Information Criteria Applied to State-Space Model Selection with Large Observation Error

Proposal for Project

School of Mathematics and Statistics
Victoria University of Wellington

Abstract

Commercial and recreational fisheries data is cheap and abundant, however, it is not the most reliable. We can model this type of data using State-Space models, which consist of two components - a measurement process, and a transition process. Commercial fisheries data is associated with inflated observation error due to factors such as recording errors, and fishing "hotspots". Hence model selection criteria such as Akaike's Information criterion (AIC) consistently fail to accurately select the correct model. AIC is composed of a biased estimator for the Kullback-Leibler distance (a measure of the model fit), and a bias correction term. AIC improved (AICi) and AIC bootstrap (AICb) are adjusted versions of AIC which approximate the bias term by Monte Carlo simulation, and bootstrap respectively. These adjusted information criteria show promising results in the literature when applied to State-Space model selection problems. Starting with the simple univariate case, their usefulness as model selection criteria for choosing between State-Space candidate models will be explored in a large observation variance setting.

Contents

1 Introduction

There are two main types of data available to fisheries scientists. The first type of data comes from scientific surveys of fish populations. Scientific surveys are meticulously planned and well designed, which leads to observations that have minimal bias. However, they are expensive to implement. The second type of data comes from commercial and recreational fishing vessels. This type of data is cheap and in abundance, however, is heavily biased. We often work with commercial and recreational fishing data, which imposes problems due to the biased observations. The large observation variances come from errors in recording observations, errors in equipment, and "hotspot" targeting. Fisheries tend to focus on these "hotspots", which are locations with an abundance of fish, meaning that observations recorded do not represent the entire population. These factors all contribute to the problem of inaccurate model selection. Model selection criteria, such as Akaike's Information Criteria (AIC) (Akaike, 1973) cannot be used to select a particular candidate model. This proposal looks at applying adjusted information criteria to state-space candidate models.

A state-space model is able to describe the behavior of many dynamic systems. They have gained popularity in recent times, as they have a general form which is intuitive in many applications. A state-space model can be used to relate an observable quantity to an unobservable latent variable of interest, through time.

The model is shown below:

$$\boldsymbol{y}_t = A\boldsymbol{x}_t + \boldsymbol{v}_t, \tag{1}$$

$$\boldsymbol{x}_t = \Phi\boldsymbol{x}_{t-1} + \boldsymbol{w}_t, \tag{2}$$

for $t = 1, ..., n$ time periods (Bengtsson & Cavanaugh (2006), Cavanaugh & Shumway (1997)).

The two equations are called the observation (measurement) equation and the state (transition) equation respectively, where $\boldsymbol{y}_t$ is an observed vector process, and $\boldsymbol{x}_t$ is an unobserved vector state process. In addition, A is a known design matrix and Φ is an unknown transition matrix. $\boldsymbol{v}_t$ and $\boldsymbol{w}_t$ are both zero-mean vector noise processes. R and Q are used to denote the covariance matrices of the observation noise process $\boldsymbol{v}_t$, and the state noise process $\boldsymbol{w}_t$ respectively. Finally, μ and Σ are used to denote the mean and covariance matrix of the initial state x_0. It is also assumed that x_0, $\boldsymbol{v}_t$ and $\boldsymbol{w}_t$ are mutually independent and multivariate normal (Cavanaugh & Shumway, 1997).

In the context of fisheries science, we might be interested in the latent variable, population abundance. We can use a state-space equation to model population abundance, through the observable catch frequency (a ratio of catch to fishing effort). We let $\boldsymbol{y}_t$ be catch frequency, and $\boldsymbol{x}_t$ be the unobservable population abundance.

2 Theoretical Framework

In 1973, Hirotugu Akaike wrote a paper on the topic of information theory, in which he formulated an information criterion, widely used today for model selection (Akaike, 1973). Akaike's Information Criterion (AIC) is an approximately unbiased estimator of the Kullback-Leibler measure of divergence. The Kullback-Leibler divergence is the distance between two distributions, and hence can be thought of as a measure of model fit.

