AIC under the Framework of Least Squares Estimation

H.T. Banks and Michele L. Joyner

Center for Research in Scientific Computation
North Carolina State University
Raleigh, NC, United States
and
Dept of Mathematics and Statistics
East Tennessee State University
Johnson City, TN 37614

May 4, 2017

Abstract

In this note we explain the use of the Akiake Information Criterion and its related model comparison indices (usually derived for maximum likelihood estimator inverse problem formulations) for use with least squares (ordinary, weighted, iterative weighted or "generalized", etc.) based inverse problem formulations. The ideas are illustrated with several examples of interest in biology.

1 Introduction and Overview of AIC

The Akaike Information Criterion (AIC) is one of the most widely used methods for choosing a "best approximating" model from several competing models given a particular data set [13, 15]. It was first developed by Akaike in 1973 [2] and expanded upon in several following papers [3, 4, 5, 6, 7, 8]. The basis of the Akaike Information Criterion relies on several assumptions. It is assumed that the given data or set of observations is a realization of a random variable which has some unknown probability distribution; however, one can draw inferences about the "true" distribution using the distribution of the data. Using this assumption, the best approximating model would be the model in which the "distance" between the estimated distribution and "true" distribution is as small as possible. Kullback-Leibler (K-L) information is a well-known measure of the "distance" between two probability distribution models. Suppose $\mathbf{Y}$ is a random variable characterized by a probability density function $p(\mathbf{y}|\boldsymbol{\theta})$ where $\boldsymbol{\theta} = (\theta_1, \theta_2, ..., \theta_k)$ is a k-dimensional parameter vector, $\boldsymbol{\theta} \in \mathbb{R}^k$, for the distribution. We assume there exists a true parameter $\boldsymbol{\theta}_0$ such that $p_0 = p(\cdot|\boldsymbol{\theta}_0)$ is the true probability density function of observations $\mathbf{Y}$. Then the K-L information between the estimated model and "true" model is given by

$$\begin{aligned} \mathcal{I}(p_0, p(\cdot, \boldsymbol{\theta})) &= \int_{\Omega_y} p_0(\mathbf{y}) \ln\left(\frac{p_0(\mathbf{y})}{p(\mathbf{y}|\boldsymbol{\theta})}\right) d\mathbf{y} \\ &= \int_{\Omega_y} p_0(\mathbf{y}) \ln\left(p_0(\mathbf{y})\right) d\mathbf{y} - \int_{\Omega_y} p_0(\mathbf{y}) \ln\left(p(\mathbf{y}|\boldsymbol{\theta})\right) d\mathbf{y} \end{aligned} \tag{1}$$

where Ω_y is the set of all possible values for $\mathbf{y}$. We know that $\mathcal{I}(p_0, p(\cdot, \boldsymbol{\theta})) = 0$ if and only if $p_0 = p(\cdot|\boldsymbol{\theta})$; therefore, a good approximation model is one in which K-L information is small. However, the K-L information quantity cannot be calculated directly as the true model p_0 is generally unknown.

Yet, the maximum likelihood estimate $\boldsymbol{\theta}_{MLE}(\mathbf{Y})$ is shown to be a natural estimator for $\boldsymbol{\theta}_0$ [5, 10, 13]. In the misspecified case (i.e., when there does not exist a "true" value $\boldsymbol{\theta}_0$ for $\boldsymbol{\theta}$ such that $p(\cdot|\boldsymbol{\theta}) \equiv p_0$), the asymptotic normality property of the maximum likelihood estimator gives that $\boldsymbol{\theta}_{MLE}(\mathbf{Y})$ is normally distributed with

$$\mathbb{E}(\boldsymbol{\theta}_{MLE}(\mathbf{Y})) = \arg\min_{\boldsymbol{\theta}\in\Omega_\theta} \mathcal{I}(p_0, p(\cdot|\boldsymbol{\theta})).$$

Furthermore, $\mathbb{E}_{\mathbf{Y}}(\mathcal{I}(p_0, p(\cdot|\boldsymbol{\theta}_{MLE}(\mathbf{Y}))) > \mathcal{I}(p_0, p(\cdot|\theta))$ [15]; therefore, $\mathbb{E}_{\mathbf{Y}}(\mathcal{I}(p_0, p(\cdot|\boldsymbol{\theta}_{MLE}(\mathbf{Y})))$ can be used to estimate the "distance" between p and p_0. Thus the best approximating model would be the one that solves

$$\min_{p\in\mathbf{P}} \mathbb{E}_{\mathbf{Y}}(\mathcal{I}(p_0, p(\cdot|\boldsymbol{\theta}_{MLE}(\mathbf{Y})))$$

where $\mathbb{P}$ is a set of candidate models. Following the derivation in [15], we can write

$$\begin{aligned}\mathbb{E}_{\mathbf{Y}}(\mathcal{I}(p_0, p(\cdot|\boldsymbol{\theta}_{MLE}(\mathbf{Y}))) &= \int_{\Omega_y} p_0(\mathbf{x}) \ln(p_0(\mathbf{x}))\, d\mathbf{x} - \mathbb{E}_{\mathbf{Y}}\left(\int_{\Omega_y} p_0(\mathbf{x}) \ln(p(\mathbf{x}|\boldsymbol{\theta}_{MLE}(\mathbf{y})))\, d\mathbf{x}\right) \\ &= \int_{\Omega_y} p_0(\mathbf{x}) \ln(p_0(\mathbf{x}))\, d\mathbf{x} - \mathbb{E}_{\mathbf{Y}}\mathbb{E}_{\mathbf{X}}\left(\ln(p(\mathbf{X}|\boldsymbol{\theta}_{MLE}(\mathbf{Y})))\right).\end{aligned}$$

Therefore,

$$\min_{p\in\mathbf{P}} \mathbb{E}_{\mathbf{Y}}(\mathcal{I}(p_0, p(\cdot|\boldsymbol{\theta}_{MLE}(\mathbf{Y})))) = \max_{p\in\mathbf{P}} \mathbb{E}_{\mathbf{Y}}\mathbb{E}_{\mathbf{X}}\left(\ln(p(\mathbf{X}|\boldsymbol{\theta}_{MLE}(\mathbf{Y})))\right).$$

Furthermore, for a large sample and "good" model, it can be shown (see [15] for details) that

$$\max_{p\in\mathbf{P}} \mathbb{E}_{\mathbf{Y}}\mathbb{E}_{\mathbf{X}}\left(\ln(p(\mathbf{X}|\boldsymbol{\theta}_{MLE}(\mathbf{Y})))\right) \approx \ln\left(\mathcal{L}(\hat{\boldsymbol{\theta}}_{MLE}|\mathbf{y})\right) - \kappa_\theta$$

where $\mathcal{L}(\hat{\boldsymbol{\theta}}_{MLE}|\mathbf{y}) = p(\mathbf{y}|\hat{\boldsymbol{\theta}}_{MLE})$ represents the likelihood of $\hat{\boldsymbol{\theta}}_{MLE}$ given sample outcomes $\mathbf{y}$ and κ_θ is the total number of estimated parameters. For historical reasons, Akaike multiplied by -2 yielding the well-known Akaike information criterion (AIC):

$$AIC = -2\ln\left(\mathcal{L}(\hat{\boldsymbol{\theta}}_{MLE}|\mathbf{y})\right) + 2\kappa_\theta. \tag{2}$$

Note that the complexity of the model, as given by the total number of parameters in the model, is considered in the AIC. Given the same level of accuracy, the simpler model is preferable to the more complex one.

