Cross-Validation Nonparametric Bootstrap Study of the Linhart-Volkers-Zucchini Out-of-Sample Prediction Error Formula for Logistic Regression Modeling

Richard M. Golden, University of Texas at Dallas, Richardson, TX 75080
Shaurabh Nandy, Foxbat Research, Dallas, TX
Vishal Patel, Foxbat Research, Dallas, TX

Abstract

Cross-validation (CV) methods are widely used to estimate out-of-sample prediction error. In big data problems, analytical formulas are attractive alternatives since CV methods are computationally expensive. If the parameter estimation time is T seconds for a data set with N records, the leave-one-out CV estimation time is TN seconds. Linhart and Volkers (1984: also see Linhart and Zucchini, 1986) showed a particular large sample analytic out-of-sample prediction error estimator was an unbiased estimator of CV estimation error for a large class of smooth empirical risk functions resulting in an estimation time of T rather than TN seconds. This theoretical result is an extension of the Takeuchi Information (Takeuchi, 1976) and Akaike Information (Akaike, 1973) Criteria. We provide easily verifiable assumptions for this theoretical result to hold. In addition, we report empirical results for logistic regression modeling that show the mean relative deviation between a nonparametric bootstrap CV estimator and the analytic out-of-sample prediction error estimator was less than 0.3% for three different data sets with respective sample sizes of $n = 583$, $n = 1728$, and $n = 4898$ records.

KEY WORDS: model selection, cross-validation, GAIC, generalized Akaike Information Criterion, AIC, Akaike Information Criterion, TIC, Takeuchi Information Criterion, bootstrap, out-of-sample prediction error

1. Introduction

Let $\mathbf{x}_1, \ldots, \mathbf{x}_n$ be a collection of n data records. $\mathcal{D}_n \equiv [\mathbf{x}_1, \ldots, \mathbf{x}_n]$ denote a data set where $\mathbf{x}_1, \ldots, \mathbf{x}_n$ are d-dimensional vectors corresponding to the n data records in $\mathcal{D}_n$. Assume, in addition, that $\mathcal{D}_n$ is a particular realization of the random vector $\tilde{\mathcal{D}}_n \equiv [\tilde{\mathbf{x}}_1, \ldots, \tilde{\mathbf{x}}_n]$ which is comprised of n independent and identically distributed observations with common Radon-Nikodým probability density $p_e : \mathcal{R}^d \rightarrow [0, \infty)$ with respect to some sigma-finite measure ν. The researcher has M probability models $\mathcal{M}_1, \ldots, \mathcal{M}_M$ where the kth probability model

$$\mathcal{M}_k \equiv \{p_k(\mathbf{x}; \boldsymbol{\theta}_k) : \boldsymbol{\theta}_k \in \Theta_k\}$$

is a set of Radon-Nikodým probability densities defined with respect to ν.

The *model selection problem* involves determining which of the models $\mathcal{M}_1, \ldots, \mathcal{M}_M$ is the most appropriate model of the data generating process which generated $\tilde{\mathcal{D}}_n$. To achieve this goal, the performance of a model with respect to a particular statistical environment is often mapped into a number called a Model Selection Criterion (MSC). Once the MSC for each of the M models is computed, then the set of models can be rank ordered or the model with the smallest MSC can be chosen.

The CVRC (Cross-Validation Risk Criterion) out-of-sample prediction error is designed to characterize the generalization performance of a fitted probability model. Specifically, given that the fitted probability model was trained on one data set $\mathcal{D}_n^1$, how will the average prediction error change when the fitted model is tested on a second data set $\mathcal{D}_n^2$?

1.1 Cross Validation Risk Model Selection Criteria

Consider the problem of evaluating the prediction error of a model as a function of sample size n. Let $\tilde{\mathbf{x}}_1, \tilde{\mathbf{x}}_2, \ldots$ be a stochastic sequence of independent and identically distributed d-dimensional random vectors. Let the first n random vectors in the stochastic sequence be denoted as $\tilde{\mathcal{D}}_n^1$, the second n random vectors in the stochastic sequence be denoted as $\tilde{\mathcal{D}}_n^2$, and so on.

Let $\hat{\boldsymbol{\theta}}_n^1$ be a strict local minimizer of the empirical risk function

$$\ell_n(\tilde{\mathcal{D}}_n^1, \cdot) = (1/n) \sum_{i=1}^{n} c(\tilde{\mathbf{x}}_i, \cdot)$$

where $c : \mathcal{R}^{d \times n} \times \Theta \to \mathcal{R}$ is the loss function. The random variable $\ell_n \left(\tilde{\mathcal{D}}_n^1, \hat{\boldsymbol{\theta}}_n^1 \right)$ is called the *training data model fit*. The training data model fit is a measure of how well the model fits the data sample $\tilde{\mathcal{D}}_n^1$. The training data model fit, $\ell_n \left(\tilde{\mathcal{D}}_n^1, \hat{\boldsymbol{\theta}}_n^1 \right)$, is a biased estimator for finite sample sizes due to an "overfitting phenomenon" which arises because the same data set was used to estimate both the parameter estimates and the fit of the model to the data.

Now suppose that an additional random sample $\tilde{\mathcal{D}}_n^2$ is available. The random variable $\ell_n \left(\tilde{\mathcal{D}}_n^2, \hat{\boldsymbol{\theta}}_n^1 \right)$ is called the *test data model fit* because the model's parameters are estimated using $\tilde{\mathcal{D}}_n^1$ yet the fit of the model is evaluated on the *out-of-sample* data sample $\tilde{\mathcal{D}}_n^2$. Furthermore, the expected value of the test data model fit provides an unbiased estimate of model fit. A natural extension of this methodology is to use K-fold cross-validation or nonparametric bootstrap simulation estimation methods to estimate the expected value of $\ell_n \left(\tilde{\mathcal{D}}_n^2, \hat{\boldsymbol{\theta}}_n^1 \right)$.

