Proceeding of the 2004 American Control Conference
Boston, Massachusetts June 30 - July 2, 2004

Analysis of the ΔAIC Statistic for Optimal Detection of Small Changes in Dynamic Systems

Jeremy S. Conner and Dale E. Seborg†
Department of Chemical Engineering
University of California,
Santa Barbara, CA 93106
jconner@engineering.ucsb.edu,
seborg@engineering.ucsb.edu †corresponding author

Wallace E. Larimore
Adaptics, Inc.
1717 Briar Ridge Road,
McLean, VA 22101
larimore@adaptics.com

***Abstract*— The Akaike Information Criterion (AIC) is often used as a measure of model accuracy. The ΔAIC statistic is defined by the difference between AIC values for two nested models. The ΔAIC statistic corresponding to a particular change detection problem has been shown to detect extremely small changes in a dynamic system as compared with traditional change detection monitoring procedures. In this paper, a theoretical analysis is developed that shows the ΔAIC is actually an optimal test for the detection of any small changes in the characteristics of a process. It is also shown that the change/no-change hypotheses are nested. This result leads to a generalized likelihood ratio test with optimal properties as well as the precise large sample distribution for the test. A simulation of a dynamic system with small changes demonstrates the precision of the distribution theory as compared with the empirical results.**

I. Overview

The problem of detecting small changes in dynamic systems is important in a number of applications. In some systems, there is a change in the dynamics over time, such as the occurrence of a small leak, a build up of soot in a boiler, or valve fouling or stiction. These changes may be small relative to the noise and disturbances in the system, requiring a significant amount of data to detect their presence. Although the changes in the relevant systems parameters may be small, the potential consequences in terms of economics, reliability, or safety may be very large. In such cases, the accurate determination of the presence of a change and the precise determination of the nature of such a change can be critical.

The nature of this small change detection problem is quite different than many detection and identification methods currently under study in the literature. Many detection methods focus on the rapid detection of large changes in a process. By their nature, the detection of small changes requires substantial amounts of data following the process change, so that this is a class of problems distinct from the rapid detection of large changes. One advantage in the problem of detecting small changes is that the need to use larger data sets facilitates the use of very powerful large sample distribution theory. Such theory leads to an optimal detection procedure for small changes in dynamic systems. This will become apparent in the development of this paper.

The ΔAIC statistic for the detection of changes or faults in dynamic systems was developed by Larimore [1], and compared with traditional failure detection methods such as CUSUM and principal component analysis by Wang et. al. [2]. Significant improvements in detection sensitivity were achieved using the ΔAIC statistic, in some cases by a factor greater than 100. The ΔAIC applies simply to both static regression models as well as complex dynamic systems. However, the major issue when using the ΔAIC has been the lack of a theory for the distribution of the test statistic that is needed to determine the probability of detection and false alarms. This paper addresses this shortcoming.

To introduce the use of ΔAIC for detecting a change, suppose that D_1 and D_2 are two data sets that are disjoint and possibly noncontiguous. Of primary interest is determining if a change has occured between the data sets D_1 and D_2. The following two hypotheses are considered:

- *No Change Hypothesis* H_n: a single model M_n is valid for the two data sets, D_1 and D_2.
- *Change Hypothesis* H_a: different models are required for the two data sets, D_1 and D_2. These models are assumed to be statistically independent and will be denoted M_1 and M_2.

Each of the models M_1, M_2, and M_n is the result of parameter estimation using the maximum likelihood (ML) method. The ML method is used because of the optimal properties of such models particularly concerning the use of likelihood ratio (LR) tests and the AIC. We will also refer to the no change hypothesis as the null hypothesis H_n because it is to be tested for rejection against the more general, alternative hypothesis H_a as in the traditional hypothesis testing terminology. It will be shown that H_n is nested in H_a, which means that the no change or null hypothesis H_n is a special case of the change, or alternative hypothesis H_a.

