PREDICTIVE INFERENCE AND THE BOOTSTRAP [1]

Yuichi Kitamura
Department of Economics
University of Wisconsin

October 16, 1999

[1]Conversations with Ken West inspired my thinking on this research. I thank Bill Brown, Frank Diebold, Ron Gallant, Wei-Yin Loh, Roberto Mariano, Mike McCracken, Per Mykland, Michael Newton, Werner Ploberger, Matt Pritsker, Gene Savin, Frank Schorfheide, Jun Shao, Jim Stock, Mark Watson, and seminar participants at the University of Wisconsin Department of Statistics, the University of Pennsylvania, the 1999 North American Summer Meetings of the Econometric Society, the 1999 NBER Summer Institute and the 1999 NBER/NSF Time Series Conference for their useful comments. I also thank Bruce Hansen, Gautam Tripathi, and Grace Wahba for discussions and encouragement. I gratefully acknowledge financial support from the National Science Foundation via grant numbers SBR-9632101 and SES-9905247.

Abstract

This paper investigates predictive inference, which can be regarded as a validation problem that uses a certain data splitting scheme. To carry out a predictive test of an econometric model, a loss function is specified and the data set is divided into the training set and the validation set. Unknown parameters of the model are estimated using the training set, and the estimated model is tested against the validation set, using the expected loss (sometimes called risk) as a criterion.

A conventional risk estimator is the validation-sample average of loss function values evaluated at the estimated parameter value. This procedure can be viewed as a variant of cross-validation (CV) that is carried out forward, therefore called forward validation (FV). A drawback of FV is that it does not explicitly incorporate parameter estimation uncertainty into risk estimation. This point is relevant to the fact that the FV (or CV) estimators often suffer from large variability caused by (small) perturbations in the data.

This paper proposes using bootstrap smoothing (BS) to remedy this drawback. The nonparametric BS algorithm repeats the following two steps: (1) a bootstrap draw of an unknown parameter is obtained by resampling the training set; (2) as in FV, the validation-sample average of the loss function values is calculated, but this time evaluated at the bootstrap draw from (1). After applying these two steps to (sufficiently many) bootstrap draws, the average of validation-sample averages is taken. Using the Hodges-Lehmann deficiency concept, this paper proves that bootstrap smoothing leads to a substantial reduction in the MSE of the risk estimators in a discrimination problem. Some connections with Bayesian methods are discussed. A small-scale simulation experiment provides evidence in favor of bootstrap smoothing.

Journal of Economic Literature Classification Numbers: C14, C52

Keywords: Bagging, Bootstrap, Cross validation, Deficiency, Prediction

PREDICTIVE INFERENCE AND THE BOOTSTRAP

by Yuichi Kitamura

1 Introduction

The idea of using out-of-sample prediction for model evaluation is not new. Many prominent theorists have proposed and investigated predictive methods. Leading examples include Dawid's (1984) "prequential analysis" and Rissanen's (1986a) "predictive least squares (PLS)." Geisser has investigated Bayesian methods for predictive analysis extensively (see, for example, Geisser, 1993). In econometrics, a series of papers and books by Fair (1980, 1984, 1994) shows that out-of-sample analysis can be a powerful tool for checking the performance of econometric models. Hendry and co-authors (Hendry, 1995, Chong and Hendry, 1986) also advocate the use of out-of-sample inference and call their procedure "forecast encompassing." See also recent papers by Swanson and White (1997) and Stock and Watson (1998) for predictive comparisons of linear and nonlinear time series models.

Recently some researchers have made important theoretical contributions that open up new possibilities for predictive inference. Diebold and Mariano (1995) merge Vuong's (1989) type-testing strategy with predictive inference, using general loss functions. Geweke (1994, 1997), Phillips and Ploberger (1995) and Phillips (1996) discuss Bayesian methods of forecast evaluation. West (1996) develops a general asymptotic theory for predictive inference. White (1997) uses a predictive test to compare multiple models that takes account of model uncertainty.

This paper treats predictive inference as a problem of model validation. That is, the researcher "trains" (i.e., estimates) a model using his/her data subset, and then "tests" or "validates" it using the rest of the data. The whole data set is divided into the training set $\{Z_i\}_{i=1}^{T}$ and the validation set $\{Z_j\}_{j=T+1}^{n+T}$. (Thus the total number of observations is $T+n$.) The main focus of the paper is on a "fixed" sample split scheme, where the data are divided at a fixed point. Other schemes, such as "rolling" and "recursive" are possible (West and McCracken, 1998) and will be mentioned in Section 6. Assume that Z_i, $i = 1, ..., T+n$ are independently drawn[1] from a sample space $\mathcal{Z}$ according to a probability law P_z. A finite dimensional vector

[1]See Section 6 for a discussion on dependent data.

of unknown parameter θ_0 is estimated by $\theta_T = \theta_T(Z_1, ..., Z_T)$ using the training data set. Let P_{θ_T} denote the probability law of θ_T.

Next, a loss function $L(z, \theta)$ needs to be specified. For example, consider a parametric predictor $\hat{y} = f(x, \theta_T)$ for a scalar y based on a vector of covariates x, and let $z = (y, x)$. A commonly-used loss function is the mean square error $L(z, \theta_T) = (\hat{y} - y)^2$. One may also consider a loss function based on the wrong sign prediction $L(z, \theta_T) = I\{\hat{y} > 0\}I\{y \leq 0\} + I\{\hat{y} \leq 0\}I\{y > 0\}$. Note that the minimizer $\hat{\theta} = \arg\min_\theta \sum_{i=1}^T L(Z_i, \theta)$ in the latter case is Manski's (1975) maximum score estimator, which has a nonstandard limit distribution (Kim and Pollard, 1990; see Horowitz, 1992, for a solution for this problem). However, our concern is model validation, and θ_T may be the least square estimator, say, and is not necessarily the minimizer of the aggregated loss function $\sum_{i=1}^T L(Z_i, \theta)$.

The task here is to evaluate the *(conditional) risk*

$$R(\theta_T) = \int L(Z, \theta_T) dP_Z.$$

Here the loss function is evaluated at $\theta = \theta_T(Z_1, ..., Z_T)$, so $R(\theta_T)$ is defined conditional on the training data. Efron (1986) terms $R(\theta_T)$ the "true error rate," in his investigation of model validation under 0-1 loss functions. Sometimes an alternative definition of risk is of interest. For example, by integrating $R(\theta_T)$ with respect to the training data, one may define the *unconditional risk* $R_T = \int \int L(Z, \theta_T) dP_Z dP_{\theta_T}$. (The latter is called the expected true error in Efron's terminology.)

An estimator for $R(\theta_T)$ (or R_T) commonly used in the literature is

$$\hat{R}^{FV} = n^{-1} \sum_j L(Z_j, \theta_T),$$

where the loss functions evaluated at the training set estimator are averaged over the validation samples. It is an analogue of the cross-validation-type estimator considered in the literature, though here the validation is only carried out "forward," thus the name "forward validation (FV) estimator" (see, e.g., Hjorth, 1982, for a discussion on $\hat{R}^{FV}$ and its variants).

Under regularity conditions, it is possible to obtain the asymptotic distribution of $\sqrt{n}(\hat{R}^{FV} - R_T)$, following West (1996) (see also McCracken, 1998). In particular, the West theorem shows how parameter estimation errors affect the asymptotic distribution. Note, however, that the uncertainty associated with parameter estimation is *not* reflected in the definition of the estimator $\hat{R}^{FV}$, as the FV estimator simply replaces the unknown θ_0 with its estimate θ_T.

One way of taking the estimation uncertainty of θ into account is to average loss functions over θ, as a Bayesian would suggest (see Section 4 for more on this point). This paper uses the (frequentist) bootstrap to achieve this. More specifically, it proposes to integrate loss functions $L(Z_i, \theta), i = T, ..., T+n$ with respect to θ using a bootstrap measure. This methodology is a predictive analogue of Efron's "leave-one-out" bootstrap, which is known to be useful in practice. In particular, it reduces the variability of a risk estimate substantially. The basic intuition behind Efron's method is straightforward: it smoothes out "kinks" of discontinuous loss functions, such as the 0-1 loss function as discussed above.

This paper makes three main contributions. The first is methodological: it proposes a new predictive risk estimator based on bootstrap smoothing. Second, it demonstrates theoretically that the above intuition about bootstrap smoothing is indeed correct in a standard discriminant analytic setting. It presents a theorem that implies that the standard estimator $\hat{R}^{FV}$ is infinitely "deficient" relative to our bootstrap-smoothed estimator in Hodges-Lehmann's (Hodges and Lehmann, 1970) sense. The theorem explains why Efron's bootstrap smoothing works. Third, it investigates the validity of the smoothing methodology in practice using a small set of sampling experiments.