Let $\boldsymbol{\theta}_n : \mathcal{R}^{d \times n} \to \Theta$ be a function defined such that

$$\boldsymbol{\theta}_n(\mathcal{D}_n) = \underset{\boldsymbol{\theta} \in \Theta}{\arg\min}\, \ell_n \left(\mathcal{D}_n, \boldsymbol{\theta} \right)$$

where Θ is some sufficiently small neighborhood of a strict local minimizer, $\boldsymbol{\theta}^*$, of the expected risk

$$\ell(\boldsymbol{\theta}) = \int \ell_n \left(\mathcal{D}_n^2, \boldsymbol{\theta} \right) p_e \left(\mathcal{D}_n^2 \right) d\nu \left(\mathcal{D}_n^2 \right).$$

Since $\tilde{\mathcal{D}}_n^1$ and $\tilde{\mathcal{D}}_n^2$ are statistically independent and using the notation $\hat{\boldsymbol{\theta}}_n^1 = \boldsymbol{\theta}_n \left(\tilde{\mathcal{D}}_n^1 \right)$ it follows that:

$$\begin{aligned} E \left\{ \ell_n \left(\tilde{\mathcal{D}}_n^2, \hat{\boldsymbol{\theta}}_n^1 \right) \right\} &= \int \ell_n \left(\mathcal{D}_n^2, \boldsymbol{\theta}_n \left(\mathcal{D}_n^1 \right) \right) p_e(\mathcal{D}_n^1) p_e(\mathcal{D}_n^2) d\nu \left(\mathcal{D}_n^1, \mathcal{D}_n^2 \right) \\ &= \int \left[\int \ell_n \left(\mathcal{D}_n^2, \boldsymbol{\theta}_n \left(\mathcal{D}_n^1 \right) \right) p_e \left(\mathcal{D}_n^2 \right) d\nu \left(\mathcal{D}_n^2 \right) \right] p_e \left(\mathcal{D}_n^1 \right) d\nu \left(\mathcal{D}_n^1 \right) \end{aligned}$$

which can be rewritten as:

$$E\left\{\ell_n\left(\tilde{\mathcal{D}}_n^2, \hat{\boldsymbol{\theta}}_n^1\right)\right\} = \int \ell\left(\boldsymbol{\theta}_n\left(\mathcal{D}_n^1\right)\right) p_e\left(\mathcal{D}_n^1\right) d\nu\left(\mathcal{D}_n^1\right) = E\left\{\ell\left(\hat{\boldsymbol{\theta}}_n^1\right)\right\}. \tag{1}$$

Equation (1) provides an explicit expression for the expected novel test data fit,

$$E\left\{\ell\left(\hat{\boldsymbol{\theta}}_n^1\right)\right\} = E\left\{\ell_n\left(\tilde{\mathcal{D}}_n^2, \hat{\boldsymbol{\theta}}_n^1\right)\right\},$$

using the training data parameter estimates $\hat{\boldsymbol{\theta}}_n^1$. The following theorem shows how to calculate the expected novel test data fit from the expected training data fit $E\left\{\hat{\ell}_n\left(\tilde{\mathcal{D}}_n^1, \hat{\boldsymbol{\theta}}_n^1\right)\right\}$. The theorem was originally stated and proved in Linhart and Volkers (1984; also see Linhart and Zucchini, 1986). The estimated generalization/overfitting bias obtained from theoretical analyses of the Akaike Information Criterion (Akaike 1973) and the Generalized Akaike Information Criterion (Takeuchi 1976) are special cases of this theorem as well.

Theorem 1.1 (Empirical Risk Overfitting Bias (Linhart and Volkers 1984; Linhart and Zucchini 1986))**.** *Assume the Empirical Risk Regularity Conditions (see Section 3.1) hold with respect to a DGP Radon-Nikodým density $p_e : \mathcal{R}^d \to [0, \infty)$, loss function $c : \mathcal{R}^d \times \mathcal{R}^q \to \mathcal{R}$, and risk function $\ell : \mathcal{R}^q \to \mathcal{R}$ defined such that for all $\boldsymbol{\theta} \in \mathcal{R}^q$:*

$$\ell(\boldsymbol{\theta}) \equiv \int c(\mathbf{x}, \boldsymbol{\theta}) p_e(\mathbf{x}) d\nu(\mathbf{x}).$$

In addition, let Θ be a closed, bounded, and convex subset of $\mathcal{R}^q$ containing a strict local minimizer, $\boldsymbol{\theta}^*$, of ℓ. Assume $\boldsymbol{\theta}^*$ is the only critical point of ℓ in Θ.

Assume $\hat{\boldsymbol{\theta}}_n \equiv \arg\min \hat{\ell}_n(\boldsymbol{\theta})$ on Θ with probability one for all $n \in \mathbb{N}$ where

$$\hat{\ell}_n(\boldsymbol{\theta}) = (1/n) \sum_{i=1}^{n} c(\tilde{\mathbf{x}}_i, \boldsymbol{\theta}).$$

Let $\hat{\mathbf{A}}_n$ and $\hat{\mathbf{B}}_n$ be defined as in (5) and (6) respectively.
Then, as $n \to \infty$,

$$E\{\ell(\hat{\boldsymbol{\theta}}_n)\} = E\{\hat{\ell}_n(\hat{\boldsymbol{\theta}}_n)\} + (1/n) \operatorname{tr}\left([\hat{\mathbf{A}}_n]^{-1} \hat{\mathbf{B}}_n\right) + o_p(1/n).$$

Definition 1.1 (Cross-Validation Risk Criterion (CVRC))**.** Let $\hat{\ell}_n$, $\hat{\boldsymbol{\theta}}_n$, $\hat{\mathbf{A}}_n$, and $\hat{\mathbf{B}}_n$ be defined as in the Empirical Risk Overfitting Bias Theorem (Theorem 1.1). The *Cross-Validation Risk Criterion* (CVRC) is defined as:

$$\text{CVRC} = \hat{\ell}_n(\hat{\boldsymbol{\theta}}_n) + (1/n) \operatorname{tr}\left([\hat{\mathbf{A}}_n]^{-1} \hat{\mathbf{B}}_n\right). \tag{2}$$

The CVRC is a consistent estimator of the expected value of $\ell(\hat{\boldsymbol{\theta}}_n)$. That is, it estimates how effectively a model fitted to training data sample $\mathbf{x}_1, \ldots, \mathbf{x}_n$ will perform on a test data sample under the assumption that the observations from both samples are independent and identically distributed with common density p_e. The more familiar GAIC (Generalized Akaike Information Criterion) also known as the TIC (Takeuchi Information Criterion) is equal to the CVRC multiplied by twice the sample size when the empirical risk function is a negative normalized log-likelihood function. Thus, the GAIC/TIC is a special case of CVRC and AIC is a special case of GAIC/TIC. Reviews of the Takeuchi Information Criterion (Takeuchi 1976) may be found in Bozdogan (2000), Claeskens and Hjort (2008), and Konishi and Kitagawa (2008).

