An Adjusted Network Information Criterion for Model Selection in Statistical Neural Network Models

Christopher Godwin Udomboso
University of Ibadan, Ibadan, Nigeria, cg.udomboso@gmail.com

Godwin Nwazu Amahia
University of Ibadan, Ibadan, Nigeria, go.amahia@ui.edu.ng

Isaac Kwame Dontwi
Kwame Nkrumah University of Science and Technology, Kumasi, Ghana, ikedontwi@hotmail.com

Follow this and additional works at: http://digitalcommons.wayne.edu/jmasm

Part of the Applied Statistics Commons, Social and Behavioral Sciences Commons, and the Statistical Theory Commons

Recommended Citation
DOI: 10.22237/jmasm/1478003040
Available at: http://digitalcommons.wayne.edu/jmasm/vol15/iss2/26

This Regular Article is brought to you for free and open access by the Open Access Journals at DigitalCommons@WayneState. It has been accepted for inclusion in Journal of Modern Applied Statistical Methods by an authorized editor of DigitalCommons@WayneState.
An Adjusted Network Information Criterion for Model Selection in Statistical Neural Network Models

Christopher Godwin Udomboso
University of Ibadan
Ibadan, Nigeria

Godwin Nwazu Amahia
University of Ibadan
Ibadan, Nigeria

Isaac K. Dontwi
Kwame Nkrumah University
of Science and Technology,
Kumasi, Ghana

In this paper, we derived and investigated the Adjusted Network Information Criterion (ANIC) criterion, based on Kullback’s symmetric divergence, which has been designed to be an asymptotically unbiased estimator of the expected Kullback-Leibler information of a fitted model. The ANIC improves model selection in more sample sizes than does the NIC.

Keywords: Statistical neural network, network information criterion, adjusted network information criterion, transfer function

Introduction

In choosing an appropriate model to characterize the sample data, it is ideal to be guided by scientific theory, as well as be well served by a data-driven selection method. Akaike (1973, 1974) introduced the Akaike information criterion, AIC, which endeavors discern the closeness of a fitted model is to the generating or true model. Akaike's work stimulated many other approaches to model selection, leading to the development of criteria such as SIC (Schwarz, 1978), BIC (Akaike, 1978), and HQ (Hannan, & Quinn 1979). Sugiura (1978) extended Akaike's original work by proposing AICc, a corrected version of AIC justified in the context of linear regression with normal errors.

The development of AICc was motivated by the need to adjust for AIC's propensity to favor high-dimensional models when the sample size is small relative to the maximum order of the models in the candidate class. Hurvich and...
Tsai (1989) show that AICc dramatically outperforms AIC in small-sample regression settings, and further extend AICc to include univariate Gaussian autoregressive models. Hurvich, Shumway, and Tsai (1990) generalize AICc to encompass univariate Gaussian autoregressive moving-average models, and Hurvich and Tsai (1993) handle the vector Gaussian autoregressive case.

The purpose of this study is to consider the selection of Statistical Neural Network model using the proposed method by Murata, Yoshizawa, and Amari (1994), which is the NIC. The NIC is observed to be sample biased, as it does not account for sample sizes. The selection of a model from a set of fitted candidate models requires objective data-driven criteria. The criterion we shall use in this study is that designed to be an asymptotically unbiased estimator of the expected Kullback-Leibler information of a fitted model (Akaike, 1973).

**Methodology**

**Adjusted Network Information Criterion (ANIC):**

We note that

\[ Y^* = HW + U \] (true model) \hspace{1cm} (1)

\[ Y^* = HW + e \] (estimated model) \hspace{1cm} (2)

Anders (1996) noted that should the network exactly map the true function \( F \), then the asymptotic relationship, \( G = 2B\sigma^2 \), so that \( tr(GB^{-1}) = 2\sigma^2 tr(I) = 2\sigma^2 k \). Thus, NIC becomes AIC as proposed by Amemiya (1980):

\[ AIC = MSE + 2\sigma^2 \frac{k}{n} \] \hspace{1cm} (3)

Therefore, in deriving an alternative NIC, we assume that the estimates network model includes the true network model, and the approach shall use the corrected AIC based on Kullback’s systematic divergence as used by Hafidi and Mkhadri (2006).

