Journal of Biopharmaceutical Statistics
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/lbps20

Resampling Methods for Model Fitting and Model Selection

G. Jogesh Babu [a]

[a] Department of Statistics, Pennsylvania State University, University Park, Pennsylvania, USA

Available online: 24 Oct 2011

To cite this article: G. Jogesh Babu (2011): Resampling Methods for Model Fitting and Model Selection, Journal of Biopharmaceutical Statistics, 21:6, 1177-1186

To link to this article: http://dx.doi.org/10.1080/10543406.2011.607749

Journal of Biopharmaceutical Statistics, 21: 1177–1186, 2011
Copyright © Taylor & Francis Group, LLC
ISSN: 1054-3406 print/1520-5711 online
DOI: 10.1080/10543406.2011.607749

RESAMPLING METHODS FOR MODEL FITTING AND MODEL SELECTION

G. Jogesh Babu
Department of Statistics, Pennsylvania State University, University Park, Pennsylvania, USA

Resampling procedures for fitting models and model selection are considered in this article. Nonparametric goodness-of-fit statistics are generally based on the empirical distribution function. The distribution-free property of these statistics does not hold in the multivariate case or when some of the parameters are estimated. Bootstrap methods to estimate the underlying distributions are discussed in such cases. The results hold not only in the case of one-dimensional parameter space, but also for the vector parameters. Bootstrap methods for inference, when the data is from an unknown distribution that may or may not belong to a specified family of distributions, are also considered. Most of the information criteria-based model selection procedures such as the Akaike information criterion, Bayesian information criterion, and minimum description length use estimation of bias. The bias, which is inevitable in model selection problems, arises mainly from estimating the distance between the "true" model and an estimated model. A jackknife type procedure for model selection is discussed, which instead of bias estimation is based on bias reduction.

Key Words: Information criterion; Jackknife; Kullback–Leibler divergence; Maximum likelihood estimation; Model selection; Nonparametric bootstrap; Parametric bootstrap.

1. INTRODUCTION

Some recently developed procedures based on resampling methods designed for model fitting and a jackknife type procedure for model selection are discussed.

A good statistical model should be simple; the fitted model should confirm to the data (goodness-of-fit) and should be easily generalizable. Occam's razor, a principle credited to the English philosopher William of Ockham (1285–1349), which essentially says that the simplest solution is usually the correct one, is the main guiding principle for statistical modeling. Occam's razor suggests that we leave off extraneous ideas to better reveal the truth. That is, select a model that adequately accommodates the data. It neither *underfits* so that it excludes key variables or effects, nor *overfits* so that it is unnecessarily complex by including extraneous explanatory variables or effects. Underfitting induces bias and overfitting induces

Received April 14, 2011; Accepted May 16, 2011

Address correspondence to G. Jogesh Babu, Department of Statistics, 326 Joab L. Thomas Building, The Pennsylvania State University, University Park, PA 16802-2111, USA; E-mail: babu@psu.edu

high variability. A model selection criterion should balance the competing objectives of conformity to the data and parsimony.

We start with considering the bootstrap approach for goodness-of-fit procedures based on the empirical distribution function (EDF). In the second part, a jackknife type method that selects a model of minimum Kullback–Leibler divergence through bias reduction is discussed.

2. MODEL FITTING

Some recently developed goodness-of-fit procedures based on bootstrap method are discussed. We start with classical distribution-free statistics. The classical procedures fail when model parameters are estimated from the data. In this article we restrict to statistics defined through empirical distribution function.

2.1. Goodness-of-Fit Statistics Based on the EDF

Among the practitioners, nonparametric goodness-of-fit procedures based on the Kolmogorov–Smirnov (K-S) statistic are popular, although other EDF-based statistics such as the Cramer–von Mises (C-vM) and Anderson–Darling (A-D) statistics have better sensitivity for some data-model differences. However, as we review later, *the goodness-of-fit probabilities derived from the K-S or other EDF statistics are usually not correct when applied in model fitting situations with estimated parameters*, or in the multivariate case.

To recall these commonly used statistics, let $X_1, \ldots, X_n$ be i.i.d. random variables having a common distribution function F. Let F_n denote the empirical distribution function of $X_1, \ldots, X_n$. That is, $F_n(a) = \frac{1}{n}\#\{1 \leq i \leq n : X_i \leq a\}$.