In this paper, we focus on models which are n-dimensional vector dynamical systems or mathematical models of the form

$$\begin{aligned}\frac{d\mathbf{x}}{dt}(t) &= \mathbf{g}(t, \mathbf{x}(t), \mathbf{q}), \\ \mathbf{x}(t_0) &= \mathbf{x}_0\end{aligned}$$

with observation process

$$\mathbf{f}(t, \mathbf{q}) = \mathcal{C}\mathbf{x}(t; \mathbf{q})$$

where $\mathcal{C}$ is the observation operator which maps $\mathbb{R}^n$ to $\mathbb{R}^m$. Some $\overline{\mathbf{x}}_0$ of the initial parameters $\mathbf{x}_0$ may be among the parameters $\boldsymbol{\theta}$ to be estimated, i.e., $\boldsymbol{\theta} = (\mathbf{q}, \overline{\mathbf{x}}_0, \sigma)$, from the data. All of the discussions below are readily extended to this case but for ease in discussion and exposition we will, without loss of generality, assume for our discussions here that the initial conditions are known.

There are a variety of techniques available for parameter estimation in models of this type, one of which is the maximum likelihood method used in the original AIC formulation; however, another popular choice

is parameter estimation in the framework of a least squares estimation problem ([10] and all the references herein). As such, our goal in this paper is to provide a concise formulation for the AIC using the least squares estimator. Depending on the statistical model for the problem, the appropriate estimation technique varies; therefore, the observations $\mathbf{Y}_j$ will be assumed to satisfy various statistical models. In each of the Sections 2, 3, and 4, we assume a statistical model with absolute error, constant weighted error, and parameter dependent weighted error, respectively. We then formulate AIC under the framework of least squares estimation for each of these statistical models. Finally in Section 5, we illustrate this approach on two experimental data sets.

2 Absolute Error Statistical Model (Ordinary Least Squares (OLS) Formulation)

In this section, we assume a statistical model

$$Y_j = \mathbf{f}(t_j, \mathbf{q}_0) + \mathcal{E}_j$$

with $\mathcal{E}_j$, $j = 1,\ 2,\ ...,\ N$ i.i.d. $\mathcal{N}(0, \sigma^2)$. Under this statistical model, $\{Y_j\}_{j=1}^N$ are independent and normally distributed random variables with mean $\mathbb{E}(Y_j) = f(t_j, \mathbf{q})$ and variance $\mathrm{Var}(Y_j) = \sigma^2$, $j = 1,\ 2,\ ...,\ N$. The probability distribution function for a normal distribution $\mathcal{N}(\mu, \sigma^2)$ is given by

$$p(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-(x-\mu)^2}{2\sigma^2}\right); \tag{3}$$

therefore, the likelihood function of $\boldsymbol{\theta} = [\mathbf{q}, \sigma]^T$ given the sample outcome $\mathbf{y}$ is

$$\begin{aligned}
\mathcal{L}(\boldsymbol{\theta}|\mathbf{y}) &= \prod_{j=1}^{N} p(y_j|f(t_j, \mathbf{q}), \sigma^2) \\
&= \prod_{j=1}^{N} \left(\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-(y_j - f(t_j, \mathbf{q}))^2}{2\sigma^2}\right)\right) \\
&= \frac{1}{\left(\sqrt{2\pi\sigma^2}\right)^N} \exp\left(\frac{-\sum_{j=1}^{N}(y_j - f(t_j, \mathbf{q}))^2}{2\sigma^2}\right).
\end{aligned}$$

Taking the natural logarithm of the above equation, we have

$$\begin{aligned}
\ln\left(\mathcal{L}(\boldsymbol{\theta}|\mathbf{y})\right) &= \ln\left(\frac{1}{\left(\sqrt{2\pi\sigma^2}\right)^N}\exp\left(\frac{-\sum_{j=1}^{N}(y_j - f(t_j,\mathbf{q}))^2}{2\sigma^2}\right)\right) \\
&= \ln\left(\frac{1}{\left(\sqrt{2\pi\sigma^2}\right)^N}\right) - \frac{\sum_{j=1}^{N}(y_j - f(t_j,\mathbf{q}))^2}{2\sigma^2} \\
&= -\ln\left(\left(\sqrt{2\pi\sigma^2}\right)^N\right) - \frac{\sum_{j=1}^{N}(y_j - f(t_j,\mathbf{q}))^2}{2\sigma^2} \\
&= -\frac{N}{2}\ln\left(2\pi\sigma^2\right) - \frac{\sum_{j=1}^{N}(y_j - f(t_j,\mathbf{q}))^2}{2\sigma^2} \\
&= -\frac{N}{2}\ln\left(2\pi\right) - N\ln\left(\sigma\right) - \frac{\sum_{j=1}^{N}(y_j - f(t_j,\mathbf{q}))^2}{2\sigma^2}.
\end{aligned} \tag{4}$$

In Eq. (2) for AIC, the log likelihood function is evaluated at the maximum likelihood estimation $\hat{\boldsymbol{\theta}}_{MLE} = (\hat{\mathbf{q}}_{MLE}, \hat{\sigma}_{MLE})$. By examining the equation for $\ln\left(\mathcal{L}(\boldsymbol{\theta}|\mathbf{y})\right)$, we note that

$$\arg\max_{\mathbf{q}\in\Omega_q}\left(\ln\left(\mathcal{L}((\mathbf{q},\sigma)|\mathbf{y})\right)\right) = \arg\min_{\mathbf{q}\in\Omega_q}\left(\sum_{j=1}^{N}(y_j - f(t_j,\mathbf{q}))^2\right)$$

The right hand side above is defined as the ordinary least squares estimate $\hat{\mathbf{q}}_{OLS}$. That is,

$$\hat{\mathbf{q}}_{OLS} = \arg\min_{\mathbf{q}\in\Omega_q}\left(\sum_{j=1}^{N}(y_j - f(t_j,\mathbf{q}))^2\right). \tag{5}$$