First, let's suppose $\boldsymbol{y} = (y_1, y_2, ..., y_n)^T$ are observed values of $\boldsymbol{Y} = (Y_1, Y_2, ..., Y_n)^T$ where the Y_i's are independent and identically distributed with unknown true density function $g(., \boldsymbol{\phi_0})$. $\boldsymbol{\phi_0} = (\phi_{01}, ..., \phi_{0p})^T$ are the unknown true values of the p-dimensional parameter vector for distribution g. Now we consider a candidate model $f_{\boldsymbol{\phi}}(.)$ and we let $\hat{\boldsymbol{\phi}}$ be the maximum likelihood estimator of $\boldsymbol{\phi_0}$ in some set Φ. In other words:

$$l(\hat{\boldsymbol{\phi}}; \boldsymbol{y}) = \sum_{i=1}^{n} log(f_{\hat{\boldsymbol{\phi}}}(y_i)) = \max_{\boldsymbol{\phi} \in \Phi} l(\boldsymbol{\phi}; \boldsymbol{y}), \tag{3}$$

where l is the log-likelihood for $\boldsymbol{Y}$.

Using the aforementioned Kullback-Leibler measure, we can measure the divergence between a candidate model, and the true density.

$$d(\boldsymbol{\phi_0}, \hat{\boldsymbol{\phi}}) = \int g(y, \boldsymbol{\phi_0}) \log\left(\frac{g(y, \boldsymbol{\phi_0})}{f_{\hat{\boldsymbol{\phi}}}(y)}\right) dy. \tag{4}$$

Using properties of expectation, observe that this can be written in the form

$$d(\boldsymbol{\phi_0}, \hat{\boldsymbol{\phi}}) = E_g[\log(g(Y, \boldsymbol{\phi_0}))] - E_g[\log(f_{\hat{\boldsymbol{\phi}}}(Y))]. \tag{5}$$

In this statement, we see that the first term is independent of the candidate model f. We can therefore measure divergence using the second term only, which is known as the expected log-likelihood. An estimate of this expected log-likelihood was proposed by Akaike to be

$$\delta(\boldsymbol{\phi_0}, \hat{\boldsymbol{\phi}}) = -2\log(f_{\hat{\boldsymbol{\phi}}}(Y)). \tag{6}$$

However, this estimator is biased, and so a bias correction term which depends on model order, and the true parameter set ϕ_0 is added:

$$B_T(p, \boldsymbol{\phi_0}) = E[\delta(\boldsymbol{\phi_0}, \boldsymbol{\phi}) - \{-2\log(f_{\hat{\boldsymbol{\phi}}}(Y))\}], \tag{7}$$

which can be approximated by $2p$, where p is the number of parameters in the candidate model. The result is an almost unbiased estimator given by

$$\frac{-2l(\hat{\boldsymbol{\phi}}; \boldsymbol{y}) + 2p}{n}, \tag{8}$$

where l is the log-likelihood function of the candidate model, p is the number of parameters in the candidate model, and n is the sample size. AIC is therefore defined as

$$AIC = -2l(\hat{\boldsymbol{\phi}}; \boldsymbol{y}) + 2p. \tag{9}$$

AIC tends to be unreliable when the candidate models are of high dimension. AIC will favor more complex models which have a large number of parameters (relative to the sample size). In order to rectify this fault, a corrected version, AICc was formulated (Sugiura, 1978) which adds a correction term to AIC:

$$AICc = AIC + \frac{2p(p+1)}{n-p-1}. \tag{10}$$

This now gives us an asymptotically unbiased estimator of $\delta(\phi_0, \phi)$. AICc is consistently shown to outperform AIC as shown in previous simulation experiments (Bengtsson & Cavanaugh, 2006). Six sets of candidate models are compiled based on state-space models. Two of the models have state processes which are univariate Gaussian autoregressive of order p, and the remaining four have state processes based on a combination of univariate Gaussian autoregressive and seasonal Gaussian processes. In all simulation sets, the correct true model is chosen by the AICc considerably more times than the AIC. However, there are still many cases where the AICc chose the incorrect model. This motivates us to explore other information criteria which extend to the state-space setting better than the AICc.

The following sections will review two information criteria which have been explored in a state-space setting. Each criterion builds on the pivotal work of Akaike.

