(1+e^{-x})$, and each logistic function, $f\{w_{jo} + \sum_{i=1}^{p} w_{ji} x_{ik}\}$, is referred to as the j^{th} hidden unit. The response function of the neural network model can thus be viewed as a logistic of these hidden unit values. In terms of further terminology, the parameters $v_0, v_1, ..., v_H$ are referred to as the connections between the hidden and output layer and each set of other parameters, $w_{j1}, w_{j2}, ..., w_{jp}$, are referred to as the connections between the inputs and hidden units, where there are *p* covariate values specific each of the *p* hidden units. This described model structure often leads to categorization of neural networks as a black box technique. None of the parameter values directly correspond to any specific main effect or interaction. Further, the degree of non-linearity cannot be explicitly determined from the number of hidden units or any easily characterized aspect of the model.

The optimal model coefficients were calculated via back-propagation (Rumelhart, et al., 1995) and the nnet routine in S-Plus (Venables & Ripley, 1997), which iteratively updates weights using a gradient descent-based algorithm. For a Bernoulli outcome, optimization is based on the minimization of the deviance (*D)*,

$$D = -2\sum_{(k=1)}^{n}[y_k log(\hat{y}_k) + (1 - y_k) log(1 - \hat{y}_k)] \quad (2)$$

with a penalty term for the sum of the squared weights (referred to as weight decay). Weight decay, represented by λ in Equation 3, is commonly utilized to improve optimization and generalization of the resulting model by minimizing the penalized likelihood (*PL*)

$$PL = D + \lambda \sum_{(j=1)}^{H}\left[v_j^2 + w_{ji}^2\right] \quad (3)$$

For this study, λ = 0.01 was utilized in all simulations based on previous experience and recommendations by Ripley (1996) established on Bayesian arguments and the range of the logistic function.

Quantifying Model Complexity through Simulations

All simulations utilized a feed-forward neural network with one hidden layer and a single binary outcome with a varying number of predictor variables. All variables were randomly generated (via S-Plus), with the predictor variables being simulated independently of the binary outcome, as to generate results under the null hypothesis of no association. Separate sets of simulations were conducted for a variety of conditions, including binary versus continuous predictors, a varying number of predictor variables, and a varying number of hidden units. For each condition, 500 data sets were each simulated, each with 2,000 observations (to approximate asymptotic results).

To quantify model complexity of the given neural network under the given set of conditions, the likelihood ratio statistic for model independence was calculated, which serves to quantify the model complexity under the null of no association between outcome and predictors. The simulations result in a distribution of likelihood ratios which should follow a chi-square distribution with the mean equal to the degrees of freedom. The mean of that distribution can then be used to quantify model complexity under the null. However, correspondence to a given chi-square distribution must be verified. In the absence of any current theoretical justification for the expected distribution, percentiles of the chi-square distribution were compared to the corresponding α-levels of the simulated distribution (of likelihood ratios). 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

$$\alpha_{0.05}^{s} = P\left(LR \geq \chi_{0.05}^{2}\left(LR\right)\right) \quad (4)$$

where *LR* represents the likelihood ratio. Simulated α-levels are then compared to the chi-square percentiles at significance levels of 0.75, 0.50, 0.25, 0.10, and 0.05. Q-Q plots are also presented to quantify agreement with the appropriate chi-square distribution.

Methods for Estimating Model Degrees of Freedom

After verifying the expected correspondence to a chi-square distribution for a given set of conditions, a new method was utilized to estimate the degrees of freedom for other sets of conditions. Since these methods vary substantially for binary versus continuous predictors, the corresponding methods are first presented separately, after their respective simulation results, and then merged into a single approach. The actual methodology is presented within the results section since these methods are intuitively motivated by the simulation results, and are thus easier to understand within that context.

Results

Simulation Results for Binary Input Variables

Results presented in Table 1 were generated using independently distributed binary inputs. All neural network models were fit using a weight decay of 0.01; for each result pertaining to binary inputs, the maximum number of terms, including all main effects and interactions, for *k* inputs equals $2^k - 1$. The number of model parameters for a model with *h* hidden units equals $h(k + 1) + (h + 1)$.

Table 1: Likelihood Ratio Statistic for All Binary Inputs

Inputs (Max # Terms)	Hidden Units	#Parameters	Likelihood Ratio		Simulated α-levels				
			Mean	Variance	0.75	0.50	0.25	0.10	0.05
2 (3)	2	9	2.86	5.81	0.75	0.52	0.26	0.06	0.03
	5	21	3.04	5.73	0.74	0.50	0.26	0.09	0.04
	10	41	3.15	5.98	0.76	0.51	0.25	0.10	0.04
3 (7)	2	11	7.04	15.90	0.75	0.49	0.22	0.09	0.04
	5	26	6.93	12.37	0.74	0.51	0.25	0.11	0.07
	10	51	7.24	13.97	0.75	0.50	0.26	0.10	0.06
4 (15)	2	13	11.94	22.05	0.74	0.50	0.25	0.11	0.08
	5	31	14.87	28.99	0.76	0.50	0.26	0.09	0.06
	10	61	14.96	31.33	0.76	0.50	0.23	0.08	0.04
5 (31)	2	15	18.36	31.03	0.75	0.50	0.26	0.13	0.08
	5	36	30.25	62.22	0.75	0.48	0.25	0.10	0.05
	10	71	31.82	69.57	0.74	0.50	0.22	0.09	0.06
6 (63)	2	17	25.07	44.05	0.71	0.49	0.28	0.14	0.07
	5	41	50.63	108.5	0.76	0.51	0.23	0.09	0.04
	10	81	63.70	147.5	0.76	0.50	0.24	0.08	0.03
7 (127)	2	19	30.92	57.98	0.74	0.50	0.26	0.10	0.05
	5	46	69.93	138.4	0.75	0.54	0.24	0.10	0.05
	10	91	117.3	245.6	0.75	0.50	0.25	0.10	0.05
8 (255)	2	21	38.75	77.43	0.74	0.51	0.25	0.08	0.04
	5	51	88.95	161.2	0.73	0.49	0.27	0.13	0.06
	10	101	168.3	318.0	0.74	0.50	0.27	0.11	0.05
9 (511)	2	23	45.76	110.9	0.79	0.51	0.20	0.06	0.02
	5	56	107.7	202.9	0.75	0.54	0.25	0.10	0.05
	10	111	214.4	394.9	0.74	0.50	0.24	0.11	0.06
10 (1023)	2	25	51.76	117.9	0.77	0.51	0.22	0.07	0.03
	5	61	126.1	248.5	0.74	0.51	0.24	0.10	0.05
	10	121	257.5	546.5	0.75	0.48	0.25	0.10	0.05
Mean Simulated α-levels					0.75	0.50	0.25	0.10	0.05