There are several approaches for comparing the two hypotheses H_n and H_a. The traditional approach is to compute the likelihood ratio statistic, while a closely related approach is to compute the difference of the AIC values for the two hypotheses. The similarities and differences between these two approaches will be discussed in more detail in following sections. The AIC statistic is an asymptotically unbiased estimator of the Kullback–Leibler information quantity and is equal to two times the negative log of the maximized likelihood function plus two times the number of estimated parameters [3], [4]:

$$AIC = -2\log L(\hat{\Theta}, \hat{\Sigma}) + 2\nu \tag{1}$$

The first term is a measure of model fit while 2ν can be viewed as a penalty term that encourages the use of parsimonious models. For comparing two hypotheses using the AIC, the values of the AIC for the respective hypotheses are differenced. Thus the ΔAIC statistic is defined as the difference of the AICs for the hypotheses H_a and H_n,

$$\begin{aligned} \Delta AIC &\triangleq AIC(H_n) - AIC(H_a) \\ &= AIC(M_n) - AIC(M_1) - AIC(M_2) \end{aligned} \tag{2}$$

where the equality follows because the AIC for the change hypothesis H_a is simply the sum of the respective AICs, from the independence of the models M_1 and M_2. Although the ΔAIC quantity is fundamental to comparisons of AIC for various hypotheses, the authors are not aware of it being applied to the comparison of the hypotheses of no change H_n and change H_a. Also we are not aware of it having been shown that these hypotheses are nested as discussed below.

In this paper, the ΔAIC statistic is analyzed in some depth to reveal the distribution theory for a number of particular cases. The ΔAIC is particularly attractive as a statistic to test for changes and faults in dynamic systems:

- First, it will be shown that for a fixed order of the various models, the ΔAIC is a nested test of the null hypothesis H_n verses the alternative hypothesis H_a of a change in the process.
- The AIC is an estimate of the Kullback–Leibler information that is a fundamental measure of model approximation. It is a measure of model disparity based on the fundamental principles of sufficiency and repeated sampling [5]–[7].
- For fixed orders of the various models, the ΔAIC is a likelihood ratio test that has optimal statistical properties as the sample size becomes large.
- Because ΔAIC is a likelihood ratio test, for large samples it is a uniformly most powerful invariant test statistic for the detection of all possible changes that might occur in the hypothesized model structure, including changes in dynamics, input and output gains, and disturbance characteristics described as a time invariant linear system.
- In using a single test statistic for determining if a change has occurred in any combination of the parameters, the generalized likelihood ratio (GLR) test is optimal. Thus the GLR statistic provides an optimal global test for any changes in the process.

In developing the distribution theory for the ΔAIC statistic, some of the extensive theory of likelihood ratio tests will be utilized. The contribution of this paper is in showing that the ΔAIC statistic does indeed fit this framework for the case of a fixed model order. Also we develop some of the specific details for showing that the problem of optimal change detection is indeed a nested problem so that the generalized likelihood ratio testing theory applies. It is hoped that a distribution theory for the ΔAIC will lead to a much greater use of it, because it is now possible to calculate confidence limits for the detection of process faults and changes.

In the development below, the multivariate regression model is discussed in Section II and the no change hypothesis is shown to be a special case of the model change hypothesis. Thus, the hypotheses are nested. The generalized likelihood ratio test and its asymptotic distribution for the nested case are discussed in Section III. The AIC is developed in Section IV and related to the likelihood ratio test to obtain the asymptotic distribution of the ΔAIC statistic for the nested case. A simulation of an ARX time–series process is given in Section V, and the observed distribution of the ΔAIC statistic is compared to the theoretical distribution.

II. Multivariate Regression and Nested Structure

In this section, the change detection problem is developed for the case of multivariate regression. It will lead to the distribution theory for the large sample case for ML estimators.