We recall that

\[ NIC = D[q, p(W)] \] \hspace{1cm} (4)
UDOMBOSO ET AL.

\[ \approx D[q, p(W_{opt})] + \frac{1}{2} \left( W - W_{opt} \right) \left( W - W_{opt} \right)' \frac{\partial^2}{\partial W_{opt}^2} D[q, p(W_{opt})]. \]  

(5)

Kullback (1968) defined the discrepancy between the true model and the estimated model as

\[ J(\theta_0, \theta) = D(\theta_0, \theta) - D(\theta_0, \theta_0) - D(\theta, \theta_0) + D(\theta, \theta) \]

(6)

where \( \theta_0 \) is the true and unknown parameter vector, \( \theta \) is the parameter vector of the candidate model. Also, \( f(Y|\theta_0) \) and \( f(Y|\theta) \) denote the densities for the true and estimates models.

Note that the second term does not depend on \( \theta \). Thus, Cavanaugh (1997, 1999), in order to discriminate among various models, proposed another form of Kullback’s symmetric divergence as

\[ K(\theta_0, \theta) = D(\theta_0, \theta) + D(\theta, \theta) \]

(7)

Given that the estimated model includes the true model, we can define the improved NIC as

\[ \text{ANIC} = D(W, W) + T \]

(8)

which is asymptotically an unbiased estimator of

\[ \Omega(d, W) = E_W \left[ N(W, W) \right] \]

(9)

where \( T \) is some value that improves the NIC, \( d \) is the dimension of \( W \), and is given as

\[ d = p + 1 \]

(10)

and \( N(W, W) \) is the NIC.
Proof:

\[ \Omega(d, W) = E_w \left\{ D(W, W) + \left[ D(W, W) - D(W, W) \right] \right\} \]  \hspace{1cm} (11)

But the true model is given as

\[ Y' = HW + U \quad U \sim N(0, \sigma^2 I_n), \]  \hspace{1cm} (12)

and the estimated model is

\[ Y^* = HW + e \]  \hspace{1cm} (13)

where \( Y^* \) is an \( n \times 1 \) observation, \( H \) is an \( n \times p \) observations, \( W = W^* \) is an \( p \times 1 \) observation. Assume that \( H \) is twice continuously differentiable in \( W \). Let \( t(\lambda) = HW \). Then, the log-likelihood of the estimated model is given as

\[ \ln f(Y^*|W) = \frac{n}{2} \ln 2\pi \sigma^2 - \frac{1}{2\sigma^2} (Y^* - t(\lambda))' (Y^* - t(\lambda)) \]  \hspace{1cm} (14)

Approach the second term of (1) by considering two hypothetical estimators \( w_1 \) and \( w_2 \), such that

\[ D(w_1, w_2) = E_{w_1} \left[ \ln f(Y^*|w_2) \right] \]  \hspace{1cm} (15)

\[ = E_{w_1} \left[ -\frac{n}{2} \ln 2\pi \sigma_{w_1}^2 - \frac{1}{2\sigma_{w_1}^2} (Y^* - t(\lambda))' (Y^* - t(\lambda)) \right] \]  \hspace{1cm} (16)

\[ = E_{w_1} \left[ -\frac{n}{2} \ln 2\pi \sigma_{w_1}^2 - \frac{1}{2\sigma_{w_1}^2} (Y^* - t(\lambda_1))' (Y^* - t(\lambda_1)) \right] \]  \hspace{1cm} (17)

\[ + \left[ (t(\lambda_1) - t(\lambda_2))' (t(\lambda_1) - t(\lambda_2)) \right] \]

\[ = -\frac{n}{2} \ln 2\pi \sigma_{w_1}^2 - \frac{1}{2\sigma_{w_1}^2} \left[ n\sigma_{w_1}^2 + (t(\lambda_1) - t(\lambda_2))' (t(\lambda_1) - t(\lambda_2)) \right]. \]  \hspace{1cm} (18)
Expand $D(W, W)$ as

$$D(W, W) = -\frac{n}{2} \ln 2\pi \sigma_w^2 - \frac{1}{2\sigma_w^2} \left[ n\sigma_w^2 + \left(t(\hat{\lambda}) - t(\lambda)\right) + \left(\hat{\lambda} - \lambda\right) \right]$$ (19)