Table 1 shows the simulated distribution of the likelihood ratio test for independence is closely followed by a chi-square distribution. In a large percentage of the cases, all of the simulated α-levels were within 1-3% of the expected percentiles. No systematic differences were evident in the results. Figures 1a and 1b illustrate two examples where: (1) the simulated distribution varied a few percent from the expected percentiles (2 inputs and 2 hidden units), and (2) the simulated distribution fell extremely close to the corresponding chi-square distribution (7 inputs and 10 hidden units). Both figures show noticeable variability at the upper end of the distribution; however, it should be noted that these few points are primarily within only the top 1% of the distribution, and thus have little effect on most of the resulting significance levels.

Figure 1a: Example Q-Q plot (with 2 Binary Inputs, 2 HUs and 0.01 WD)
Illustrating Greater Variability from the Expected Chi-square Distribution

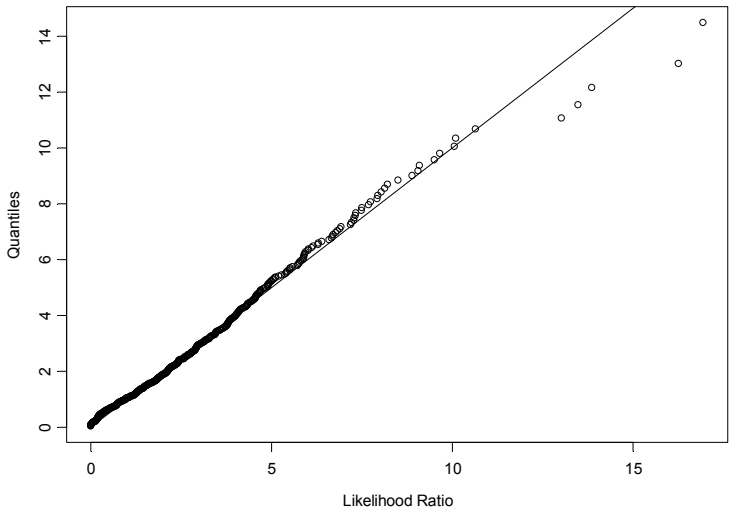


Figure 1b: Example Q-Q plot (with 7 Binary Inputs, 10 HUs and 0.01 WD)
Illustrating a Close Correspondence with the Expected Chi-square Distribution

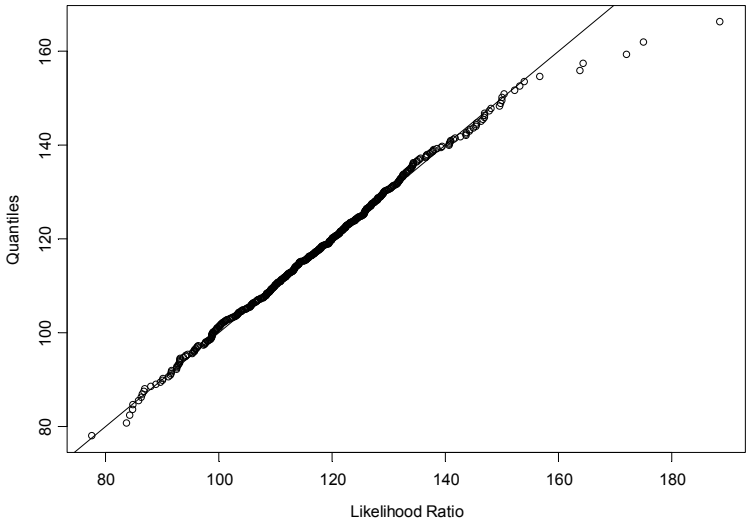


Table 2 shows additional simulation results for other cases (with at most 10 inputs) where the number of model parameters (p) is less than the maximum number of terms (m). Results again showed that the simulated α-levels were very close the expected percentiles. Q-Q plots for these cases (not shown here) resulted in similar findings as displayed in Figures 1a and 1b. No additional simulations are shown here for two, three or four inputs because these cases all corresponded to models where $p > m$, and will therefore produce a known degrees of freedom (3, 7, and 15 for 2, 3 and 4 inputs, respectively). Simulations were conducted, but not shown here, to verify that such results would hold for 4 hidden units (since p was only slightly greater than m in the case of 3 hidden units); results verified the expected finding (of 15 degrees of freedom). Other combinations leading to a known degrees of freedom (with $p > m$) were also excluded from the table (including 5 inputs with 6-9 hidden units and 6 inputs with 8-9 hidden units).

Estimating Degrees of Freedom for Binary Input Variables

The above results indicate that the model degrees of freedom for strictly binary inputs appear to intuitively depend on two factors:

(1) the maximum number of possible main effects and interactions = $(2^k - 1)$, and
(2) the number of model parameters = $h(k + 1) + (h + 1)$.