2 Methodology

Let δ_z denote the probability measure degenerated at z. $\hat{P}_Z = T^{-1} \sum_{i=1}^{T} \delta_{Z_i}$ is then the empirical measure of the training samples. Let θ_T^* be the bootstrap version of θ_T, and $P_{\theta_T}^*$ its (bootstrap) measure (conditional on the training samples). That is, $\theta_T^* = \theta_T(Z_1^*, ..., Z_T^*)$ is obtained from the bootstrap samples $\{Z_i^*\}_{i=1}^T \sim_{IID} \hat{P}_Z$.[2] We propose the following bootstrap-averaged version of $\hat{R}^{FV}$:

$$\hat{R}^{BS} = n^{-1} \sum_j \int L(Z_j, \theta_T^*) dP_{\theta_T}^*. \tag{2.1}$$

(2.1) is evaluated most conveniently by simulation: simply draw $\theta_T^{*(b)}, b = 1, ..., B$ by resampling the training data repeatedly and take the average $B^{-1} n^{-1} \sum_b \sum_j L(Z_j, \theta_T^{*(b)})$. B is the number of bootstrap replications, and is chosen by the researcher. $\hat{R}^{BS}$ is similar to Efron's "leave-one-out" bootstrap estimator (Efron, 1982, 1983, 1986, Efron and Tibshirani, 1997), which

[2] An alternative resampling algorithm should be used for a dependent data set. See Section 6.

smoothes cross-validation estimates by applying the bootstrap only to training data (for each validation). Efron recommends the bootstrap in this situation on the grounds that a cross-validation estimator tends to have large variance. The bootstrap often reduces it drastically. In a wide variety of areas, such as discriminant analysis and statistical pattern recognition, the use of Efron's method is dominant.

The estimator $\hat{R}^{BS}$ is also closely related to the recent methodology in machine learning, called "bagging" (short for "bootstrap averaging"), advocated by Breiman (1996). Bagging aims to reduce the variance of a predictor by averaging it over training observations. The intuition behind bagging is that when the predictor is sensitive to the realization of training samples, its variance is reduced by first perturbing the training observations using the bootstrap and then averaging over the perturbed series. To put the current problem in Breiman's framework, consider the following problem of predicting "future" losses: for a given sequence $Z_j, j = T + 1, ...T+n$, one wishes to predict $L(Z_j, \theta_0)$ where θ_0 is unknown, based on the training observations $\{Z_i\}_{i=1}^{T}$. $\hat{R}^{BS}$ is simply the sample average of the "bagged" loss function predictors.

The superior performances of both Efron's method and bagging have been documented in the literature, mainly using Monte Carlo simulations. Therefore, noting the connection between $\hat{R}^{BS}$ and the methods by Efron and Breiman, it is reasonable to expect that the bootstrapped prediction risk estimator would outperform the conventional estimator $\hat{R}^{FV}$. Indeed, the theoretical result in Section 3 shows that this conjecture is born out in a prediction problem based on a standard classifier. Simulation results in Section 5 also provide some support for this conjecture.

3 Asymptotic Theory

Efron's papers on model validation cited above focus on errors of predictors that take binary values. His problem is essentially the same as the sign prediction exercise mentioned in Section 1. For a discontinuous loss function, a cross-validation estimator of expected error rates is also discontinuous. He emphasizes that the bootstrap is particularly useful in this situation, as it smoothes out discontinuity. This intuitive logic also applies to our predictive problem: when $L(z, \theta)$ is discontinuous, bootstrap smoothing is expected to provide a significant improvement for the same reasons.

The efficacy of bootstrap smoothing has been investigated extensively in the statistics

literature. These studies differ from ours in that they focus on "sample-reuse methods" such as cross-validation and the leave-one-out bootstrap. Simulation studies, such as Gong (1986) and Efron and Tibshirani (1997), indicate that bootstrap smoothing methods (e.g., Efron's leave-one-out) tend to outperform cross-validation. In spite of the overwhelming simulation evidence documented in the literature, however, little attempts has been made to clarify the theoretical properties of bootstrap smoothing.[3] This section addresses this issue. It provides a new and rigorous justification for bootstrap smoothing in the context of predictive inference. We use a theoretical framework that involves the analysis of higher order expansions and apply the Hodges-Lehmann asymptotic relative deficiency concept. This enables us to make a meaningful comparison between the naive forward-validation risk estimator and its bootstrap version.

To investigate the efficacy of the bootstrap smoothing, we follow the literature cited above and consider a prediction problem in a standard discriminant analytic framework. Discriminant analysis is one of the primary concerns of the series of Efron's papers cited above. See McLachlan (1992) for a comprehensive discussion of discriminant analysis, and Amemiya (1985), Amemiya and Powell (1983), Maddala (1983), and Sargent (1993) for its use in econometrics. This section focuses on a specific and relatively simple model that appears to be suited for our purpose. If our concern were usual first order asymptotic efficiency, we could handle much more general models than considered here. Such an approach, however, typically implies that methods with/without bootstrap smoothing are asymptotically equivalent, and therefore fails to explain the difference in the empirical performance of the two methods. The model considered here is representative in the literature, and also it has a simple structure that makes higher-order expansions tractable. We investigate alternative prediction problems in Section 5 using simulations.

The rest of this section deals with the following classification problem, similar to ones considered by Davison and Hall (1992), Randles (1982) and Sedransk and Okamoto (1971). The goal of the classification procedure is to predict the outcome (choice) Y when a value of the "covariate" X is given. Y takes either the value of 0 or 1. The joint distribution of Y and X is specified as follows. With probability $\frac{1}{2}$, $Y = 0$ and X is drawn according to a probability law P_{X_0}. Similarly, with probability $\frac{1}{2}$, $Y = 1$ and X obeys P_{X_1}. P_{X_0} and P_{X_1} have means μ_0 and μ_1, and without loss of generality, assume that $\mu_0 < \mu_1$. Also assume that they have a common variance $\sigma^2 < 0$. The researcher observes a sequence of independent draws of $Z = (Y, X)$ from

[3]Notable exceptions include the analyses by Hall (1995) and Davison and Hall (1992).

this data generating mechanism. This is a fairly standard setup used in the literature.

Suppose the researcher observes T training data that are used to construct a classification rule. T_0 (T_1) of them have $Y = 0$ ($Y = 1$). This set is denoted by $\mathbf{X_0} = \{X_{01}, X_{02}, ..., X_{0T_0}\}$ ($\mathbf{X_1} = \{X_{11}, X_{12}, ..., X_{1T_1}\}$), $T_0 + T_1 = T$. Define $\bar{X}_0 = T_0^{-1} \sum_{t=1}^{T_0} X_{0t}$ and $\bar{X}_1 = T_1^{-1} \sum_{s=1}^{T_1} X_{1s}$. Let $\hat{y}(x) = \frac{1}{2}(\bar{X}_0 + \bar{X}_1) - x$. The well-known Fisher classifier is then given by

$$\hat{Y}(x) = I\{\hat{y}(x) \leq 0\}. \tag{3.2}$$

In the notation used in Section 1, $\theta_T = (\bar{X}_0 + \bar{X}_1)/2$, and $\theta_0 = (\mu_0 + \mu_1)/2$. We follow the literature (see, for example, Lachenbruch and Mickey, 1968, and Page, 1985) and consider the following risk function:

$$\begin{aligned} R(\theta_T) &= \int I\{\hat{y}(X_1) > 0\} dP_{X_1} \\ &= \int I\{\theta_T > X_1\} dP_{X_1}, \end{aligned} \tag{3.3}$$

that is, the probability of predicting $\hat{Y} = 0$ when the true outcome/choice is $Y = 1$. This conditional risk function is implied by the loss function $L(z, \theta_T) = 2y(1 - \hat{Y}(x))$. Due to the symmetry of the problem, the analysis of the probability of obtaining an incorrect prediction $\hat{Y} = 1$ when the truth is $Y = 0$ is nearly identical and thus omitted.