Expanding $t(\hat{\lambda})$ in order one at $\hat{\lambda} = \lambda$,

$$t(\hat{\lambda}) \approx t(\lambda) + \frac{\partial t}{\partial \lambda} (\hat{\lambda} - \lambda)$$ (20)

This results in

$$D(W, W) \equiv -\frac{1}{2} \left\{ n \ln 2\pi \sigma_w^2 + \frac{1}{\sigma_w^2} \left[ 2\sigma_w^2 + \left(t(\lambda) + \frac{\partial t}{\partial \lambda} (\hat{\lambda} - \lambda) - t(\lambda) \right) \right] \right\}$$ (21)

$$= -\frac{1}{2} \left\{ n \ln 2\pi \sigma_w^2 + \frac{1}{\sigma_w^2} \left[ 2\sigma_w^2 + (\hat{\lambda} - \lambda) \left( \frac{\partial t}{\partial \lambda} + \frac{\partial t}{\partial \lambda} \right) \right] \right\}$$ (22)

Similarly,

$$D(W, W) = -\frac{n}{2} \ln 2\pi \sigma_w^2 - \frac{1}{2\sigma_w^2} \left[ n\sigma_w^2 + \left(t(\hat{\lambda}) - t(\lambda)\right) + \left(\hat{\lambda} - \lambda\right) \right]$$ (23)

$$= -\frac{1}{2} \left\{ n \ln 2\pi \sigma_w^2 + n \right\}$$ (24)

Thus, the second term of (11) becomes
\[ D(W, W) - D(W, W) \approx -\frac{1}{2} \left\{ n \ln 2\pi \sigma_w^2 + \frac{1}{2\sigma_w^2} \left[ n \sigma_w^2 + (\hat{\lambda} - \lambda) \left( \frac{\partial t}{\partial \hat{\lambda}} \right)' \left( \frac{\partial t}{\partial \hat{\lambda}} \right) \right] - n \ln 2\pi \sigma_w^2 \right\} \] (25)

\[ = -\frac{1}{2} \left\{ n \ln 2\pi \sigma_w^2 + n \frac{\sigma_w^2}{\sigma_w^2} \right\} \]

\[ = -\frac{1}{2} \left\{ n \left[ \ln 2\pi \sigma_w^2 - n \ln 2\pi \sigma_w^2 + \frac{\sigma_w^2}{\sigma_w^2} \right] \right\} \]

\[ = -\frac{1}{2} \left\{ n \left[ \ln \frac{2\pi \sigma_w^2}{2\pi} + \frac{\sigma_w^2}{\sigma_w^2} \right] \right\} \]

\[ = -\frac{1}{2} \left\{ n \left[ \ln \frac{\sigma_w^2}{\sigma_w^2} + \frac{\sigma_w^2}{\sigma_w^2} \right] \right\} \]

The distribution of
\[ \frac{2\sigma_w^2}{\sigma_w^2} \sim \chi_{n-p} \]

and

\[ \frac{1}{\sigma_w^2} \left[ \hat{\lambda} - \lambda \right] \left[ \frac{\hat{t}}{\partial \lambda} \right] \left[ \frac{\hat{t}}{\partial \lambda} \right] (\hat{\lambda} - \lambda) \sim \chi_p \]

Therefore,

\[ D(\mathbf{W}, \mathbf{W}) - D(\mathbf{W}, \mathbf{W}) = -\frac{1}{2} \left\{ n \ln \frac{\sigma_w^2}{\sigma_w} + (n - p) + p - n \right\} \]

\[ = -\frac{1}{2} \left\{ n \ln \frac{\sigma_w^2}{\sigma_w} \right\} \]

Taking expectation, the above becomes

\[ E\left[ D(\mathbf{W}, \mathbf{W}) - D(\mathbf{W}, \mathbf{W}) \right] \approx -\frac{1}{2} E \left\{ n \ln \frac{\sigma_w^2}{\sigma_w} \right\} \]