In cases where the number of model parameters is sufficient to fit all main effects and interactions, the degrees of freedom is equal to that maximum number of terms. For example, regardless of the number of hidden units, the degrees of freedom (*df*) are approximately 3.0 for two binary inputs and approximately 7.0 for three binary inputs. For four binary inputs, two hidden units (and subsequently 13 parameters) are insufficient to fit all 15 terms and result in approximately 12 *df*.

In such cases, where the number of model parameters is less than the maximum number of terms, the *df* is generally in between (or at least very close to) the number of model parameters (p) and the maximum number of terms (m). Exactly where the *df* falls depends on how close the number of model parameters is to the maximum number of terms. In general, the ratio of degrees of freedom by number of model parameters may be expressed as a function of $m - p$. To produce a linear relationship, it is more convenient (with binary inputs) to express df/p as a function of $\log_2(m\text{-}p)$. The simulated degrees of freedom from Table 1 was used to derive a relationship, and Figure 2 shows a plot of the simulated data from Table 1 (with 2, 5 or 10 hidden units) overlaid with the linear regression line

$$df/p = 0.6643 + 0.1429 \times log_2(m - p). \quad (5)$$

Figure 2 shows a general trend between the difference in $m - p$ and the degrees of freedom (divided by the number of parameters), but also illustrates some variability between the simulated values and the subsequent estimates. To evaluate the significance of these discrepancies, the estimated *df* were compared to the simulated distribution of the likelihood ratio statistic (for model independence). Results are shown in Table 3.

Results indicate that the estimated *df* usually approximates the simulated value within an absolute error of a few percent. For example, most of the conditions (11 of 16) result in a 5% significance level between 0.03 and 0.07; the largest discrepancy is an absolute difference of 0.04 from the true 5% level. The 10% significance level corresponds to somewhat larger errors, with the estimated p-values as high as 0.17 and as low as 0.02. The 75th, 50th and 25th percentiles showed similar findings with occasionally substantial discrepancies.

The above rule for estimating the *df*, based on the previously fit linear regression of df/p as a function of $log_2(m - p)$, was also evaluated with respect to its adequacy to predict model complexity for new cases (with 3, 4, or 6-9 hidden units). Figure 3 shows a plot of these additional simulated data overlaid with the linear same regression line $df/p = 0.6643 + 0.1429 \cdot log_2(m - p)$.

Figure 3 shows the trend between the difference in $m - p$ and the *df* (divided by the number of parameters), but again illustrates variability between the simulated values and the

Table 2: Additional Simulated Likelihood Ratio Statistics with All Binary Inputs

Inputs (Max # Terms)	Hidden Units	#Parameters	Likelihood Ratio		Simulated α-levels				
			Mean	Variance	0.75	0.50	0.25	0.10	0.05
5 (31)	3	22	23.84	47.66	0.73	0.50	0.26	0.10	0.06
	4	29	28.69	54.80	0.74	0.50	0.26	0.12	0.06
6 (63)	3	25	34.57	67.28	0.74	0.49	0.25	0.10	0.06
	4	33	42.81	96.51	0.77	0.51	0.24	0.08	0.03
	6	49	55.86	109.8	0.74	0.50	0.26	0.09	0.05
	7	57	59.60	110.0	0.77	0.49	0.26	0.11	0.05
7 (127)	3	28	45.93	92.58	0.75	0.50	0.23	0.10	0.05
	4	37	58.06	111.6	0.75	0.50	0.24	0.10	0.05
	6	55	82.27	148.3	0.75	0.49	0.26	0.11	0.06
	7	64	92.84	175.6	0.74	0.51	0.27	0.10	0.05
	8	73	102.5	189.5	0.75	0.51	0.25	0.09	0.06
	9	82	111.7	224.0	0.76	0.51	0.24	0.10	0.06
8 (255)	3	31	54.90	101.0	0.75	0.50	0.26	0.11	0.07
	4	41	73.02	148.2	0.75	0.52	0.23	0.08	0.04
	6	61	107.8	223.0	0.75	0.49	0.24	0.09	0.05
	7	71	124.8	258.3	0.76	0.50	0.25	0.10	0.03
	8	81	139.7	238.2	0.71	0.52	0.28	0.12	0.06
	9	91	155.0	268.0	0.73	0.52	0.24	0.13	0.08
9 (511)	3	34	65.13	135.0	0.77	0.50	0.24	0.10	0.05
	4	45	87.02	179.6	0.76	0.51	0.25	0.09	0.04
	6	67	131.4	228.8	0.73	0.51	0.27	0.10	0.06
	7	78	152.3	286.6	0.74	0.50	0.25	0.10	0.06
	8	89	171.8	338.5	0.74	0.51	0.26	0.11	0.05
	9	100	194.7	303.6	0.72	0.50	0.27	0.14	0.08
10 (1023)	3	37	75.5	163.5	0.76	0.51	0.25	0.08	0.03
	4	49	100.9	190.8	0.73	0.52	0.26	0.10	0.05
	6	73	152.7	297.1	0.77	0.50	0.24	0.10	0.05
	7	85	178.5	341.9	0.74	0.51	0.24	0.09	0.06
	8	97	204.8	430.0	0.77	0.51	0.24	0.08	0.04
	9	109	230.0	425.7	0.74	0.52	0.25	0.10	0.06
Mean Simulated α-levels					0.75	0.51	0.25	0.10	0.05

Figure 2: Plot of the Degrees of Freedom for Binary Inputs (2, 5, and 10 Hidden Units) as a Function of the Difference between Maximum Number of Terms and Number of Parameters

