

11-1-2016

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

Christopher Godwin Udombosu

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

Udombosu, Christopher Godwin; Amahia, Godwin Nwazu; and Dontwi, Isaac Kwame (2016) "An Adjusted Network Information Criterion for Model Selection in Statistical Neural Network Models," *Journal of Modern Applied Statistical Methods*: Vol. 15: Iss. 2, Article 26.

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 administrator 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

Dr. Udomboso is a Lecturer in the Department of Statistics. Email him at: cg.udomboso@gmail.com. Professor Amahia is in the Department of Statistics. Email at: go.amahia@ui.edu.ng. Professor Dontwi is in the Department of Mathematics. Email him at: ikedontwi@hotmail.com.

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

$$\mathbf{Y}^* = \mathbf{H}\mathbf{W} + \mathbf{U} \quad (\text{true model}) \quad (1)$$

$$\mathbf{Y}^* = \mathbf{H}\mathbf{W} + \mathbf{e} \quad (\text{estimated model}) \quad (2)$$

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):

$$AIC = MSE + 2\sigma^2 \frac{k}{n} \quad (3)$$

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

$$NIC \equiv D\left[q, p(\mathbf{W})\right] \quad (4)$$

$$\cong D[q, p(\mathbf{W}_{opt})] + \frac{1}{2}(\mathbf{W} - \mathbf{W}_{opt})'(\mathbf{W} - \mathbf{W}_{opt}) \frac{\partial^2}{\partial \mathbf{W}_{opt}^2} D[q, p(\mathbf{W}_{opt})]. \quad (5)$$

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

$$J(\boldsymbol{\theta}_0, \boldsymbol{\theta}) = [D(\boldsymbol{\theta}_0, \boldsymbol{\theta}) - D(\boldsymbol{\theta}_0, \boldsymbol{\theta}_0)] - [D(\boldsymbol{\theta}, \boldsymbol{\theta}_0) - D(\boldsymbol{\theta}, \boldsymbol{\theta})] \quad (6)$$

where $\boldsymbol{\theta}_0$ is the true and unknown parameter vector, $\boldsymbol{\theta}$ is the parameter vector of the candidate model. Also, $f(\mathbf{Y}|\boldsymbol{\theta}_0)$ and $f(\mathbf{Y}|\boldsymbol{\theta})$ denote the densities for the true and estimates models.

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

$$K(\boldsymbol{\theta}_0, \boldsymbol{\theta}) = [D(\boldsymbol{\theta}_0, \boldsymbol{\theta})] + [D(\boldsymbol{\theta}, \boldsymbol{\theta}_0) - D(\boldsymbol{\theta}, \boldsymbol{\theta})] \quad (7)$$

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

$$ANIC = D(\mathbf{W}, \mathbf{W}) + T \quad (8)$$

which is asymptotically an unbiased estimator of

$$\Omega(d, \mathbf{W}) = E_{\mathbf{W}} [N(\mathbf{W}, \mathbf{W})] \quad (9)$$

where T is some value that improves the NIC, d is the dimension of \mathbf{W} , and is given as

$$d = p + 1 \quad (10)$$

and $N(\mathbf{W}, \mathbf{W})$ is the NIC.

Proof:

$$\Omega(d, \mathbf{W}) = E_{\mathbf{W}} \left\{ D(\mathbf{W}, \mathbf{W}) + \left[D(\mathbf{W}, \mathbf{W}) - D(\mathbf{W}, \mathbf{W}) \right] \right\} \quad (11)$$

But the true model is given as

$$\mathbf{Y}^* = \mathbf{H}\mathbf{W} + \mathbf{U} \quad \mathbf{U} \sim N(0, \sigma^2 I_n), \quad (12)$$

and the estimated model is

$$\mathbf{Y}^* = \mathbf{H}\mathbf{W} + \mathbf{e} \quad (13)$$

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

$$\ln f(\mathbf{Y}^* | \mathbf{W}) = \frac{n}{2} \ln 2\pi\sigma^2 - \frac{1}{2\sigma^2} (\mathbf{Y}^* - t(\lambda))' (\mathbf{Y}^* - t(\lambda)) \quad (14)$$