The three commonly used statistics based on EDF mentioned already are:

Kolmogorov–Smirnov (K-S):

$$\sup_x |F_n(x) - F(x)|$$

Cramer–von Mises (C-vM):

$$\int (F_n(x) - F(x))^2 \, dF(x)$$

Anderson–Darling (A-D):

$$\int \frac{(F_n(x) - F(x))^2}{F(x)(1 - F(x))} \, dF(x)$$

Here F is the model distribution function, and "sup" means the supremum. The K-S statistic is most sensitive to large-scale differences in location (i.e., median value) and shape between the model and data. The C-vM statistic is effective for both large-scale and small-scale differences in distribution shape. Both of these measures are relatively insensitive to differences near the tails of the distribution F. This deficiency is addressed by the A-D statistic, a weighted version of the C-vM statistic to emphasize differences near the tails.

All these statistics are distribution-free as long as F is continuous. In particular, the probability distribution H of the K-S statistic given by

$$H(y) = P\Big(\sup_x |F_n(x) - F(x)| \leq y\Big)$$

is free from F. Consequently, the confidence bands for the "unknown" distribution F can be obtained from standard tables of K-S, C-vM, or A-D probabilities, which depend only on the number of data points and the chosen significance level.

Model fitting requires complete specification of the underlying postulated distribution function F. It is not enough to know the shape of the curve; it requires complete knowledge of the parameters. If the model family truly represents the underlying phenomenon, the fitted parameters give insights into proposed physical theory for the underlying population. For example, when assessing whether the data are from a normal distribution, the two parameters specifying the location and scale (mean and variance) are needed to compare it with the empirical distribution function. Practitioners in applied fields often use these statistics, by first estimating the parameters and then using the same data to fit the model. When some parameters are estimated, the *goodness-of-fit probabilities derived from the K-S or other EDF-based statistics are usually not correct* (Lilliefors, 1969). In fact, the distributions (such as H) of the statistics mentioned earlier, when parameters are estimated, depend in a complicated way on the unknown parameters and the shape of the unknown distribution F. The K-S probabilities are not valid unless the parameters of the model are derived independently of the data set at hand: e.g., from some previous data sets or from prior modeling considerations.

2.1.1. Failure of the Multivariate Case.

The failure of the K-S and other EDF-based statistics, when two or more dimensions are present, was demonstrated by Simpson (1951) using the following simple example.

Let (X_1, Y_1) be a data point from a bivariate distribution F on the unit square. Simpson showed that if F_1 denotes the EDF of (X_1, Y_1), then

$$P\big(|F_1(x, y) - F(x, y)| < .72, \text{ for all } x, y\big) \begin{cases} > 0.065 & \text{if } F(x, y) = xy^2 \\ < 0.058 & \text{if } F(x, y) = xy(x + y)/2 \end{cases}$$

Thus, the distribution of the K-S statistic varies with the unknown continuous F and hence is not distribution-free when two or more dimensions are present. The K-S statistic still is a measure of "distance" between the data and model. However, the distribution of the statistic is intractable without a detailed calculation for each case under consideration. Several methodological studies in the astronomical literature and other fields discuss the two-dimensional K-S statistic, ignoring the fact that it is no longer distribution-free.

2.2. Bootstrap

Fortunately, there is an alternative to the erroneous use of K-S procedure, although it requires a numerically intensive calculation for each dataset. It is based on bootstrap resampling (Efron, 1979; Chernick, 2007), a data-based Monte Carlo method that has been mathematically shown to be valid under a very wide range of situations (Babu and Rao, 1993).