Therefore, the maximum log likelihood estimate for $\mathbf{q}$ is the same as the ordinary least squares estimate; in other words, $\hat{\mathbf{q}}_{MLE} = \hat{\mathbf{q}}_{OLS}$. To find the maximum likelihood estimate $\hat{\sigma}_{MLE}$ of σ, we evaluate

$$\left.\frac{\partial\ln\left(\mathcal{L}(\boldsymbol{\theta}|\mathbf{y})\right)}{\partial\sigma}\right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_{MLE}} = 0.$$

Taking the partial of Eq. (4) with respect to σ, we have

$$\frac{\partial\ln\left(\mathcal{L}(\boldsymbol{\theta}|\mathbf{y})\right)}{\partial\sigma} = -\frac{N}{\sigma} + \frac{\sum_{j=1}^{N}(y_j - f(t_j,\mathbf{q}))^2}{\sigma^3}.$$

Therefore,

$$\left.\frac{\partial \ln\left(\mathcal{L}(\boldsymbol{\theta}|\mathbf{y})\right)}{\partial \sigma}\right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_{MLE}} = 0$$

gives

$$\sigma^2_{MLE} = \frac{1}{N}\sum_{j=1}^{N}(y_j - f(t_j, \mathbf{q}_{MLE}))^2.$$

Substituting $\hat{\boldsymbol{\theta}}_{MLE} = (\hat{\mathbf{q}}_{MLE}, \hat{\sigma}_{MLE})$ into Eq. (4) gives

$$\begin{aligned}
\ln\left(\mathcal{L}(\boldsymbol{\theta}_{MLE}|\mathbf{y})\right) &= -\frac{N}{2}\ln(2\pi) - N\ln(\hat{\sigma}_{MLE}) - \frac{\sum_{j=1}^{N}(y_j - f(t_j, \hat{\mathbf{q}}_{MLE}))^2}{2\hat{\sigma}^2_{MLE}} \\
&= -\frac{N}{2}\ln(2\pi) - N\ln\left(\left(\frac{\sum_{j=1}^{N}(y_j - f(t_j, \hat{\mathbf{q}}_{MLE}))^2}{N}\right)^{1/2}\right) - \frac{\sum_{j=1}^{N}(y_j - f(t_j, \hat{\mathbf{q}}_{MLE}))^2}{2\frac{1}{N}\sum_{j=1}^{N}(y_j - f(t_j, \hat{\mathbf{q}}_{MLE}))^2} \\
&= -\frac{N}{2}\left(\ln(2\pi) + 1\right) - \frac{N}{2}\ln\left(\frac{\sum_{j=1}^{N}(y_j - f(t_j, \hat{\mathbf{q}}_{MLE}))^2}{N}\right).
\end{aligned}$$

We recall $\hat{\mathbf{q}}_{MLE} = \hat{\mathbf{q}}_{OLS}$. Furthermore, κ_θ, the total number of estimated parameters, is given by $\kappa_{\mathbf{q}} + 1$ where $\kappa_{\mathbf{q}}$ is the total number of model parameters since there is only one statistical parameter σ. Substituting everything into Eq. (2), we have

$$\text{AIC} = N\left(\ln(2\pi) + 1\right) + N\ln\left(\frac{\sum_{j=1}^{N}(y_j - f(t_j, \hat{\mathbf{q}}_{OLS}))^2}{N}\right) + 2(\kappa_{\mathbf{q}} + 1).$$

The constant term will be the same across all models; therefore, the formula for AIC under a constant variance statistical model is given by

$$\text{AIC}_{OLS} = N\ln\left(\frac{\sum_{j=1}^{N}(y_j - f(t_j, \hat{\mathbf{q}}_{OLS}))^2}{N}\right) + 2(\kappa_{\mathbf{q}} + 1). \tag{6}$$

3 Constant Weighted Error Statistical Model (Weighted Least Squares (WLS) Formulation)

In this section, we assume a statistical model

$$Y_j = f(t_j, \mathbf{q}_0) + w_j\mathcal{E}_j \tag{7}$$

where w_j are known weights and $\mathcal{E}_j$, $j = 1, 2, ..., N$ are i.i.d. $\mathcal{N}(0, \sigma^2)$. Under this statistical model, $\{Y_j\}_{j=1}^N$ are independent and normally distributed random variables with mean $\mathbb{E}(Y_j) = f(t_j, \mathbf{q})$ and variance $\mathrm{Var}(Y_j) = w_j^2\sigma^2$, $j = 1, 2, ..., N$. Using Eq. (3) for the probability distribution function of a normal distribution, the likelihood function of $\boldsymbol{\theta} = [\mathbf{q}, \sigma]^T$ given the sample outcome $\mathbf{y}$ for this statistical model is

$$\begin{aligned}
\mathcal{L}(\boldsymbol{\theta}|\mathbf{y}) &= \prod_{j=1}^{N} p(y_j|f(t_j,\mathbf{q}), w_j^2\sigma^2) \\
&= \prod_{j=1}^{N} \left(\frac{1}{\sqrt{2\pi w_j^2\sigma^2}} \exp\left(\frac{-(y_j - f(t_j,\mathbf{q}))^2}{2w_j^2\sigma^2} \right) \right) \\
&= \frac{1}{\left(\sqrt{2\pi\sigma^2}\right)^N} \prod_{j=1}^{N} (w_j^{-1}) \exp\left(-\frac{1}{2\sigma^2} \sum_{j=1}^{N} w_j^{-2}(y_j - f(t_j,\mathbf{q}))^2 \right).
\end{aligned}$$

Taking the natural logarithm of the above equation, we have

$$\begin{aligned}
\ln\left(\mathcal{L}(\boldsymbol{\theta}|\mathbf{y})\right) &= \ln\left(\frac{1}{\left(\sqrt{2\pi\sigma^2}\right)^N} \prod_{j=1}^{N} (w_j^{-1}) \exp\left(-\frac{1}{2\sigma^2} \sum_{j=1}^{N} w_j^{-2}(y_j - f(t_j,\mathbf{q}))^2 \right) \right) \\
&= \ln\left(\frac{1}{(\sqrt{2\pi\sigma^2})^N} \right) + \ln\left(\prod_{j=1}^{N} (w_j^{-1}) \right) - \frac{1}{2\sigma^2} \sum_{j=1}^{N} w_j^{-2}(y_j - f(t_j,\mathbf{q}))^2 \\
&= \ln\left(\frac{1}{(\sqrt{2\pi\sigma^2})^N} \right) + \sum_{j=1}^{N} \left(\ln(w_j^{-1}) \right) - \frac{1}{2\sigma^2} \sum_{j=1}^{N} w_j^{-2}(y_j - f(t_j,\mathbf{q}))^2 \\
&= -\ln\left((\sqrt{2\pi\sigma^2})^N \right) - \sum_{j=1}^{N} \ln(w_j) - \frac{1}{2\sigma^2} \sum_{j=1}^{N} w_j^{-2}(y_j - f(t_j,\mathbf{q}))^2 \\
&= -\frac{N}{2}\ln(2\pi) - N\ln(\sigma) - \sum_{j=1}^{N} \ln(w_j) - \frac{1}{2\sigma^2} \sum_{j=1}^{N} w_j^{-2}(y_j - f(t_j,\mathbf{q}))^2.
\end{aligned} \tag{8}$$