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

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

$$= E_{w_1} \left[-\frac{n}{2} \ln 2\pi\sigma_2^2 - \frac{1}{2\sigma_2^2} (Y^* - t(\lambda_2))' (Y^* - t(\lambda_2)) \right] \quad (16)$$

$$= E_{w_1} \left[-\frac{n}{2} \ln 2\pi\sigma_2^2 - \frac{1}{2\sigma_2^2} (Y^* - t(\lambda_1))' (Y^* - t(\lambda_1)) \right. \\ \left. + (t(\lambda_1) - t(\lambda_2))' (t(\lambda_1) - t(\lambda_2)) \right] \quad (17)$$

$$= -\frac{n}{2} \ln 2\pi\sigma_2^2 - \frac{1}{2\sigma_2^2} \left[n\sigma_1^2 + (t(\lambda_1) - t(\lambda_2))' (t(\lambda_1) - t(\lambda_2)) \right]. \quad (18)$$

Expand $D(\mathbf{W}, \mathbf{W})$ as

$$D(\mathbf{W}, \mathbf{W}) = -\frac{n}{2} \ln 2\pi\sigma_w^2 - \frac{1}{2\sigma_w^2} \left[n\sigma_w^2 + \left(t(\hat{\lambda}) - t(\lambda) \right)' \left(t(\hat{\lambda}) - t(\lambda) \right) \right] \quad (19)$$

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

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

This results in

$$D(\mathbf{W}, \mathbf{W}) \cong -\frac{1}{2} \left\{ \frac{n \ln 2\pi\sigma_w^2 + \frac{1}{\sigma_w^2} \left[2\sigma_w^2 + \left[t(\lambda) + \frac{\partial t}{\partial \hat{\lambda}} (\hat{\lambda} - \lambda) - t(\lambda) \right]' \left[t(\lambda) + \frac{\partial t}{\partial \hat{\lambda}} (\hat{\lambda} - \lambda) - t(\lambda) \right] \right]}{\sigma_w^2} \right\} \quad (21)$$

$$= -\frac{1}{2} \left\{ n \ln 2\pi\sigma_w^2 + \frac{1}{\sigma_w^2} \left[2\sigma_w^2 + (\hat{\lambda} - \lambda) \left[\frac{\partial t}{\partial \hat{\lambda}} \right]' \left[\frac{\partial t}{\partial \hat{\lambda}} \right] (\hat{\lambda} - \lambda) \right] \right\} \quad (22)$$

Similarly,

$$D(\mathbf{W}, \mathbf{W}) = -\frac{n}{2} \ln 2\pi\sigma_w^2 - \frac{1}{2\sigma_w^2} \left[n\sigma_w^2 + \left(t(\hat{\lambda}) - t(\hat{\lambda}) \right)' \left(t(\hat{\lambda}) - t(\hat{\lambda}) \right) \right] \quad (23)$$

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

Thus, the second term of (11) becomes

$$D(\mathbf{W}, \mathbf{W}) - D(\mathbf{W}, \mathbf{W}) \cong -\frac{1}{2} \left\{ \begin{array}{l} n \ln 2\pi\sigma_{\mathbf{w}}^2 + \\ \frac{1}{2\sigma_{\mathbf{w}}^2} \left[n\sigma_{\mathbf{w}}^2 + (\hat{\lambda} - \lambda) \left[\frac{\partial t}{\partial \hat{\lambda}} \right]' \right] \\ \left[\frac{\partial t}{\partial \hat{\lambda}} \right] (\hat{\lambda} - \lambda) \end{array} \right\} - n \ln 2\pi\sigma_{\mathbf{w}}^2 \quad (25)$$

$$= -\frac{1}{2} \left\{ \begin{array}{l} n \ln 2\pi\sigma_{\mathbf{w}}^2 + n \frac{\sigma_{\mathbf{w}}^2}{\sigma_{\mathbf{w}}^2} + \\ \frac{1}{\sigma_{\mathbf{w}}^2} \left[(\hat{\lambda} - \lambda) \left[\frac{\partial t}{\partial \hat{\lambda}} \right]' \left[\frac{\partial t}{\partial \hat{\lambda}} \right] (\hat{\lambda} - \lambda) \right] \end{array} \right\} - n \ln 2\pi\sigma_{\mathbf{w}}^2 - n \quad (26)$$

$$= -\frac{1}{2} \left\{ \begin{array}{l} n \left[\ln 2\pi\sigma_{\mathbf{w}}^2 - n \ln 2\pi\sigma_{\mathbf{w}}^2 + \frac{\sigma_{\mathbf{w}}^2}{\sigma_{\mathbf{w}}^2} \right] \\ + \frac{1}{\sigma_{\mathbf{w}}^2} \left[(\hat{\lambda} - \lambda) \left[\frac{\partial t}{\partial \hat{\lambda}} \right]' \left[\frac{\partial t}{\partial \hat{\lambda}} \right] (\hat{\lambda} - \lambda) \right] \end{array} \right\} - n \quad (27)$$