We now outline the methodology underlying the bootstrap procedure. Let $\{F_\theta : \theta \in \Theta\}$ be a family of continuous candidate distributions parameterized by θ. We want to test whether the univariate data set $X_1, \ldots, X_n$ comes from $F = F_{\theta_0}$ for some fixed $\theta_0 \in \Theta$. The K-S, C-vM, and A-D statistics (and a few other goodness-of-fit tests) are continuous functionals of the stochastic process

$$Y_n(x; \hat{\theta}_n) = \sqrt{n}\big(F_n(x) - F_{\hat{\theta}_n}(x)\big)$$

indexed by x. As defined earlier, F_n denotes the EDF of $X_1, \ldots, X_n$, $\hat{\theta}_n = \theta_n(X_1, \ldots, X_n)$ is an estimator of θ derived from the data set, and $F_{\hat{\theta}_n}$ is the model evaluated at the estimated parameters. For example, if $\{F_\theta : \theta \in \Theta\}$ denotes the Gaussian family with $\theta = (\mu, \sigma^2)$, then $\hat{\theta}_n$ can be taken as $(\overline{X}_n, s_n^2)$, where $\overline{X}_n$ is the sample mean and s_n^2 is the sample variance based on the data $X_1, \ldots, X_n$.

To estimate the distribution of the goodness-of-fit statistics for a model where the parameters have been estimated from the data, the bootstrap can be used in two different ways: the *parametric bootstrap* and the *nonparametric bootstrap*. The parametric bootstrap may be familiar to many practitioners as a well-established technique of creating fake data sets (simulations) from the parametric model by Monte Carlo methods (see Press et al. (1997)). The nonparametric bootstrap, in contrast, is a particular Monte Carlo realization of the observed EDF using a "random selection with replacement" procedure.

We now outline the asymptotics underlying these procedures. Let $\widehat{F}_n$ be an estimator of F, based on $X_1, \ldots, X_n$. In order to bootstrap, we generate data $X_1^*, \ldots, X_n^*$ from the estimated population $\widehat{F}_n$ and then construct $\hat{\theta}_n^* = \theta_n(X_1^*, \ldots, X_n^*)$ using the same functional form.

2.3. Parametric Bootstrap

The bootstrapping procedure is called parametric if $\widehat{F}_n = F_{\hat{\theta}_n}$; that is, we generate data $X_1^*, \ldots, X_n^*$ from the model assuming the estimated parameter values $\hat{\theta}_n$. For example, if F_θ is Gaussian with $\theta = (\mu, \sigma^2)$ and if $\hat{\theta}_n = (\overline{X}_n, s_n^2)$, then $\theta_n^* = (\overline{X}_n^*, s_n^{*2})$, where $X_1^*, \ldots, X_n^*$ are i.i.d. random variables generated from Gaussian distribution with mean $\overline{X}_n$, and variance s_n^2. The process

$$Y_n^P(x) = \sqrt{n}\big(F_n^*(x) - F_{\theta_n^*}(x)\big)$$

and the sample process

$$Y_n(x; \hat{\theta}_n) = \sqrt{n}\big(F_n(x) - F_{\hat{\theta}_n}(x)\big)$$

both converge to the same Gaussian process Y for almost all samples. Consequently,

$$L_n = \sqrt{n}\, \sup_x |F_n(x) - F_{\hat{\theta}_n}(x)| \quad \text{and} \quad L_n^* = \sqrt{n}\, \sup_x |F_n^*(x) - F_{\theta_n^*}(x)|$$

both have the same limiting distribution. For the K-S statistic, the critical values of L_n can be derived as follows: Construct B resamples based on the parametric model (generally $B \approx 1000$ should suffice), arrange the resulting L_n^* values in increasing order to obtain 90 or 99 percentile points for getting 90% or 99% critical values. This procedure replaces the incorrect use of the standard K-S probability tabulation.

2.4. Nonparametric Bootstrap

The nonparametric bootstrap involving resamples from the EDF,

$$\begin{aligned} Y_n^N(x) &= \sqrt{n}\big(F_n^*(x) - F_{\theta_n^*}(x)\big) - B_n(x) \\ &= \sqrt{n}\big(F_n^*(x) - F_n(x) + F_{\hat{\theta}_n}(x) - F_{\theta_n^*}(x)\big) \end{aligned}$$

is operationally easy to perform but requires an additional step of bias correction,

$$B_n(x) = \sqrt{n}(F_n(x) - F_{\hat{\theta}_n}(x))$$

Conditioned on the original data, $\{B_n(x)\}$ is a sequence of constants. The sample process Y_n and the bias corrected nonparametric process Y_n^N conditioned on the original data converge to the same Gaussian process Y. In particular,

$$L_n = \sqrt{n} \sup_x |F_n(x) - F_{\hat{\theta}_n}(x)| \quad \text{and} \quad J_n^* = \sup_x |\sqrt{n}\left(F_n^*(x) - F_{\theta_n^*}(x)\right) - B_n(x)|$$

both have the same limiting distribution. The critical values of the distribution of L_n can then be derived as in the case of parametric bootstrap. For the details on the regularity conditions under which these results hold, see Babu and Rao (2004).