The multivariate regression model

$$y_i = \Theta u_i + e_i \quad ; \quad \Sigma = \mathcal{E}(e_i e_i^T) \tag{3}$$

$$Y = \Theta U + E \tag{4}$$

over some specified set of measurements, for example $i = 1, \ldots, N$, will be considered below with several variations. Here $\mathcal{E}$ is the population average or expectation operation, Y is the $(p \times N)$ measurement matrix with the ith measurements as the p-dimensional column vector y_i, and E is the $(p \times N)$ measurement error matrix with ith measurement error vector e_i. It is assumed that e_i is normally distributed, independent of e_j for $j \neq i$, and has covariance matrix Σ. The $(p \times q)$ matrix Θ is the unknown parameter matrix, and U is the $(q \times N)$ regressor matrix with the ith column u_i. The dimensions of Θ and U in the discussion below will depend on the particular model under consideration.

For the no change hypothesis H_n, which is the null hypothesis of a single model valid for both data sets D_1 and D_2, the subscript "n" will be used. Thus the unknown parameter matrix Θ_n is $(p \times q)$, and the regressor matrix

U_n is $(q \times N)$. The multivariate regression model for the no change hypothesis H_n is then

$$y_i = \Theta_n u_i + e_i \quad ; \quad \Sigma_n = \mathcal{E}(e_i e_i^T) \tag{5}$$

for $i = 1, \dots, N$.

Under the change, or alternative hypothesis, H_a, suppose that the two data sets D_1 and D_2 are distributed independently with samples (Y_1, U_1) and (Y_2, U_2) of sample sizes N_1 and N_2 respectively with $N = N_1 + N_2$, and that U in Eq. 4 is correspondingly partitioned as $U = (U_1 \; U_2)$. If the $(p \times q)$ parameter matrix Θ is estimated separately as Θ_1 and Θ_2 for each of the data sets, D_1 and D_2 respectively, then two independent regression models are obtained

$$y_i = \Theta_1 u_i + e_i \quad ; \quad \Sigma_1 = \mathcal{E}(e_i e_i^T) \tag{6}$$

for $i = 1, \dots, N_1$ that is data set D_1, and

$$y_i = \Theta_2 u_i + e_i \quad ; \quad \Sigma_2 = \mathcal{E}(e_i e_i^T) \tag{7}$$

for $i = N_1 + 1, \dots, N$ that is data set D_2. Thus the model for the alternative hypothesis is given in the form of (4) by,

$$\begin{bmatrix} Y_1 & 0 \\ 0 & Y_2 \end{bmatrix} = \begin{bmatrix} \Theta_1 & 0 \\ 0 & \Theta_2 \end{bmatrix} \begin{bmatrix} U_1 & 0 \\ 0 & U_2 \end{bmatrix} + \begin{bmatrix} E_1 & 0 \\ 0 & E_2 \end{bmatrix} \tag{8}$$

and the parameter matrix for the model M_a can be represented as:

$$\Theta_a = \begin{bmatrix} \Theta_1 & 0 \\ 0 & \Theta_2 \end{bmatrix} \tag{9}$$

Now it is shown that the no change hypothesis H_n is a special case of the change hypothesis H_a when the parameter matrices Θ_n, Θ_1, and Θ_2 all have the same dimension, *i.e.* the model orders are all the same for M_n, M_1, and M_2. To show this, the log likelihood function for H_a is expressed with the values Θ_n and Σ_n substituted for Θ_1,Θ_2 and Σ_1,Σ_2, respectively. The resulting log likelihood function is

$$\begin{aligned} &\log L(Y_1|U_1; \Theta_1 = \Theta_n, \Sigma_1 = \Sigma_n) \\ &\qquad + \log L(Y_2|U_2; \Theta_2 = \Theta_n, \Sigma_2 = \Sigma_n) \\ &= \frac{N_1}{2} \log|\Sigma_n| + \frac{1}{2} \sum_{i=1}^{N_1} (y_i - \Theta_n u_i)^T \Sigma_n^{-1} (y_i - \Theta_n u_i) \\ &+ \frac{N_2}{2} \log|\Sigma_n| + \frac{1}{2} \sum_{i=N_1+1}^{N} (y_i - \Theta_n u_i)^T \Sigma_n^{-1} (y_i - \Theta_n u_i) \\ &= \frac{N}{2} \log|\Sigma_n| + \frac{1}{2} \sum_{i=1}^{N} (y_i - \Theta_n u_i)^T \Sigma_n^{-1} (y_i - \Theta_n u_i) \\ &= \log L(Y|U; \Theta_n, \Sigma_n) \end{aligned} \tag{10}$$

This function is precisely the log likelihood for H_n.