Next, the estimation of the above risk function is considered. The FV and BS risk estimators are constructed using a set of validation samples. Due to the above definition of the risk, it is enough to consider validation samples with $Y = 1$. Suppose n of such samples $\{X_j\}_{j=1}^n$ are available. The FV estimator for $R(\theta_T)$ is given by

$$\hat{R}^{FV} = n^{-1} \sum_{j=1}^{n} I\{\theta_T > X_j\}. \tag{3.4}$$

The bootstrap-smoothed version of $\hat{R}^{FV}$ is obtained by integrating it against an appropriate bootstrap measure. As with other bootstrap risk estimators considered in the classification literature, we apply an ordinary nonparametric bootstrap technique. That is, a bootstrap version of $\bar{X}_0$ is obtained by resampling $\mathbf{X_0}$ with replacement and forming the sample mean $\bar{X}_0^*$. Obtain $\bar{X}_1^*$ in a similar manner, and let $\theta_T^* = \bar{X}_0^* + \bar{X}_1^*$. This is the bootstrap version of the training estimator θ_T. Let $P_{\theta_T}^*$ denote the probability law of θ_T^* (conditional on the training samples $\mathbf{X_0}$ and $\mathbf{X_1}$). Define the bootstrap-smoothed risk estimator as follows:

$$\hat{R}^{BS} = n^{-1} \sum_{j=1}^{n} \int I\{\theta_T^* > X_j\} dP_{\theta_T}^*. \tag{3.5}$$

Our comparison of the two risk estimators is based on their mean square errors (MSE). For a risk estimator $\hat{R}$, define $\text{MSE}(\hat{R}) = \text{E}[(\hat{R} - R(\theta_T))^2]$, where the expectation is taken with respect to both the training samples and the validation samples. Theorem 1 below is the basis of our analysis. It derives asymptotic expansions for $\text{MSE}(\hat{R}^{FV})$ and $\text{MSE}(\hat{R}^{BS})$ under mild regularity conditions. Theorem 2 compares the MSE of the two estimators using the Hodges-Lehmann deficiency measure.

Let F_{X_1} denote the CDF corresponding to P_{X_1} and let $f_{X_1}(x)$ denote the density of F_{X_1} at x if it exists. The proofs of the theorems can be found in the Appendix.

Theorem 1: *Suppose that* $\int X_0^4 dP_{X_0} < \infty$, $\int X_1^4 dP_{X_1} < \infty$ *and that there exists a constant* $A > 0$ *independent of* θ *and* x *such that* $|F_{X_1}(x) - F_{X_1}(\theta) - (x-\theta) f_{X_1}(\theta)| \leq A(x-\theta)^2$ *for all* θ *in a neighborhood* $\mathcal{N}$ *of* θ_0 *and all* $x \in \mathbb{R}$. *Then if* $T/n^{2/3} \to \infty$ *as* $n \to \infty$,

$$\text{MSE}(\hat{R}^{FV}) - \text{MSE}(\hat{R}^{BS}) = \frac{\sigma f_{X_1}(\theta_0) C}{n\sqrt{T}} + o\left(\frac{1}{n\sqrt{T}}\right),$$

where $C = \frac{1}{\sqrt{\pi}} = 0.5641....$

Remark 1: The first assumption is a mild moment condition. The second assumption imposes a certain degree of smoothness on $F_{X_1}(x)$. The latter is satisfied, for example, if $(\partial/\partial x) f_{X_1}$ exists and is uniformly bounded; simply let $A = \sup_x (\partial/\partial x) f_{X_1}(x)/2$. Our proof of Theorem 1 uses approximation techniques which go back to Azzalini (1981), Reiss (1981) and Falk (1983). These papers are concerned with a quite different problem from ours — the kernel type estimation of smooth cumulative distribution functions. Nevertheless, mathematically our problem shares some similarities with theirs.

Indeed, Glick (1978) proposes the kernel-smoothed estimation of the prediction risk in discriminant analysis. The original Glick estimator smoothes the "apparent error rate", which is not predictive as the entire data set is used for both training and validation. However, it is possible to apply kernel smoothing to cross-validation estimators (Devroye, Györfi and Lugosi, 1991, p.550). Let us modify the FV estimator (3.4) using a smooth kernel function $K : \mathbb{R} \to [-M, M]$ for some finite $M \geq 1$. Let $K(\cdot)$ be continuous and satisfy $\lim_{x \to -\infty} K(x) = 0$

and $\lim_{x\to\infty} K(x) = 1$. Let h be the usual bandwidth such that $h \to 0$ as $n \to \infty$. The "kernel-smoothed FV" estimator is[4]

$$\hat{R}^{FV}_{\text{kern}} = n^{-1} \sum_{j=1}^{n} K\left(\frac{\theta_T - X_j}{h}\right). \tag{3.6}$$

Glick's (1978) motivation for the kernel-smoothed risk estimator is variance reduction. Note the similarity between (3.6) and the expression of $\hat{R}^{BS}$ given in the Appendix (Equation 8.4). In the latter expression, a "kernel" implied by the bootstrap measure emerges. In practice, the performance of kernel smoothed risk estimators is sensitive to the choice of the bandwidth h (McLachlan, 1992, Chapter 10). In bootstrap smoothing, the researcher does not face this vexing problem; effectively, the bootstrap lets data choose the "bandwidth". Devroye, Györfi and Lugosi (1991, Chapter 31) point out further theoretical problems associated with kernel-smoothing of risk estimators. Also, the kernel-based approach is conceptually not attractive, in that it alters the loss function and the risk function the researcher is interested in.[5]

Remark 2: The theorem immediately implies that $\hat{R}^{FV}$ is deficient relative to $\hat{R}^{BS}$ in the sense of Hodges-Lehmann (Hodges and Lehmann, 1970). Let $k(n)$ be the smallest integer such that $\text{MSE}(\hat{R}^{FV}) \leq \text{MSE}(\hat{R}^{BS})$. The Hodges-Lehmann measure of deficiency is defined to be $d(n) = k(n) - n$. Put loosely, it is the number of observations "wasted" for $\hat{R}^{FV}$ to achieve a level of accuracy that is comparable to $\hat{R}^{BS}$. By Theorem 1, it is straightforward to see that

$$d(n) = k(n) - n \to \infty, \tag{3.7}$$

i.e., $\hat{R}^{FV}$ is infinitely deficient relative to $\hat{R}^{BS}$. Indeed, we can say more about the deficiency measure:

Theorem 2: *Under the same conditions as in Theorem 1,*

$$\frac{\sqrt{T}d(n)}{n} = \frac{\sigma f_{X_1}(\theta_0) C}{F_{X_1}(\theta_0)(1 - F_{X_1}(\theta_0))} + o(1).$$

[4]Note some resemblances between Glick's variance reduction technique and Horowitz's (1992) smoothed maximum score estimation method, which also uses a kernel of this type to smooth out indicator functions that appear in Manski's original version of the maximum score estimator (Manski, 1975).

[5]Pawlak (1989) studies a non-predictive risk estimator with kernel smoothing, and derives a relation similar to ours. His method of proof is quite different though, reflecting the difference between his problem and ours. He also assumes normality of covariates, which in unnecessary in our investigation.

Remark 3: Theorem 2 provides a theoretical justification for bootstrap smoothing initiated by Efron. The ordinary first order asymptotic analysis fails to detect the difference between the naive forward validation estimator and its bootstrap-smoothed version. The theorem overcomes this difficulty by using the Hodges-Lehmann approach and establishes the fact that bootstrap smoothing delivers an asymptotic gain. In fact, the deficiency measure diverges to infinity. This result is in accordance with the simulation result which will be reported shortly in Section 5.

Remark 4: A closer look at Theorem 2 reveals some more interesting points. For a given validation sample size n, as the training sample size T gets larger, the deficiency measure $d(n)$ tends to get smaller (though note the growth condition $T/n^{2/3} \to \infty$). This is natural to expect. Recall bootstrap smoothing is directed to reduce the estimation errors associated with variations in the training samples. If the size of the training set is much larger than that of the validation set, the estimation error is dominated by validation data sampling variations, and therefore the bootstrap would have little effect. The flip side of this fact is that bootstrap smoothing potentially has a substantial impact when the size of the training data set is (relatively) small and there are large variations in θ_T. As a benchmark, consider a case where the growth rates of the two sizes are "balanced", i.e., $T \propto n$. The deficiency measure then grows at a fairly rapid rate of $n^{1/2}$. (Note that the original Hodges-Lehmann paper mainly investigates situations where $\lim d_n \to d < \infty$.)

Remark 5: Bootstrap smoothing is originally proposed in the context of cross-validation. It is of interest to establish an asymptotic relative deficiency result as shown above in such settings. That would imply a Hodges-Lehmann deficiency result of cross-validation relative to the leave-one-out-bootstrap. It appears that the result obtained by Pawlak (1988) about the relative merit of the kernel smoothed risk estimator is useful for such an investigation.

4 Bayesian Interpretation

One of the advantages of Bayesian approaches is that they provide a natural and unified framework for model comparison and predictions. Geweke (1994), for example, discusses Bayesian model comparisons with emphasis on predictive methods. He suggests conducting a Bayesian inference by splitting samples into a training set and a validation set (though he does not use

this terminology). In particular, his equation (3) uses a posterior density obtained from a training data set (and a prior density) to integrate the likelihood, thereby yielding the "predictive likelihood". In this way parameter uncertainty can be incorporated in the process of model validation.