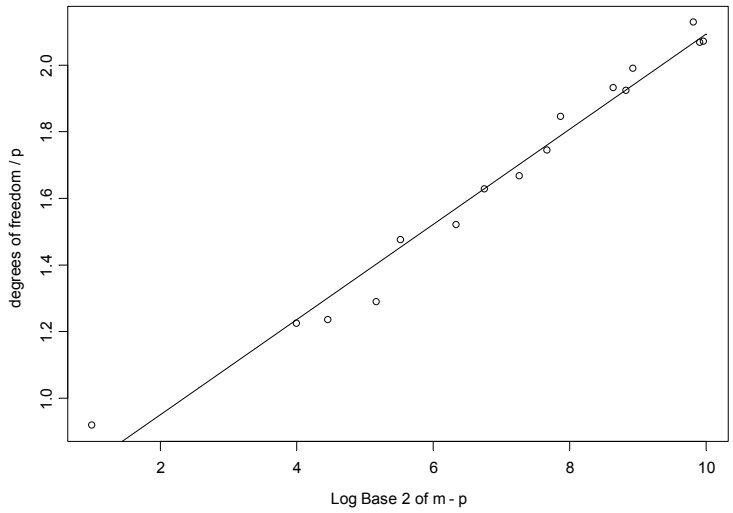
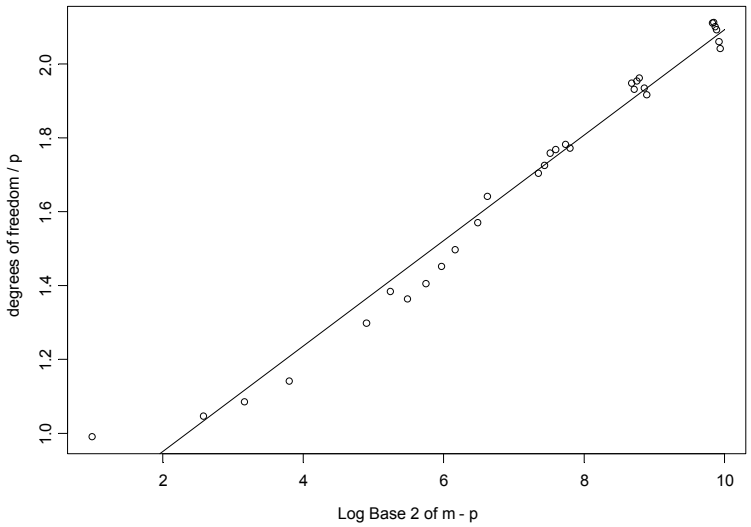


Table 3: Comparison of Estimated to Simulated Degrees of Freedom with All Binary Inputs

Max # of Terms (Inputs)	# of Parameters	Hidden Units	Simulated *df*	Estimated *df*	Simulated α-levels using the Estimated *df*				
					0.75	0.50	0.25	0.10	0.05
15 (4)	13	2	11.94	10.49	0.66	0.41	0.19	0.07	0.05
31 (5)	15	2	18.36	18.54	0.74	0.49	0.25	0.12	0.07
63 (6)	17	2	25.07	24.71	0.73	0.51	0.30	0.15	0.08
	41	5	50.63	53.36	0.67	0.40	0.16	0.05	0.02
127 (7)	19	2	30.92	30.96	0.74	0.50	0.26	0.10	0.05
	46	5	69.93	72.23	0.69	0.46	0.18	0.07	0.03
	91	10	117.3	127.7	0.50	0.25	0.09	0.02	0.01
255 (8)	21	2	38.75	37.57	0.78	0.56	0.30	0.11	0.05
	51	5	88.95	89.80	0.71	0.47	0.25	0.11	0.06
	101	10	168.3	172.0	0.67	0.42	0.21	0.07	0.03
511 (9)	23	2	45.76	44.63	0.82	0.56	0.24	0.08	0.03
	56	5	107.7	107.9	0.75	0.53	0.24	0.10	0.05
	111	10	214.4	210.8	0.80	0.57	0.30	0.15	0.08
1023 (10)	25	2	51.76	52.21	0.76	0.49	0.21	0.07	0.03
	61	5	126.1	126.9	0.72	0.49	0.23	0.09	0.05
	121	10	257.5	250.2	0.84	0.61	0.36	0.17	0.09
Mean Estimated α-levels					0.72	0.57	0.24	0.10	0.05

Figure 3: Plot of the Degrees of Freedom for Binary Inputs (3, 4, and 6-9 Hidden Units) as a Function of the Difference between Maximum Number of Terms and Number of Parameters



subsequent estimates. Further, results do not show any systematic difference from the previous set of findings (as graphed in Figure 2). To evaluate the significance of these discrepancies, the estimated *df* were compared to the simulated distribution of the likelihood ratio statistic (for model independence). Results are shown in Table 4.

Results again indicate that the estimated degrees of freedom usually approximated the simulated value within an absolute error of a few percent. For example, most of the conditions (20 of 30) resulted in a 5% significance level between 0.03 and 0.07; with two exceptions, the largest discrepancy is an absolute difference of 0.04 from the true 5% level. The 10% significance level, however, again corresponds to somewhat larger errors, with the estimated p-values being as high as 0.34 and as low as 0.04; most results (19 of 30), however, were between 0.07 and 0.13. The 75^{th}, 50^{th} and 25^{th} percentiles showed similar findings with occasionally higher discrepancies.

The above results identify some complications and discrepancies that arise when using this method to estimate the model *df* for strictly binary inputs. First, the subsequent simulations show only a fair degree of correspondence between the predicted and simulated *df*. The majority of conditions led to percentiles within an absolute difference of a few percent, but other conditions led to more substantial discrepancies. Secondly, the established rules under this method led to some logical inconsistencies in the predicted *df*. For example, with 5 inputs, the predicted *df* for 3 hidden units (24.58) is actually larger than that predicted for 4 hidden units (23.41). This apparent contradiction arises from the fact that the *df* are essentially predicted by scaling the number of parameters by a function of the difference between the maximum number of terms and the number of model parameters. While this approach has some intuitive appeal - and generally leads to an increase in the degrees of freedom as the number of hidden units increases (for a given number of input variables) - no guarantee exists that this pattern will hold universally.