Definition 1.2 (Generalized Akaike Information Criterion (GAIC))**.** Let $\hat{\ell}_n$, $\hat{\boldsymbol{\theta}}_n$, $\hat{\mathbf{A}}_n$, and $\hat{\mathbf{B}}_n$ be defined as in the Empirical Risk Overfitting Bias Theorem. In addition, assume that $\hat{\ell}_n$ is a negative normalized log-likelihood function defined with respect to some probability model $\mathcal{M}$ and data generating density p_e. The *Generalized Akaike Information Criterion* (GAIC) is defined as:

$$\text{GAIC} = 2n\hat{\ell}_n(\hat{\boldsymbol{\theta}}_n) + 2\,\text{tr}\left([\hat{\mathbf{A}}_n]^{-1}\hat{\mathbf{B}}_n\right). \tag{3}$$

In the special case where the probability model $\mathcal{M}$ is correctly specified with respect to the data generating density p_e (i.e., $p_e \in \mathcal{M}$), then the GAIC reduces as a special case to the popular Akaike Information Criterion (AIC) (Akaike 1973, 1974; Stone 1977; Bozdogan 2000; Claeskens and Hjort 2008; Konishi and Kitagawa 2008).

Definition 1.3 (Akaike Information Criterion (AIC))**.** Let $\hat{\ell}_n$, $\hat{\boldsymbol{\theta}}_n$, and q be defined as in the Empirical Risk Overfitting Bias Theorem. In addition, assume that $\hat{\ell}_n$ is a negative normalized log-likelihood function defined with respect to some probability model $\mathcal{M}$ and data generating density p_e. Assume, in addition, that $\mathcal{M}$ is correctly specified with respect to p_e. Then, the *Akaike Information Criterion* (AIC) is defined as:

$$\text{AIC} = 2n\hat{\ell}_n(\hat{\boldsymbol{\theta}}_n) + 2q. \tag{4}$$

2. Simulation Studies

The purpose of these data analyses was to empirically investigate whether the estimated CVRC out-of-sample prediction error bias obtained using the CVRC formula was comparable in magnitude to the estimated out-of-sample prediction error bias obtained using a nonparametric bootstrap sampling methodology.

2.1 Data Sets

Three data sets ("car", n=1728; "liver", n= 583; "white wine", n=4898) were downloaded from the UCI machine learning data repository (archive.ics.uci.edu/ml/) and prepared for logistic regression modeling by rescaling numerical predictors and removing redundant predictors.

2.2 Methods

The following nonparametric bootstrap sampling methodology was used to evaluate the performance of the CVRC out-of-sample prediction error for each of the three data sets.

- **Step 1.** Select one of the three training data sets. Assume that training set consists of n records.
- **Step 2.** Compute the in-sample prediction error using training data.
- **Step 3.** Use CVRC formula to estimate out-of-sample prediction error using training data.
- **Step 4.** Let K = 1.

- **Step 5.** Sample with replacement n times from original training data set to create bootstrap training data set of n records.
- **Step 6.** Estimate the bootstrap in-sample prediction error on the bootstrap training data set.
- **Step 7.** Estimate the bootstrap out-of-sample prediction error on a new bootstrap test data set using bootstrap training data set parameters.
- **Step 8.** Let K = K + 1. Stop if $K > K_{max}$; otherwise Go To Step 5.

2.3 Data Analyses

The number of pairs of bootstrap data sets generated for each of the three data sets was approximately $K_{max} = 550$ where the first member of the pair was designated as a "bootstrap" training data set and the second member of the pair was designated as a "bootstrap" test data set. The bootstrap prediction error bias was then estimated for each bootstrap sample pair by subtracting the bootstrap in-sample prediction error in Methods Step 6 from the bootstrap out-of-sample prediction error in Methods Step 7. Next, the first K bootstrap prediction error biases were then averaged to obtain the bootstrap prediction error bias after parameter estimation was completed using training data from the first K bootstrap data sets. The bootstrap out-of-sample prediction error was then computed by adding the in-sample prediction error (Methods Step 2) to the bootstrap prediction error bias.

This procedure was repeated five times using five different random number seeds so that the sampling with replacement algorithm would generate five different bootstrap sampling outcomes for each of the three data sets. Thus, results from 15 distinct simulation runs were obtained.

2.4 Results

Figures 1, 2, 3, 4, and 5 show for the CAR data set (n = 1728) the in-sample prediction error (Training Data) computed in Methods Step 2 and CVRC out-of-sample prediction error computed in Methods Step 3 (Analytic Formula). These quantities are not functionally dependent upon the number of bootstrap data sets. In addition, the average bootstrap out-of-sample prediction error (Bootstrap Sampling) was computed by adding the in-sample prediction error on the training data obtained in Methods Step 2 to the bootstrap estimator of the out-of-sample prediction error bias. For the CAR data set, the mean relative deviation between the CVRC out-of-sample prediction error and the bootstrap out-of-sample prediction error was 0.2% across the five CAR simulation runs after all bootstrap data sets were used.

Figures 6, 7, 8, 9, and 10 show for the Liver data set (n = 583) that the mean relative deviation was 0.2% between the CVRC out-of-sample prediction error (Analytic Formula) and the bootstrap out-of-sample prediction error (Bootstrap Sampling) when the bootstrap out-of-sample prediction error incorporated all bootstrap data sets.

Figures 11, 12, 13, 14, and 15 show for the white wine data set (n = 4898) that the mean relative deviation was 0.04% between the CVRC out-of-sample prediction error (Analytic Formula) and the bootstrap out-of-sample prediction error (Bootstrap Sampling) when the bootstrap out-of-sample prediction error incorporated all bootstrap data sets.