Our bootstrap smoothing is, *prima facie*, similar to such a Bayesian method of model validation, if we could regard the bootstrap distribution of the training data estimator θ_T as a posterior distribution. In what follows we show that this is more than just a formal similarity. In a related study, Rao and Tibshirani (1997) use bootstrapping for model averaging and claim that their method implicitly achieves Bayesian model averaging with a diffuse prior. Their argument is based on the connection between the bootstrap and Bayesian methods, as studied by Efron (1979, 1981) and Rubin (1981). This connection is also valid in the current analysis, as explained below.

Consider again the general model validation problem discussed in Section 1. Let $\mathbf{Z}_{\text{train}}$ denote the training data set $\{Z_i\}_{i=1}^{T}$. If a prior for the unknown parameter θ is available, the likelihood function would imply the posterior density $p(\theta|\mathbf{Z}_{\text{train}})$. Define $\hat{R}(\theta) = n^{-1}\sum_{j=T+1}^{T+n} L(x_j, \theta)$, the empirical risk of using a parameter value θ, calculated for the validation samples $\{Z_j\}_{j=T+1}^{T+n}$. It is then natural to use $\int \hat{R}(\theta)p(\theta|\mathbf{Z}_{\text{train}})d\theta$ as a risk estimate. We interpret $\hat{R}^{BS}$ as such a Bayesian risk estimate, implied by a particular diffuse prior. Our argument relies on results reported in the studies by Efron and Rubin cited above.

Suppose the sample space $\mathcal{Z}$ is discrete; if not, partition $\mathcal{Z}$ into sufficiently many categories. Let C denote the number of categories. Define $f_c = \text{Prob}\{Z_i \in \text{Category } c\}$, and its empirical version $\hat{f}_c = T^{-1}\#\{Z_i \in \text{Category } c, i = 1, ..., T\}$ for $c = 1, ..., C$. Define $1 \times C$ vectors $\mathbf{f} = (f_1, ..., f_C)$, $\hat{\mathbf{f}} = (\hat{f}_1, ..., \hat{f}_C)$ and $\mathbf{1} = (1, ..., 1)$. Assume $\theta = \theta(\mathbf{f})$ is smooth in $\mathbf{f}$. Suppose the prior distribution for $\mathbf{f}$ is symmetric Dirichlet with parameter a:

$$\mathbf{f} \sim \text{Di}_C(a\mathbf{1}), \tag{4.8}$$

i.e., the density of $\mathbf{f}$ is proportional to $\Pi_{c=1}^{C} f_c^{a-1}$. Then the posterior distribution conditional on the training data is:

$$\mathbf{f}|\hat{\mathbf{f}} \sim \text{Di}_C(a\mathbf{1} + T\hat{\mathbf{f}}). \tag{4.9}$$

Take the limit $a \to 0$ of (4.9) so that it corresponds to a diffuse prior:

$$\mathbf{f}|\hat{\mathbf{f}} \sim \mathrm{Di}_C(T\hat{\mathbf{f}}). \tag{4.10}$$

(4.10) is well-approximated by the (bootstrap) multinomial distribution $\mathbf{f}^*|\hat{\mathbf{f}} \sim T^{-1}\mathrm{Mult}(T, \hat{\mathbf{f}})$. Therefore the bootstrap distribution of θ_T^* can be regarded as an approximation of the posterior distribution of θ under the diffuse Dirichlet prior. With this interpretation, $\hat{R}^{BS}$ is regarded as an approximation of the posterior mean of $\hat{R}(\theta)$ with a prior that reflects the researcher's ignorance.

5 Numerical Examples

In this section we compare the performance of $\hat{R}^{FV}$ and $\hat{R}^{BS}$ in predictive validation problems using Monte Carlo experiments. The first set of simulations uses a design which can be viewed as a predictive version of Efron's experiments on logistic regression (1983, 1986). Results using a design that closely follows the discrimination analytic problem considered in Section 3 will be reported in a later version of the paper.

Efron reports comprehensive simulation results on the performance of CV, leave-one-out-bootstrap, and other types of risk estimators using logistic regression. We basically follow his experimental design, but use forward (predictive) validation. In this design, independent random vectors $Z_i = (Y_i, X_i')', i = 1, ..., T+n$ are drawn using the rule stated below. Y_i is distributed as

$$Y_i = \begin{cases} 0 & \text{with probability } \frac{1}{2} \\ 1 & \text{with probability } \frac{1}{2}, \end{cases} \tag{5.11}$$

and for each Y_i, a $(k-1)$-variate normal vector is generated according to

$$\underline{X_i}|Y_i \sim \mathrm{N}(\mu(Y_i), I_{k-1}), \tag{5.12}$$

where the conditional mean $\mu(Y_i)$ is either $((0.5 - Y_i), 0, ..., 0)'$ or a $(k-1)$ vector of zeros. Define $X_i = (1, \underline{X_i}')'$ and $\theta_0 = (0, 1, 0, ..., 0)$ or $(0, ..., 0)$. It is easy to see that $\mathrm{Prob}\{Y_i = 1|X_i\} = \frac{\exp(X_i'\theta_0)}{1+\exp(X_i'\theta_0)}$, and therefore the simulated observations $Z_i = (Y_i, X_i')', i = 1, ..., T+n$ can be interpreted as data from a logit model. Using the first T observations, the (unknown) true vector θ_0 is estimated by the logit maximum likelihood estimator, denoted by θ_T. The following prediction rule based on the cutoff point 0.5 is used:

$$\hat{Y}_i = \begin{cases} 1 & \text{if } \hat{\text{Prob}}\{Y_i = 1\} > 0.5 \\ 0 & \text{otherwise} \end{cases} \tag{5.13}$$

where

$$\hat{\text{Prob}}\{Y_i = 1\} = \frac{\exp(X_i'\theta_T)}{1 + \exp(X_i'\theta_T)}.$$

Validation is carried out by applying this rule to the last n observations, with a loss function $L(Z_i, \theta_T) = \hat{Y}_i(1 - Y_i) + (1 - \hat{Y}_i)Y_i$. The naive risk estimator $\hat{R}^{FV}$ is, as discussed before, simply the sample average of $L(Z_i, \theta_T)$ over the validation samples. To compute the bootstrap-smoothed risk estimator $\hat{R}^{BS}$, the logit MLE θ_T is resampled using the standard nonparametric bootstrap. The number of bootstrap replications is 100.

The dimension k of the covariate X_i is either 3 or 5, and $\theta_0 = (0, 0, 0)'$, $(0, 1, 0)$, $(0, 1, 0, 0, 0)'$ or $(0, 0, 0, 0, 0)$). A "fixed" data splitting scheme is used, with $T = 15$ or 30 and $n = 15, 30$ or 50. 200 Monte Carlo replications are simulated per experiment. We report the root mean square errors (RMSE) of risk estimators for the conditional risk $R(\theta_T)$ and the unconditional risk R_T (thus the former and the latter are labeled "conditional RMSE" and "RMSE", respectively).

Simulation results are reported in Tables 1-4. Comparisons of $\hat{R}^{FV}$ and $\hat{R}^{BS}$ show that bootstrap smoothing uniformly reduces RMSEs. For some parameterizations the reduction can be dramatic, as large as 44 percent. As expected, the bootstrap-smoothed estimator $\hat{R}^{BS}$ sometimes has a small "pessimistic" bias (see Tables 2 and 4). This well-known bias stems from the fact that the bootstrap typically uses only a subset of the whole training sample — about $100(1 - e^{-1})\%$ — thereby inflating the corresponding risk estimator to some extent. As can be seen in the simulation results, this bias has a minimal impact on RMSE. The benefit of bootstrap smoothing via variance reduction dominates the bias effect.