In Eq. (2) for AIC, the log likelihood function is evaluated at the maximum likelihood estimation $\hat{\boldsymbol{\theta}}_{MLE} = (\hat{\mathbf{q}}_{MLE}, \hat{\sigma}_{MLE})$. By examining the equation for $\ln\left(\mathcal{L}(\boldsymbol{\theta}|\mathbf{y})\right)$, we note that

$$\arg\max_{\mathbf{q}\in\Omega_q} \left(\ln\left(\mathcal{L}((\mathbf{q},\sigma)|\mathbf{y}) \right) \right) = \arg\min_{\mathbf{q}\in\Omega_q} \left(\sum_{j=1}^{N} w_j^{-2}(y_j - f(t_j,\mathbf{q}))^2 \right).$$

The right hand side above is defined as the weighted least squares estimate $\hat{\mathbf{q}}_{WLS}$. That is,

$$\hat{\mathbf{q}}_{WLS} = \arg\min_{\mathbf{q}\in\Omega_q} \left(\sum_{j=1}^{N} w_j^{-2}(y_j - f(t_j,\mathbf{q}))^2 \right). \tag{9}$$

Therefore, the maximum log likelihood estimate for $\mathbf{q}$ when the statistical model has constant weighted variance is the same as the weighted least squares estimate; in other words, $\hat{\mathbf{q}}_{MLE} = \hat{\mathbf{q}}_{WLS}$. To find the maximum likelihood estimate $\hat{\sigma}_{MLE}$ of σ, we evaluate

$$\left.\frac{\partial \ln\left(\mathcal{L}(\boldsymbol{\theta}|\mathbf{y})\right)}{\partial \sigma}\right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_{MLE}} = 0.$$

Taking the partial of Eq. (8) with respect to σ, we have

$$\frac{\partial \ln\left(\mathcal{L}(\boldsymbol{\theta}|\mathbf{y})\right)}{\partial \sigma} = -\frac{N}{\sigma} + \frac{\sum_{j=1}^{N} w_j^{-2}(y_j - f(t_j, \mathbf{q}))^2}{\sigma^3}.$$

Therefore,

$$\left.\frac{\partial \ln\left(\mathcal{L}(\boldsymbol{\theta}|\mathbf{y})\right)}{\partial \sigma}\right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_{MLE}} = 0$$

gives

$$\sigma^2_{MLE} = \frac{1}{N}\sum_{j=1}^{N} w_j^{-2}(y_j - f(t_j, \mathbf{q}_{MLE}))^2.$$

Substituting $\hat{\boldsymbol{\theta}}_{MLE} = (\hat{\mathbf{q}}_{MLE}, \hat{\sigma}_{MLE})$ into Eq. (8) gives

$$\begin{aligned}
\ln\left(\mathcal{L}(\boldsymbol{\theta}_{MLE}|\mathbf{y})\right) &= -\frac{N}{2}\ln(2\pi) - N\ln(\hat{\sigma}_{MLE}) - \sum_{j=1}^{N}\ln(w_j) - \frac{\sum_{j=1}^{N} w_j^{-2}(y_j - f(t_j, \hat{\mathbf{q}}_{MLE}))^2}{2\hat{\sigma}^2_{MLE}} \\
&= -\frac{N}{2}\ln(2\pi) - N\ln\left(\left(\frac{\sum_{j=1}^{N} w_j^{-2}(y_j - f(t_j, \hat{\mathbf{q}}_{MLE}))^2}{N}\right)^{1/2}\right) \\
&\quad - \sum_{j=1}^{N}\ln(w_j) - \frac{\sum_{j=1}^{N} w_j^{-2}(y_j - f(t_j, \hat{\mathbf{q}}_{MLE}))^2}{2\frac{1}{N}\sum_{j=1}^{N} w_j^{-2}(y_j - f(t_j, \hat{\mathbf{q}}_{MLE}))^2} \\
&= -\frac{N}{2}\left(\ln(2\pi) + 1\right) - \sum_{j=1}^{N}\ln(w_j) - \frac{N}{2}\ln\left(\frac{\sum_{j=1}^{N} w_j^{-2}(y_j - f(t_j, \hat{\mathbf{q}}_{MLE}))^2}{N}\right).
\end{aligned}$$

We recall $\hat{\mathbf{q}}_{MLE} = \hat{\mathbf{q}}_{WLS}$ for this statistical model, and as in Section 2, $\kappa_\theta = \kappa_{\mathbf{q}} + 1$. Substituting this information into Eq. (2), we have

$$\text{AIC} = N\left(\ln(2\pi) + 1\right) + 2\sum_{j=1}^{N}\ln(w_j) + N\ln\left(\frac{\sum_{j=1}^{N} w_j^{-2}(y_j - f(t_j, \hat{\mathbf{q}}_{WLS}))^2}{N}\right) + 2(\kappa_{\mathbf{q}} + 1).$$

The two constant terms will be the same across all models; therefore, the formula for AIC under a weighted variance statistical model is given by

$$\mathrm{AIC}_{WLS} = N \ln \left(\frac{\sum_{j=1}^{N} w_j^{-2} (y_j - f(t_j, \hat{\mathbf{q}}_{WLS}))^2}{N} \right) + 2(\kappa_{\mathbf{q}} + 1). \tag{10}$$

4 Parameter Dependent Weighted Error Statistical Model (Iterative Reweighted Weighted Least Squares (IRWLS) Formulation)

A method motivated by the WLS (as we have presented it above) is the Iterative Reweighted Weighted Least Squares (IRWLS) (also called simply the Iterative Weighted Least Squares (IWLS) [10, 16, 17, 21] or the "Generalized" Least Squares (GLS))

$$Y_j = f(t_j, \mathbf{q}_0) + w_j(\mathbf{q}_0)\mathcal{E}_j = f(t_j, \mathbf{q}_0) + f^{\gamma}(t_j, \mathbf{q}_0)\mathcal{E}_j \tag{11}$$

and can be motivated by examining the special weighted least squares estimate

$$\hat{\mathbf{q}}_{IRWLS} = \arg\min_{\mathbf{q} \in \Omega_q} \left(\sum_{j=1}^{N} w_j^{-2} (y_j - f(t_j, \mathbf{q}))^2 \right) \tag{12}$$

for $w_j(\mathbf{q}) = f^{\gamma}(t_j; \mathbf{q})$. By definition the corresponding iterative procedure is given by:

1. Solve for the initial estimate $\hat{\mathbf{q}}^{(0)}$ obtained using the OLS minimization (5). Set $l = 0$.
2. Form the weights $\hat{w}_j^{(l)} = f^{\gamma}(t_j; \hat{\mathbf{q}}^{(l)})$.
3. Re-estimate $\mathbf{q}$ to obtain $\hat{\mathbf{q}}^{(l+1)}$ defined by

$$\hat{\mathbf{q}}^{(l+1)} = \arg\min_{\mathbf{q} \in \Omega_q} \left(\sum_{j=1}^{N} (\hat{w}_j^{(l)})^{-2} [y_j - f(t_j; \mathbf{q})]^2 \right). \tag{13}$$

4. Set $l = l + 1$ and return to step 2. Terminate the process and set $\hat{\mathbf{q}}_{IRWLS} = \hat{\mathbf{q}}^{(l+1)}$ when two of the successive estimates are sufficiently close.

We note that the above procedure is equivalent to finding the weighted least squares for a statistical model (7) for a sequence of weights $w_j = \hat{w}_j^{(l)}$. Recalling the arguments up through (9) we thus have $\hat{\mathbf{q}}^{(l+1)} = \hat{\mathbf{q}}_{WLS}^{(l+1)} = \hat{\mathbf{q}}_{MLE}(\hat{w}_j^{(l)})$ for the sequence of weights and hence one has the arguments of the previous section leading to

$$\mathrm{AIC}_{WLS}^{(l+1)} = N \ln \left(\frac{\sum_{j=1}^{N} w_j^{-2} (y_j - f(t_j, \hat{\mathbf{q}}_{WLS}^{l+1}))^2}{N} \right) + 2(\kappa_{\mathbf{q}} + 1), \tag{14}$$

for the weights $w_j = \hat{w}_j^{(l)}$. Thus under reasonable conditions [17], if the process enumerated above is continued a sufficient number of times, then $\hat{w}_j{}^{(l)} \to f^{\gamma}(t_j; \hat{\mathbf{q}}_{IRWLS})$ and thus we may establish

$$\mathrm{AIC}_{IRWLS} \approx N \ln \left(\frac{\sum_{j=1}^{N} w_j^{-2} (y_j - f(t_j, \hat{\mathbf{q}}_{WLS}^{M}))^2}{N} \right) + 2(\kappa_{\mathbf{q}} + 1), \tag{15}$$

where $w_j = \hat{w}_j^M \approx f^{\gamma}(t_j; \hat{\mathbf{q}}_{IRWLS})$ where M is the number of times the process is enumerated.

5 Model Comparison Examples

In this section, we use an appropriate least squares formulation of the AIC to compare a set of candidate models in two examples. In the first example, we compare decay models for the size distribution of aggregates in amyloid fibril formulation in which the data has been shown to have an absolute error statistical model [20], i.e., we will use the ordinary least squares formulation of the AIC in the model comparison. In the second example we compare growth models for longitudinal data collected from algae growth which has a parameter dependent weighted error statistical model [11], i.e, the iterative reweighted weighted least squares (IRWLS) formulation of the AIC is necessary. In both data sets, the sample size is small; therefore, the modification of AIC for small sample sizes is first discussed in Section 5.1. In addition, Section 5.2 focuses on Aikake weights which are used as a means for judging the relative strength of the 'best' model (the one with the smallest AIC value). We then apply these techniques to the amyloid fibril data and algae growth data in Sections 5.3 and 5.4, respectively.

5.1 AIC for Small Sample Size

In the original formulation of AIC, it is assumed that the sample size is sufficiently large; thus, if the sample size is not large enough relative to the number of parameters which must be estimated, the AIC may perform poorly. In [15], it was suggested that the AIC only be used if the sample size N is at least 40 times as large as the total number of estimated parameters, i.e. $N/\kappa_\theta \geq 40$. However, in many cases, only a small sample can be collected. In the case of small sample sizes, Sugiura [23] proposed the AIC_c for scalar linear regression models which was later extended by Hurvich and Tsai [18] for a scalar non-linear regression model and by Bedrick and Tsai [12] in the case of multivariate observations. In the derivation of AIC_c in [23], it was assumed that the measurement errors $\mathcal{E}_j$, $j = 1,\ 2,\ ...,\ N$ were independent and identically distributed with $\mathcal{E}_j \sim \mathcal{N}(0, \sigma^2)$. In this case, the penalty term $2\kappa_\theta$ in the AIC formula (Eq. (2)) is modified by a correction term $\dfrac{N}{N - \kappa_\theta - 1}$,

$$\mathrm{AIC}_c = -2 \ln \left(\mathcal{L}(\hat{\boldsymbol{\theta}}_{MLE} | \mathbf{y}) \right) + 2\kappa_\theta \frac{N}{N - \kappa_\theta - 1}.$$

We can rewrite this in terms of the original AIC by noticing that the new penalty term can be written as the old penalty term plus an additional term:

$$2\kappa_\theta \frac{N}{N - \kappa_{\boldsymbol{\theta}} - 1} = 2\kappa_{\boldsymbol{\theta}} + \frac{2\kappa_{\boldsymbol{\theta}} (\kappa_{\boldsymbol{\theta}} + 1)}{N - \kappa_{\boldsymbol{\theta}} - 1}.$$

Therefore,

$$\mathrm{AIC}_c = \mathrm{AIC} + \frac{2\kappa_{\boldsymbol{\theta}} (\kappa_{\boldsymbol{\theta}} + 1)}{N - \kappa_{\boldsymbol{\theta}} - 1}. \tag{16}$$

If we derive the AIC_c under the ordinary least squares formulation as in Section 2 and substituting $\kappa_\theta = \kappa_{\mathbf{q}}+1$, we have a similar equation in which

$$\text{AIC}_{OLS_c} = \text{AIC}_{OLS} + \frac{2(\kappa_{\mathbf{q}}+1)(\kappa_{\mathbf{q}}+2)}{N-\kappa_{\mathbf{q}}}. \tag{17}$$

In [14], it was shown that if the error was not too far skewed from the normal distribution, the same modification factor was sufficient for small data sets. In the case of the weighted least squares formulation, we have $Y_j = \mathcal{N}(0, w_j^2\sigma^2)$; therefore, we can formulate the AIC_c in a similar method for this case since the error for each Y_j is normally distributed with only varying weights for the variance of each data point. We obtain the formula

$$\text{AIC}_{WLS_c} = \text{AIC}_{WLS} + \frac{2(\kappa_{\mathbf{q}}+1)(\kappa_{\mathbf{q}}+2)}{N-\kappa_{\mathbf{q}}}. \tag{18}$$