$$= -\frac{1}{2} \left\{ \begin{array}{l} n \left[\ln \frac{2\pi\sigma_{\mathbf{w}}^2}{2\pi\sigma_{\mathbf{w}}^2} + \frac{\sigma_{\mathbf{w}}^2}{\sigma_{\mathbf{w}}^2} \right] + \\ \frac{1}{\sigma_{\mathbf{w}}^2} \left[(\hat{\lambda} - \lambda) \left[\frac{\partial t}{\partial \hat{\lambda}} \right]' \left[\frac{\partial t}{\partial \hat{\lambda}} \right] (\hat{\lambda} - \lambda) \right] \end{array} \right\} - n \quad (28)$$

$$= -\frac{1}{2} \left\{ \begin{array}{l} n \left[\ln \frac{\sigma_{\mathbf{w}}^2}{\sigma_{\mathbf{w}}^2} + \frac{\sigma_{\mathbf{w}}^2}{\sigma_{\mathbf{w}}^2} \right] + \\ \frac{1}{\sigma_{\mathbf{w}}^2} \left[(\hat{\lambda} - \lambda) \left[\frac{\partial t}{\partial \hat{\lambda}} \right]' \left[\frac{\partial t}{\partial \hat{\lambda}} \right] (\hat{\lambda} - \lambda) \right] \end{array} \right\} - n \quad (29)$$

The distribution of

$$\frac{2\sigma_{\mathbf{w}}^2}{\sigma_{\mathbf{w}}^2} \sim \chi_{n-p}$$

and

$$\frac{1}{\sigma_{\mathbf{w}}^2} \left[(\hat{\lambda} - \lambda) \left[\frac{\partial t}{\partial \hat{\lambda}} \right]' \left[\frac{\partial t}{\partial \hat{\lambda}} \right] (\hat{\lambda} - \lambda) \right] \sim \chi_p$$

Therefore,

$$D(\mathbf{W}, \mathbf{W}) - D(\mathbf{W}, \mathbf{W}) = -\frac{1}{2} \left\{ n \ln \frac{\sigma_{\mathbf{w}}^2}{\sigma_{\mathbf{w}}^2} + (n-p) + p-n \right\} \quad (30)$$

$$= -\frac{1}{2} \left\{ n \ln \frac{\sigma_{\mathbf{w}}^2}{\sigma_{\mathbf{w}}^2} \right\} \quad (31)$$

Taking expectation, the above becomes

$$E \left[D(\mathbf{W}, \mathbf{W}) - D(\mathbf{W}, \mathbf{W}) \right] \cong -\frac{1}{2} E \left\{ n \ln \frac{\sigma_{\mathbf{w}}^2}{\sigma_{\mathbf{w}}^2} \right\} \quad (32)$$

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,

$$E \left[\ln \chi_{df} \right] = \ln df - \frac{1}{df} + o \left[\frac{1}{(df)^2} \right] \quad (33)$$

where df is degrees of freedom. Write

$$E \left[n \ln \frac{\sigma_{\mathbf{W}}^2}{\sigma_{\mathbf{W}}} \right] = nE \left[\ln \frac{n\sigma_{\mathbf{W}}^2}{\sigma_{\mathbf{W}}} \right] - n \ln n \quad (34)$$

By Bickel and Doksum (1977) relation, and according to Cavanaugh (1997, 1999),

$$E \left[n \ln \frac{\sigma_{\mathbf{W}}^2}{\sigma_{\mathbf{W}}} \right] = n \left\{ \ln(n-p) - \frac{1}{n-p} + o \left[\frac{1}{(n-p)^2} \right] \right\} - n \ln n \quad (35)$$