Bickel and Doksum (1977) noted that by taking a second order expansion of \( \ln \chi_{df} \) about \( df \) and evaluating the expectation of the result, the following relation ensues,

\[ E\left[ \ln \chi_{df} \right] = \ln df - \frac{1}{df} + o \left( \frac{1}{(df)^2} \right) \]

where \( df \) is degrees of freedom. Write
By Bickel and Doksum (1977) relation, and according to Cavanaugh (1997, 1999),

\[
E \left[ n \ln \frac{\sigma_w^2}{\sigma^2_w} \right] = nE \left[ \ln \frac{n\sigma_w^2}{\sigma^2_w} \right] - n \ln n
\]  

(34)

The first-order expansion of \( \ln (n - p) \) is

\[
\ln (n - p) = \ln n - \frac{p}{n} + o \left( \frac{p}{n} \right)^2
\]

(36)

Thus,

\[
E \left[ n \ln \frac{\sigma_w^2}{\sigma^2_w} \right] = n \left\{ \ln n - \frac{p}{n} + o \left( \frac{p}{n} \right)^2 - \frac{1}{n - p} + o \left( \frac{1}{(n - p)^2} \right) \right\} - n \ln n
\]

(37)

\[
\approx \left\{ p + \frac{n}{n - p} \right\}
\]

(38)

\[
= - \left\{ \frac{np - p^2 + n}{n - p} \right\}
\]

(39)

Putting this result back in (32),

\[
E \left[ D(W,W) - D(W,W) \right] \approx - \frac{1}{2} \left\{ - \left[ \frac{np - p^2 + n}{n - p} \right] \right\}
\]

(40)

\[
= \frac{np - p^2 + n}{2(n - p)}
\]

(41)

Thus, the alternative NIC becomes
ANIC = NIC + \frac{np - p^2 + n}{2(n - p)} \quad (42)

which is a correction for the biased NIC.

**Results**

**Illustrative Examples:**

The following illustrations demonstrate the power of the adjusted network information criterion in accounting for sample size. Anders (1996) proposed a statistical neural network model given as

\[
y = f(X, w) + u
\]

where \( y \) is the dependent variable, \( X = (x_0 = 1, x_1, \ldots, x_l) \) is a vector of independent variables, \( w = (\alpha, \beta, \gamma) \) is the network weight: ‘\( \alpha \)' is the weight of the input unit, ‘\( \beta \)' is the weight of the hidden unit, and ‘\( \gamma \)' is the weight of the output unit, and \( u_i \) is the stochastic term that is normally distributed (that is, \( u_i \sim N(0, \sigma^2 I_n) \)).

\( f(X, w) \) is the artificial neural network function, expressed as

\[
f(X, w) = \alpha X + \sum_{h=1}^{H} \beta_h g \left( \sum_{i=0}^{I} \gamma_{hi} x_i \right).
\]

where \( g(\cdot) \) is the transfer function.

The proposed convoluted form of the artificial neural network function used in this study is

\[
f(X, w) = \alpha X + \sum_{h=1}^{H} \beta_h \left[ g_1 \left( \sum_{i=0}^{I} \gamma_{hi} x_i \right) g_2 \left( \sum_{i=0}^{I} \gamma_{hi} x_i \right) \right].
\]

and thus, the form of the statistical neural network model proposed is

\[
y = \alpha X + \sum_{h=1}^{H} \beta_h \left[ g_1 \left( \sum_{i=0}^{I} \gamma_{hi} x_i \right) g_2 \left( \sum_{i=0}^{I} \gamma_{hi} x_i \right) \right] + u_i u_f,
\]

419
where \( y \) is the dependent variable, \( \bm{X} = (x_0 = 1, x_1, \ldots, x_I) \) is a vector of independent variables, \( \bm{w} = (\alpha, \beta, \gamma) \) is the network weight: \( \alpha \) is the weight of the input unit, \( \beta \) is the weight of the hidden unit, and \( \gamma \) is the weight of the output unit, \( u_i \) and \( u_j \) are the stochastic terms that are normally distributed (that is, \( u_i, u_j \sim N(0, \sigma^2 I_n) \)), and \( g_1(.) \) and \( g_2(.) \) are the transfer functions.