2.5. Confidence Limits Under Misspecification of Model Family

We now address the problem of comparing best-fit models derived for different model families: e.g., the Pareto distribution versus exponential model fits. Essentially, we are asking, "How far away" is the unknown distribution underlying the observed data set from the hypothesized family of models?

Let the original data set $X_1, \ldots, X_n$ come from an unknown distribution G. G may or may not belong to the family $\{F_\theta : \theta \in \Theta\}$. Let F_{θ_0} be the specific model in the family that is "closest" to G where proximity is based on the Kullback–Leibler (KL) divergence (Kullback and Leibler, 1951) D_{KL}, also known as information divergence, or relative entropy. $D_{KL}[g; f]$ for probability densities g and f is defined as

$$D_{KL}[g; f] := \int g(x) \log g(x)\, dx - \int g(x) \log f(x)\, dx \tag{1}$$

D_{KL} is always non negative and it is a non symmetric measure of the "distance" between two probability densities f and g. This arises naturally due to maximum likelihood arguments.

If the maximum likelihood estimator $\hat{\theta}_n$ tends to θ_0, then the stochastic process

$$U_n(x; \hat{\theta}_n) = \sqrt{n}\big(F_n(x) - F_{\hat{\theta}_n}(x)\big) - \sqrt{n}\big(G(x) - F_{\theta_0}(x)\big)$$

converges weakly to a Gaussian process U (Babu and Rao, 2003). In this (nonparametric bootstrap) case, both

$$Y_n^N(x) = \sqrt{n}\big(F_n^*(x) - F_{\theta_n^*}(x)\big) - \sqrt{n}\big(F_n(x) - F_{\hat{\theta}_n}(x)\big)$$

and U_n, converge to the same Gaussian process. For the K-S statistic, for any $0 < \alpha < 1$,

$$P\big(\sqrt{n}\sup_x |F_n(x) - F_{\hat{\theta}_n}(x) - (G(x) - F_{\theta_0}(x))| \leq C_\alpha^*\big) - \alpha \to 0 \quad \text{as } n \to \infty,$$

where C_α^* is the α-th quantile of

$$\sup_x |\sqrt{n}\,\big(F_n^*(x) - F_{\theta_n^*}(x)\big) - \sqrt{n}\big(F_n(x) - F_{\hat{\theta}_n}(x)\big)|$$

Note that the preceding expression is completely dependent on the data $X_1, \ldots, X_n$. This provides an estimate of the distance between the true distribution and the family of distributions under consideration (Babu and Bose, 1988).

3. MODEL SELECTION

The traditional maximum likelihood paradigm, as applied to statistical modeling, provides a mechanism for estimating the unknown parameters of a model having a specified dimension and structure. As the complexity of the model is increased, the model becomes more capable of adapting to the characteristics of the data. Thus, the selection of the fitted model that maximizes the empirical likelihood invariably leads to the selection of the most complex model in the candidate family. Model selection based on the likelihood principle, therefore, requires an extension of the traditional likelihood paradigm.

Grounding in the concept of entropy, Akaike in his seminal 1973 paper considered a framework in which the model dimension is also unknown, and must therefore be determined from the data Akaike (1973, 1974). Akaike proposed an information criterion that is now popularly known as *Akaike's information criterion* (AIC), a framework wherein both model estimation and selection could be simultaneously accomplished. Today, the AIC continues to be the most widely known and used model selection tool among practitioners.

The popular model selection methods such as the AIC, Bayesian information criterion (BIC), and minimum description length (MDL) have been developed, using estimation of bias. The bias arises from estimating the distance between an unknown true model and an estimated model. A recently developed model selection based on bias reduction through a jackknife type procedure (Lee et al., 2012) is discussed here. The jackknife-based procedure is, in particular, applicable to problems of selecting a model from separated families, especially when the true model is unknown. In this issue, Sauerbrei et al. (2011) discuss bootstrap model selection for high-dimensional data and Gunter et al. (2011) describe approaches to model selection using the criterion of choosing variables that provide qualtitative interactions between the variable and treatment with respect to outcome, particularly with respect to medical applications.