Given that the iterative reweighted weighted least squares formulation is an iterative process using updated weights in the weighted least squares formulation, we also have

$$\text{AIC}_{IRWLS_c} = \text{AIC}_{IRWLS} + \frac{2(\kappa_{\mathbf{q}}+1)(\kappa_{\mathbf{q}}+2)}{N-\kappa_{\mathbf{q}}}. \tag{19}$$

5.2 Comparison of AIC or AIC_c Values for Model Selection

In general, one wants the AIC to be as small as possible; however, as Burnham and Anderson [15] stated, "It's not the absolute size of the AIC value, it is the relative values, and particularly the AIC differences (Δ_i), that are important." Given a set of candidate models and the values of AIC or AIC_c for these models, it is easy to order the values of AIC from least to greatest; however, one often wants to know how much more likely the 'best' model is compared to the next best model. As such, Akaike weights are important in the comparison of models. To define weights, we first define AIC differences $\Delta_i(\text{AIC})$ [15, 24],

$$\Delta_i(\text{AIC}) = \text{AIC}_i - \text{AIC}_{min},$$

where AIC_{min} denotes the minimum calculated AIC value across all candidate models and the term AIC refers to either the original AIC, AIC_c or other variations of the AIC. Note that if Model l is the model with the minimum AIC value, then $\Delta_l = 0$. Akaike [9] indicates that the likelihood of model i given data set $\mathbf{y}$ is proportional to $\exp\left(-\frac{1}{2}\Delta_i\right)$; therefore, we can use this value as an indication of the relative strength of evidence for each candidate model. Normalizing the relative likelihoods, we obtain the Akaike weights $w_i(AIC)$ [15, 24],

$$w_i(\text{AIC}) = \frac{\exp\left(-\frac{1}{2}\Delta_i(\text{AIC})\right)}{\sum_{k=1}^{K}\exp\left(-\frac{1}{2}\Delta_k(\text{AIC})\right)} \tag{20}$$

where K is the number of candidate models. We note that the weights of all candidate models sum to 1, so the weight gives a probability that each model is the best model. Furthermore, the evidence ratio

$$\frac{w_i(\text{AIC})}{w_j(\text{AIC})} \tag{21}$$

indicates how much more likely model i is compared to model j. In addition, if there are two models, say models i and j, which have the largest and second largest weights respectively, then the normalized probability

$$\frac{w_i(\text{AIC})}{w_i(\text{AIC}) + w_j(\text{AIC})} \tag{22}$$

indicates the probability of model i over model j [24]. We now apply these techniques to analyze two experimental data sets.

5.3 Amyloid fibril data

The first example involves the size and distribution of amyloid fibrils. Many diseases, such as Alzheimers, Huntingtons and Prion diseases (e.g. mad cow) are related to aggregations of proteins which exhibit an abnormal folding [22]. These protein aggregates are called amyloids and have been the focus of many recent studies [19, 20, 25, 26]. In the paper by Prigent et. al [19], the size distribution of aggregates in amyloid fibril formation was studied. Two samples of fibrils were collected, one of size 531 and one of size 95. The frequency of each fribil size (monomers) was obtained and converted to proportions and the sizes were divided by 10,000 which was assumed to be a theoretical maximum. We refer the reader to [19] for full details about the experimental process.

In this section we consider five candidate models for the fibril data: exponential distribution, Weibull distribution, Gamma distribution, logistic decay and Gompertz decay models. The exponential model has only two parameters which must be estimated while each of the other models involve three parameters. The exponential distribution probability density is given by $E(x;\lambda) = \lambda e^{-\lambda x}$, but we consider the model

$$E(x; A, \lambda) = A\lambda e^{-\lambda x} \tag{23}$$

where A is an additional parameter. For modeling purposes, we also add an additional parameter to the Weibull distribution model and simplify the parameter estimation by considering $\tilde{\lambda} = \frac{1}{\lambda}$ is the traditional equation for the probability density. Therefore, we consider the model

$$W(x; A, \tilde{\lambda}, k) = Ak\tilde{\lambda}(\tilde{\lambda}x)^{k-1}e^{-(\tilde{\lambda}k)^k}. \tag{24}$$

The traditional formula for the probability density function of the gamma distribution is defined as

$$G(x; k, \theta) = \frac{x^{k-1}e^{-x/\theta}}{\theta^k\Gamma(k)} \quad \text{for } x > 0 \text{ and } k, \theta > 0.$$

Again, we add an additional parameter A and simplify the formula by allowing $\theta = \frac{1}{\lambda}$. Therefore, we consider the Gamma distribution model

$$G(x; A, k, \lambda) = A\frac{\lambda^k}{\Gamma(k)}x^{k-1}e^{-\lambda x}. \tag{25}$$

The two final models are the logistic decay model,

$$L(x; a, b, c) = \frac{c}{1 + ae^{-bx}}, \tag{26}$$

and the Gompertz decay model,

$$D(x; A, \lambda, k) = A\exp\left(-\lambda e^{-kx}\right). \tag{27}$$

Using ordinary least squares parameter estimation and the formula for AIC_{OLS_c} in Eq. (17), we obtain the fitted models in Figure 1. Table 1 gives the AIC_{OLS_c} values and weights for each model.

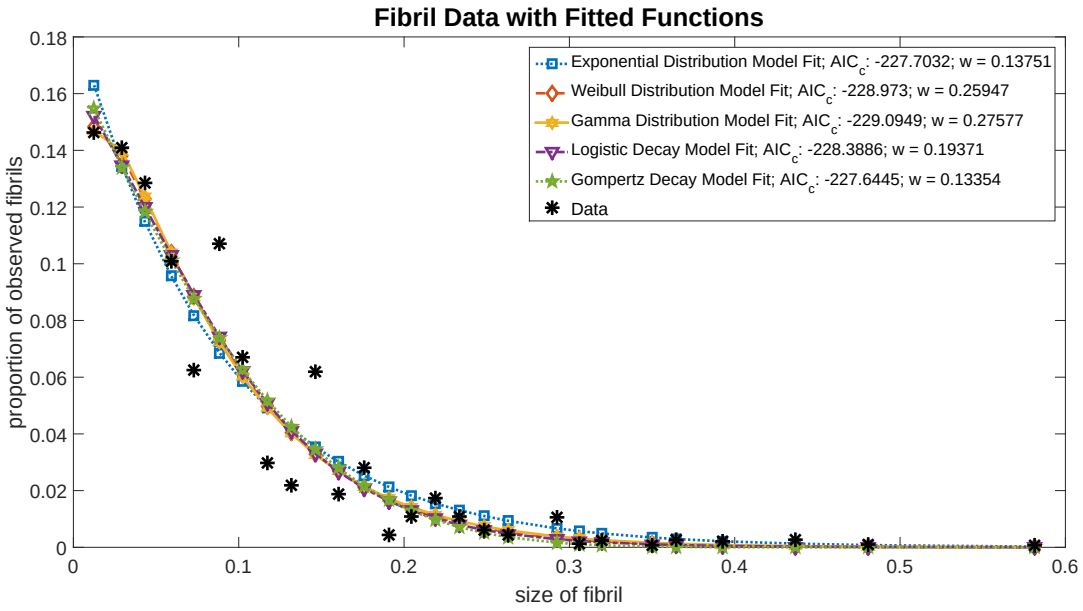

Figure 1: This figure shows experimental fibril data along with the fitted models given by Eqs. (23) - (27). For each model, the AIC_c (Eq. (17)) and the model weight (Eq. (20)) are given.