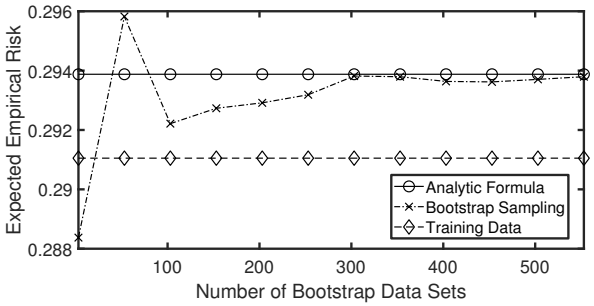

Figure 1: *Prediction error as a function of bootstrap data sets for CAR data set* ($n = 1728$) *(1st Random Seed).*

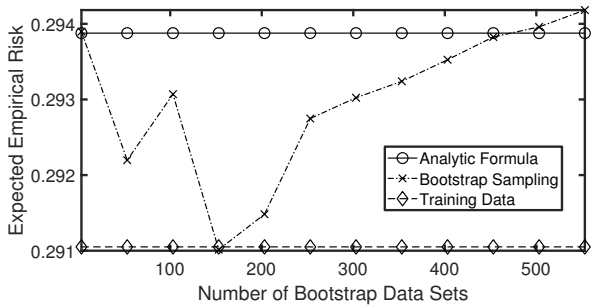

Figure 2: *Prediction error as a function of bootstrap data sets for CAR data set* ($n = 1728$) *(2nd Random Seed).*

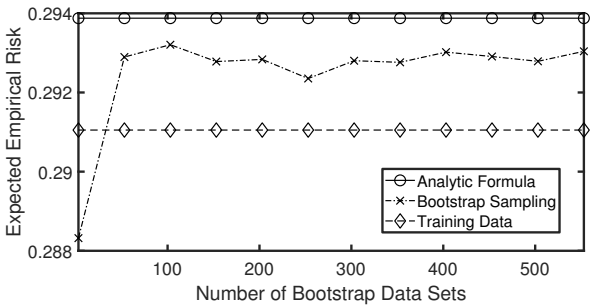

Figure 3: *Prediction error as a function of bootstrap data sets for CAR data set* ($n = 1728$) *(3rd Random Seed).*

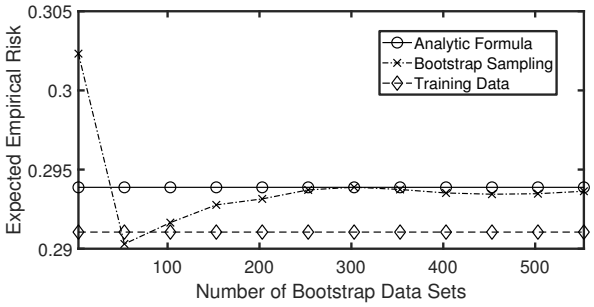

Figure 4: *Prediction error as a function of bootstrap data sets for CAR data set* ($n = 1728$) *(4th Random Seed).*

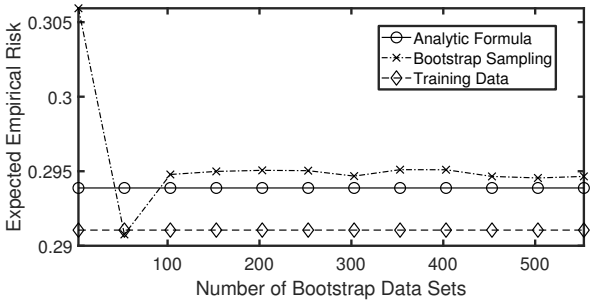

Figure 5: *Prediction error as a function of bootstrap data sets for CAR data set* ($n = 1728$) *(5th Random Seed).*

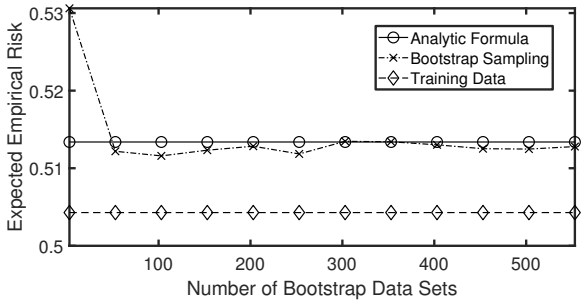

Figure 6: *Prediction error as a function of bootstrap data sets for LIVER data set* ($n = 583$) *(1st Random Seed).*

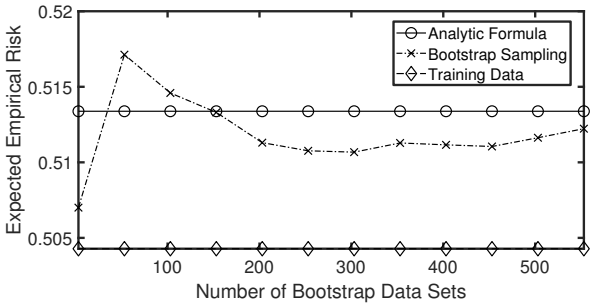

Figure 7: *Prediction error as a function of bootstrap data sets for LIVER data set* ($n = 583$) *(2nd Random Seed).*

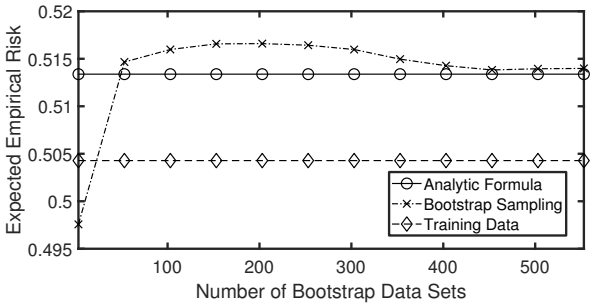

Figure 8: *Prediction error as a function of bootstrap data sets for LIVER data set* ($n = 583$) *(3rd Random Seed).*

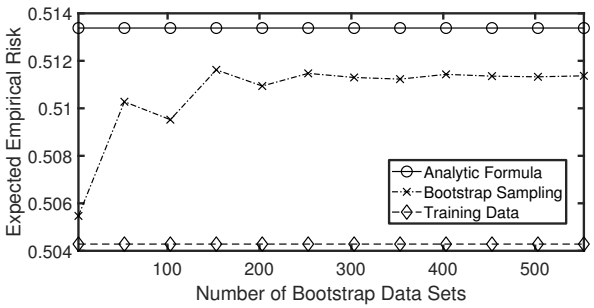

Figure 9: *Prediction error as a function of bootstrap data sets for LIVER data set* ($n = 583$) *(4th Random Seed).*