From the above analysis, Theorem 1 follows:

Theorem 1. Nested Model Structure. If $U = (U_1, U_2)$, so the matrix dimensions of Θ_n, Θ_1, and Θ_2 are identical, then the parameters (Θ_n, Σ_n) under the no change hypothesis H_n lie in the subspace of the parameter space $(\Theta_1, \Theta_2, \Sigma_1, \Sigma_2)$ for the change hypothesis H_a defined by the constraints, $\Theta_1 = \Theta_2$ and $\Sigma_1 = \Sigma_2$.

The maximum likelihood estimators for the model (3) obtained by maximization of the likelihood function are developed in Anderson [8] as

$$\hat{\Theta} = YU^T(UU^T)^{-1} \; ; \; \hat{\Sigma} = (YY^T - \hat{\Theta}UU^T\hat{\Theta}^T)/N \tag{11}$$

For the case of the change model M_a that consists of model M_1 for dataset D_1 and model M_2 for dataset D_2, the log likelihood function is as in (10) except that the parameter values are not constrained to be the same for both datasets. Then the two likelihood expressions for each dataset, D_1 and D_2, are maximized separately in maximizing the sum. Thus there is no difficulty in considering a concatenation of two models that involve separate parameters (Θ_1, Σ_1) or (Θ_2, Σ_2) associated with the respective disjoint datasets D_1 and D_2.

The ML estimators have the optimum statistical properties asymptotically in large samples under regularity conditions of:

- Unbiased parameter estimates referred to as consistent estimates.
- Minimum variance estimates referred to as efficient estimators relative to the Cramer-Rao lower bound.

The ML estimators are used extensively below in both the GLR tests that are also called maximum likelihood ratio tests as well as in the computation of the AIC, which uses the logarithm of the maximized likelihood function.

III. Likelihood Ratio Tests

A traditional approach in statistics for testing nested hypotheses as in Theorem 1 is to use generalized likelihood ratio tests. We compare two models, model M_n under the null hypothesis H_n and model M_a under the alternative hypothesis H_a, identified from the same dataset of length N, but allow for concatenated submodels as in Theorem 1. The models have ν_n and ν_a parameters, respectively. In this section, the null hypothesis model M_n is assumed to be a subset of the alternative hypothesis model M_a. In other words, model M_n is *nested* in model M_a, and $\nu_n \leq \nu_a$. Let λ denote the *generalized likelihood ratio* that is also sometimes called the maximum likelihood ratio:

$$\lambda(\hat{\Theta}_n, \hat{\Theta}_a) \triangleq \frac{L_n(\hat{\Theta}_n)}{L_a(\hat{\Theta}_a)} \tag{12}$$

The maximized likelihood functions $L_n(\hat{\Theta}_n)$ and $L_a(\hat{\Theta}_a)$ are for models M_n and M_a. In particular in the discussion, the null hypothesis H_n and the alternative hypothesis H_a are nested by Theorem 1. To satisfy the regularity conditions for the asymptotic large sample results, we will consider the situation where the sample sizes N_1 and N_2 increase without bound in a fixed rational proportion $r = n_r/m_r$ of the form $N_1/N_2 = r$ where n_r and m_r are integers.

First consider the case that the null hypothesis H_n is true. The *log likelihood ratio statistic,* $-2\log\lambda(\hat{\Theta}_n, \hat{\Theta}_a)$, can be used to test the null hypothesis that the additional parameters in model M_a are not significantly different from zero. Asymptotically for large sample size, the log likelihood ratio statistic for the test of additional parameters in nested models has been shown to follow a χ^2 distribution with $\nu_a - \nu_n$ degrees of freedom [9].