The choice of the transfer functions used was based on preliminary investigations of the fifteen (15) transfer functions which uses hidden neurons that included 2, 5, 10, 50, and 100 at 1000 iterations. Best performances came from Hyperbolic Tangent transfer function (TANH), Hyperbolic Tangent Sigmoid transfer function (TANSIG), and Symmetric Saturating Linear transfer function (SATLINS), respectively. Similarly, further investigation was conducted on the choice of convolution, and it was found out that best performance was obtained in the convolution of the Symmetric Saturating Linear transfer function and the Hyperbolic Tangent transfer function (SATLINS_TANH), followed by the convolution of the Symmetric Saturating Linear transfer function and the Hyperbolic Tangent Sigmoid transfer function (SATLINS_TANSIG). The data used for the analyses used in this research were split into two – 2 and 3. The hidden neurons used include 2, 5, 10, 20, 40, 60, 80, and 100, while the sample sizes include 10, 20, 40, 60, 80, 100, 125, 150, 175, 200, 250, 300, and 400.

Based on two (2) variables, it is shown in Table 1 that the values of NIC across samples, while Table 2 shows the values of ANIC across the samples. It is shown in Table 3 that the sample points at which the values of NIC and ANIC are low in each heterogeneous models in comparison to the root (homogeneous) models.

**Table 1. Model Selections across Samples based on NIC (2 Variables)**

<table>
<thead>
<tr>
<th>NIC</th>
<th>( n = )</th>
<th>10</th>
<th>20</th>
<th>40</th>
<th>60</th>
<th>80</th>
<th>100</th>
<th>125</th>
<th>150</th>
<th>175</th>
<th>200</th>
<th>250</th>
<th>300</th>
<th>400</th>
</tr>
</thead>
<tbody>
<tr>
<td>SATLINS</td>
<td>0.0038</td>
<td>0.0026</td>
<td>0.0239</td>
<td>0.0021</td>
<td>0.0002</td>
<td>0.0007</td>
<td>0.0013</td>
<td>0.0011</td>
<td>0.0044</td>
<td>0.0039</td>
<td>0.0012</td>
<td>0.0031</td>
<td>0.0068</td>
<td></td>
</tr>
<tr>
<td>TANH</td>
<td>0.0054</td>
<td>0.0217</td>
<td>0.0016</td>
<td>0.0006</td>
<td>0.0113</td>
<td>0.0003</td>
<td>0.0005</td>
<td>0.0021</td>
<td>0.0023</td>
<td>0.0021</td>
<td>0.0017</td>
<td>0.0029</td>
<td>0.0045</td>
<td></td>
</tr>
<tr>
<td>TANSIG</td>
<td>0.0031</td>
<td>0.0120</td>
<td>0.0017</td>
<td>0.0047</td>
<td>0.0023</td>
<td>0.0003</td>
<td>0.0113</td>
<td>0.0011</td>
<td>0.0038</td>
<td>0.0024</td>
<td>0.0017</td>
<td>0.0052</td>
<td>0.0044</td>
<td></td>
</tr>
<tr>
<td>SATLINS_TANH</td>
<td>0.0066</td>
<td>0.0227</td>
<td>0.0028</td>
<td>0.0008</td>
<td>0.0110</td>
<td>0.0001</td>
<td>0.0007</td>
<td>0.0004</td>
<td>0.0011</td>
<td>0.0024</td>
<td>0.0024</td>
<td>0.0023</td>
<td>0.0037</td>
<td></td>
</tr>
<tr>
<td>SATLINS_TANSIG</td>
<td>0.0049</td>
<td>0.0125</td>
<td>0.0056</td>
<td>0.0010</td>
<td>0.0013</td>
<td>0.0003</td>
<td>0.0018</td>
<td>0.0019</td>
<td>0.0050</td>
<td>0.0039</td>
<td>0.0007</td>
<td>0.0041</td>
<td>0.0043</td>
<td></td>
</tr>
</tbody>
</table>
Table 2. Model Selections across Samples based on ANIC (2 Variables)