2.1 AICi

It is noted that the bias correction term (7) does not depend on ϕ_0 asymptotically (Bengtsson & Cavanaugh, 2006). Because of this, it is possible to accurately estimate the bias correction term through a Monte Carlo simulation (as shown in Bengtsson et al.). AICi takes the following form:

$$AICi = -2\log f_{\hat{\phi}}(Y) + \frac{1}{M}\sum_{j=1}^{M}\left[d(\phi_s, \hat{\phi_j}) - \{-2\log f_{\hat{\phi}_j}(Y)\}\right], \tag{11}$$

where d is defined as (5), ϕ_s is some arbitrary, but convenient choice of parameter in place of ϕ_0 and $\hat{\phi_1}, ..., \hat{\phi_M}$ are estimates of the true parameter ϕ_0, of a candidate model for iterations $1, ..., M$.

In order to compute AICi, the bias correction term needs only to be computed once for each model order. To compute the bias approximation, a candidate model is initially fit to the observations from a model with parameter ϕ_s. From this fitted model, a set of M observations is generated, to which M candidate models of the same structure are fitted. The term is then calculated using the formula above.

This information criterion was compared to a range of model selection criterion in the simulation study described in Bengtsson & Cavanaugh (2006). AICi was compared to Schwarz information criterion (SIC), final prediction error (FPE), Hannan-Quin (HQ), and Bayesian information criterion (BIC) as well as AIC and AICc. It was shown to outperform all other criterion, as it selected the correct model the most times in all simulation sets.

2.2 AICb

Similar to Bengtsson et al., Cavanaugh & Shumway (1997) proposed a bootstrap based estimator for the bias correction term (7). AICb was shown to have the following form:

$$AICb = -2\log f_{\hat{\phi}}(Y) + 2\left\{\frac{1}{N}\sum_{i=1}^{N} -2\log\frac{f_{\hat{\phi}*_i}(Y)}{f_{\hat{\phi}}(Y)}\right\} \tag{12}$$

In order to calculate the correction term, N bootstrap replicates, $\hat{\phi}_1^*, ..., \hat{\phi}_N^*$, of $\hat{\phi}$ from the candidate model are computed via non-parametric approach.

A simulation experiment was performed, using two different models. The first model used was a univariate Gaussian autoregressive process of order p, and the second was an autoregressive state-space model (again, with order p). AICb was found to perform the best out of weighted-average information criterion (WIC), AIC, AICc, FPE, HQ, BIC and SIC, however, it only performs slightly better than AICc. It is advantageous in the state-space setting, vastly out-performing AIC in three ways. It provides a considerably less biased estimate of $d(\phi_0, \hat{\phi})$, than AIC, has a higher success rate when data is generated from a known finite dimensional model, and it does not overfit, as exhibited by AIC (and many other criterion).

3 Methodology

For this project, we will focus on fitting the simple univariate state-space models of the following form:

$$y_t = m_t x_t + v_t, \tag{13}$$

$$x_t = \phi_1 x_{t-1} + \phi_2 x_{t-2} + ... + \phi_p x_{t-p} + w_t, \tag{14}$$

for $t = 1, ..., n$ time periods. Here, $\phi_1, ..., \phi_p$ are the transition coefficients, and we let σ_R^2 and σ_Q^2 be the observation variance and the transition variance respectively.

Fitting state-space models not only involves the estimation of the unknown parameters $\phi_1, ..., \phi_p, \sigma_R^2$ and σ_Q^2, but also the unobserved latent variables of interest x_t. We can estimate x_t via the Kalman filter equations, while the rest of the parameters can be estimated via maximum likelihood techniques.

3.1 Kalman Filter

Prediction of x_t from the observed data in state-space modeling is of concern. Because we are looking at modeling in a frequentist sense, we rely on a recursive process called the Kalman filter process (Shumway & Stoffer, 1982). This involves calculating conditional expectations $E(x_t|Y^{t-1})$, and $E(x_t|Y^t)$, where Y^t is the observed data up until time t. These quantities are regarded as the best linear estimates for x_t. The calculation of these quantities involves the process of initializing, prediction and correction. The prediction and correction steps are repeated for each time step. At this point, applying the Kalman filter will need more time than has been alloted, and so for the purpose of this project, we will let the conditional expectations be the values used to generate the observed y_t's in the simulation.