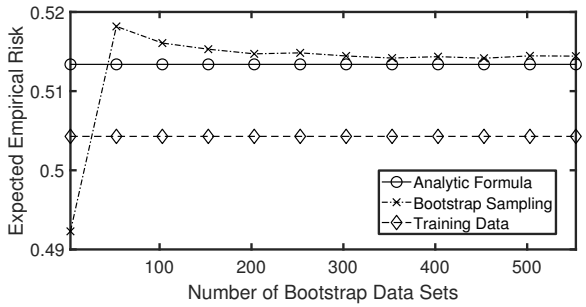

Figure 10: *Prediction error as a function of bootstrap data sets for LIVER data set* ($n = 583$) *(5th Random Seed).*

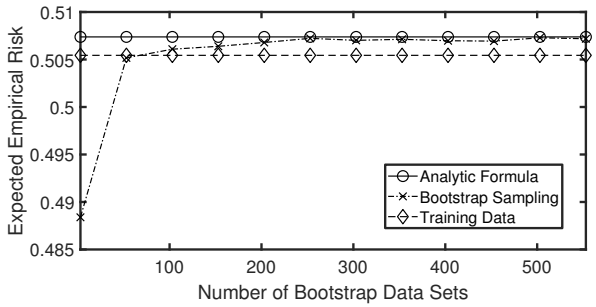

Figure 11: *Prediction error as a function of bootstrap data sets for WHITE-WINE data set* ($n = 4898$) *(1st Random Seed).*

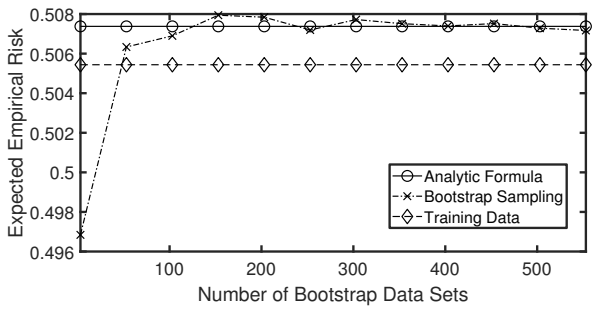

Figure 12: *Prediction error as a function of bootstrap data sets for WHITE-WINE data set* ($n = 4898$) *(2nd Random Seed).*

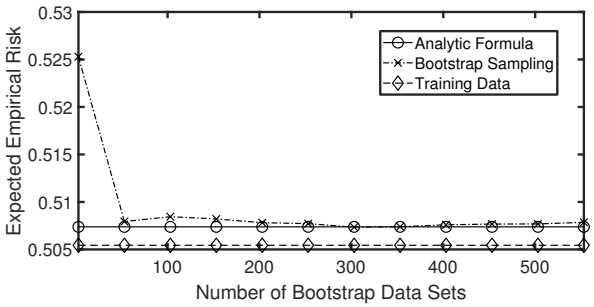

Figure 13: *Prediction error as a function of bootstrap data sets for WHITE-WINE data set* $(n = 4898)$ *(3rd Random Seed).*

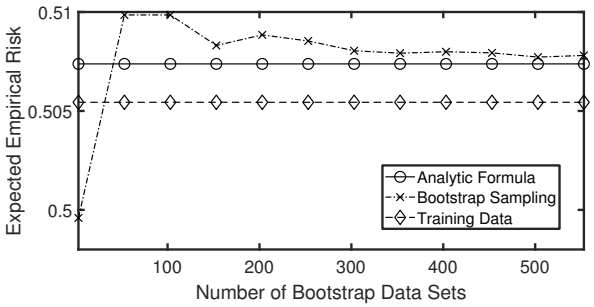

Figure 14: *Prediction error as a function of bootstrap data sets for WHITE-WINE data set* $(n = 4898)$ *(4th Random Seed).*

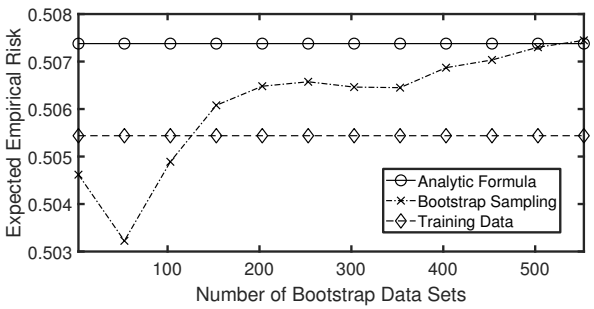

Figure 15: *Prediction error as a function of bootstrap data sets for WHITE-WINE data set* $(n = 4898)$ *(5th Random Seed).*

2.5 Discussion

These simulation results support the consideration of the CVRC out-of-sample prediction error formula in logistic regression modeling applications involving large data sets where bootstrap simulation methods for estimating out-of-sample prediction error are not practical. Moreover, these simulation results also illustrate the applicability of the general theory. Future research in this area should use similar simulation methodologies to empirically evaluate the use of the CVRC out-of-sample prediction error formula for other applications such as linear regression, multinomial logistic regression, Gaussian mixture models, and structural equation modeling applications.

3. Technical Details of Theorems and Proofs

As previously noted, the theorems and proofs presented here are due to Linhart and Volkers (1984) and Linhart and Zucchini (1986) (also see analyses by Akaike 1973 and Takeuchi 1976). However, the statements of these theorems and the choices of specific methods for presenting the regularity assumptions have been modified to encourage more wide-spread use in engineering practice. In addition, details of supporting mathematical arguments are provided here.

3.1 Regularity Assumptions

Let $\Omega \subseteq \mathcal{R}^d$. Let $\Theta \subseteq \mathcal{R}^q$. The function $c : \Omega \times \Theta \rightarrow \mathcal{R}$ is called a *continuously differentiable random function* on Θ if $c(\mathbf{x}, \dot{)}$ is continuously differentiable on Θ for each $\mathbf{x} \in \Omega$ and $c(\cdot, \boldsymbol{\theta})$ is Borel measurable for each $\boldsymbol{\theta} \in \Theta$. Note that if f is a continuous function, then f is a Borel measurable function.