<table>
<thead>
<tr>
<th>Model</th>
<th>10</th>
<th>20</th>
<th>40</th>
<th>60</th>
<th>80</th>
<th>100</th>
<th>125</th>
<th>150</th>
<th>200</th>
<th>250</th>
<th>300</th>
<th>400</th>
</tr>
</thead>
<tbody>
<tr>
<td>SATLINS</td>
<td>1.6217</td>
<td>1.5581</td>
<td>1.5500</td>
<td>1.5154</td>
<td>1.5130</td>
<td>1.5091</td>
<td>1.5069</td>
<td>1.5048</td>
<td>1.5046</td>
<td>1.5034</td>
<td>1.5061</td>
<td>1.5080</td>
</tr>
<tr>
<td>TANH</td>
<td>1.6073</td>
<td>1.5261</td>
<td>1.5224</td>
<td>1.5154</td>
<td>1.5015</td>
<td>1.5065</td>
<td>1.5083</td>
<td>1.5078</td>
<td>1.5064</td>
<td>1.5046</td>
<td>1.5045</td>
<td>1.5178</td>
</tr>
<tr>
<td>TANSIG</td>
<td>1.5627</td>
<td>1.5185</td>
<td>1.5199</td>
<td>1.5093</td>
<td>1.5193</td>
<td>1.5164</td>
<td>1.5080</td>
<td>1.5063</td>
<td>1.5050</td>
<td>1.5076</td>
<td>1.5055</td>
<td>1.6102</td>
</tr>
<tr>
<td>SATLINS_TANH</td>
<td>1.6025</td>
<td>1.5245</td>
<td>1.5215</td>
<td>1.5149</td>
<td>1.5012</td>
<td>1.5085</td>
<td>1.5071</td>
<td>1.5071</td>
<td>1.5059</td>
<td>1.5099</td>
<td>1.5201</td>
<td>1.5252</td>
</tr>
<tr>
<td>SATLINS_TANSIG</td>
<td>1.5257</td>
<td>1.5260</td>
<td>1.5151</td>
<td>1.5120</td>
<td>1.5089</td>
<td>1.5074</td>
<td>1.5062</td>
<td>1.5039</td>
<td>1.5066</td>
<td>1.5056</td>
<td>1.5047</td>
<td>1.5961</td>
</tr>
</tbody>
</table>

Table 3. Sample points at which NIC and ANIC are low in each heterogeneous model in comparison to the root models (2 Variables)

<table>
<thead>
<tr>
<th>Model</th>
<th>Sample Size n</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>NIC</td>
</tr>
<tr>
<td>SATLINS_TANH</td>
<td>100,150,175,400</td>
</tr>
<tr>
<td>SATLINS_TANSIG</td>
<td>100,250,400</td>
</tr>
</tbody>
</table>

Figure 1. Graph of NIC based on Sample Sizes (2 Variables)
Correspondingly based on two (2) variables, Figure 1 is the graph of NIC across samples, while Figure 2 is the graph of ANIC across samples. The models in ANIC are almost parallel between sample number 10 and 150 inclusive.

Similarly, based on three (3) variables, Table 4 shows the values of NIC across samples, while Table 5 shows the values of ANIC across the samples. Table 6 shows the sample points at which the values of NIC and ANIC are low in each heterogeneous models in comparison to the root (homogeneous) models.

**Table 4. Model Selections across Samples based on NIC (3 Variables)**

<table>
<thead>
<tr>
<th>NIC</th>
<th>10</th>
<th>20</th>
<th>40</th>
<th>60</th>
<th>80</th>
<th>100</th>
<th>125</th>
<th>150</th>
<th>175</th>
<th>200</th>
<th>250</th>
<th>300</th>
<th>400</th>
</tr>
</thead>
<tbody>
<tr>
<td>SATLINS</td>
<td>0.4682</td>
<td>0.0306</td>
<td>0.0196</td>
<td>0.0363</td>
<td>0.0210</td>
<td>0.0561</td>
<td>0.0090</td>
<td>0.0166</td>
<td>0.0154</td>
<td>0.0139</td>
<td>0.0203</td>
<td>0.0230</td>
<td>0.0436</td>
</tr>
<tr>
<td>TANH</td>
<td>0.3184</td>
<td>1.0532</td>
<td>0.0301</td>
<td>0.0350</td>
<td>0.0197</td>
<td>0.0158</td>
<td>0.0141</td>
<td>0.0228</td>
<td>0.0154</td>
<td>0.0213</td>
<td>0.0196</td>
<td>0.0225</td>
<td>0.0736</td>
</tr>
<tr>
<td>TANSIG</td>
<td>0.3115</td>
<td>0.1102</td>
<td>0.0216</td>
<td>0.0537</td>
<td>0.0160</td>
<td>0.0189</td>
<td>0.0213</td>
<td>0.0173</td>
<td>0.0254</td>
<td>0.0165</td>
<td>0.0206</td>
<td>0.0489</td>
<td></td>
</tr>
<tr>
<td>SATLINS_TANH</td>
<td>0.3540</td>
<td>0.0274</td>
<td>0.0245</td>
<td>0.0159</td>
<td>0.0193</td>
<td>0.0137</td>
<td>0.0159</td>
<td>0.0471</td>
<td>0.0159</td>
<td>0.0192</td>
<td>0.0112</td>
<td>0.0179</td>
<td>0.0462</td>
</tr>
<tr>
<td>SATLINS_TANSIG</td>
<td>0.0517</td>
<td>0.0784</td>
<td>0.0601</td>
<td>0.0198</td>
<td>0.0201</td>
<td>0.0282</td>
<td>0.0193</td>
<td>0.0206</td>
<td>0.0180</td>
<td>0.0176</td>
<td>0.0143</td>
<td>0.0192</td>
<td>0.1375</td>
</tr>
</tbody>
</table>
Table 5. Model Selections across Samples based on ANIC (3 Variables)