6 Discussion

(1) Previous sections treat IID samples, as some theoretical papers on model validation and predictive inference do; see, for example, Breiman (1996, 1998), Efron (1983, 1986), Harris (1989) and Rissanen (1986a), to name a few. In many prediction problems it is important to be able to deal with dependent data, however. The theoretical analysis presented in Section 3 does not provide a clear guidance about what type of bootstrap should be used if dependence is

Table 1: $\theta_0 = (0,0,0)$'

Sample sizes/Risk	Estimator	Mean	RMSE	Bias
$T = 15$, $n = 15$, $R_T = 0.5000$	$\hat{R}^{FV}$	0.4977	0.1347	-0.0023
	$\hat{R}^{BS}$	0.4954	0.0835	-0.0046
		(1 - RMSE(BS)/RMSE(FV))×100 = 38.0 (%)		
$T = 15$, $n = 30$, $R_T = 0.5000$	$\hat{R}^{FV}$	0.4928	0.0910	-0.0072
	$\hat{R}^{BS}$	0.4959	0.0524	-0.0041
		(1 - RMSE(BS)/RMSE(FV))×100 = 42.4 (%)		
$T = 15$, $n = 50$, $R_T = 0.5000$	$\hat{R}^{FV}$	0.4999	0.0792	-0.0001
	$\hat{R}^{BS}$	0.4989	0.0494	-0.0011
		(1 - RMSE(BS)/RMSE(FV))×100 = 37.7 (%)		
$T = 30$, $n = 15$, $R_T = 0.5000$	$\hat{R}^{FV}$	0.4933	0.1200	-0.0067
	$\hat{R}^{BS}$	0.4909	0.0733	-0.0091
		(1 - RMSE(BS)/RMSE(FV))×100 = 38.9 (%)		
$T = 30$, $n = 30$, $R_T = 0.5000$	$\hat{R}^{FV}$	0.5027	0.0929	0.0027
	$\hat{R}^{BS}$	0.5049	0.0544	0.0049
		(1 - RMSE(BS)/RMSE(FV))×100 = 41.4 (%)		
$T = 30$, $n = 50$, $R_T = 0.5000$	$\hat{R}^{FV}$	0.5072	0.0694	0.0072
	$\hat{R}^{BS}$	0.5030	0.0387	0.0030
		(1 - RMSE(BS)/RMSE(FV))×100 = 44.2 (%)		

Table 2: $\theta_0 = (0,1,0)$'

Sample sizes/Unconditional Risk	Estimator	Mean	RMSE	Bias	Conditional RMSE
$T = 15$, $n = 15$, $R_T = 0.3662$	$\hat{R}^{FV}$	0.3593	0.1353	-0.0068	0.1230
	$\hat{R}^{BS}$	0.3939	0.1011	0.0277	0.0986
	(1 - RMSE(BS)/RMSE(FV))×100 = 25.3 (%)				
	(1 - Cond.RMSE(BS)/Cond.RMSE(FV))×100 = 19.8(%)				
$T = 15$, $n = 30$, $R_T = 0.3662$	$\hat{R}^{FV}$	0.3612	0.1071	-0.0050	0.0905
	$\hat{R}^{BS}$	0.3897	0.0806	0.0236	0.0728
	(1 - RMSE(BS)/RMSE(FV))×100 = 24.7(%)				
	(1 - Cond.RMSE(BS)/Cond.RMSE(FV))×100 = 19.5 (%)				
$T = 15$, $n = 50$, $R_T = 0.3662$	$\hat{R}^{FV}$	0.3602	0.0848	-0.0060	0.0634
	$\hat{R}^{BS}$	0.3891	0.0719	0.0230	0.0557
	(1 - RMSE(BS)/RMSE(FV))×100 = 15.2 (%)				
	(1 - Cond.RMSE(BS)/Cond.RMSE(FV))×100 = 12.1 (%)				
$T = 30$, $n = 15$, $R_T = 0.3406$	$\hat{R}^{FV}$	0.3373	0.1207	-0.0033	0.1146
	$\hat{R}^{BS}$	0.3621	0.0975	0.0215	0.0932
	(1 - RMSE(BS)/RMSE(FV))×100 = 19.3 (%)				
	(1 - Cond.RMSE(BS)/Cond.RMSE(FV))×100 = 18.7(%)				
$T = 30$, $n = 30$, $R_T = 0.3406$	$\hat{R}^{FV}$	0.3430	0.0963	0.0024	0.0869
	$\hat{R}^{BS}$	0.3659	0.0832	0.0253	0.0759
	(1 - RMSE(BS)/RMSE(FV))×100 = 13.6(%)				
	(1 - Cond.RMSE(BS)/Cond.RMSE(FV))×100 = 12.7 (%)				
$T = 30$, $n = 50$, $R_T = 0.3406$	$\hat{R}^{FV}$	0.3399	0.0737	-0.0007	0.0684
	$\hat{R}^{BS}$	0.3585	0.0625	0.0179	0.0568
	(1 - RMSE(BS)/RMSE(FV))×100 = 15.2 (%)				
	(1 - Cond.RMSE(BS)/Cond.RMSE(FV))×100 = 16.9 (%)				

Table 3: θ_0 = (0,0,0,0,0)'

Sample sizes/Risk	Estimator	Mean	RMSE	Bias
$T = 15$, $n = 15$, $R_T = 0.5000$	$\hat{R}^{FV}$	0.4993	0.1215	-0.0007
	$\hat{R}^{BS}$	0.5028	0.0716	0.0028
		(1 - RMSE(BS)/RMSE(FV))×100 = 41.1 (%)		
$T = 15$, $n = 30$, $R_T = 0.5000$	$\hat{R}^{FV}$	0.4970	0.0841	-0.0030
	$\hat{R}^{BS}$	0.4957	0.0512	-0.0043
		(1 - RMSE(BS)/RMSE(FV))×100 = 39.1 (%)		
$T = 15$, $n = 50$, $R_T = 0.5000$	$\hat{R}^{FV}$	0.4945	0.0708	-0.0055
	$\hat{R}^{BS}$	0.4950	0.0403	-0.0050
		(1 - RMSE(BS)/RMSE(FV))×100 = 43.1 (%)		

Table 4: θ_0 = (0,1,0,0,0)'

Sample sizes/Unconditional Risk	Estimator	Mean	RMSE	Bias	Conditional RMSE
$T = 15$, $n = 15$, $R_T = 0.3976$	$\hat{R}^{FV}$	0.4067	0.1346	0.0090	0.1241
	$\hat{R}^{BS}$	0.4291	0.0966	0.0314	0.0914
		(1 - RMSE(BS)/RMSE(FV))×100 = 28.8 (%)			
		(1 - Cond.RMSE(BS)/Cond.RMSE(FV))×100 = 26.4(%)			
$T = 15$, $n = 30$, $R_T = 0.3976$	$\hat{R}^{FV}$	0.4022	0.0950	0.0045	0.0810
	$\hat{R}^{BS}$	0.4270	0.0739	0.0294	0.0731
		(1 - RMSE(BS)/RMSE(FV))×100 = 22.2(%)			
		(1 - Cond.RMSE(BS)/Cond.RMSE(FV))×100 = 9.8 (%)			
$T = 15$, $n = 50$, $R_T = 0.3976$	$\hat{R}^{FV}$	0.3929	0.1014	-0.0047	0.0746
	$\hat{R}^{BS}$	0.4211	0.0738	0.0235	0.0597
		(1 - RMSE(BS)/RMSE(FV))×100 = 27.2 (%)			
		(1 - Cond.RMSE(BS)/Cond.RMSE(FV))×100 = 19.9 (%)			

present. Nevertheless, the naive bootstrap that ignores dependence is apparently not appropriate for incorporating parameter uncertainty into risk estimates. If the dependence of the process $\{Z_i\}_{i=1}^{T+n}$ is modeled in the prediction method under consideration, such as an ARMA model, then it is natural to use "model-based resampling" (Davison and Hinkley, 1997, Chapter 8.2.2) to obtain the bootstrap measure $P^*_{\theta_T}$. If one wishes to treat the dependence nonparametrically, a convenient solution is to use the blockwise bootstrap (Carlstein, 1986, Künsch, 1989). This procedure resamples blocks of the data in place of raw observations. Either of these approaches yields an appropriate $P^*_{\theta_T}$ that can be used in the definition (2.1).

(2) The theoretical analyses presented in the previous sections focus on a fixed sample split scheme (between training samples and validation samples). Other schemes for predictive inference are also used in practice. For example, Lo and MacKinlay (1997) use a "rolling window" estimation scheme to study the predictability of asset returns.[6] Section 2 of West and McCracken (1998) provides a useful summary of alternative splitting schemes.

The bootstrap smoothing method applies to various sample splitting schemes, with minor changes in its definition. To fix ideas, consider Rissanen's PLS or predictive MDL (minimum description length) (Rissanen, 1986a, 1986b, 1989) or Dawid's prequential analysis (Dawid, 1984), which splits samples recursively. This means that every time the $(t+1)$-th loss function is calculated, the unknown parameter is re-estimated, using training samples up to the t-th observation. Let N be the total sample size. The FV recursive risk estimator is given by

$$\hat{R}^{FV}_{\text{recursive}} = N^{-1} \sum_{t=0}^{N} L(Z_{t+1}, \theta_t). \tag{6.14}$$

It is straightforward to define a bootstrap-smoothed version of (6.14). Let $\hat{P}_{Zt} = t^{-1} \sum_{i=1}^{t} \delta_{Z_i}, t = 1, ..., N$ be the empirical measure of partial samples. The bootstrap version of $\theta_t, t = 1, ..., N$ is $\theta^*_t = \theta_t(Z_{1t}, ..., Z_{tt})$, obtained from the triangular array $\{Z^*_{it}\}_{i=1}^{t} \sim_{IID} \hat{P}_{Zt}$. Let $P^*_{\theta_t}$ be the probability law of θ^*_t conditional on $\{Z^*_{it}\}_{i=1}^{t}$. The BS recursive risk estimator is

$$\hat{R}^{BS}_{\text{recursive}} = N^{-1} \sum_{t=0}^{N} \int L(Z_{t+1}, \theta^*_t) dP^*_{\theta_t}. \tag{6.15}$$

[6]Their study is particularly relevant to our analysis of bootstrap smoothing, since one of their loss functions is based on a Henriksson-Merton type sign statistic (Merton, 1981, Henriksson and Merton, 1981), which is discontinuous in the observations.