A random function $c : \Omega \times \Theta \rightarrow \mathcal{R}$ is *dominated by an integrable function* on Θ with respect to a Radon-Nikodým density $p : \Omega \rightarrow [0, \infty)$ with sigma-finite measure ν if (i) there exists a function $D : \Omega \rightarrow \mathcal{R}$ such that $|c(\mathbf{x}, \boldsymbol{\theta})| \leq D(\mathbf{x})$ for all $\boldsymbol{\theta} \in \Theta$ and for all $\mathbf{x}$ in the support of p_e, and (ii) $\int D(\mathbf{x})p(\mathbf{x})d\nu(\mathbf{x})$ is finite.

For example, if c is continuous in both arguments and $p : \Omega \rightarrow [0, 1]$ is a probability mass function on a finite sample space Ω then these conditions are sufficient to ensure c is dominated by an integrable function on Θ with respect to p.

- **A1 : I.I.D. Data Generating Process.** Let the data sample $\tilde{\mathcal{D}}_n \equiv [\tilde{\mathbf{x}}_1, \ldots, \tilde{\mathbf{x}}_n]$ be a sequence of independent and identically distributed d-dimensional random vectors with common Radon-Nikodým density $p_e : \mathcal{R}^d \rightarrow [0, \infty)$ defined with respect to sigma-finite measure ν.

- **A2: Smooth Loss Function.** Assume the function $c : \mathcal{R}^d \times \mathcal{R}^q \rightarrow \mathcal{R}$ is a twice continuously differentiable random function on $\mathcal{R}^q$.

- **A3: Expectations are Finite.** Let $\mathbf{g}(\mathbf{x}, \boldsymbol{\theta}) = [dc(\mathbf{x}, \boldsymbol{\theta})/d\boldsymbol{\theta}]^T$. Assume the functions c, $\mathbf{g}\mathbf{g}^T$, and $d^2c/d\boldsymbol{\theta}^2$ are dominated by integrable functions on $\mathcal{R}^q$ with respect to p_e.

- **A4: Construct a Restricted Parameter Space Θ containing a Local Minimizer.** Let $\boldsymbol{\theta}^*$ be a strict local minimizer of

$$\ell(\boldsymbol{\theta}) = \int c(\mathbf{x}, \boldsymbol{\theta})p_e(\mathbf{x})d\nu(\mathbf{x})$$

on $\mathcal{R}^q$. Assume Θ is a closed, bounded, and convex subset of $\mathcal{R}^q$ defined such that $\boldsymbol{\theta}^*$ is in the interior of Θ. In addition, assume $\boldsymbol{\theta}^*$ is the only critical point of ℓ in Θ.

- **A5: Parameter Estimation Procedure.** Assume $\hat{\boldsymbol{\theta}}_n \equiv \arg\min \hat{\ell}_n(\boldsymbol{\theta})$ on Θ with probability one for all $n \in \mathbb{N}$ where

$$\hat{\ell}_n(\boldsymbol{\theta}) = (1/n)\sum_{i=1}^{n} c(\tilde{\mathbf{x}}_i, \boldsymbol{\theta}).$$

- **A6: Hessian Positive Definite at Minimizer.** Let $\mathbf{A}$ be the Hessian of ℓ. Assume $\mathbf{A}^* \equiv \mathbf{A}(\boldsymbol{\theta}^*)$ is positive definite.

- **A7: Outer Product Gradient Positive Definite at Minimizer.** Let the matrix-valued function $\mathbf{B} : \mathcal{R}^q \to \mathcal{R}^{q\times q}$ be defined such that for all $\boldsymbol{\theta} \in \mathcal{R}^q$:

$$\mathbf{B}(\boldsymbol{\theta}) = \int g(\mathbf{x}, \boldsymbol{\theta}) g(\mathbf{x}, \boldsymbol{\theta})^T p_e(\mathbf{x}) d\nu(\mathbf{x}).$$

  Assume $\mathbf{B}^* \equiv \mathbf{B}(\boldsymbol{\theta}^*)$ is positive definite.

Let $\tilde{\mathbf{g}}_i^T \equiv \mathbf{g}(\tilde{\mathbf{x}}_i, \hat{\boldsymbol{\theta}}_n)$. Let the notation

$$\hat{\mathbf{A}}_n \equiv (1/n)\sum_{i=1}^{n} \nabla^2 \tilde{c}_i(\hat{\boldsymbol{\theta}}_n), \tag{5}$$

$$\hat{\mathbf{B}}_n \equiv (1/n)\sum_{i=1}^{n} \tilde{\mathbf{g}}_i \tilde{\mathbf{g}}_i^T, \tag{6}$$

and

$$\hat{\mathbf{C}}_n \equiv \hat{\mathbf{A}}_n^{-1} \hat{\mathbf{B}}_n \hat{\mathbf{A}}_n^{-1}. \tag{7}$$

3.2 Proofs of Key Theorems

Lemma 3.1 (GAIC Lemma)**.** *Let $\Theta \subseteq \mathcal{R}^q$. Assume the Regularity Conditions of the previous section hold with respect to a DGP Radon-Nikodym density $p_e : \mathcal{R}^d \to [0, \infty)$, loss function $c : \mathcal{R}^d \times \Theta \to \mathcal{R}$, and empirical risk function $\hat{\ell}_n : \mathcal{R}^{dn} \times \Theta \to \mathcal{R}$ defined such that for all $\boldsymbol{\theta} \in \Theta$:*

$$\hat{\ell}_n(\boldsymbol{\theta}) = (1/n)\sum_{i=1}^{n} c(\tilde{\mathbf{x}}_i, \boldsymbol{\theta}).$$

Let $\hat{\boldsymbol{\theta}}_n$ be the unique strict global minimizer of the restriction of $\hat{\ell}_n$ to Θ. Then,

$$E\{(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^*)^T \mathbf{A}^* (\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^*)\} = \frac{\mathrm{tr}\left([\mathbf{A}^*]^{-1}\mathbf{B}^*\right)}{n}. \tag{8}$$

Proof of GAIC Lemma.