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

$$\ln(n-p) = \ln n - \frac{p}{n} + o \left(\frac{p}{n} \right)^2 \quad (36)$$

Thus,

$$E \left[n \ln \frac{\sigma_{\mathbf{W}}^2}{\sigma_{\mathbf{W}}} \right] = n \left\{ \ln n - \frac{p}{n} + o \left(\frac{p}{n} \right)^2 - \frac{1}{n-p} + o \left[\frac{1}{(n-p)^2} \right] \right\} - n \ln n \quad (37)$$

$$\cong - \left\{ p + \frac{n}{n-p} \right\} \quad (38)$$

$$= - \left\{ \frac{np - p^2 + n}{n-p} \right\} \quad (39)$$

Putting this result back in (32),

$$E \left[D(\mathbf{W}, \mathbf{W}) - D(\mathbf{W}, \mathbf{W}) \right] \cong - \frac{1}{2} \left\{ - \left[\frac{np - p^2 + n}{n-p} \right] \right\} \quad (40)$$

$$= \frac{np - p^2 + n}{2(n-p)} \quad (41)$$

Thus, the alternative NIC becomes

$$\text{ANIC} = \text{NIC} + \frac{np - p^2 + n}{2(n - p)} \quad (42)$$

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

$$\mathbf{y} = f(\mathbf{X}, \mathbf{w}) + \mathbf{u} \quad (43)$$

where \mathbf{y} is the dependent variable, $\mathbf{X} = (\mathbf{x}_0 \equiv \mathbf{1}, \mathbf{x}_1, \dots, \mathbf{x}_I)$ is a vector of independent variables, $\mathbf{w} = (\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma})$ is the network weight: ' $\boldsymbol{\alpha}$ ' is the weight of the input unit, ' $\boldsymbol{\beta}$ ' is the weight of the hidden unit, and ' $\boldsymbol{\gamma}$ ' is the weight of the output unit, and \mathbf{u}_i is the stochastic term that is normally distributed (that is, $\mathbf{u}_i \sim N(0, \sigma^2 I_n)$).

$f(\mathbf{X}, \mathbf{w})$ is the artificial neural network function, expressed as

$$f(\mathbf{X}, \mathbf{w}) = \boldsymbol{\alpha}\mathbf{X} + \sum_{h=1}^H \boldsymbol{\beta}_h g\left(\sum_{i=0}^I \boldsymbol{\gamma}_{hi} x_i\right), \quad (44)$$

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

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

$$f(\mathbf{X}, \mathbf{w}) = \boldsymbol{\alpha}\mathbf{X} + \sum_{h=1}^H \boldsymbol{\beta}_h \left[g_1\left(\sum_{i=0}^I \boldsymbol{\gamma}_{hi} x_i\right) g_2\left(\sum_{i=0}^I \boldsymbol{\gamma}_{hi} x_i\right) \right], \quad (45)$$

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

$$\mathbf{y} = \boldsymbol{\alpha}\mathbf{X} + \sum_{h=1}^H \boldsymbol{\beta}_h \left[g_1\left(\sum_{i=0}^I \boldsymbol{\gamma}_{hi} x_i\right) g_2\left(\sum_{i=0}^I \boldsymbol{\gamma}_{hi} x_i\right) \right] + \mathbf{u}_i \mathbf{u}_j, \quad (46)$$

ADJUSTED NETWORK INFORMATION FOR SNN MODEL SELECTION

where y is the dependent variable, $\mathbf{X} = (\mathbf{x}_0 \equiv \mathbf{1}, \mathbf{x}_1, \dots, \mathbf{x}_I)$ is a vector of independent variables, $\mathbf{w} = (\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma})$ is the network weight: ‘ $\boldsymbol{\alpha}$ ’ is the weight of the input unit, ‘ $\boldsymbol{\beta}$ ’ is the weight of the hidden unit, and ‘ $\boldsymbol{\gamma}$ ’ 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 $g_1(\cdot)$ and $g_2(\cdot)$ 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)

	NIC												
$n =$	10	20	40	60	80	100	125	150	175	200	250	300	400
SATLINS	0.0038	0.0026	0.0239	0.0021	0.0002	0.0007	0.0013	0.0011	0.0044	0.0039	0.0012	0.0031	0.0068
TANH	0.0054	0.0217	0.0016	0.0006	0.0113	0.0003	0.0005	0.0021	0.0023	0.0021	0.0017	0.0029	0.0045
TANSIG	0.0031	0.0120	0.0017	0.0047	0.0023	0.0003	0.0113	0.0011	0.0038	0.0024	0.0017	0.0052	0.0044
SATLINS_TANH	0.0066	0.0227	0.0028	0.0008	0.0110	0.0001	0.0007	0.0004	0.0011	0.0024	0.0024	0.0023	0.0037
SATLINS_TANSIG	0.0049	0.0125	0.0056	0.0010	0.0013	0.0003	0.0018	0.0019	0.0050	0.0039	0.0007	0.0041	0.0043

Table 2. Model Selections across Samples based on ANIC (2 Variables)