To set up the framework for general model selection, let D denote the observed data and let $M_1, \ldots, M_k$ denote the models for D under consideration. For each model M_j, let $f(D \mid \theta_j; M_j)$ and $L(\theta_j) = \log f(D \mid \theta_j; M_j)$ denote the likelihood and log-likelihood respectively, where θ_j is a p_j-dimensional parameter vector. Here $f(\cdot \mid \theta_j; M_j)$ denotes the probability density function generating data D.

The model M_1 is said to be *nested* in M_2 if some coordinates of θ_1 are fixed, i.e., $\theta_2 = (\alpha, \gamma)$ and $\theta_1 = (\alpha, \gamma_0)$, where γ_0 is some known fixed constant vector. In this case, the largest likelihood achievable by M_2 will *always* be larger than that achievable by M_1. This suggests that adding a penalty on "larger" models would achieve a balance between overfitting and underfitting. This leads to the so-called *penalized likelihood approach*.

The AIC finds a model that minimizes the Kullback–Leibler divergence D_{KL} defined in (1). The term $\int g(x) \log g(x) dx$ in D_{KL} is the same for all the models M_j, when the data come from a distribution with density g. The AIC chooses a model that maximizes the estimate of $E_g[\log f(D \mid \hat{\theta}_j; M_j)]$, from which bias becomes unavoidable. Here $\hat{\theta}_j$ is the maximum likelihood estimator of θ_j based on D, the i.i.d. random variables $X_1, \ldots, X_n$. Since g is unknown, the bias is estimated by replacing the distribution corresponding to g with the empirical distribution function. This leads to the definition of the AIC for model M_j as $2L(\hat{\theta}_j) - 2p_j$. The term $2L(\hat{\theta}_j)$ is known as the *goodness-of-fit* term, and $2p_j$ is known as the *penalty* term. The penalty term increase as the complexity of the model grows. The AIC selects the model M_i if $i = \text{argmax}_j(2L(\hat{\theta}_j) - 2p_j)$. That is, the AIC attempts to find the model that best explains the data with a minimum of free parameters.

The AIC can be used to compare nested as well as non-nested models. One of the disadvantages of the AIC is the requirement of large samples especially in complex modeling frameworks. In addition, it is not consistent, in the sense that if p_0 is the correct number of parameters, and $\hat{p} = p_i$ $(i = \text{argmax}_j(2L(\hat{\theta}_j) - 2p_j))$, then $\lim_{n\to\infty} P(\hat{p} > p_0) > 0$. That is, even if we have a very large number of observations, $\hat{p}$ does not approach the "true value".

The BIC, sometimes called the *Schwarz Bayesian criterion*, is another popular model selection criteria. Unlike the AIC, the BIC defined as

$$2L(\hat{\theta}_j) - p_j \log n \tag{2}$$

is consistent. It is derived by giving all the models under consideration equal weights, i.e., equal prior probabilities to all the models under consideration. This leads to selecting the model with highest marginal likelihood. The marginal likelihood, expressed as an integral, is approximated using Laplace's method. This in turn leads to expression (2). Another model evaluation criterion is based on the concept of *minimum description length* (MDL) in transmitting a set of data by coding using a family of probability models. It turns out that the MDL is same as $-\frac{1}{2}\times$ BIC, so these two are equivalent. Like the AIC, the models need not be nested to be compared using the BIC.

Maximum entropy is similar to maximizing the log-likelihood in AIC; however, MDL measures the complexity of a model, which contributes as a penalty term. On the other hand, BIC was developed from the idea of choosing a model of most probable posterior distribution. Unlike the AIC, other criteria such as the BIC and MDL are consistent but parsimonious. Their applications are restrictive depending on conditions. Conditions under which these two criteria are mathematically justified are often ignored in practice. Some practitioners apply them even in situations where they *should not be* applied. The AIC penalizes free parameters less strongly than does the Schwarz's BIC. See Konishi and Kitagawa (2008) for a detailed exposition on model selection methodology. See also Burnham and Anderson (2002).