As in the original leave-one-out method for cross-validation estimators, the bootstrap is applied to the training set, which varies with the validation data index $t+1$. In calculating $\hat{R}^{FV}_{\text{recursive}}$ and $\hat{R}^{BS}_{\text{recursive}}$, both θ_t and θ^*_t may behave erratically for small t's, or they may not even be defined. In such a case, it is wise to take the summations starting from a small (but fixed) integer.

7 Conclusion

The main goal of this paper is to develop a new risk estimator and investigate its accuracy. There are various potential uses of the risk estimator. For example, it is routine to rank competing models based on risk estimates. Model selection based on the new risk estimator is an important topic for future research. One may also wish to test the statistical significance of the difference between the risk estimates computed for competing models.

As demonstrated above, bootstrap smoothing reduces the variability of risk estimators. Though the technique may tend to yield somewhat pessimistic (i.e., upward biased) results, the theoretical analysis in Section 3 and the simulation study in Section 5 suggest that bootstrapping improves MSEs. The increase in bias is dwarfed by the reduction in variance. The impact of the bias may be smaller still when carrying out model comparisons. If two risk estimators are compared, they will be biased in the same direction, so the biases would partially offset each other.

8 Appendix: Proofs of Theorems

Proof of Theorem 1: It is useful to use the notation E_{train} (E_{valid}) to denote the expectation with respect to the training samples (validation samples). The usual notation $\Phi(\cdot)$ and $\phi(\cdot)$ signify the CDF and the density of the standard normal distribution.

Let $\hat{P}_{X_1}$ and $\hat{F}_{X_1}(x)$ be the empirical measure of the validation samples and their CDF. Also let $F_{X_1}(x)$ denote the CDF that corresponds to P_{X_1}. First note that

$$\begin{aligned} \mathrm{MSE}(\hat{R}^{FV}) &= \mathrm{E}_{train}\mathrm{E}_{valid}[(\hat{P}_{X_1}\{\theta_T > X_1\} - P_{X_1}\{\theta_T > X_1\})^2] && (8.1)\\ &= \mathrm{E}_{train}\mathrm{E}_{valid}[(\hat{F}_{X_1}(\theta_T) - F_{X_1}(\theta_T))^2] && (8.2)\\ &= \mathrm{E}_{train}[n^{-1}F_{X_1}(\theta_T)(1 - F_{X_1}(\theta_T))]. && (8.3) \end{aligned}$$

Next, let $u_T^* = -\sqrt{T}(\theta_T^* - \theta_T)/\sigma$, i.e., the bootstrap version of estimation error (with certain normalization). Let $G_T^*(u)$ denote the CDF of u_T^* (conditional on the training samples). Then

$$\begin{aligned} \hat{R}^{BS} &= n^{-1}\sum_{j=1}^{n}\int I\{\theta_T - T^{-1/2}u_T^* > X_j\}dG_T^*(u_T^*)\\ &= n^{-1}\sum_{j=1}^{n} G_T^*((\theta_T - X_j)/(T^{-1/2}\sigma))\\ &= \tilde{F}_{X_1}(\theta_T), \text{say.} && (8.4) \end{aligned}$$

The last notation reflects the fact that, conditional on the training set, $\hat{R}^{BS}$ can be regarded as a "kernel smoothed" version of $\hat{F}_{X_1}$, with "kernel" G_T^* and "bandwidth" $T^{-1/2}\sigma$. Note that, by the consistency of the bootstrap distribution, G_T^* converges to the standard normal distribution. Define the n-fold product measure for validation samples $P_{X_1}^n = P_{X_1} \otimes P_{X_1} \otimes \cdots \otimes P_{X_1}$, so that

$$\mathrm{MSE}(\hat{R}^{BS}) = \mathrm{E}_{train}\int (\tilde{F}_{X_1}(\theta_T) - F_{X_1}(\theta_T))^2 dP_{X_1}^n.$$

To evaluate the integral in the above equation (conditional on the training data), define

$$\mu_T(\theta_T) = \int G_T^*\left(\frac{\theta_T - X_1}{T^{-1/2}\sigma}\right) dP_{X_1},$$

$$v_T(\theta_T) = \int G_T^{*\,2}\left(\frac{\theta_T - X_1}{T^{-1/2}\sigma}\right) dP_{X_1} - F_{X_1}(\theta_T)$$

and

$$b_T(\theta_T) = \mu_T(\theta_T) - F_{X_1}(\theta_T).$$

Note that $b_T(\theta_T)$ is the bias term of the "kernel smoothed estimator" $\tilde{F}_{X_1}(\theta_T)$ of the CDF F_{X_1} at θ_T. Then, conditional on the training data $\mathbf{X_1}$ and $\mathbf{X_2}$,

$$\begin{aligned}
\int (\tilde{F}_{X_1}(\theta_T) &- F_{X_1}(\theta_T))^2 dP^n_{X_1} = \int \tilde{F}^2_{X_1}(\theta_T) dP^n_{X_1} - 2F_{X_1}(\theta_T) \int \tilde{F}_{X_1}(\theta_T) dP^n_{X_1} + F^2_{X_1}(\theta_T) \\
&= \int \left(\frac{1}{n} \sum_{i=1}^{n} G^*_T \left(\frac{\theta_T - X_1}{T^{-1/2}\sigma} \right) \right)^2 dP^n_{X_1} - 2F_{X_1}(\theta_T) \int \left(\frac{1}{n} \sum_{i=1}^{n} G^*_T \left(\frac{\theta_T - X_1}{T^{-1/2}\sigma} \right) \right) dP^n_{X_1} + F^2_{X_1}(\theta_T) \\
&= \frac{1}{n} \int G^{*\,2}_T \left(\frac{\theta_T - X_1}{T^{-1/2}\sigma} \right) dP^n_{X_1} + \frac{n-1}{n} \mu^2_T(\theta_T) - 2F_{X_1}\mu_T(\theta_T) + F^2_{X_1}(\theta_T) \\
&= \frac{1}{n} \int G^{*\,2}_T \left(\frac{\theta_T - X_1}{T^{-1/2}\sigma} \right) dP^n_{X_1} + \frac{n-1}{n} \mu^2_T(\theta_T) - \frac{2(n-1)}{n} F_{X_1}(\theta_T)\mu_T(\theta_T) + \frac{n-1}{n} F^2_{X_1}(\theta_T) \\
&\quad -\frac{1}{n} F_{X_1}(\theta_T) - \frac{2}{n} F_{X_1}(\theta_T)\mu_T(\theta_T) + \frac{2}{n} F^2_{X_1}(\theta_T) + \frac{1}{n} F_{X_1}(\theta_T) - \frac{1}{n} F^2_{X_1}(\theta_T) \\
&= \frac{1}{n} \int G^{*\,2}_T \left(\frac{\theta_T - X_1}{T^{-1/2}\sigma} \right) dP^n_{X_1} - \frac{1}{n} F_{X_1}(\theta_T) + \frac{n-1}{n} (\mu_T(\theta_T) - F_{X_1}(\theta_T))^2 \\
&\quad -\frac{2}{n} F_{X_1}(\theta_T)(\mu_T(\theta_T) - F_{X_1}(\theta_T)) + \frac{1}{n} F_{X_1}(\theta_T)(1 - F_{X_1}(\theta_T)) \\
&= \frac{1}{n} v_T(\theta_T) + \frac{n-1}{n} b^2_T(\theta_T) - \frac{2}{n} F_{X_1}(\theta_T) b_T(\theta_T) + \frac{1}{n} F_{X_1}(\theta_T)(1 - F_{X_1}(\theta_T)) \qquad (8.5)
\end{aligned}$$