3.2 Derivation of the Maximum Likelihood Estimators

At this point, we also need to estimate the parameters of the candidate models, namely, $\phi_1, ..., \phi_p, \sigma_R^2$ and σ_Q^2. We first note that the observations, y_t, in equation (17) are distributed as follows:

$$y_t \overset{iid}{\sim} N(m_t x_t, \sigma_R^2), \quad t = 1, ..., T$$

The joint density for the measurement process is therefore defined as the product of densities for each individual y_t:

$$f(y; x, \sigma_R^2) = \prod_{t=1}^{T} f(y_t; x, \sigma_R^2) = \prod_{t=1}^{T} \left[(2\pi\sigma_R^2)^{-\frac{1}{2}} \exp\left\{ -\frac{(y_t - m_t x_t)^2}{2\sigma_R^2} \right\} \right] \tag{15}$$

Similarly, the unobserved latent variables, x_t, in equation (18) are distributed as follows:

$$\begin{aligned} x_t &\overset{iid}{\sim} N(\phi_1 x_{t-1} + ... + \phi_p x_{t-p}, \sigma_Q^2), \quad t = p, ..., T, \\ x_{p-1} &\overset{iid}{\sim} N(\phi_1 x_{p-2} + ... + \phi_{p-1} x_0, \sigma_Q^2), \\ &\vdots \\ x_1 &\overset{iid}{\sim} N(\phi_1 x_0, \sigma_Q^2), \\ x_0 &\overset{iid}{\sim} N(\mu_0, \sigma_0^2), \end{aligned}$$

Therefore, the joint density for the state process can be defined as:

$$\begin{aligned} f(x; \phi^p, \sigma_Q^2, \mu_0, \sigma_0^2) &= \prod_{t=p}^{T} f(x_t; \phi^p, \sigma_Q^2) f(x_{p-1}; \phi^{p-1}, \sigma_Q^2) ... f(x_1; \phi_1, \sigma_Q^2) f(x_0; \mu_0, \sigma_0^2), \\ &= T(2\pi\sigma^2 Q)^{-\frac{1}{2}} (2\pi\sigma_0^2)^{-\frac{1}{2}} \exp\left\{ -\frac{1}{2\sigma_Q^2} \left[\sum_{t=p}^{T} (x_t - (\phi_1 x_{t-1} + ... + \phi_p x_{t-p}))^2 + \right.\right. \\ &\left.\left. (x_t - (\phi_1 x_{p-2} + ... + \phi_{p-1} x_0))^2 + ... + (\phi_1 x_0)^2 \right] \right\} \exp\left\{ -\frac{1}{2\sigma_0^2} (x_0 - \mu_0)^2 \right\}, \end{aligned} \tag{16}$$

where $\phi^p = \{\phi_1, ..., \phi_p\}$.

Therefore, the overall likelihood, $L(\Phi)$, is expressed as:

$$\begin{aligned} L(\Phi) = T(2\pi\sigma_R^2)^{-\frac{1}{2}} \exp\left\{ \sum_{t=1}^{T} -\frac{(y_t - m_t x_t)^2}{2\sigma_R^2} \right\} T(2\pi\sigma^2 Q)^{-\frac{1}{2}} (2\pi\sigma_0^2)^{-\frac{1}{2}} \\ \exp\left\{ -\frac{1}{2\sigma_Q^2} \left[\sum_{t=p}^{T} (x_t - (\phi_1 x_{t-1} + ... + \phi_p x_{t-p}))^2 + \right.\right. \\ \left.\left. (x_t - (\phi_1 x_{p-2} + ... + \phi_{p-1} x_0))^2 + ... + (\phi_1 x_0)^2 \right] \right\} \exp\left\{ -\frac{1}{2\sigma_0^2} (x_0 - \mu_0)^2 \right\} \end{aligned} \tag{17}$$

By taking the natural log of equation (17), and maximizing with respect to each parameter $\phi_1, \dots \phi_p, \sigma_Q^2, \sigma_R^2$, we arrive at a set of maximum likelihood estimators.