Due to this, some corrections are therefore needed for predicting the model *df* in these scenarios. To do so, when a decrease is observed with an increase in hidden units, it is possible to simply take the average of the previous result with the next number of hidden units. For example, for the case of 5 inputs with 4 hidden units, the previous result (24.58 for 3 hidden units) would be averaged with the next result (31 for 5 hidden units) to obtain 27.79, which is much closer to the simulated result of

Table 4: Comparison of Estimated to Simulated Degrees of Freedom with Binary Inputs

Max #of Terms (Inputs)	# of Parameters	(Hidden Untis)	Simulated *df*	Estimated *df*	Simulated α-levels using the Estimated *df*				
					0.75	0.50	0.25	0.10	0.05
31 (5)	22	3	23.84	24.58	0.69	0.46	0.22	0.08	0.05
	29	4	28.69	23.41	0.91	0.77	0.55	0.34	0.21
63 (6)	25	3	34.57	35.36	0.71	0.45	0.22	0.08	0.05
	33	4	42.81	45.06	0.69	0.42	0.17	0.05	0.02
	49	6	55.86	59.21	0.63	0.38	0.17	0.05	0.03
	57	7	59.60	58.92	0.79	0.52	0.28	0.13	0.06
127 (7)	28	3	45.93	45.13	0.78	0.53	0.26	0.11	0.06
	37	4	58.06	58.90	0.73	0.47	0.22	0.09	0.04
	55	6	82.27	85.03	0.68	0.41	0.20	0.07	0.04
	64	7	92.84	97.18	0.63	0.38	0.17	0.05	0.02
	73	8	102.5	108.5	0.61	0.35	0.14	0.04	0.02
	82	9	111.7	118.8	0.59	0.33	0.12	0.04	0.02
255 (8)	31	3	54.90	55.18	0.74	0.49	0.25	0.11	0.06
	41	4	73.02	72.59	0.76	0.54	0.24	0.09	0.05
	61	6	107.8	106.77	0.78	0.52	0.26	0.11	0.06
	71	7	124.8	123.50	0.78	0.53	0.28	0.11	0.04
	81	8	139.7	139.96	0.70	0.51	0.27	0.12	0.06
	91	9	155.0	156.13	0.71	0.50	0.23	0.11	0.07
511 (9)	34	3	65.13	65.82	0.75	0.48	0.22	0.09	0.04
	45	4	87.02	86.89	0.76	0.51	0.26	0.09	0.04
	67	6	131.4	128.7	0.79	0.58	0.33	0.14	0.09
	78	7	152.3	149.4	0.79	0.57	0.31	0.14	0.08
	89	8	171.8	170.0	0.77	0.55	0.29	0.13	0.06
	100	9	194.7	190.5	0.78	0.58	0.34	0.20	0.12
1023 (10)	37	3	75.5	77.16	0.72	0.46	0.21	0.06	0.02
	49	4	100.9	102.1	0.71	0.49	0.23	0.09	0.04
	73	6	152.7	151.7	0.78	0.53	0.26	0.11	0.06
	85	7	178.5	176.4	0.77	0.55	0.28	0.11	0.07
	97	8	204.8	201.0	0.82	0.59	0.31	0.12	0.07
	109	9	230.0	225.6	0.80	0.60	0.32	0.14	0.08
Mean Estimated α-levels					0.74	0.50	0.25	0.11	0.06

28.69 (and provides better correspondence to the simulated distribution of likelihood ratios). The other example arises with 6 inputs and 7 hidden units, where the estimated value of 58.92 would be replaced by 61.11 (which is slightly further away from the simulated result of 59.6).

Simulation Results for Continuous Input Variables

Results shown in Table 5 were generated using independently distributed continuous inputs from a standard normal distribution. All neural network models were fit using a weight decay of 0.01. Table 5 shows that the simulated distribution of the likelihood ratio test for independence again closely followed a chi-square distribution. In a large percentage of the cases, all of the simulated α-levels were within a few percent of the expected percentiles. No systematic differences were evident in the results. Figures 4a and 4b show two examples where: (1) the simulated distribution varied a few percent from the expected percentiles (2 inputs and 2 hidden units), and (2) the simulated distribution fell extremely close to the corresponding chi-square distribution (2 inputs and 10 hidden units). Both figures show noticeable variability in at least one extreme end of the distribution; however, these few points have little effect on the resulting significance levels that would be of practical interest.

Table 6 shows additional results for other cases with 3 or 8 hidden units and up to 10 inputs. Results again showed simulated α-levels very close to the expected percentiles. Q-Q plots (not shown here) showed similar findings as in Figures 4a and 4b.

Estimating Degrees of Freedom for Continuous Input Variables

As opposed to the case of binary inputs, the degrees of freedom (*df*) for continuous input variables do not have any specific limiting value. Therefore, it was assumed that the *df* would be a continuous (and probably linear) function of the number of hidden units. Further, it was assumed that the result would increase by some constant amount with an increase in the number of input variables. Using the results in Table 5, the relationship $df = h \times [3 \times (k-2) + 5]$ is obtained, which appears to hold well across those results (with 2, 5, and 10 hidden units). Since the specific values from Table 5 were not used to derive this relationship (other than observing the general trend), subsequent results combine simulations from Tables 5 and 6 (i.e., 2-10 inputs and 2, 3, 5, 8 and 10 hidden units). Figure 5 shows the relationship between the simulated and estimated *df* from the results in Tables 5 and 6. The plot illustrates a close correspondence between the simulated and estimated results, especially for smaller degrees of freedom.

Results in Table 7 show somewhat greater variability in the *df* and subsequent significance levels. Only the 5% significance level showed no systematic error, with most of the simulations giving a result (for 5% significance) within 2% of the correct level (e.g., between 3% and 7%). The variability in significance levels can be attributed to either the difference between the simulated and estimated *df* and/or the variability from a chi-square distribution. In most cases, the estimated *df* was at least slightly higher than the simulated result.

