

Comparison of Model Selection for Regression

Vladimir Cherkassky

cherkass@ece.umn.edu

Yunqian Ma

myq@ece.umn.edu

*Department of Electrical and Computer Engineering, University of Minnesota,
Minneapolis, MN 55455, U.S.A.*

We discuss empirical comparison of analytical methods for model selection. Currently, there is no consensus on the best method for finite-sample estimation problems, even for the simple case of linear estimators. This article presents empirical comparisons between classical statistical methods—Akaike information criterion (AIC) and Bayesian information criterion (BIC)—and the structural risk minimization (SRM) method, based on Vapnik-Chervonenkis (VC) theory, for regression problems. Our study is motivated by empirical comparisons in Hastie, Tibshirani, and Friedman (2001), which claims that the SRM method performs poorly for model selection and suggests that AIC yields superior predictive performance. Hence, we present empirical comparisons for various data sets and different types of estimators (linear, subset selection, and k -nearest neighbor regression). Our results demonstrate the practical advantages of VC-based model selection; it consistently outperforms AIC for all data sets. In our study, SRM and BIC methods show similar predictive performance. This discrepancy (between empirical results obtained using the same data) is caused by methodological drawbacks in Hastie et al. (2001), especially in their loose interpretation and application of SRM method. Hence, we discuss methodological issues important for meaningful comparisons and practical application of SRM method. We also point out the importance of accurate estimation of model complexity (VC-dimension) for empirical comparisons and propose a new practical estimate of model complexity for k -nearest neighbors regression.

1 Introduction and Background ---

We consider standard regression formulation under a general setting for predictive learning (Vapnik, 1995; Cherkassky & Mulier, 1998; Hastie, Tibshirani, & Friedman, 2001). The goal is to estimate unknown real-valued function in the relationship,

$$y = g(\mathbf{x}) + \varepsilon, \quad (1.1)$$

where ε is independent and identically distributed (i.i.d.) zero-mean random error (noise), \mathbf{x} is a multidimensional input, and y is a scalar output. The estimation is made based on a finite number of samples (training data): (\mathbf{x}_i, y_i) , $(i = 1, \dots, n)$. The training data are i.i.d. generated according to some (unknown) joint probability density function (pdf),

$$p(\mathbf{x}, y) = p(\mathbf{x})p(y | \mathbf{x}). \quad (1.2)$$

The unknown function in equation 1.1 is the mean of the output conditional probability (regression function) $g(\mathbf{x}) = \int yp(y | \mathbf{x}) dy$.

A learning method (or estimation procedure) selects the "best" model $f(\mathbf{x}, \omega_0)$ from a set of approximating functions (or possible models) $f(\mathbf{x}, \omega)$ parameterized by a set of parameters $\omega \in \Omega$. The quality of an approximation is measured by the loss or discrepancy measure $L(y, f(\mathbf{x}, \omega))$. An appropriate loss function for regression is the squared error

$$L(y, f(\mathbf{x}, \omega)) = (y - f(\mathbf{x}, \omega))^2. \quad (1.3)$$

The squared-error loss, equation 1.3, is commonly used for model selection comparisons. The set of functions $f(\mathbf{x}, \omega)$, $\omega \in \Omega$ supported by a learning method may or may not contain the regression function $g(\mathbf{x})$. Thus, learning is the problem of finding the function $f(\mathbf{x}, \omega_0)$ (regressor) that minimizes the prediction risk functional,

$$R(\omega) = \int (y - f(\mathbf{x}, \omega))^2 p(\mathbf{x}, y) d\mathbf{x} dy, \quad (1.4)$$

using only the training data. This risk functional measures the accuracy of the learning method's predictions of unknown target function $g(\mathbf{x})$.

For a given parametric model (with fixed number of parameters), the model parameters are estimated by minimizing the empirical risk:

$$R_{emp}(\omega) = \frac{1}{n} \sum_{i=1}^n (y_i - f(\mathbf{x}_i, \omega))^2. \quad (1.5)$$

The problem of model selection (complexity control) arises when a set of possible models $f(\mathbf{x}, \omega)$ consists of (parametric) models of varying complexity. Then the problem of regression estimation requires optimal selection of model complexity (i.e., the number of parameters) in addition to parameter estimation via minimization of empirical risk (see equation 1.5).

Analytic model selection criteria attempt to estimate (unknown) prediction risk, equation 1.4, as a function of (known) empirical risk, equation 1.5, penalized by some measure of model complexity. Then one selects the model complexity model corresponding to the smallest (estimated) risk. It is important to point out that for finite-sample problems, the goal of model selection is distinctly different from the goal of accurate estimation of prediction risk.