The maximum likelihood estimator for the observation variance, σ_R^2, is shown below.

$$\hat{\sigma}_R^2 = \frac{1}{T}\sum_{t=1}^{T}(y_t - m_t x_t)^2 \tag{18}$$

The maximum likelihood estimator for the transition parameters, ϕ_j, is shown below.

$$\begin{aligned}\hat{\phi}_j = \frac{1}{\sum_{t=j}^{T} x_{t-j}^2}\Bigg[\sum_{t=j}^{T} x_t x_{t-j} - \phi_1 \sum_{t=j}^{T} x_{t-1}x_{t-j} - \dots - \phi_{j-1}\sum_{t=j}^{T} x_{t-(j-1)}x_{t-j} \\ - \phi_{j+1}\sum_{t=j+1}^{T} x_{t-(j+1)}x_{t-j} - \dots - \phi_p \sum_{t=p}^{T} x_{t-p}x_{t-j}\Bigg]\end{aligned} \tag{19}$$

The maximum likelihood estimator for the state variance, σ_Q^2, is shown below.

$$\hat{\sigma}_Q^2 = \frac{1}{T}\left[\sum_{t=p}^{T}(x_t - (\phi_1 x_{t-1} + \dots + \phi_p x_{t-p}))^2 + (x_t - (\phi_1 x_{p-2} + \dots + \phi_{p-1} x_0))^2 + \dots + (\phi_1 x_0)^2\right] \tag{20}$$

By programming these estimators in **R** we can easily fit the candidate state-space models, and hence compute their log-likelihood values.

3.3 Simulation

To test the characteristics of AICi and AICb, a simulation study will be carried out. The process to be followed is shown below:

1. Generate values for x_t, using state equation (15), and suitable values for $\phi_1, \dots, \phi_p$, σ_Q^2, μ_0, and σ_0^2.
2. Generate values for y_t, using measurement equation (14), and suitable values for m_t, and σ_R^2.
3. Estimate x_t from the observed y_t values using the Kalman Filter Equations.
4. Fit candidate state-space models, with state equations of order one through ten.
5. Compute AIC, AICc, AICi, and AICb for each candidate model.
6. Repeat steps 1 to 5 many times.

The number of times that each candidate model is selected by an information criterion will be recorded in a table. From this, a bar chart can be constructed to visually show the behavior of each information criterion.

What follows is a made up example where a simulation was conducted to investigate the behavior of AIC and AICc in a simple linear regression case. Ten models differing by number of predictors were fit to data generated from a three predictor model. The table and plots that follow give an example of the visual summaries to be included in the project.

Candidate Model	AIC	AICc
1	0	0
2	28	91
3	492	797
4	119	79
5	89	24
6	67	2
7	95	6
8	110	1

Table 1: Model selection frequencies using AIC and AICc for observations with small variation.

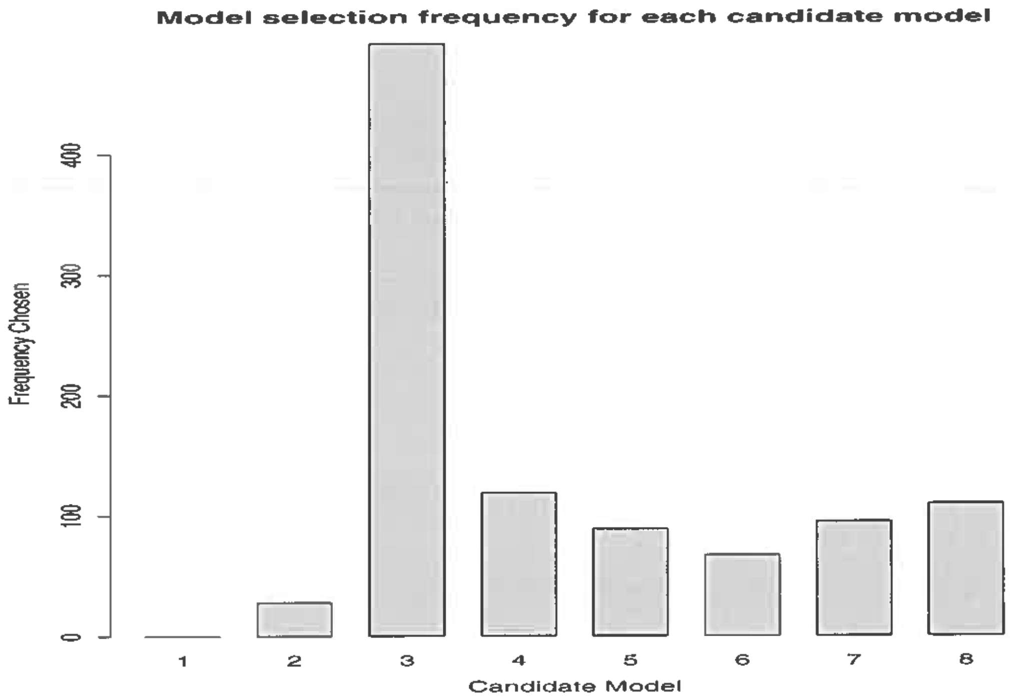

Figure 1: Model selection frequency using AIC when observations have small variation.