Simulation Results for both Binary and Continuous Input Variables

Table 8 shows results for both binary and continuous input variables. For each of these simulations, the number of hidden units was kept constant at 2, the number of continuous inputs was specified as 2, 5, or 10, and the number of binary inputs was varied between 2 and 10. The degrees of freedom (*df*) in parentheses in the first two columns of the table are the estimated values for model complexity (as described in the previous sections of this report). The additional *df* (column 5) gives the difference between the simulated *df* (when combining a given number of continuous and binary inputs) and the sum of estimated *df* (totaled from columns 1, 2 and 3).

The results in Table 8 illustrate several key issues. First, the simulation results show substantially more variability than predicted by the chi-square distribution, which is most likely a consequence of sub-optimal results from the minimization (of the deviance) routine in S-Plus. Secondly, a definite trend exists between the

number of continuous and binary variables and the additional *df* gained (or lost) when combining a given number of (continuous and binary) input variables.

At this point, the observed trends could be used to derive estimates of model complexity for the cases in Table 8 and for other cases with larger numbers of hidden units and other combinations of continuous and binary inputs (as done previously when separately considering continuous or binary inputs). However, the lack of correspondence with the chi-square distribution, and the subsequent need for improved model fitting (e.g., more global optimization procedures) would invalidate any subsequent findings. Therefore, modifications of the S-Plus procedures need to be pursued for these cases before any specific rules can be effectively formulated for the case of both continuous and binary inputs.

Table 5: Likelihood Ratio Statistic for Model Independence with Continuous Inputs

Inputs	Hidden Units	# Parameters	Likelihood Ratio		Simulated α-levels				
			Mean	Variance	0.75	0.50	0.25	0.10	0.05
2	2	9	10.8	25.0	0.74	0.51	0.25	0.08	0.03
	5	21	26.2	60.5	0.75	0.51	0.23	0.08	0.04
	10	41	49.8	99.4	0.75	0.50	0.25	0.09	0.05
3	2	11	16.6	33.6	0.74	0.51	0.26	0.10	0.04
	5	26	41.0	71.2	0.74	0.52	0.26	0.11	0.06
	10	51	82.5	171.2	0.76	0.50	0.24	0.10	0.05
4	2	13	22.2	40.0	0.77	0.53	0.25	0.09	0.04
	5	31	55.4	111.0	0.75	0.51	0.27	0.10	0.04
	10	61	115.3	216.9	0.75	0.49	0.26	0.11	0.06
5	2	15	27.7	57.9	0.77	0.51	0.25	0.07	0.03
	5	36	69.3	131.4	0.75	0.51	0.26	0.09	0.04
	10	71	144.8	293.4	0.75	0.50	0.24	0.10	0.05
6	2	17	33.4	65.4	0.76	0.54	0.27	0.08	0.04
	5	41	83.0	164.3	0.74	0.50	0.25	0.10	0.05
	10	81	176.7	341.1	0.74	0.53	0.24	0.09	0.05
7	2	19	38.7	100.1	0.77	0.51	0.21	0.07	0.02
	5	46	98.3	202.0	0.75	0.50	0.27	0.08	0.04
	10	91	205.2	375.8	0.75	0.51	0.25	0.11	0.05
8	2	21	44.8	101.1	0.78	0.52	0.23	0.08	0.04
	5	51	112.5	220.8	0.74	0.49	0.24	0.11	0.06
	10	101	239.0	476.9	0.75	0.51	0.25	0.10	0.04
9	2	23	49.9	142.2	0.79	0.53	0.19	0.05	0.02
	5	56	127.4	239.9	0.74	0.48	0.26	0.11	0.06
	10	111	269.0	487.6	0.73	0.50	0.27	0.11	0.05
10	2	25	54.6	166.1	0.80	0.49	0.19	0.05	0.03
	5	61	140.8	280.2	0.76	0.49	0.24	0.12	0.06
	10	121	299.5	546.4	0.75	0.51	0.26	0.10	0.04
Mean Simulated α-levels					0.75	0.51	0.25	0.09	0.04

Table 6: Additional Simulated Likelihood Ratio Statistics with Continuous Inputs

Inputs	Hidden Units	# Parameters	Likelihood Ratio		Simulated α-levels				
			Mean	Variance	0.75	0.50	0.25	0.10	0.05
2	3	13	15.9	35.4	0.76	0.51	0.23	0.08	0.04
	8	33	40.3	82.8	0.76	0.51	0.23	0.10	0.04
3	3	16	24.4	42.0	0.74	0.51	0.29	0.10	0.05
	8	41	65.9	135.7	0.73	0.49	0.25	0.11	0.06
4	3	19	32.9	60.6	0.73	0.53	0.27	0.10	0.05
	8	49	90.5	171.1	0.75	0.49	0.25	0.10	0.06
5	3	22	41.2	85.7	0.76	0.53	0.26	0.07	0.04
	8	57	115.2	200.1	0.73	0.51	0.26	0.11	0.06
6	3	25	48.9	84.4	0.73	0.51	0.29	0.13	0.05
	8	65	139.3	231.3	0.72	0.49	0.29	0.13	0.06
7	3	28	57.0	100.7	0.75	0.52	0.24	0.12	0.06
	8	73	160.9	299.5	0.73	0.49	0.27	0.11	0.06
8	3	31	64.9	140.9	0.75	0.50	0.25	0.10	0.04
	8	81	187.3	376.1	0.75	0.50	0.24	0.10	0.05
9	3	34	74.3	158.7	0.78	0.47	0.24	0.09	0.05
	8	89	211.4	421.9	0.75	0.50	0.25	0.11	0.05
10	3	37	81.4	171.6	0.76	0.50	0.22	0.09	0.06
	8	97	235.5	392.1	0.74	0.50	0.27	0.13	0.06
Mean Simulated α-levels					0.75	0.50	0.25	0.10	0.05