$$E\{(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^*)^T \mathbf{A}^* (\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^*)\} = \mathrm{tr}(E\left\{[\mathbf{A}^*]^{1/2}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^*)\left([\mathbf{A}^*]^{1/2}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^*)\right)^T\right\}$$

$$= \mathrm{tr}([\mathbf{A}^*]^{1/2} E\left\{(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^*)(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^*)^T\right\}[\mathbf{A}^*]^{1/2}). \tag{9}$$

Using the assumed regularity assumptions and the asymptotic properties of M-estimators:

$$E\left\{(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^*)(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^*)^T\right\} = (1/n)[\mathbf{A}^*]^{-1}\mathbf{B}^*[\mathbf{A}^*]^{-1}$$

and substitute this relation into (9) to obtain:

$$E\{(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^*)^T\mathbf{A}^*(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^*)\} = (1/n)[\mathbf{A}^*]^{1/2}[\mathbf{A}^*]^{-1}\mathbf{B}^*[\mathbf{A}^*]^{-1}[\mathbf{A}^*]^{1/2}. \quad (10)$$

Now use the result that the trace operator is invariant under cyclic permutations to obtain the relation:

$$\operatorname{tr}\left([\mathbf{A}^*]^{1/2}[\mathbf{A}^*]^{-1}\mathbf{B}^*[\mathbf{A}^*]^{-1}[\mathbf{A}^*]^{1/2}\right) = \operatorname{tr}\left([\mathbf{A}^*]^{-1}\mathbf{B}^*\right). \quad (11)$$

Substitute (11) into the right-hand size of (10) to complete the proof. □

Proof of Main Theorem. The proof follows the approach of Konishi and Kitagawa (2008; pp. 55-58) and Linhart and Zucchini (1986, Appendix A).

$$\ell(\hat{\boldsymbol{\theta}}_n) = \hat{\ell}_n(\hat{\boldsymbol{\theta}}_n) + \left(\ell(\hat{\boldsymbol{\theta}}_n) - \ell(\boldsymbol{\theta}^*)\right) + \left(\ell(\boldsymbol{\theta}^*) - \hat{\ell}_n(\boldsymbol{\theta}^*)\right) + \left(\hat{\ell}_n(\boldsymbol{\theta}^*) - \hat{\ell}_n(\hat{\boldsymbol{\theta}}_n)\right). \quad (12)$$

Step 1: Parameter Estimation Error with True Risk

Let $\delta_A(\lambda_n) \equiv \mathbf{A}(\ddot{\boldsymbol{\theta}}_n(\lambda_n)) - \mathbf{A}(\boldsymbol{\theta}^*)$. The second term, $\ell(\hat{\boldsymbol{\theta}}_n) - \ell(\boldsymbol{\theta}^*)$ on the right-hand side of (12) corresponds to the effects of parameter estimation error using the true risk function. To estimate the value of $\ell(\hat{\boldsymbol{\theta}}_n) - \ell(\boldsymbol{\theta}^*)$, expand ℓ about $\boldsymbol{\theta}^*$ and evaluate at $\hat{\boldsymbol{\theta}}_n$ to obtain:

$$\ell(\hat{\boldsymbol{\theta}}_n) = \ell(\boldsymbol{\theta}^*) + \nabla\ell(\boldsymbol{\theta}^*) + (1/2)(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^*)\mathbf{A}^*(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^*) + \tilde{R}_n \quad (13)$$

where

$$\tilde{R}_n \equiv (1/2)(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^*)\delta_A(\lambda_n)(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^*),$$

and $\ddot{\boldsymbol{\theta}}_n(\lambda_n) \equiv \boldsymbol{\theta}^* + \lambda_n(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^*)$ with $\lambda_n \in [0, 1]$. Using the standard M-estimation result that $\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^*)$ converges in distribution to a random vector it follows that $\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^*) = O_p(1)$. And since $\mathbf{A}$ is continuous and $\hat{\boldsymbol{\theta}}_n \to \boldsymbol{\theta}^*$ with probability one implies that $\mathbf{A}(\hat{\boldsymbol{\theta}}_n) \to \mathbf{A}^*$ with probability one as $n \to \infty$, it follows that $\delta_A(\tilde{\lambda}_n) = o_p(1)$ for n sufficiently large. Thus,

$$\tilde{R}_n = O_p(n^{-1/2})o_p(1)O_p(n^{-1/2}) = o_p(1/n).$$

Now take the expectation of (13), using the GAIC Lemma Equation (8), and noting that $\nabla\ell(\boldsymbol{\theta}^*)$ vanishes one obtains:

$$E\{\ell(\hat{\boldsymbol{\theta}}_n)\} = \ell(\boldsymbol{\theta}^*) + (1/2)(1/n)\operatorname{tr}\left([\mathbf{A}^*]^{-1}\mathbf{B}^*\right) + o_p(1/n). \quad (14)$$

Step 2: Empirical Risk Approximation Error

The third term on the right-hand side of (12), $\ell(\boldsymbol{\theta}^*) - \hat{\ell}_n(\boldsymbol{\theta}^*)$, estimates the effects of estimating the true risk function ℓ using the empirical risk function $\hat{\ell}_n$ and evaluating at a true local risk minimizer $\boldsymbol{\theta}^*$. Since $\hat{\ell}_n$ is dominated by an integrable function

$$E\{\hat{\ell}_n(\boldsymbol{\theta}^*)\} = \ell(\boldsymbol{\theta}^*). \quad (15)$$

Step 3: Combined Parameter and Risk Function Error

The fourth term on the right-hand side of (12), $\hat{\ell}_n(\boldsymbol{\theta}^*) - \hat{\ell}_n(\hat{\boldsymbol{\theta}}_n)$, corresponds to the effects of parameter estimation error with respect to the empirical risk function. To estimate the value of $\hat{\ell}_n(\boldsymbol{\theta}^*) - \hat{\ell}_n(\hat{\boldsymbol{\theta}}_n)$, expand $\hat{\ell}_n$ in a second order Taylor expansion about $\boldsymbol{\theta}^*$ to obtain:

$$\hat{\ell}_n(\hat{\boldsymbol{\theta}}_n) = \hat{\ell}_n(\boldsymbol{\theta}^*) + (\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^*)^T \nabla \hat{l}_n(\boldsymbol{\theta}^*) + \frac{(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^*)^T \nabla^2 \hat{\ell}_n(\ddot{\boldsymbol{\theta}}_n(\lambda))(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^*)}{2} \quad (16)$$

where for some $\lambda \in [0, 1]$:

$$\ddot{\boldsymbol{\theta}}_n(\lambda) \equiv \hat{\boldsymbol{\theta}}_n + \lambda(\hat{\boldsymbol{\theta}}_n - \hat{\boldsymbol{\theta}}_n).$$