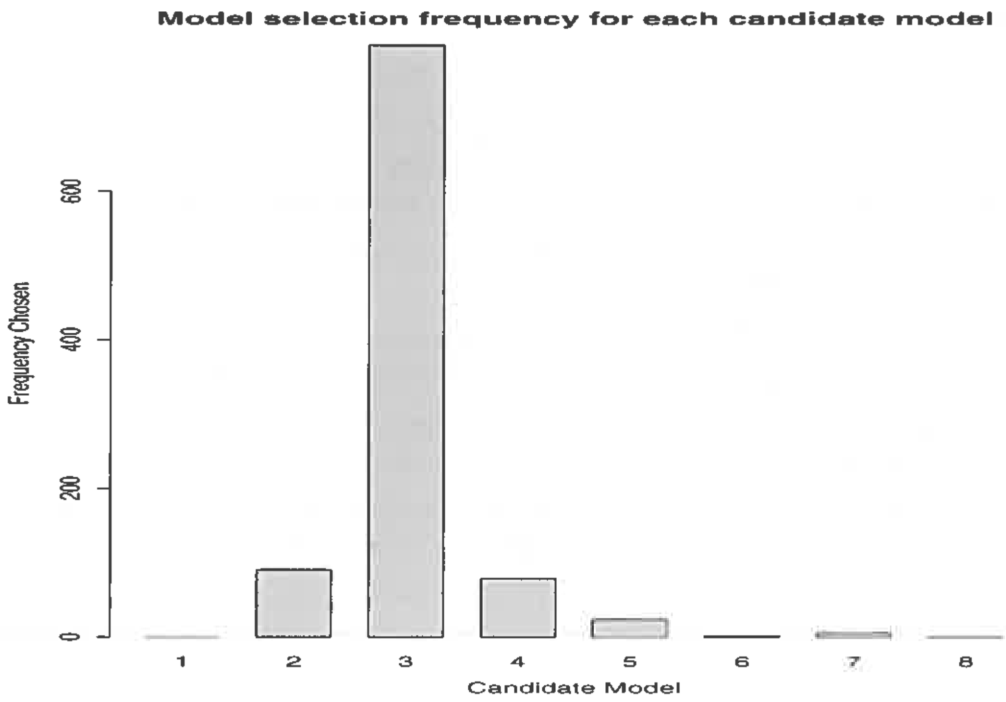

Figure 2: Model selection frequency using AICc when observations have small variation.

4 Research Goals

Thus far, this project proposal has made reference to relevant literature about state-space models, and relatively new information criteria. We have derived maximum likelihood estimators which can be used to fit univariate state-space models. We have also seen how to apply our two information criteria to the fitted candidate models, using Monte Carlo simulations, and bootstrap methods. In the next few days, we hope to accomplish the following research objectives:

1. Fit several univariate state-space candidate models, using the Kalman filter equations, and maximum likelihood techniques
2. Compute AIC, AICc, AICi, and AICb

The overall objective is to explore the model selection characteristics of the two adjusted information criteria in a large variance situation. Characteristics that will be investigated include how often the "correct" model is selected, and whether there is a tendency to select models of a particular order. In order to do this, the observation variance in the state-space setting will be increased, to investigate the performance of AICi, and AICb (compared to the popular AIC, and AICc).

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. *Second International Symposium on Information Theory*, , 267–281.

Bengtsson, T. & Cavanaugh, J. E. (2006). An improved akaike information criterion for state-space model selection. *Computational Statistics & Data Analysis*, **50**(10), 2635–2654.

Cavanaugh, J. E. & Shumway, R. H. (1997). A bootstrap variant of aic for state-space model selection. *Statistica Sinica*, **7**(2), 473–496.

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.

Shumway, R. H. & Stoffer, D. S. (1982). An approach to time series smoothing and forecasting using the em algorithm. *Journal of time series analysis*, **3**(4), 253–264.

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