The last line of Equation (8.5) provides a decomposition of the conditional MSE of $\hat{R}^{BS}$. Its last term has the first order contribution. The behavior of the second and third terms can be determined by obtaining a bound for the bias term $b_T(\theta_T)$. To this end, use integration by parts for the Lesbegue-Stieltjes integral of distribution functions (Shiryaev, 1984, Theorem 11) and a change of variable $z = (\theta_T - x)/(T^{-1/2}\sigma)$ to obtain

$$\begin{aligned}
\mu_T(\theta_T) &= \int_{-\infty}^{\infty} G^*_T \left(\frac{\theta_T - x}{T^{-1/2}\sigma} \right) dF_{X_1}(x) \\
&= \int_{-\infty}^{\infty} F_{X_1}(x) dG^*_T \left(\frac{\theta_T - x}{T^{-1/2}\sigma} \right) \\
&= \int_{-\infty}^{\infty} F_{X_1}(\theta_T - T^{-1/2}\sigma z) dG^*_T(z). \qquad (8.6)
\end{aligned}$$

Also, by the definition of G^*_T,

$$\int_{-\infty}^{\infty} z dG^*_T(z) = \int \sqrt{T}(\theta^*_T - \theta_T)/\sigma dP^*_{\theta_T}$$

$$
\begin{aligned}
&= -\frac{\sqrt{T}}{\sigma}(\theta_T - \theta_T) \\
&= 0. \qquad (8.7)
\end{aligned}
$$

(8.7) is a useful fact: the first order moment of the "kernel" G_T^* is zero even though G_T^* is not guaranteed to be symmetric around $(0, 0.5)$ in finite samples. By (8.6), (8.7) and the theorem assumptions,

$$
\begin{aligned}
|\mu_T(\theta_T) - F_{X_1}(\theta_T)| &= \left| \int_{-\infty}^{\infty} (F_{X_1}(\theta_T - T^{-1/2}\sigma z) - F_{X_1}(\theta_T)) dG_T^*(z) \right| \\
&\leq \left| \int_{-\infty}^{\infty} \left\{ (F_{X_1}(\theta_T - T^{-1/2}\sigma z) - F_{X_1}(\theta_T) - T^{-1/2}\sigma z f_{X_1}(\theta_T)) \right\} dG_T^*(z) \right| \\
&\quad + T^{-1/2}\sigma f_{X_1}(\theta_T) \left| \int_{-\infty}^{\infty} z dG_T^*(z) \right| \\
&\leq A \int_{-\infty}^{\infty} (T^{-1/2}\sigma z)^2 dG_T^*(z) \qquad (8.8)
\end{aligned}
$$

for θ_T in $\mathcal{N}$ defined in the theorem. The last integral can be rewritten in terms of the bootstrap variance estimator $\mathrm{Var}^*(\theta_T^*)$. To see this, use (8.7) to obtain

$$
\begin{aligned}
\int_{-\infty}^{\infty} z^2 dG_T^*(z) &= \int_{-\infty}^{\infty} \left(z - \int_{-\infty}^{\infty} z dG_T^* \right)^2 dG_T^* \\
&= \mathrm{Var}^*[\sqrt{T}(\theta_T^* - \theta_T)/\sigma] \\
&= \frac{T}{\sigma^2}\mathrm{Var}^*(\theta_T^*). \qquad (8.9)
\end{aligned}
$$

By (8.8) and (8.9),

$$
|b_T(\theta_T)| = |\mu_T(\theta_T) - F_{X_1}(\theta_T)| \leq T^{-1}A(T\mathrm{Var}^*(\theta_T^*)) \qquad (8.10)
$$

for $\theta_T \in \mathcal{N}$.

Finally, to see the behavior of the first term of (8.5), apply integration by parts and the change of variable as before to $v_T(\theta_T)$ for $\theta_T \in \mathcal{N}$:

$$
\begin{aligned}
v_T(\theta_T) &= \int_{-\infty}^{\infty} {G_T^*}^2 \left(\frac{\theta_T - X_1}{T^{-1/2}\sigma} \right) dP_{X_1} - F_{X_1}(\theta_T) \\
&= -\int_{-\infty}^{\infty} F_{X_1}(x) d{G_T^*}^2 \left(\frac{\theta_T - X_1}{T^{-1/2}\sigma} \right) + {G_T^*}^2(-\infty)F_{X_1}(\infty) - {G_T^*}^2(\infty)F_{X_1}(-\infty) - F_{X_1}(\theta_T) \\
&= \int_{-\infty}^{\infty} F_{X_1}(\theta_T - T^{-1/2}\sigma z) d{G_T^*}^2(z) - F_{X_1}(\theta_T)
\end{aligned}
$$

$$
\begin{aligned}
&= \int_{-\infty}^{\infty} F_{X_1}(\theta_T - T^{-1/2}\sigma z) dG_T^{*\,2}(z) - F_{X_1}(\theta_T) \\
&= \int_{-\infty}^{\infty} \left(F_{X_1}(\theta_T) - T^{-1/2}\sigma z f_{X_1}(\theta_T) \right) dG_T^{*\,2}(z) + O(T^{-1}) \int_{-\infty}^{\infty} z^2 dG_T^{*\,2}(z) - F_{X_1}(\theta_T) \\
&= -T^{-1/2}\sigma f_{X_1}(\theta_T) \int_{-\infty}^{\infty} z dG_T^{*\,2}(z) + O(T^{-1}) \int_{-\infty}^{\infty} z^2 dG_T^{*\,2}(z). \qquad (8.11)
\end{aligned}
$$

Now notice

$$
\begin{aligned}
\lim_{\alpha\to\infty} \sup_T \int_{\{z^2>\alpha\}} z^2 dG_T^{*\,2}(z) &\leq 2 \lim_{\alpha\to\infty} \sup_T \int_{\{z^2>\alpha\}} z^2 \sup_z G_T^*(z) dG_T^*(z) \\
&= \lim_{\alpha\to\infty} \sup_T \int_{\{z^2>\alpha\}} z^2 dG_T^*(z) \\
&= 0, \qquad (8.12)
\end{aligned}
$$

where the last equality holds because $u_T^{*\,2}$ is uniformly integrable under the stated assumptions (recall G_T^* is the CDF of u_T^*). Now consider a sequence of random variables $Z_T, T = 1, 2, ...$, such that the CDF of Z_T is given by $G_T^{*\,2}(z)$. Then by (8.12), Z_T^2 and Z_T are uniformly integrable. Also note that $\sup |G_T^{*\,2}(z) - \Phi^2(z)| \leq 2 \sup |G_T^*(z) - \Phi(z)| \to_p 0$. Using, for example, Theorem 5.4 in Billingsley (1968), conclude that

$$
\int_{-\infty}^{\infty} z^2 dG_T^{*\,2}(z) \to_p \int_{-\infty}^{\infty} z^2 d\Phi^2(z) \qquad (8.13)
$$

and

$$
\begin{aligned}
\int_{-\infty}^{\infty} z dG_T^{*\,2}(z) &\to_p \int_{-\infty}^{\infty} z d\Phi^2(z) \\
&= 2 \int_{-\infty}^{\infty} z\phi(z)\Phi(z) dz \\
&= 2 \int_{-\infty}^{\infty} \phi^2(z) dz \\
&= \frac{1}{\sqrt{\pi}}. \qquad (8.14)
\end{aligned}
$$

Equations (8.5), (8.10), (8.11), (8.13), (8.14) and the fact that $\theta_T \in \mathcal{N}$ with probability approaching to one imply that

$$
\mathrm{MSE}(\hat{R}^{BS}) = \mathrm{E}_{train} \left[\int (\tilde{F}_{X_1}(\theta_T) - F_{X_1}(\theta_T))^2 dP_{X_1}^n \right]
$$

$$
\begin{aligned}
&= -n^{-1}T^{-1/2}2\sigma f_{X_1}\int_{-\infty}^{\infty} z\phi(z)\Phi(z)dz + O(n^{-1}T^{-1}) + \frac{n-1}{n}O(T^{-1})\mathrm{E}_{train}\left[(T\mathrm{Var}^*(\theta_T^*))^2\right] \\
&\quad -2n^{-1}O(T^{-1})\mathrm{E}_{train}\left[F_{X_1}T\mathrm{Var}^*(\theta_T^*)\right] + n^{-1}E_{train}[F_{X_1}(\theta_T)(1-F_{X_1}(\theta_T))] \\
&= n^{-1}E_{train}[F_{X_1}(\theta_T)(1-F_{X_1}(\theta_T))] - n^{-1}T^{-1/2}\frac{\sigma f_{X_1}(\theta_0)}{\sqrt{\pi}} + o(n^{-1}T^{-1/2}).
\end{aligned}
$$

The result follows. □

Proof of Theorem 2: Expand $F_{X_1}(\theta_T)(1-F_{X_1}(\theta_T))$ around θ_0 using the assumption, and note that $\mathrm{E}_{train}(\theta_T-\theta_0)=0$ to deduce that $\mathrm{E}_{train}[F_{X_1}(\theta_T)(1-F_{X_1}(\theta_T))] = F_{X_1}(\theta_0)(1-F_{X_1}(\theta_0)) + O(T^{-1})$. Then argue as in the proof of Theorem of Falk (1984). □

References

Amemiya, T. (1985). *Advanced Econometrics*. Harvard University Press: Cambridge.