Now consider the case where the null hypothesis H_n is false, *i.e.*, a single model is not valid for both datasets. We still require model M_n to be nested within model M_a, but instead of testing that the additional parameters are zero, we are testing if their estimated values are significant. Wald [10] has shown that the log likelihood ratio statistic follows a noncentral χ^2 distribution if the additional model structure is significant (that is, the null hypothesis H_n is false). Let $\chi^2(\nu_a - \nu_n, \delta^2)$ denote a noncentral χ^2 distribution with $\nu_a - \nu_n$ degrees of freedom and *noncentrality parameter* δ^2. If the noncentrality parameter δ^2 is nonzero, then the probability of rejecting the null hypothesis H_a is increased. This will be illustrated in Section V with a simulation example.

Asymptotically for large samples, GLR tests are uniformly most powerful invariant tests [11]. The invariance property derives from the asymptotic property of ML estimators, that transformation of the data by scaling, rotation, or translation of the data produces a corresponding transformation on the parameter estimates to leave the distributional properties unchanged. As a result for large samples, such transformations on the data leave the GLR statistic invariant so that decisions are not affected by these transformations. This guarantees that among such invariant tests, the GLR test is the optimal single test of all possible changes that may potentially occur in the process. This property guarantees that no other single invariant test has lower probabilities of errors than the GLR test. In the next section the AIC will be discussed. A derivation linking the ΔAIC statistic to GLR testing will be presented.

IV. Akaike Information Criterion

In this section, the concept of the AIC is developed starting with the Kullback–Leibler information. Unless otherwise noted, the asymptotic large sample behavior of the AIC will be primarily discussed.

A natural starting point for the AIC is the use of the K–L information as the natural measure of model approximation. Based on the fundamental statistical principles of sufficiency and repeated sampling, it has been shown that the K–L information gives the natural measure of statistical model approximation [5]–[7]. This result applies to a very general class of problems including finite sample size and arbitrary probability distributions. In many of the papers of Akaike, arguments involving entropy or information were used, although no fundamental justification for the use of information measures was given.

Adoption of the K–L information as the measure of model approximation gives a very clear justification for the AIC. The K–L information [4], [12] between the estimated model $f_{\hat{\Theta},\hat{\Sigma}}(x)$ and the true model $f_*(x)$ is given by:

$$I\left(f_*, f_{\hat{\Theta},\hat{\Sigma}}\right) \triangleq \int f_*(x) \log \frac{f_*(x)}{f_{\hat{\Theta},\hat{\Sigma}}(x)} dx \qquad (13)$$

Asymptotically for large samples, the AIC is an unbiased estimator of K–L information, so:

$$AIC \triangleq \mathcal{E}_{\hat{\Theta},\hat{\Sigma}}\left[I\left(f_*, f_{\hat{\Theta},\hat{\Sigma}}\right)\right] \qquad (14)$$

where $\mathcal{E}_{\hat{\Theta},\hat{\Sigma}}$ is the expectation taken with respect to the random variables $(\hat{\Theta}, \hat{\Sigma})$ of estimated parameters.

The value of the AIC can be shown to be equal to two times the negative log of the maximized likelihood function plus two times the number of estimated parameters, ν [3], [4]. Repeating Eq. 1:

$$AIC = -2 \log L(\hat{\Theta}, \hat{\Sigma}) + 2\nu \qquad (15)$$

The number of estimated parameters is equal to the number pq of elements of Θ plus the number $p(p+1)/2$ of unique elements of the symmetric matrix Σ. Thus $\nu = pq + p(p+1)/2$. The AIC value for a dataset of N independent observations and the regression model of Eq. 5 is:

$$AIC = N\left(\log(2\pi) + 1\right) + N \log|\hat{\Sigma}| + 2pq + p(p+1) \quad (16)$$

Note that for the special case where the covariance matrix has a known diagonal structure, only the diagonal elements of Σ, need be estimated, and $\nu = pq + p$.