		ANIC													
<i>n</i> =	10	20	40	60	80	100	125	150	175	200	250	300	400		
SATLINS	1.6217	1.5581	1.5500	1.5154	1.5130	1.5107	1.5091	1.5069	1.5048	1.5046	1.5043	1.5051	1.5080		
TANH	1.6073	1.5261	1.5224	1.5154	1.5015	1.5095	1.5083	1.5078	1.5064	1.5046	1.5045	1.5178	1.5248		
TANSIG	1.5627	1.5185	1.5199	1.5093	1.5099	1.5193	1.5164	1.5080	1.5063	1.5050	1.5076	1.5056	1.6102		
SATLINS_TANH	1.6025	1.5245	1.5215	1.5149	1.5012	1.5091	1.5080	1.5059	1.5071	1.5039	1.5201	1.5252	1.5119		
SATLINS_TANSIG	1.5257	1.5260	1.5151	1.5120	1.5089	1.5074	1.5062	1.5039	1.5066	1.5056	1.5047	1.5961	1.5706		

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

Model	Sample Size <i>n</i>	
	NIC	ANIC
SATLINS_TANH	100,150,175,400	10,20,40,60,80,100,125,150,200
SATLINS_TANSIG	100,250,400	10,40,80,100,125,150

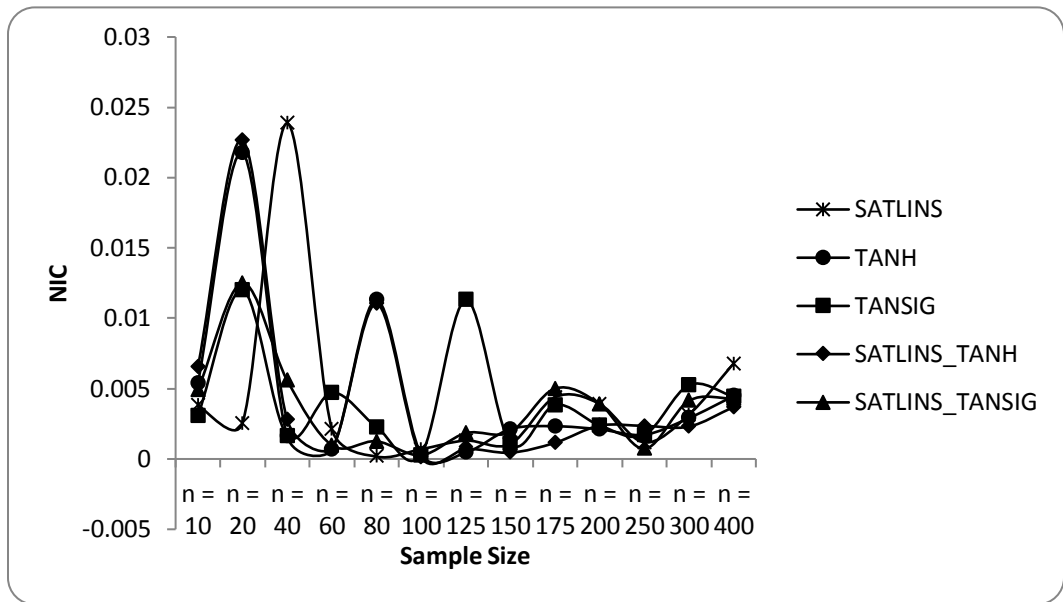


Figure 1. Graph of NIC based on Sample Sizes (2 Variables)