Amemiya, T. and J. L. Powell (1983). A comparison of the logit model and normal discriminant analysis when the independent variables are binary. In *Studies in Econometric, Time Series and Multivariate Statistics* (S. Karlin, T. Amemiya and L. A. Goodman, eds.) 3-30, Academic Press: New York.

Azzalini, A. (1981). A note on the estimation of a distribution function and quantiles by a kernel method. *Biometrika* 68, 326-328.

Billingsley, P. (1968). *Convergence of Probability Measures*. Wiley: New York.

Breiman, L. (1996). Bagging predictors. *Machine Learning* 26, 123-140.

Breiman, L. (1998). Arcing classifiers. *Annals of Statistics* 26, 801-824.

Carlstein, E. (1986). The use of subseries values for estimating the variance of a general statistic from a stationary sequence. *Annals of Statistics* 14, 1171-1179.

Chong, Y. Y. and Hendry, David F., (1986). Econometric evaluation of linear macroeconomic models, *Review of Economic Studies*, 53, 671-690.

Davison, A. C. and P. Hall (1992). On the bias and variability of bootstrap and cross-validation estimates of error rate in discrimination problems. *Biometrika* 79, 279-284.

Davison, A. C. and D. V. Hinkley (1997). *Bootstrap Methods and Their Application*. Cambridge University Press: Cambridge.

Dawid, A. P. (1984) Statistical theory. The prequential approach. *Journal of the Royal Statistical Society Series A* 147 (1984), no. 2, 278–292.

Devroye, L. Györfi, L. and Lugosi, G. (1991). *A Probabilistic Theory of Pattern Recognition*. Springer: New York.

Diebold, Francis X., and Mariano, Roberto S., (1995). Comparing predictive accuracy. *Journal of Business and Economic Statistics* 13, 253-263.

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

Efron, Bradley (1982). *The jackknife, the bootstrap and other resampling plans*. CBMS-NSF Regional Conference Series in Applied Mathematics, 38. Society for Industrial and Applied Mathematics (SIAM), Philadelphia.

Efron, Bradley (1983). Estimating the error rate of a prediction rule: improvement on cross-validation. *Journal of the American Statistical Association* 78, no. 382, 316–331.

Efron, Bradley (1986). How biased is the apparent error rate of a prediction rule? *Journal of the American Statistical Association* 81, no. 394, 461–470.

Efron, Bradley (1992). Jackknife-after-bootstrap standard errors and influence functions. *Journal of the Royal Statistical Society* Series B 54, no. 1, 83–127.

Efron, Bradley and Tibshirani, Robert (1997). Improvements on cross-validation: the .632+ bootstrap method. *Journal of the American Statistical Association* 92, no. 438, 548–560.

Fair, Ray C. (1980) Estimating the expected predictive accuracy of econometric models. *International Economic Review* 21, 355-378.

Fair, Ray C. (1984) *Specification, Estimation, and Analysis of Macroeconometric Models*, Cambridge: Harvard University Press.

Fair, Ray C. (1994) *Testing Macroeconometric Models*. Cambridge: Harvard University Press.

Falk, M. (1983). Relative efficiency and deficiency of kernel type estimators of smooth distribution functions. *Statistica Neerlandica* 37, 73-83.

Falk, M. (1984). Relative efficiency of kernel type estimators of quantiles. *Annals of Statistics* 12, 261-268.

Geisser, Seymour (1993). *Predictive inference. An introduction. Monographs on Statistics and Applied Probability* 55. Chapman and Hall, New York.

Geweke, J. (1994). Bayesian comparison of econometric models. Department of Economics, University of Minnesota, mimeo.

Geweke, J. (1997). Posterior simulators in econometrics. Kreps, D. M. and Wallis, K. F. (eds). *Advances in economics and econometrics: Theory and applications*: Seventh World Congress. 129-165.

Glick, N. (1978). Additive estimators for probabilities of correct classification. *Pattern Recognition* 10, 211-222.

Gong, Gail (1986). Cross-validation, the jackknife, and the bootstrap: Excess error estimation in forward logistic regression. *Journal of the American Statistical Association* 81, 108-113.

Hall, P. (1995). On the biases of error estimators in prediction problems. *Statistics and Probability Letters* 24, 257-262.

Harris, I. R. (1989). Predictive fit for natural exponential families. *Biometrika* 76, 675-684.

Hendry, David F. (1995). *Dynamic Econometrics*. Oxford: Oxford University Press.

Henriksson, R. and Merton, R. (1981). On market timing and investment performance, II: Statistical procedures for evaluating forecasting skills. *Journal of Business* 54, 513-533.

Hjorth, Urban (1982). Model selection and forward validation. *Scandinavian Journal of Statistics* 9, 95-105.

Horowitz, Joel L. (1992). A smoothed maximum score estimator for the binary response model. *Econometrica* 60, no. 3, 505–531.

Hodges, J. L. and Lehmann, E. L. (1970). Deficiency. *Annals of Mathematical Statistics* 41, 783-801.

Kim, Jeankyung and Pollard, David (1990). Cube root asymptotics. *Annals of Statistics*. 18, no. 1, 191–219.

Künsch, H. R. (1989). The jackknife and the bootstrap for general stationary observations. *Annals of Statistics* 17, 1217-1241.

Lachenbruch, P. and Mickey, M. R. (1968). Estimation of error rates in discriminant analysis. *Technometrics* 10, 1-11.

Lo, Andrew W. and MacKinlay, A. Craig (1997). Maximizing predictability in the stock and bond markets. *Macroeconomic Dynamics* 1, 102-134.

Maddala, G. S. (1983). *Limited-Dependent and Qualitative Variables in Econometrics*. Cambridge University Press: Cambridge.

Manski, Charles F. (1975) Maximum score estimation of the stochastic utility model of choice. *Journal of Econometrics* 3, no. 3, 205–228.

McCracken, Michael W. (1998). Robust out-of-sample inference. Louisiana State University, mimeo.

McLachlan, G. J. (1992). *Discriminant Analysis and Statistical Pattern Recognition*. Wiley: New York.

Merton, R. (1981). On market timing and investment performance, II: An equilibrium theory of value of market forecasts. *Journal of Business* 54, 363-406.

Page, J. T. (1985). Error-rate estimation in discriminant analysis. *Technometrics* 27, 189-198.

Pawlak, M. (1988). On the asymptotic properties of smoothed estimators of the classification error rate. *Pattern Recognition* 21, 515-524.

Phillips, P. C. B. (1996). Econometric model determination. *Econometrica* 64, 763-812.

Phillips, P. C. B. and Ploberger, W. (1996). An asymptotic theory of Bayesian inference for time series. *Econometrica* 64, 381-412.

Randles, Ronald H. (1982). On the asymptotic normality of statistics with estimated parameters. *Annals of Statistics* 10, no. 2, 462–474.

Rao and Tibshirani (1997) The out-of bootstrap method for model averaging and selection, University of Toronto, mimeo.

Reiss, R. -D. (1981). Nonparametric estimation of smooth distribution functions. *Scandinavian Journal of Statistics* 8, 116-119.

Rissanen, Jorma (1986a). A predictive least square principle. *IMA Journal of Mathematical Control and Information* 3, 211-222.

Rissanen, Jorma (1986b). Stochastic complexity and modeling. *Annals of Statistics*. 14, no. 3, 1080–1100.

Rissanen, Jorma (1989). *Stochastic complexity in statistical inquiry*. World Scientific Series in Computer Science, 15. World Scientific Publishing Co., Inc., Teaneck, NJ.

Rubin, D. (1981). The Bayesian bootstrap. *Annals of Statistics* 9, 130-134.

Sargent, T. J. (1993). *Bounded Rationality in Macroeconomics*. Oxford University Press: Oxford.

Sedransk, N. and Okamoto, M. (1971). Estimation of the probability of misclassification for a linear discriminant function in the univariate normal case. *Annals of the Institute of Statistical Mathematics* 23, 413-435.

Shiryaev, A. N. (1984). *Probability*. Springer: New York.

Stock, James H. and Watson, Mark W. (1998). A comparison of linear and nonlinear models for forecasting macroeconomic time series. Mimeo.

Swanson, N. R. and White, H. (1997). A model selection approach to real-time macroeconomic forecasting using linear models and artificial neural networks, *Review of Economics and Statistics* 79, 540-550.

Vuong, Q. H. (1989). Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica* 57, 307 - 333.

West, Kenneth D. (1996). Asymptotic inference about predictive ability. *Econometrica* 64, 1067-1084.

West, K. D. and McCracken, M. W. (1998). Regression-based Tests of Predictive Ability. *International Economic Review* 39, 817-840.