Figure 5: Plot of the Estimated by Simulated Degrees Of Freedom for Continuous Inputs and 2, 5 and 10 Hidden Units

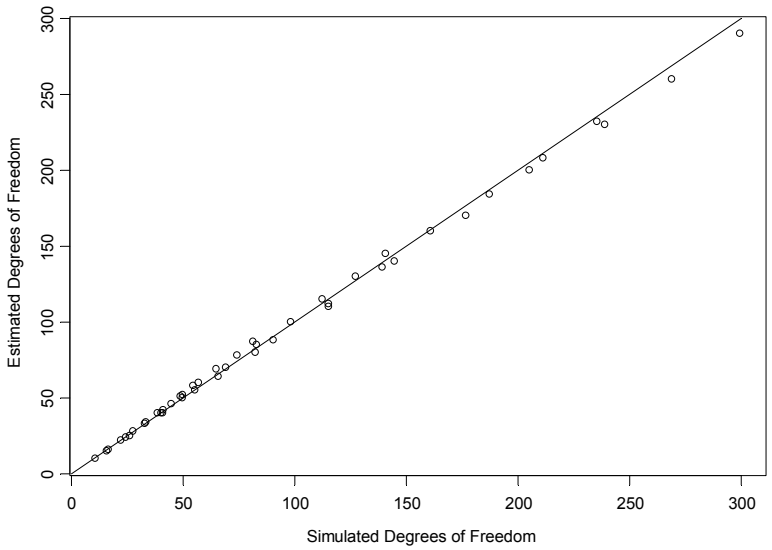


Table 7: Estimated and Simulated Degrees of Freedom with Continuous Inputs

Inputs	Hidden Units	# Parameters	Simulated *df*	Estimated *df*	Simulated α-levels using the Estimated *df*				
					0.75	0.50	0.25	0.10	0.05
2	2	9	10.8	10	0.79	0.58	0.32	0.11	0.05
	3	13	15.9	15	0.80	0.57	0.28	0.11	0.05
	5	21	26.2	25	0.80	0.58	0.28	0.11	0.06
	8	33	40.3	40	0.77	0.53	0.24	0.11	0.04
	10	41	49.8	50	0.75	0.49	0.25	0.08	0.05
3	2	11	16.6	16	0.77	0.55	0.30	0.13	0.06
	3	16	24.4	24	0.76	0.54	0.31	0.12	0.06
	5	26	41.0	40	0.77	0.57	0.30	0.13	0.07
	8	41	65.9	64	0.78	0.55	0.31	0.15	0.08
	10	51	82.5	80	0.81	0.58	0.30	0.14	0.07
4	2	13	22.2	22	0.77	0.54	0.26	0.10	0.04
	3	19	32.9	33	0.73	0.53	0.27	0.10	0.05
	5	31	55.4	55	0.76	0.52	0.28	0.10	0.05
	8	49	90.5	88	0.81	0.56	0.31	0.13	0.09
	10	61	115.3	110	0.85	0.62	0.39	0.20	0.11
5	2	15	27.7	28	0.76	0.49	0.24	0.07	0.03
	3	22	41.2	42	0.73	0.49	0.23	0.07	0.03
	5	36	69.3	70	0.73	0.49	0.25	0.09	0.04
	8	57	115.2	112	0.79	0.59	0.34	0.16	0.09
	10	71	144.8	140	0.83	0.62	0.34	0.16	0.09
6	2	17	33.4	34	0.73	0.50	0.24	0.07	0.03
	3	25	48.9	51	0.65	0.43	0.22	0.09	0.03
	5	41	83.0	85	0.68	0.44	0.20	0.07	0.03
	8	65	139.3	136	0.79	0.60	0.36	0.18	0.08
	10	81	176.7	170	0.84	0.67	0.36	0.17	0.10
7	2	19	38.7	40	0.72	0.45	0.17	0.05	0.01
	3	28	57.0	60	0.65	0.40	0.16	0.07	0.04
	5	46	98.3	100	0.71	0.46	0.23	0.06	0.03
	8	73	160.9	160	0.75	0.50	0.29	0.13	0.06
	10	91	205.2	200	0.83	0.61	0.34	0.17	0.08
8	2	21	44.8	46	0.74	0.47	0.19	0.06	0.03
	3	31	64.9	69	0.63	0.36	0.15	0.05	0.01
	5	51	112.5	115	0.69	0.43	0.19	0.08	0.04
	8	81	187.3	184	0.80	0.57	0.30	0.14	0.07
	10	101	239.0	230	0.86	0.67	0.39	0.19	0.09
9	2	23	49.9	52	0.73	0.45	0.14	0.03	0.01
	3	34	74.3	78	0.68	0.36	0.16	0.05	0.02
	5	56	127.4	130	0.69	0.41	0.21	0.09	0.04
	8	89	211.4	208	0.80	0.57	0.31	0.14	0.07
	10	111	269.0	260	0.84	0.65	0.41	0.20	0.11
10	2	25	54.6	58	0.70	0.37	0.11	0.03	0.01
	3	37	81.4	87	0.61	0.34	0.12	0.04	0.02
	5	61	140.8	145	0.68	0.40	0.17	0.07	0.04
	8	97	235.5	232	0.79	0.57	0.32	0.17	0.09
	10	121	299.5	290	0.86	0.66	0.40	0.19	0.10
Mean Simulated α-levels					0.76	0.52	0.27	0.11	0.05