Table 1: Comparison of AIC_{OLS_c} Values for each Candidate Model for the Fibril Size Data

Model	AIC_{OLS_c}	w_i
Exponential	-227.70	0.138
Weibull	-228.97	0.259
Gamma	-229.09	0.276
Logistic	-228.39	0.194
Gompertz	-227.64	0.133

As indicated in Table 1, the Gamma distribution model has the lowest AIC_{OLS_c} value and therefore would be considered the 'best' of the candidate models for this data, closely followed by the Weibull model. However, heuristically speaking, using the evidence ratio in Eq. (21), the Gamma model is only 1.1 times more likely to be the best model in terms of the Kullback-Leibler discrepancy than the Weibull model with a normalized probability of only 0.52 (using Eq. (22)). This low normalized probability is evidenced in Figure 1 where all the fitted curves are practically identical. We note that the only discernable difference in curve fits is with the exponential distribution model, the model with the greatest AIC_{OLS_c} value out of the candidate models. However, comparing the normalized probability of the Gamma distribution model to the exponential distribution model, there is still only a 0.59 probability of the Gamma distribution model as the preferred model over the exponential distribution model. We now do a similar comparison for algae growth data.

5.4 Algae growth data

In a paper by Banks et. al. [11], longitudinal data was collected from four replicate population experiments with green algae, formally known as *Raphidocelis subcapitata*. The authors were concerned with the growth dynamics of the algae as this is the major feeding source for *Daphnia magna* [1] in experimental settings. In ecology, *D. magna* can be thought of as the modern day "canary in the mine shaft", because changes in the daphnia population can alert ecologists to changing dynamics in the environment.

In this paper, we compare three different dynamical population models for algae growth using the AIC: logistic model, Bernoulli model, and Gompertz model. The logistic model is given by

$$\frac{dx}{dt} = rx(t)\left(1 - \frac{x(t)}{K}\right), \quad x(0) = x_0 \tag{28}$$

where r is the growth rate and K is the carrying capacity for the population. The Bernoulli model contains one additional parameter β and is given by

$$\frac{dx}{dt} = rx(t)\left(1 - \left(\frac{x(t)}{K}\right)^{\beta}\right), \quad x(0) = x_0.$$

Note that the logistic growth model is obtained from the Bernoulli growth model by setting β equal to 1. In standard form, the parameters K and β are are found jointly in the denominator causing a problem with identifiability. To address this issue, we let $\tilde{K} = K^{\beta}$ and instead consider the model

$$\frac{dx(t)}{dt} = rx(t)\left(1 - \frac{(x(t))^{\beta}}{\tilde{K}}\right), \quad x(0) = x_0 \tag{29}$$

where K can be obtained from $\tilde{K}$ using $K = \tilde{K}^{(1/\beta)}$. The final model considered is the Gompertz model,

$$\frac{dx(t)}{dt} = \kappa x(t) \log\left(\frac{K}{x(t)}\right), \quad x(0) = x_0, \tag{30}$$

where K is the carrying capacity as in the other two models and κ scales the time. We note that both the logistic and Gompertz models contain only two parameters while the Bernoulli model contains three parameters.

In terms of modeling the algae data, it is demonstrated in the paper by Banks et. al. [11] that the appropriate statistical model for this data is a parameter dependent weighted error statistical model with $\gamma = 1$ in Equation (11). Therefore, we use AIC_{IRWLS_c} in Equation (19) to compare models as this is a small data set with only 36 data points for each of the four replicates. The results for each replicate are given in Table 2 with the fitted models and data for replicate 1 plotted in Figure 2.

As shown in Table 2, there is minimal difference across the four replicates and in each case, the smallest AIC value is given by the Gompertz model followed closely by the Bernoulli model. Recall that the Gompertz model has 2 parameters; whereas the Bernoulli model has three; therefore, although the two curves are lying on top of one another in Figure 2, the Bernoulli model is penalized more by the extra parameter. If we heuristically compare the Gompertz and Bernoulli models using the evidence ratio in Equation (21) and the normalized probability in Equation (22), we see that the Gompertz model is only 1.03 times more likely with a normalized probability of only 0.51 (only slightly more than equal probability). Therefore, either the Gompertz or the Bernoulli model appears to be a good model of the candidate models examined.

Table 2: Comparison of AIC_{IRWLS_c} Values for each Candidate Model for the Algae Data

	Replicate 1		Replicate 2		Replicate 3		Replicate 4	
Model	AIC_{IRWLS_c}	w_i	AIC_{IRWLS_c}	w_i	AIC_{IRWLS_c}	w_i	AIC_{IRWLS_c}	w_i
logistic	-127.42	2.7e-05	-117.07	1.5e-07	-117.02	1.5e-07	-104.24	2.5e-10
Bernoulli	-147.05	0.492	-147.05	0.492	-147.05	0.492	-147.05	0.492
Gompertz	-147.11	0.508	-147.11	0.508	-147.11	0.508	-147.11	0.508