Since all analytic model selection criteria are based on certain assumptions (notably, asymptotic analysis and linearity), it is important to perform empirical comparisons in order to understand their practical usefulness in settings when these assumptions may not hold. Recently, Cherkassky and Mulier (1998), Cherkassky, Shao, Mulier, and Vapnik (1999), and Shao, Cherkassky, and Li (2000) showed such empirical comparisons for several analytic model selection methods for linear and penalized linear estimators and concluded that model selection based on structural risk minimization (SRM) is superior (to other methods). Later, Cherkassky and Shao (2001) and Cherkassky and Kilts (2001) successfully applied SRM-based model selection to wavelet signal denoising. Similarly, Chapelle, Vapnik, and Bengio (2002) suggest an analytic model selection approach for small-sample problems based on Vapnik-Chervonenkis (VC) theory and claim superiority of the VC approach using empirical comparisons. These findings contradict a widely held opinion that VC generalization bounds are too conservative for practical model selection (Bishop, 1995; Ripley, 1996; Duda, Hart, & Stork, 2001). Recently, Hastie et al. (2001) presented empirical comparisons for model selection and concluded that "SRM performs poorly overall."

This article is intended to clarify the current state of affairs regarding the practical usefulness of SRM model selection. In addition, we address important methodological issues related to empirical comparisons in general and meaningful application of SRM model selection in particular. The article is organized as follows. Section 2 describes classical model selection criteria and the VC-based approach used for empirical comparisons. Section 3 describes empirical comparison for low-dimensional data sets using linear estimators and k -nearest neighbors method. Section 4 describes empirical comparisons for high-dimensional data sets taken from Hastie et al. (2001) using k -nearest neighbor and linear subset selection regression. Conclusions are presented in section 5.

2 Analytical Model Selection Criteria

In general, analytical estimates of (unknown) prediction risk R_{est} as a function of (known) empirical risk R_{emp} take one of the following forms,

$$R_{est}(d) = R_{emp}(d) \cdot r(d, n) \quad \text{or} \quad R_{est}(d) = R_{emp}(d) + r(d/n, \sigma^2),$$

where $r(d, n)$ is often called the penalization factor, which is a monotonically increasing function of the ratio of model complexity (degrees of freedom) d to the number of samples n .

In this article, we discuss three model selection methods. The first two are representative statistical methods: Akaike information criterion (AIC)

and Bayesian information criterion (BIC),

$$AIC(d) = R_{emp}(d) + \frac{2d}{n} \hat{\sigma}^2 \quad (2.1)$$

$$BIC(d) = R_{emp}(d) + (\ln n) \frac{d}{n} \hat{\sigma}^2, \quad (2.2)$$

where d is the number of free parameters (of a linear estimator) and $\hat{\sigma}^2$ denotes an estimate of noise variance in equation 1.1. Both AIC and BIC are derived using asymptotic analysis (i.e., a large sample size). In addition, AIC assumes that the correct model belongs to the set of possible models. In practice, however, AIC and BIC are often used when these assumptions do not hold. When using AIC or BIC for practical model selection, one is faced with two issues.

The first issue is the estimation and meaning of (unknown) noise variance. When using a linear estimator with d parameters, the noise variance can be estimated from the training data as

$$\hat{\sigma}^2 = \frac{n}{n-d} \cdot \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2. \quad (2.3)$$

Then one can use equation 2.3 in conjunction with AIC or BIC in one of two possible ways. Under the first approach, one estimates noise via equation 2.3 for each (fixed) model complexity (Cherkassky et al., 1999; Chapelle et al., 2002). Thus, different noise estimates are used in AIC or BIC expression for each (chosen) model complexity. This leads to the following form of AIC known as final prediction error (FPE) (Akaike, 1970):

$$FPE(d) = \frac{1 + d/n}{1 - d/n} R_{emp}(d). \quad (2.4)$$

Under the second approach, one first estimates noise via equation 2.3 using a high-variance/low-bias estimator, and then this noise estimate is plugged into AIC or BIC expressions 2.1 or 2.2 to select optimal model complexity. In this article, we use the latter approach since it has been used in empirical comparisons by Hastie et al. (2001); however, there seems to be no consensus on which approach is best for practical model selection. Further, even though one can obtain an estimate of noise variance (as outlined above), the very interpretation of noise becomes difficult for practical problems when the set of possible models does not contain the true target function. In this case, it is not clear whether the notion of noise refers to a discrepancy between admissible models and training data or reflects the difference between the true target function and the training data as in formulation 1.1. In particular, noise estimation becomes impossible when there is significant mismatch

between an unknown target function and an estimator. Unfortunately, all data sets used in Hastie et al. (2001) are of this sort. For example, they use k -nearest neighbor regression to estimate discontinuous target functions, even though it is well known that kernel estimators are intended for smooth target functions (Härdle, 1995). Hence, all empirical comparisons presented in this article assume that for AIC and BIC methods, the variance of additive noise in equation 1.1 is known. This helps to avoid ambiguity with respect to using different strategies for noise estimation and gives an additional competitive advantage to AIC/BIC versus SRM.