To compare two models, let the AIC values for models M_n and M_a be denoted by AIC_n and AIC_a. We will use the ΔAIC statistic to compare these two models. The ΔAIC statistic, defined in Eq. 2 is:

$$\Delta AIC = AIC_n - AIC_a \qquad (17)$$

For a nested hypothesis test, the ΔAIC can be related to the GLR statistic. In the case of a nested model comparison, the theoretical probability distribution of ΔAIC depends on whether the null hypothesis is true or false.

For the case where the null hypothesis H_n is true, from the GLR discussion it can be shown that the ΔAIC statistic, Eq. 2, follows a $\chi^2(\nu_a - \nu_n)$ distribution shifted by a constant, $2(\nu_a - \nu_n)$, for regression models and maximum likelihood estimation [4], [13]. Starting from the result of Wilks [9],

$$-2 \log \lambda(\hat{\Theta}_n, \hat{\Theta}_a) \sim \chi^2(\nu_a - \nu_n) \qquad (18)$$

$$-2 \log \frac{L_n(\hat{\Theta}_n)}{L_a(\hat{\Theta}_a)} \sim \chi^2(\nu_a - \nu_n) \qquad (19)$$

$$-2(\log L_n(\hat{\Theta}_n) - \log L_a(\hat{\Theta}_a)) \sim \chi^2(\nu_a - \nu_n) \qquad (20)$$

To count parameters, the number of outputs, p, is always fixed. In the simplest case of Theorem 1, each model M_n, M_1, or M_2 has $pq + p(p+1)/2$ parameters so that the number of additional parameters in M_a is $\nu_a - \nu_n = pq + p(p+1)/2$. In more general cases with the models still nested, $\nu_a - \nu_n = p(q_1 + q_2 - q_n) + p(p+1)/2$. In any nested case, using (15), (17), and (20) gives:

TABLE I

Significance Level for the Test $\Delta AIC - 2(\nu_a - \nu_n) \geq 0$.

Degrees of freedom, $\nu_a - \nu_n$	1	2	3	4	5	8	11	16	20
Significance level, α	0.157	0.135	0.112	0.092	0.075	0.042	0.024	0.010	0.005

$$\Delta AIC + 2(\nu_a - \nu_n) \sim \chi^2(\nu_a - \nu_n) \tag{21}$$

For the case where the null hypothesis H_n is false, using the result of Wald [10] and taking expected values of Eq. 21 gives,

$$E[\Delta AIC] + 2(\nu_a - \nu_n) = E[\chi^2(\nu_a - \nu_n, \delta^2)] \tag{22}$$

Using the fact that the expected value of a noncentral χ^2 distribution, $\chi^2(\nu, \delta^2)$ is equal to $\nu + \delta^2$,

$$E[\Delta AIC] = -(\nu_a - \nu_n) + E[\delta^2] \tag{23}$$

Or, solving for the expected value of the noncentrality parameter,

$$E[\delta^2] = E[\Delta AIC] + (\nu_a - \nu_n) \tag{24}$$

Thus an unbiased estimator of the noncentrality parameter is [13]:

$$\hat{\delta}^2 = \Delta AIC + (\nu_a - \nu_n) \tag{25}$$

The ΔAIC is a GLR test where the probability of rejection α of the null hypothesis is a function of the number of additional parameters. Because the test statistic $\Delta AIC - 2(\nu_a - \nu_n)$ is the GLR condition for rejecting the null hypothesis, then under large sample theory, Table I shows the probability, α, of rejecting the null hypothesis H_n when it is true. α is also called the significance level of the test.

The reason why the α level adjusts with the number of additional parameters is because the shape of a $\chi^2(\nu)$ distribution changes when the number of degrees of freedom, ν, increases. The automatic adjustment of the α level with the number of additional parameters deals with one of the major issues in using GLR tests with few or many additional parameters: the need that it take into account the number of additional parameters being estimated. Of course, by choosing a criterion different from $\Delta AIC - 2(\nu_a - \nu_n) \geq 0$, the α values in Table I can be changed.