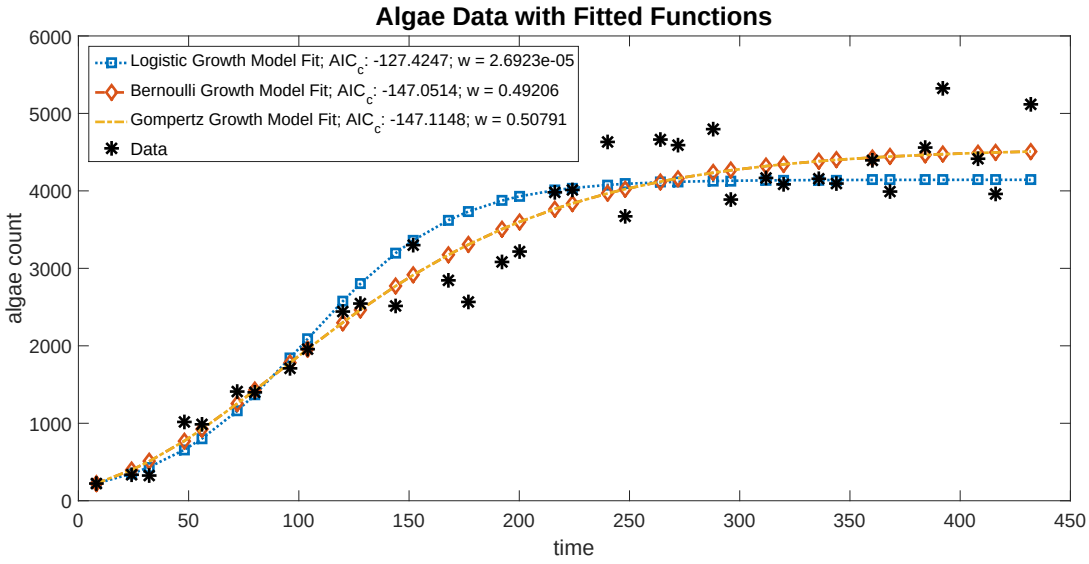

Figure 2: This figure shows experimental algae data along with the fitted models given by Eqs. (28) - (30). For each model, the AIC_c (Eq. (11)) and the model weight (Eq. (20)) are given.

6 Conclusions

To summarize, given the traditional form of the AIC, we derived a concise formulation applicable when parameter estimation is performed in the framework of a least squares estimation problem. We derived the formulation under three different types of statistical models: one which utilizes ordinary least squares (OLS), another for weighted least squares (WLS) and a final formulation using iterative reweighted weighted least squares (IRWLS). Finally, we illustrated the effectiveness of the formulation using two experimental data sets, one which required ordinary least squares estimation and one which required iterative reweighted weighted least squares estimation. In both cases, we could identify 'best' models from the candidate models and heuristically discuss the normalized probability of one model choice over another model choice.

Acknowledgements

This research was supported in part by the National Institute on Alcohol Abuse and Alcoholism under grant number 1R01AA022714-01A1, and in part by the Air Force Office of Scientific Research under grant number AFOSR FA9550-15-1-0298.

References

[1] Kaska Adoteye, H.T. Banks, Karissa Cross, Stephanie Eytcheson, kevin Flores, gerald A. LeBlanc, Timothy Nguyen, Chelsea Ross, Emmaline Smith, Michale Stemkovski, and Sarah Stokely. Statistical validation of structured population models for *Daphnia magna*. *Mathematical Biosciences*, 266:73–84, 2015.

[2] Hirotugu Akaike. Information theory as an extension of the maximum likelihood. Proceedings of IEEE International Symposium on Information Theory, 1973.

[3] Hirotugu Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–723, 1974.

[4] Hirotugu Akaike. Canonical correlation analysis of time series and the use of an information criterion. *Mathematics in Science and Engineering*, 126:27–96, 1976.

[5] Hirotugu Akaike. On entropy maximization principle. *Application of Statistics*, 1977.

[6] Hirotugu Akaike. On newer statistical approaches to parameter estimation and structure determination. In *International Federation of Automatic Control*, volume 3, pages 1877–1884, 1978.

[7] Hirotugu Akaike. Likelihood of a model and information criteria. *Journal of Econometrics*, 16(1):3–14, 1981.

[8] Hirotugu Akaike. Modern development of statistical methods. *Trends and Progress in System Identification*, 1:169, 1981.

[9] Hirotugu Akaike. Information measures and model selection. *Bulletin of the International Statistical Institute*, 50(1):277–291, 1983.

[10] H. T. Banks, Shuhua Hu, and W. Clayton Thompson. *Modeling and Inverse Problems in the Presence of Uncertainty*. CRC Press, 2014.

[11] H.T. Banks, Elizabeth Collins, Kevin Flores, Prayag Pershad, Michael Stemkovski, and Lyric Stephenson. Standard and proportional error model comparison for logistic growth of green algae (*Raphidocelis subcapiala*). *Applied Mathematical Letters*, 64:213–222, 2017.

[12] Edward J. Bedrick and Chih-Ling Tsai. Model selection for multivariate regression in small samples. *Biometrics*, pages 226–231, 1994.

[13] Hamparsum Bozdogan. Model selection and Akaike's Information Criterion (AIC): The general theory and its analytical extensions. *Psychometrika*, 52(3):345–370, 1987.

[14] Kenneth P Burnham, David R Anderson, and Gary C White. Evaluation of the kullback-leibler discrepancy for model selection in open population capture-recapture models. *Biometrical Journal*, 36(3):299–315, 1994.

[15] K.P. Burnham and D.R. Anderson. *Information and Likelihood Theory: A Practical Information-Theoretic Approach*. Springer-Verlag New York, 2002.

[16] Raymond J. Carroll and David Ruppert. *Transformation and Weighting in Regression*, volume 30. CRC Press, 1988.

[17] Marie Davidian and David Gilitanan. *Nonlinear Models for Repaeated Measurement Data*, volume 62. CRC Press, 1998.

[18] Clifford M Hurvich and Chih-Ling Tsai. Regression and time series model selection in small samples. *Biometrika*, pages 297–307, 1989.

[19] Stéphanie Prigent, Annabelle Ballesta, Frédérique Charles, Natacha An efficient kinetic model for assemblies of amyloid fibrils and its application to polyglutamine aggregation. *PLoS One*, 7(11):e43273, 2012.

[20] Stéphanie Prigent, Wafaâ Haffaf, H. T. Banks, M Hoffmann, Human Rezaei, and Marie Doumic. Size distribution of amyloid fibrils. Mathematical models and experimental data. *International Journal of Pure and Applied Mathematics*, 93(6):845–878, 2014.

[21] G.A.F. Seber and C.J. Wild. *Nonlinear Regression*. J. Wiley and Sons, 2003.

[22] Jean D Sipe and Alan S Cohen. Review: history of the amyloid fibril. *Journal of Structural Biology*, 130(2-3):88–98, 2000.

[23] Nariaki Sugiura. Further analysts of the data by Akaike's Information Criterion and the finite corrections: Further analysts of the data by Akaike's. *Communications in Statistics-Theory and Methods*, 7(1):13–26, 1978.

[24] Eric-Jan Wagenmakers and Simon Farrell. AIC model selection using Akaike weights. *Psychonomic Bulletin & Review*, 11(1):192–196, 2004.

[25] Wei-Feng Xue, Steve W Homans, and Sheena E Radford. Systematic analysis of nucleation-dependent polymerization reveals new insights into the mechanism of amyloid self-assembly. *Proceedings of the National Academy of Sciences*, 105(26):8926–8931, 2008.

[26] Wei-Feng Xue, Steve W Homans, and Sheena E Radford. Amyloid fibril length distribution quantified by atomic force microscopy single-particle image analysis. *Protein Engineering Design and Selection*, 22:489-496, 2009.