The second issue is estimation of model complexity. AIC and BIC use the number of free parameters (for linear estimators), but it is not clear what is a good measure of complexity for other types of estimators (e.g., penalized linear, subset selection, k -nearest neighbors). There have been several suitable generalizations of model complexity (known as effective degrees-of-freedom; Bishop, 1995; Cherkassky & Mulier, 1998; Hastie et al., 2001).

The third model selection method used in this article is based on SRM, which provides a very general and powerful framework for model complexity control. Under SRM, a set of possible models forms a nested structure, so that each element (of this structure) represents a set of models of fixed complexity. Hence, a structure provides a natural ordering of possible models according to their complexity. Model selection amounts to choosing an optimal element of a structure using VC generalization bounds. For regression problems, we use the following VC bound (Vapnik, 1998; Cherkassky & Mulier, 1998; Cherkassky et al., 1999):

$$R(h) \leq R_{emp}(h) \left(1 - \sqrt{p - p \ln p + \frac{\ln n}{2n}} \right)_+^{-1}, \quad (2.5)$$

where $p = h/n$ and h is a measure of model complexity (called the VC-dimension). The practical form of the VC-bound, equation 2.5, is a special case of the general analytical bound (Vapnik, 1995) with appropriately chosen practical values of theoretical constants. (See Cherkassky et al., 1999, for detailed derivation of equation 2.5 from the general bounds developed in VC-theory.)

According to the SRM method, model selection amounts to choosing the model minimizing the upper bound on prediction risk (see equation 2.5). Note that under the SRM approach, we do not need to estimate noise variance. The bound, equation 2.5, has been derived under very general assumptions (e.g., finite-sample setting, nonlinear estimation). However, it requires an accurate estimation of the VC-dimension. Here we are faced with the same problem as estimating effective degrees of freedom (DoF) in the AIC or BIC method: the VC-dimension coincides with the number of free parameters for linear estimators but is (usually) hard to obtain for other types of estimators.

This discussion suggests the following commonsense methodology for empirical comparisons of AIC, BIC, and SRM. First, perform comparisons for linear estimators, for which model complexity can be accurately estimated for all methods. Then perform comparisons for other types of estimators using either advanced methods for estimating model complexity (Vapnik, Levin, & Cun, 1994; Shao et al., 2000) or crude heuristic estimates of model complexity. The latter approach is pursued in this article for k -nearest neighbor regression (where accurate estimates of model complexity are not known).

3 Empirical Comparisons for Linear Estimators and k -Nearest Neighbors

We describe first experimental comparisons of AIC, BIC, and SRM for linear estimators using comparison methodology and data sets from Cherkassky et al. (1999). It is important to obtain meaningful comparisons for model selection with linear estimators first, since in this case the model complexity (DoF in AIC/BIC and VC-dimension in SRM) can be analytically estimated. Later (in section 4) we show comparisons using data sets from Hastie et al. (2001) for (nonlinear) estimators with crude estimates of the model complexity used for model selection criteria.

First, we describe the experimental setup and comparison metrics and then show empirical comparison results.

3.1 Experimental Setup

3.1.1 *Target Functions Used.* The following target functions were used:

Sine-squared function: $g_1(x) = \sin^2(2\pi x)$ $x \in [0, 1]$

Discontinuous piecewise polynomial function:

$$g_2(x) = \begin{cases} 4x^2(3 - 4x) & x \in [0, 0.5] \\ (4/3)x(4x^2 - 10x + 7) - 3/2 & x \in (0.5, 0.75] \\ (16/3)x(x - 1)^2 & x \in (0.75, 1] \end{cases}$$

Two-dimensional *sinc* function:

$$g_3(\mathbf{x}) = \sin \sqrt{x_1^2 + x_2^2} / \sqrt{x_1^2 + x_2^2} \quad \mathbf{x} \in [-5, 5]^2$$

In addition, the training and test samples were uniformly distributed in the input domain.