Regardless of their theoretical and historical background, the BIC, MDL, and other modified model selection criteria that are similar to the AIC are popular among practitioners. Despite the simplicity of the AIC, a few drawbacks were reported: the tendency of picking an overfitted model (Hurvich and Tsai, 1989), and the lack of consistency in selecting the correct model (McQuarrie et al., 1997). As remedies for these shortcomings of the AIC, several information criteria with improved bias estimation were proposed. In order to adjust the inconsistency of the AIC, especially for small sample size, Hurvich and Tsai (1989) and McQuarrie et al. (1997) introduced the corrected Akaike information criterion (AICc) and the unbiased Akaike information criterion (AICu), respectively. The penalty terms in their information criteria estimate the bias more consistently so that the AICc and AICu select the correct model more often than does the AIC. To relax the assumptions on bias estimation, other information criteria were presented, such as the generalized information criterion (GIC; Konishi and Kitagawa, 1996), and the information complexity (ICOMP; Bozdogan, 2000).

3.1. Jackknife Information Criterion

Lee et al. (2012) developed a jackknife information criterion (JIC) based on the KL divergence measure as a competing model selection criterion. To describe the jackknife estimator of the log likelihood, let $X_1, \ldots, X_n$ be i.i.d. random variables with a common unknown density g and $\mathcal{M} = \{f_\theta : \theta \in \Theta\}$ is a class of candidate models. Let the log-likelihood, $\log f_\theta(X_i)$, of X_i be denoted by $l_i(\theta)$. The log-likelihood function of all the observations and the log-likelihood function without the ith observation are denoted by $L(\theta) = \sum_{i=1}^n l_i(\theta)$ and $L_{-i}(\theta) = \sum_{j\neq i} l_j(\theta)$, respectively.

If $\mathcal{M}$ contains g, the parameter value θ_g that maximizes $E_g[\log f_\theta(X_i)]$ satisfies $D_{KL}[g; f_{\theta_g}] = 0$. However, g is likely to be unknown so that the estimate of the density f_{θ_g} that minimizes the KL divergence is sought as a surrogate model of g. The jackknife estimator of the log likelihood is defined as

$$J_n = nL(\hat{\theta}_n) - \sum_{i=1}^n L_{-i}(\hat{\theta}_{-i})$$

where $\hat{\theta}_n$ and $\hat{\theta}_{-i}$ are the maximum likelihood estimators of θ, based on $X_1, \ldots, X_n$ and on $\{X_j : j \neq i; 1 \leq j \leq n\}$, respectively. Under some regularity conditions, Lee, Babu, and Rao established that

$$J_n = nL(\hat{\theta}_n) - \sum_{i=1}^n L_{-i}(\hat{\theta}_{-i}) = L(\theta_g) + \frac{1}{n}\sum_{i\neq j} \nabla l_i(\theta_g)^T \Lambda^{-1} \nabla l_j(\theta_g) + \epsilon_n$$

and

$$J_n = L(\theta_g) + O(\log\log n) \quad \text{a.s.},$$

where $\Lambda = -E_g[\nabla^2_\theta l_i(\theta_g)]$ and ϵ_n is a random variable satisfying $\lim_{n\to\infty} E_g[\epsilon_n] = 0$ and $\|\epsilon_n\| \xrightarrow{a.s} 0$. Since

$$E_g\left(\sum_{i\neq j} \nabla l_i(\theta_g)^T \Lambda^{-1} \nabla l_j(\theta_g)\right) = 0$$

J_n provides an asymptotically unbiased estimator of the log likelihood function. This is in contrast to the cross-validation and the bootstrap estimators of the log-likelihood (Shao, 1993; Chung et al., 1996; Shibata, 1997). Therefore, the model selection criterion based on the jackknife method is expected to perform differently for large sample sizes, compared to the model selection criteria based on other resampling procedures.

Since the popular information criteria-based model selection procedures involve estimators of expected log likelihood multiplied by −2, the jackknife information criterion JIC can be defined as

$$JIC = -2J_n = -2nL(\hat{\theta}_n) + 2\sum_{i=1}^{n} L_{-i}(\hat{\theta}_{-i})$$

Adapting from the data-oriented penalty by Rao and Wu (1989) and Bai et al. (1999), the adjusted JIC (JICa) selects a model consistently, which is defined as

$$JICa = JIC + C_n,$$

with C_n satisfying $\frac{C_n}{n} \to 0$ and $\frac{C_n}{\log\log n} \to \infty$. This is especially useful when the true model is unspecified or misspecified.