Table 8: Likelihood Ratios with Continuous and Binary Inputs and 2 Hidden Units

Continuous Inputs (*df*)	Binary Inputs	*df*	Mean Likelihood Ratio	Additional *df*	Simulated α-levels				
					0.75	0.50	0.25	0.10	0.05
2 (10)	2	3.0	19.7	6.7	0.73	0.52	0.27	0.10	0.05
	3	7.0	24.8	7.8	0.76	0.51	0.26	0.08	0.04
	4	10.5	31.0	10.5	0.75	0.54	0.24	0.08	0.03
	5	18.5	36.7	8.2	0.78	0.53	0.25	0.06	0.03
	6	24.7	42.4	7.7	0.78	0.54	0.22	0.06	0.02
	7	31.0	47.7	6.7	0.77	0.51	0.22	0.06	0.02
	8	37.6	54.0	6.8	0.78	0.49	0.22	0.07	0.02
	9	44.6	58.4	3.8	0.81	0.49	0.19	0.06	0.02
	10	52.2	64.1	1.9	0.80	0.47	0.18	0.04	0.02
5 (28)	2	3.0	38.0	7.0	0.76	0.51	0.23	0.06	0.02
	3	7.0	43.6	8.6	0.79	0.53	0.20	0.05	0.02
	4	10.5	47.9	9.4	0.80	0.47	0.20	0.05	0.02
	5	18.5	52.9	6.4	0.80	0.50	0.18	0.06	0.02
	6	24.7	59.5	6.8	0.82	0.48	0.17	0.06	0.02
	7	31.0	64.3	5.3	0.84	0.44	0.18	0.04	0.02
	8	37.6	69.5	3.9	0.79	0.47	0.19	0.05	0.01
	9	44.6	73.6	1.0	0.80	0.46	0.19	0.05	0.02
	10	52.2	79.1	-1.1	0.82	0.45	0.17	0.04	0.01
10 (58)	2	3.0	64.4	3.4	0.81	0.46	0.18	0.05	0.02
	3	7.0	69.4	4.4	0.82	0.47	0.18	0.04	0.01
	4	10.5	72.2	3.7	0.80	0.50	0.18	0.05	0.02
	5	18.5	78.1	1.5	0.78	0.47	0.19	0.06	0.03
	6	24.7	83.0	0.3	0.79	0.46	0.16	0.06	0.02
	7	31.0	89.1	0.1	0.78	0.48	0.17	0.04	0.02
	8	37.6	94.7	-1.1	0.78	0.46	0.17	0.05	0.01
	9	44.6	98.6	-4.0	0.79	0.46	0.20	0.05	0.01
	10	52.2	103.3	-6.9	0.79	0.45	0.15	0.04	0.01
Mean Simulated α-levels					0.79	0.49	0.20	0.06	0.02

Categorical Input Variables

Additional simulations were conducted for categorical variables. In theory, a categorical variable with 3 levels should produce fewer degrees of freedom than 2 binary inputs, since the 3 levels would be coded as 2 binary inputs, but would not have an interaction between the 2 levels. Simulations (not shown here) provided evidence of this type of relationship, but simulation results differed substantially from the expected chi-square distribution. Therefore, as in the cases of both binary and continuous inputs, further work on these types of data are being delayed until a better optimization routine can be implemented with S-Plus or with another programming language.

Conclusions

One issue that was not addressed was the correlation of the input variables. All simulations were run with independently generated data. Comparing the current findings to previous analyses with some overlap (Landsittel, et al., 2003) indicates that the

degrees of freedom (*df*) may be somewhat lower with moderately correlated data, which is somewhat intuitive since correlated variables, to some degree, add less information than independently distributed variables. Rather than consider this further complication, it was decided that all simulations should use independent data as a starting point, and the effects of correlation should be addressed separately in a future manuscript.

Another limitation encountered here was the failure of the S-Plus routine to achieve acceptably optimal results in minimizing the deviance. Because the simulations with only binary, or only continuous inputs led to close correspondence with the chi-square distribution (which allows the use of the mean as the *df*), it would be expected that this would hold for models with both binary and continuous inputs. The failure to achieve this result is most likely a function of the (only locally optimal) routines. Future work will address this point through investigating other optimization routines (e.g., genetic algorithms), and incorporating those routines into the current approaches and methodology.

To the best of our knowledge, these studies are the first to use *df* under the null as a measure of model complexity. Unlike generalized linear models or other standard regression methods, the model complexity may vary substantially for different data sets. In terms of the general applicability of this approach, the complexity under the null may provide a more appropriate penalty for subsequent use in model selection in many scenarios, as higher complexity may be desirable if the true underlying association is highly non-linear. In contrast to a measure such as the generalized *df*, where the complexity tends to increase substantially when fit to data with some observed association, the complexity under the null only penalizing the model for incorrectly fitting non-linearity when none exists. Using an AIC-type statistic with generalized or effective *df*, for example, would highly penalize the neural network model for accurately fitting a highly non-linear association, and likely make it very difficult to select an adequately complex model.

Despite these limitations, the results contribute significantly to our understanding of neural network model complexity by providing explicit equations to quantify complexity under a range of scenarios. Once improved methods are implemented to better optimize more complex models (where there was significant variability from the expected chi-square distribution), the derived equations for *df* can be tested across a much wider range of models. Assuming results hold for other scenarios (to be tested after achieving more global optimization), the estimated *df* can be implemented in practice for model selection via AIC or BIC statistics. Such approaches would serve as a favorable alternative to any of the ad-hoc approaches currently being utilized in practice.

Acknowledgements

The author would like to acknowledge Mr. Dustin Ferris, who, as an undergraduate student in Mathematics at Duquesne University, conducted many of the initial simulations used to generate some of the results in this manuscript, and Duquesne University, for the Faculty Development Fund award, that included funding of Mr. Ferris's stipend.

References

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

Hastie, T., & Tibshirani, R. (1990). *Generalized additive models*. NY: Chapman and Hall.

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. (2003). Null distribution of the likelihood ratio statistic with feed-forward neural networks. *Journal of Modern and Applied Statistical Methods*, *1*(*2*), 333-342.

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*, 847-854. San Mateo, CA: Morgan Kaufmann.

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*, 9-14. North Holland: Elsevier Science Publishers.

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

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

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*. NY: Springer-Verlag.

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.