ADJUSTED NETWORK INFORMATION FOR SNN MODEL SELECTION

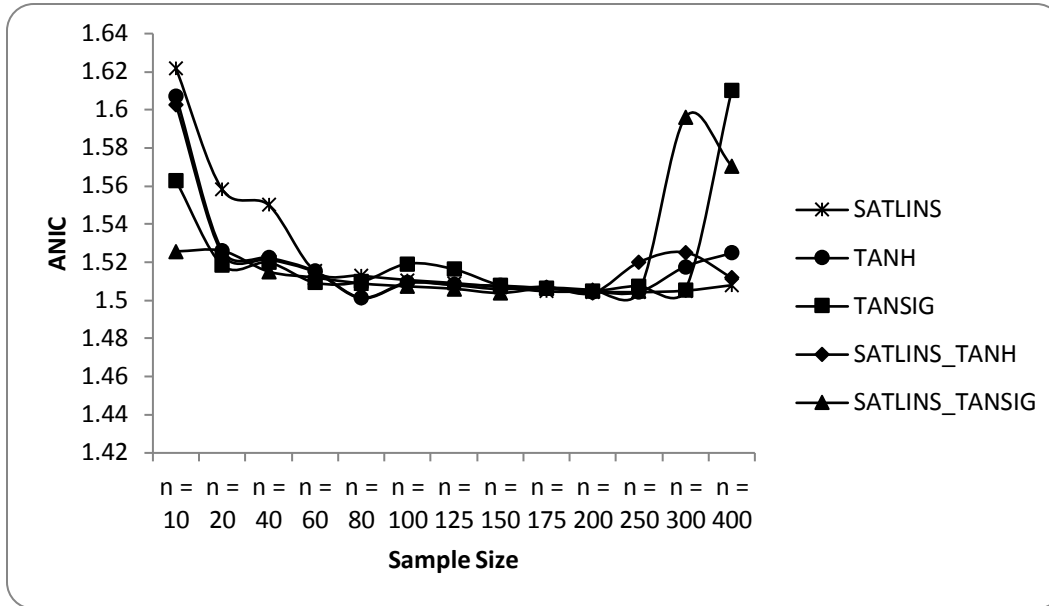


Figure 2. Graph of ANIC based on Sample Sizes (2 Variables)

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)

	NIC													
n =	10	20	40	60	80	100	125	150	175	200	250	300	400	
SATLINS	0.4682	0.0306	0.0196	0.0363	0.0210	0.0561	0.0090	0.0166	0.0154	0.0139	0.0203	0.0230	0.0436	
TANH	0.3184	1.0532	0.0301	0.0350	0.0197	0.0158	0.0141	0.0228	0.0154	0.0213	0.0195	0.0225	0.0736	
TANSIG	0.3115	0.1102	0.0216	0.0537	0.0160	0.0189	0.0149	0.0213	0.0173	0.0254	0.0165	0.0206	0.0489	
SATLINS_TANH	0.3540	0.0274	0.0245	0.0159	0.0193	0.0137	0.0159	0.0471	0.0159	0.0192	0.0112	0.0179	0.0462	
SATLINS_TANSIG	0.0517	0.0784	0.0601	0.0198	0.0201	0.0282	0.0193	0.0206	0.0180	0.0176	0.0143	0.0192	0.1375	

Table 5. Model Selections across Samples based on ANIC (3 Variables)

		ANIC												
$n =$		10	20	40	60	80	100	125	150	175	200	250	300	400
SATLINS		2.1172	2.1083	2.0572	2.0269	2.0405	2.0684	2.0209	2.0230	2.0215	2.0186	2.0227	2.0229	2.0349
TANH		2.4044	3.1075	2.0372	2.0144	2.0388	2.0276	2.0142	2.0199	2.0177	2.0238	2.0203	2.0109	2.0039
TANSIG		2.0076	2.1847	2.0383	2.0748	2.0338	2.0344	2.0248	2.0159	2.0216	2.0234	2.0156	2.0145	2.0223
SATLINS_TANH		2.2510	2.0752	2.0464	2.0383	2.0349	2.0261	2.0243	1.9935	2.0207	2.0258	2.0116	2.0170	1.9995
SATLINS_TANSIG		2.1847	2.1356	2.0093	2.0413	2.0368	2.0312	2.0248	2.0223	2.0168	2.0140	2.0086	2.0192	1.8820

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

Model	Sample Size n	
	NIC	ANIC
SATLINS_TANH	20,60,80,100,250,300	20,100,150,250,400
SATLINS_TANSIG	60,250,300	40,175,200,250,400