<table>
<thead>
<tr>
<th>Model</th>
<th>ANIC</th>
<th>Sample Size</th>
<th>NIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>SATLINS</td>
<td>2.1172</td>
<td>10, 20</td>
<td>2.1083</td>
</tr>
<tr>
<td>TANH</td>
<td>2.0444</td>
<td>2.0372</td>
<td>2.0144</td>
</tr>
<tr>
<td>TANSIG</td>
<td>2.0276</td>
<td>2.0338</td>
<td>2.0248</td>
</tr>
<tr>
<td>SATLINS_TANH</td>
<td>2.0209</td>
<td>175, 200</td>
<td>2.0142</td>
</tr>
<tr>
<td>SATLINS_TANSIG</td>
<td>2.0199</td>
<td>200, 250</td>
<td>2.0199</td>
</tr>
</tbody>
</table>

Table 6. Sample points at which NIC and ANIC are low in each heterogeneous model in comparison to the root models (3 Variables)

<table>
<thead>
<tr>
<th>Model</th>
<th>Sample Size n</th>
<th>NIC</th>
<th>ANIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>SATLINS_TANH</td>
<td>20, 60, 80, 100, 250, 300</td>
<td></td>
<td>20, 100, 150, 250, 400</td>
</tr>
<tr>
<td>SATLINS_TANSIG</td>
<td>60, 250, 300</td>
<td></td>
<td>40, 175, 200, 250, 400</td>
</tr>
</tbody>
</table>

Figure 3. Graph of NIC based on Sample Sizes (3 Variables)
Correspondingly based on three (3) variables, Figure 3 is the graph of NIC across samples, while Figure 4 is the graph of ANIC across samples. The models in ANIC became almost parallel from around sample number 20 and 40 up till sample number 400.

A test shows significant difference between the homogeneous and heterogeneous models ($p < 0.05$). Rates of selection for the heterogeneous models are respectively 72.9%, and 72.1% using NIC, against 66.9%, 55.9% and 65.1% respectively for the homogeneous models, while with ANIC the heterogeneous models have rates of selection respectively as 66.9% and 66.8%, against 66.7%, 66.2%, and 66.6 for the respective homogeneous models. The results of the ANIC demonstrate the high precision of SNN models at large samples.
Conclusion

An ANIC criterion was derived, based on Kullback’s symmetric divergence, for model selection in some Statistical Neural Network models. The analyses show that on a general note, the ANIC improves model selection in more sample sizes than does the NIC. Because neural network is a data-driven model, then more attention should be paid to the sample size when determining the model to be selected.

References

ADJUSTED NETWORK INFORMATION FOR SNN MODEL SELECTION


criterion determining the number of hidden units for artificial neural network
models. *IEEE Transactions on Neural Networks, 5*(6), 865-872.
doi: 10.1109/72.329683

Statistics, 6*(2), 461-464. doi: 10.1214/aos/1176344136

Sugiura, N. (1978). Further analysis of the data by Akaike's information
criterion and the finite corrections. *Communications in Statistics - Theory and
Methods, 7*(1), 13-26. doi: 10.1080/03610927808827599