Finally, we discuss a small sample version of the AIC derived by Hurvich and Tsai [14]. The corrected AIC value, AIC_c, is of particular use when the sample size is small relative to the number of estimated parameters. AIC_c is asymptotically equivalent to AIC for large samples, and provides an asymptotically unbiased estimator of K–L information. The small sample bias correction for AIC using N data points is:

$$AIC_c = -2\log L(\hat{\Theta}, \hat{\Sigma}) + 2\nu \left(\frac{N}{N - \nu - (p+1)/2} \right) \tag{26}$$

The AIC_c has a small sample correction factor multiplying the 2ν penalty term that appears in the AIC. This factor approaches one for large samples. Asymptotically, $\Delta AIC_c \rightarrow \Delta AIC$. The distribution of the likelihood ratio (18) is much more complicated in the small sample case, but the bias of the AIC is corrected as in the AIC_c.

In the above discussion, the nested case has been discussed for comparison with the GLR test. Further, the AIC applies to the comparison of a multitude of hypotheses, not just two as in the GLR test. In complex processes such as dynamic systems, there are a multitude of models and hypotheses because typically the state order or ARX order is unknown and must be estimated from the data.

V. Simulation Example

A simple simulation example is used to confirm the theoretical result for the distribution of the ΔAIC statistic. The following ARX model was simulated by specifying the input, u to be a zero–mean gaussian process with unit variance:

$$\begin{aligned} y(t) = 0.2y(t-1) + 0.1y(t-2) - 0.7u(t) + 3u(t-1) \\ + 1.2u(t-2) - 0.15u(t-3) + e(t) \end{aligned} \tag{27}$$

The unmeasured noise, $e(t)$ was a zero–mean gaussian process with a variance of 0.1. One thousand sets of two 500 point data series were generated ($N_1 = N_2 = 500, N = 1000$), and the ΔAIC statistic was calculated as described in Section IV, assuming the correct model order is known. A histogram of the calculated ΔAIC values and the theoretical $\chi^2(8, 0)$ distribution are shown in Figure 1. Next, the simulation was repeated, making a small change of ± 0.01 to each model parameter in Eq. 27 for the second dataset. Using the same method to calculate the ΔAIC statistic, a new histogram was produced and is shown in Figure 2. It is clear that the ΔAIC statistic does not follow the theoretical $\chi^2(8, 0)$ distribution when a small process change is present for the second dataset. When a process change occurs, the null hypothesis, H_n is false and the ΔAIC statistic follows a $\chi^2(8, \delta^2)$ distribution, where an estimate of the noncentrality parameter is given in Eq. 25, $\hat{\delta}^2 = \Delta AIC + (\nu_a - \nu_n)$.

VI. Summary

The theoretical probability distribution of ΔAIC has been derived for the nested case based on mild assumptions. It was shown that under suitable regularity conditions on the estimated parameters, and assuming independence of the two datasets, ΔAIC follows a χ^2 distribution, shifted by $2(\nu_a - \nu_n)$, and with $\nu_a - \nu_n$ degrees of freedom. The ΔAIC can also be calculated for the case of two noncontiguous datasets. In such a case, the comparison is made between a model obtained from the entire dataset and

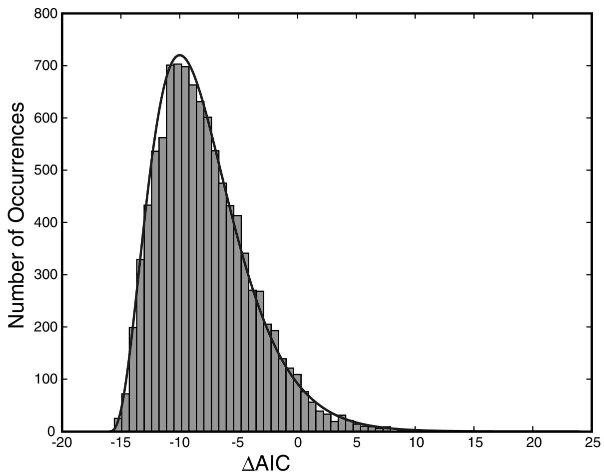

Fig. 1. Histogram of Calculated ΔAIC Values and Theoretical $\chi^2(8)$ Distribution When no Process Change is Present.