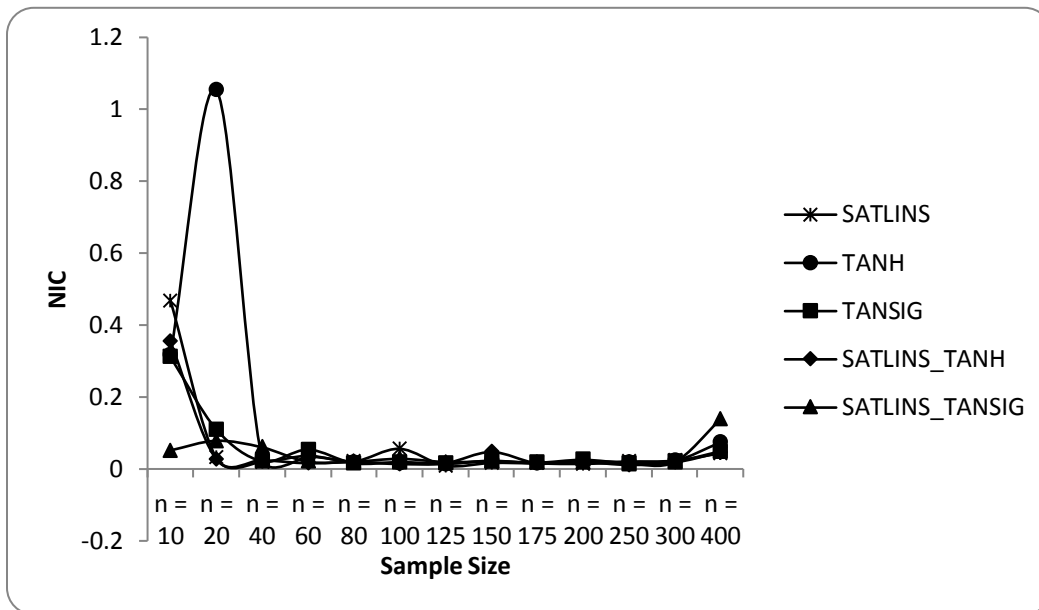


Figure 3. Graph of NIC based on Sample Sizes (3 Variables)

ADJUSTED NETWORK INFORMATION FOR SNN MODEL SELECTION

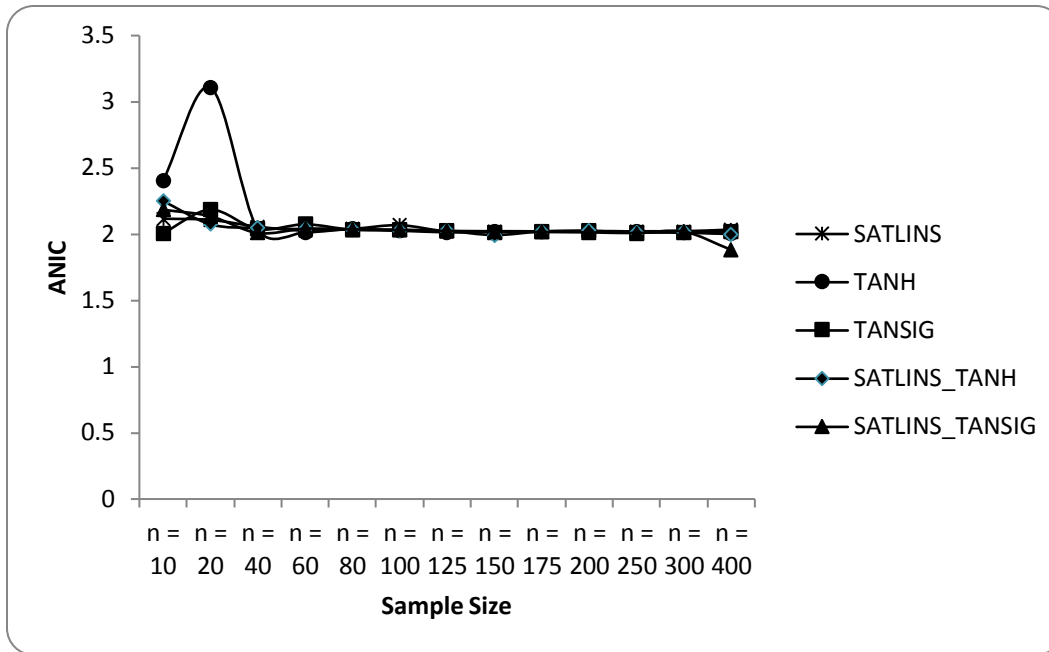


Figure 4. Graph of ANIC based on Sample Sizes (3 Variables)

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.