ACKNOWLEDGMENT

The work was supported in part by National Science Foundation grant AST-1047586.

REFERENCES

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

Akaike, H. (1974). A new look at statistical model identification. *IEEE Transactions on Automatic Control* 19:716–723.

Babu, G. J., Bose, A. (1988). Bootstrap confidence intervals. *Statistics and Probability Letters* 7:151–160.

Babu, G. J., Rao, C. R. (1993). Bootstrap methodology. In: Rao, C. R. ed. *Handbook of Statistics, Computational Statistics*. Vol. 9. Amsterdam: Elsevier Science Publishers, pp. 627–659.

Babu, G. J., Rao, C. R. (2003). Confidence limits to the distance of the true distribution from a misspecified family by bootstrap. *Journal of Statistical Planning and Inference* 115(2):471–478.

Babu, G. J., Rao, C. R. (2004). Goodness-of-fit tests when parameters are estimated. *Sankhyā* 66(1):63–74.

Bai, Z. D., Rao, C. R., Wu, Y. (1999). Model selection with data-oriented penalty. *Journal of Statistical Planning and Inference* 77:103–117.

Bozdogan, H. (2000). Akaike's information criterion and recent developments in information complexity. *Journal of Mathemathical Psychology* 44(1):62–91.

Burnham, K. P., Anderson, D. R. (2002). *Model Selection and Inference, A Practical Information-Theoretic Approach*. 2nd ed. New York: Springer-Verlag.

Chernick, M. R. (2007). *Bootstrap Methods–A Guide for Practitioners and Researchers*. 2nd ed. New York: Wiley Interscience.

Chung, H., Lee, K., Koo, J. (1996). A note on bootstrap model selection criterion. *Statatistics & Probability Letters* 26:35–41.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *Annals of Statistics* 7(1):1–26.

Gunter, L., Zhu, J., Murphy, S. (2011). Variable selection for qualitative interactions in personalized medicine while controlling the family-wise error rate. *Journal of Biopharmaceutical Statistics* 21(6):1063–1078.

Hurvich, C. M., Tsai, C. (1989). Regression and time series model selection in small samples. *Biometrika* 76(2):297–307.

Konishi, S., Kitagawa, G. (1996). Generalised information criteria in model selection. *Biometrika* 83(4):875–890.

Konishi, S., Kitagawa, G. (2008). *Information Criteria and Statistical Modeling*. New York: Springer Series in Statistics.

Kullback, S., Leibler, R. A. (1951). On information and sufficiency. *Annals of Mathematical Statistics* 22(1):79–86.

Lee, H., Babu, G. J., Rao, C. R. (2012). A jackknife type approach to statistical model selection. *Journal of Statistical Planning and Inference* 142(1):301–311.

Lilliefors, H. W. (1969). On the Kolmogorov–Smirnov test for the exponential distribution with mean unknown. *Journal of the American Statistical Association* 64(325):387–389.

McQuarrie, A., Shumway, R., Tsai, C. (1997). The mdel selection criterion AICu. *Statistics & Probability Letters* 34:285–292.

Press, W. H., Teukolsky, S. A., Vetterling, W. T., Flannery, B. P. (1997). *Numerical Recipes in C: The Art of Scientific Computing*. Cambridge: Cambridge University Press.

Rao, C. R., Wu, Y. (1989). A strongly consistent procedure for model selection in a regression problem. *Biometrika* 76:369–374.

Sauerbrei, W., Boulesteix, A.-L., Binder, H. (2011). Stability investigations of multivariable regression models derived from low and high dimensional data. *Journal of Biopharmaceutical Statistics* 21(6):1206–1231.

Shao, J. (1993). Linear model selection by cross-validation. *Journal of the American Statistical Association* 88:486–494.

Shibata, R. (1997). Bootstrap estimate of Kullback–Leibler information for model selection. *Statistica Sinica* 7:375–394.

Simpson, P. B. (1951). Note on the estimation of a bivariate distribution function. *Annals of Mathematical Statistics* 22:476–478.