Now expand $\nabla \hat{\ell}_n$ in a first order Taylor expansion about $\boldsymbol{\theta}^*$ to obtain:

$$\nabla \hat{\ell}_n(\hat{\boldsymbol{\theta}}_n) = \nabla \hat{\ell}_n(\boldsymbol{\theta}^*) + \nabla^2 \hat{\ell}_n(\ddot{\boldsymbol{\theta}}_n(\boldsymbol{\eta})(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^*) \quad (17)$$

where $\ddot{\boldsymbol{\theta}}_n(\boldsymbol{\eta})$ is a vector defined such that its kth element

$$\ddot{\theta}_{n,k}(\eta_k) \equiv \hat{\theta}_{n,k} + \eta_k(\theta_k^* - \hat{\theta}_{n,k})$$

where $\eta_k \in [0, 1]$, θ_k^* is the kth element of $\boldsymbol{\theta}^*$, and $\hat{\theta}_{n,k}$ is the kth element of $\hat{\boldsymbol{\theta}}_n$.

Since, by definition, $\nabla \hat{\ell}_n(\hat{\boldsymbol{\theta}}_n)$ is a vector of zeros, (17) becomes:

$$\nabla \hat{\ell}_n(\boldsymbol{\theta}^*) = -\nabla^2 \hat{\ell}_n \left(\ddot{\boldsymbol{\theta}}_n^{\boldsymbol{\eta}} \right) (\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^*). \quad (18)$$

Substitute (18) into (16) to obtain:

$$\hat{\ell}_n(\hat{\boldsymbol{\theta}}_n) = \hat{\ell}_n(\boldsymbol{\theta}^*) - (\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^*)^T \nabla^2 \hat{l}_n \left(\ddot{\boldsymbol{\theta}}_n^{\boldsymbol{\eta}} \right) (\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^*) + (1/2)(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^*)^T \nabla^2 \hat{\ell}_n \left(\ddot{\boldsymbol{\theta}}_n^{\boldsymbol{\lambda}} \right) (\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^*). \quad (19)$$

Since $\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^*) = O_p(1)$ and since $\nabla^2 \hat{\ell}_n \left(\ddot{\boldsymbol{\theta}}_n^{\boldsymbol{\lambda}} \right) = \hat{\mathbf{A}}_n + o_{a.s.}(1)$, these relations can be used in (19) to obtain:

$$\hat{\ell}_n(\boldsymbol{\theta}^*) - \hat{\ell}_n(\hat{\boldsymbol{\theta}}_n) = \frac{1}{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^*)^T \mathbf{A}^* (\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^*) + o_p(1/n). \quad (20)$$

Now take the expectation of (20), using the GAIC Lemma Equation (8), to obtain:

$$E\{\hat{\ell}_n(\boldsymbol{\theta}^*)\} - E\{\hat{\ell}_n(\hat{\boldsymbol{\theta}}_n)\} + \frac{1}{2n} \operatorname{tr}\left([\mathbf{A}^*]^{-1} \mathbf{B}^* \right) + o_p(1/n). \quad (21)$$

Equation (20) corresponds to the fourth term on the right-hand side of (12).

Step 4: Calculate Empirical Risk Bias

Now take the expectation of both sides of (12) to obtain:

$$E\left\{\ell(\hat{\boldsymbol{\theta}}_n)\right\} = E\left\{\hat{\ell}_n(\hat{\boldsymbol{\theta}}_n)\right\} + E\left\{\left(\ell(\hat{\boldsymbol{\theta}}_n)\right) - \ell(\boldsymbol{\theta}^*)\right\} + E\left\{\left(\ell(\boldsymbol{\theta}^*) - \hat{\ell}_n(\boldsymbol{\theta}^*)\right)\right\} + E\left\{\left(\hat{\ell}_n(\boldsymbol{\theta}^*) - \hat{\ell}_n(\hat{\boldsymbol{\theta}}_n)\right)\right\}. \quad (22)$$

Now substitute the results of (14), (15), and (21) into (22) to obtain:

$$E\{\ell(\hat{\boldsymbol{\theta}}_n)\} = E\{\hat{\ell}_n(\hat{\boldsymbol{\theta}}_n)\} + \frac{1}{2n} \operatorname{tr}\left([\mathbf{A}^*]^{-1} \mathbf{B}^* \right) + 0 + \frac{1}{2n} \operatorname{tr}\left([\mathbf{A}^*]^{-1} \mathbf{B}^* \right) + o_p(1/n). \quad (23)$$

The conclusion of the theorem then follows by substituting $\mathbf{A}^* = \hat{\mathbf{A}}_n + o_p(1)$ and $\mathbf{B}^* = \hat{\mathbf{B}}_n + o_p(1)$ into (23). □

REFERENCES

Akaike, H. 1973. "Information theory and an extension of the maximum likelihood principle." In *2nd International Symposium on Information Theory*, pp. 267–281. Akadémiai Kiadó.

Akaike, H. 1974. "A new look at the statistical model identification." *IEEE Transactions on Automatic Control 19*(6), 716–723.

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

Claeskens, G. and N. L. Hjort 2008. *Model Selection and Model Averaging*. Cambridge Series in Statistics and Probabilistic Mathematics. New York: Cambridge University Press.

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

Linhart, H. and P. Volkers 1984. "Asymptotic criteria for model detection." *OR Spektrum 6*, 6.

Linhart, H. and W. Zucchini 1986. *Model Selection*. Wiley Series in Probability and Statistics. Hoboken, NJ: Wiley.

Stone, M. 1977. "An asymptotic equivalence of choice of model by cross-validation and Akaike's criterion." *Journal of the Royal Statistical Society. Series B (Methodological) 39*(1), 44–47.

Takeuchi, K. 1976. "Distribution of information statistics and a criterion of model fitting for adequacy of models." *Mathematical Sciences 153*, 12–18.