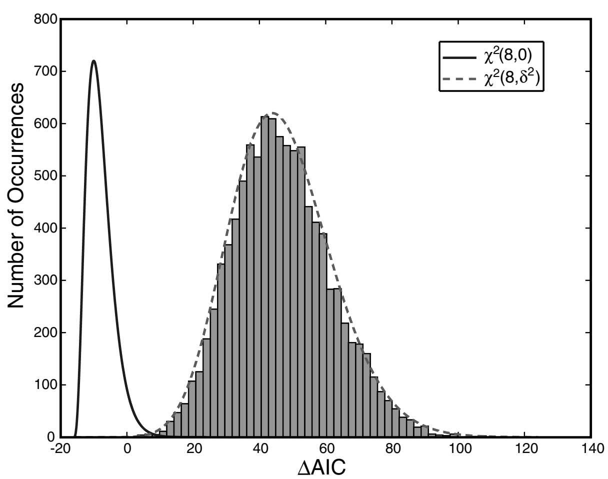

Fig. 2. Histogram of Calculated ΔAIC Values and Theoretical $\chi^2(8)$ and $\chi^2(8, \delta^2)$ Distributions When a Process Change is Present.

from a model obtained by combining likelihood functions for the two independent models for each subset of data. The properties of ΔAIC also are valid for the case of time–series data, where the regressors consist of past process inputs and outputs. A simple numerical simulation example was used to show the distribution of the ΔAIC statistic for an ARX model. The simulation results agree closely with the theoretical distributions.

REFERENCES

[1] W. E. Larimore, "Optimal reduced rank modeling, prediction, monitoring, and control using canonical variate analysis," in *Proc. IFAC ADCHEM 97 Sympos.*, Banff, Canada, June 1997, pp. 61–66.

[2] Y. Wang, D. E. Seborg, and W. E. Larimore, "Process monitoring using canonical variate analysis and principal component analysis," in *Proc. IFAC ADCHEM 97 Sympos.*, Banff, Canada, June 1997, pp. 523–528.

[3] Y. Sakamoto, M. Ishiguro, and G. Kitagawa, *Akaike Information Criterion Statistics*. Tokyo: KTK Scientific Publishers, 1986.

[4] K. P. Burnham and D. R. Anderson, *Model Selection and Multimodel Inference, A Practical Information–Theoretic Approach*, 2^{nd} *ed.* New York: Springer–Verlag, 2002.

[5] W. E. Larimore, "Predictive inference, sufficiency, entropy, and an asymptotic likelihood principle," *Biometrika*, vol. 70, pp. 175–181, 1983.

[6] ——, "Statistical optimality and canonical variate analysis system identification," *Signal Processing*, vol. 52, pp. 131–144, 1996.

[7] W. E. Larimore and R. K. Mehra, "The problem of overfitting data," *Byte*, vol. 10, pp. 167–180, 1985.

[8] T. W. Anderson, *An Introduction to Multivariate Statistical Analysis*. New York: Wiley, 1984.

[9] S. S. Wilks, "The large–sample distribution of the likelihood ratio for testing composite hypotheses," *Annals of Math. Statistics*, vol. 9, no. 1, pp. 60–62, March 1938.

[10] A. Wald, "Tests of statistical hypotheses concerning several parameters when the number of observations is large," *Trans. Amer. Math. Soc.*, vol. 54, no. 3, pp. 426–482, 1943.

[11] D. R. Cox and D. V. Hinkley, *Theoretical Statistics*. New York: Chapman and Hall, 1974.

[12] S. Kullback, *Information Theory and Statistics*. New York: Wiley, 1959.

[13] W. E. Larimore, "Accuracy confidence bands including the bias of model under-fitting," in *Proc. American Control Conf.*, San Francisco, CA, 1993, pp. 1995–1999.

[14] C. M. Hurvich and C.-L. Tsai, "Regression and time series model selection in small samples," *Biometrika*, vol. 76, no. 2, pp. 297–307, June 1989.