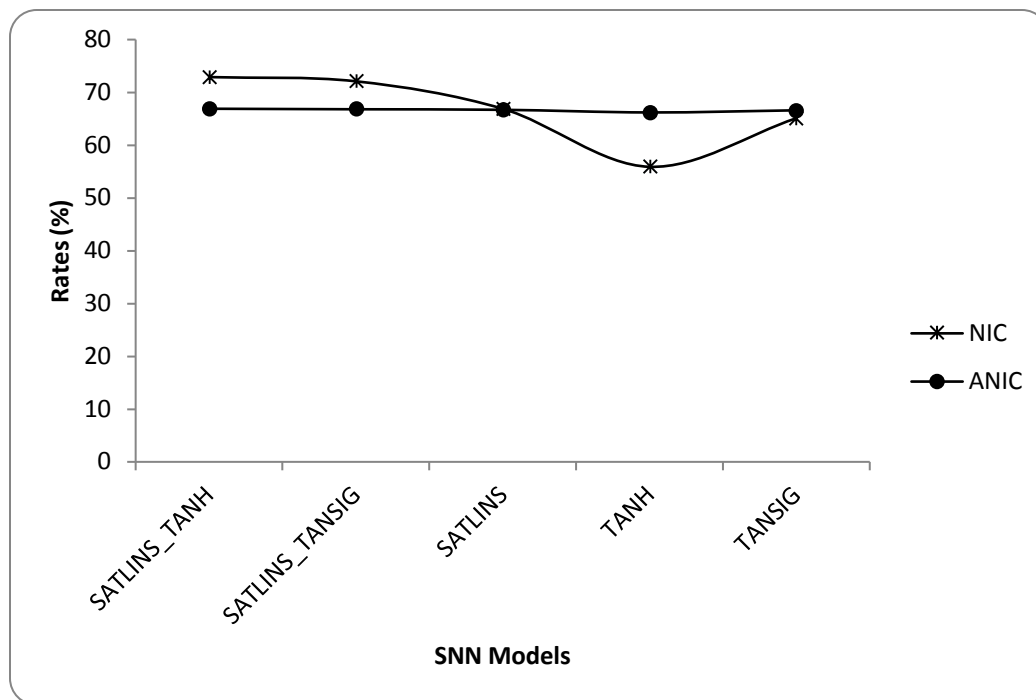


Figure 5. Overall Rates of Efficiency and Selection of the SNN Models: SATLINS_TANH, SATLINS_TANSIG, SATLINS, TANH, TANSIG

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

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. N. Petrov & F. Csaki, (Eds.), *Second International Symposium Information Theory* (pp. 267-281). Budapest: Akademia Kiado.

- Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6), 716-723.
doi: 10.1109/TAC.1974.1100705
- Akaike, H. (1978). A Bayesian analysis of the minimum AIC procedure. *Annals of the Institute of Mathematical Statistics*, 30(1), 9-14.
- Amemiya, T. (1980). Selection of regressors. *International Economic Review*, 21(2), 331-354.
- Anders, U. (1996). Statistical model building for neural networks. 6th *International AFIR Colloquium*. Nürnberg, Germany.
- Bickel, P. J., & Doksum, K. A. (1977). *Mathematical statistics: Basic ideas and selected topics*. New Jersey, NJ: Prentice-Hall.
- Cavanaugh, J. E. (1997). Unifying the derivations of the Akaike and corrected Akaike information criteria. *Statistics & Probability Letters*, 33(2), 201-208. doi: 10.1016/S0167-7152(96)00128-9
- Cavanaugh, J. E. (1999). A large-sample model selection criterion based on Kullback's symmetric divergence. *Statistics & Probability Letters*, 42(4), 333-343. doi: 10.1016/S0167-7152(98)00200-4
- Hafidi, B., & Mkhadri, A. (2006). A corrected Akaike criterion based on Kullback's symmetric divergence: applications in time series, multiple and multivariate regression. *Computational Statistics & Data Analysis*, 50(6), 1524-1550. doi: 10.1016/j.csda.2005.01.007
- Hannan, E., & Quinn, B. (1979). The determination of the order of an autoregression. *Journal of the Royal Statistical Society. Series B (Methodological)*, 41(2), 190-195.
- Hurvich, C. M., Shumway, R., & Tsai, C.-L. (1990). Improved estimators of Kullback–Leibler information for autoregressive model selection in small samples. *Biometrika*, 77(4), 709-719. doi: 10.1093/biomet/77.4.709
- Hurvich, C. M., & Tsai, C.-L. (1989). Regression and time series model selection in small samples. *Biometrika*, 76(2), 297-307.
doi: 10.1093/biomet/76.2.297
- Hurvich, C. M., & Tsai, C.-L. (1993). Corrected Akaike information criterion for vector autoregressive model selection. *Journal of Time Series Analysis*, 14(3), 271-279. doi: 10.1111/j.1467-9892.1993.tb00144.x
- Kullback, S. (1968). *Information theory and statistics*. New York, NY: Dover.

Murata, N., Yoshizawa, S., & Amari, S. (1994). Network information 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

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of 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