3.1.2 *Estimators Used.* The linear estimators include polynomial and trigonometric estimators, that is,

$$\text{Algebraic polynomials: } f_m(x, \omega) = \sum_{i=0}^{m-1} \omega_i x^i + \omega_0 \quad (3.1)$$

$$\text{Trigonometric expansion: } f_m(x, \omega) = \sum_{i=0}^{m-1} \omega_i \cos(ix) + \omega_0 \quad (3.2)$$

For linear estimators, the number of parameters (DoF) is used as a measure of model complexity (VC-dimension) for all model selection methods.

In the *k-nearest neighbors* regression, the unknown function is estimated by taking a local average of k training samples nearest to the estimation point. In this case, an estimate of effective DoF or VC-dimension is not known, even though sometimes the ratio n/k is used to estimate model complexity (Hastie et al., 2001). However, this estimate appears too crude and can be criticized using both commonsense and theoretical arguments, as discussed next. With the *k-nearest neighbors* method, the training data can be divided into n/k neighborhoods. If the neighborhoods were nonoverlapping, then one could fit one parameter in each neighborhood (leading to an estimate $d = n/k$). However, the neighborhoods are, in fact, overlapping, so that a sample point from one neighborhood affects regression estimates in an adjacent neighborhood. This suggests that a better estimate of DoF has the form $d = n/(c * k)$ where $c > 1$. The value of (unknown) parameter c is unknown but (hopefully) can be determined empirically or using additional theoretical arguments. Using the ratio n/k to estimate model complexity is inconsistent with the main result in VC-theory (that the VC-dimension of any estimator should be finite). Indeed, asymptotically (for large n), the ratio n/k grows without bound. On the other hand, asymptotic theory for *k-nearest neighbor* estimators (Härdle, 1995) provides asymptotically optimal k -values (when n is large), namely, $k \sim n^{4/5}$. This suggests the following (asymptotic) dependency for DoF: $d \sim \frac{n}{k} * \frac{1}{n^{1/5}}$. This (asymptotic) formula is clearly consistent with commonsense expression $d = n/(c^*k)$ with $c > 1$. We found that a good practical estimate of DoF can be found empirically by assuming the dependency

$$d = \text{const} * n^{4/5} / k \quad (3.3)$$

and then empirically “tuning” the value of $\text{const} = 1$ using a number of data sets. This leads to the following empirical estimate for DoF:

$$d = \frac{n}{k} * \frac{1}{n^{1/5}}. \quad (3.4)$$

Prescription 3.4 is used as an estimate of DoF and VC-dimension for all model selection comparisons presented later in this article. We point out that

equation 3.4 is a crude practical measure of model complexity. However, it is certainly better than an estimate $d = n/k$ used in Hastie et al. (2001) because using expression 3.4 instead of $d = n/k$ actually improves the prediction accuracy of all model selection methods (AIC, BIC, and SRM) for all data sets used in our comparisons (including those used in Hastie et al. (2001)). Also, proposed estimate 3.4 is consistent with DoF estimates provided by asymptotic theory and commonsense arguments.

3.1.3 Noise Estimation. All comparisons use the true noise variance for AIC and BIC methods. Hence, AIC and BIC have an additional competitive advantage versus the SRM method.

3.2 Experimental Procedure. A training set of fixed size (n) is generated using standard regression formulation 1.1 using a target function with y -values corrupted by additive gaussian noise. The x -values of training data follow random uniform distribution in the input domain. Then the model (of fixed complexity) is estimated using training data (via least-squares fitting for linear estimators or k -nearest neighbors regression). Model selection is performed by (least-squares) fitting of models of different complexity and selecting an optimal model corresponding to the smallest prediction risk as estimated by a given model selection criterion. Prediction risk (accuracy) of the model chosen by a model selection method is then measured (experimentally). The *prediction risk* is measured as the mean squared error (MSE) between a model (estimated from training data) and the true values of target function $g(\mathbf{x})$ for independently generated test inputs:

$$R_{pred} = \frac{1}{n_{test}} \sum_{i=1}^{n_{test}} (g(\mathbf{x}_i) - \hat{y}_i)^2.$$

The above experimental procedure is repeated 100 times using 100 different random realizations of n training samples (from the same statistical distribution), and the empirical distribution of the prediction risk (observed for each model selection method) is displayed using standard box plot notation, with marks at 95%, 75%, 50%, 25%, and 5% of the empirical distribution.

3.2.1 Experiment 1. The training data are generated using the sine-squared target function corrupted by gaussian noise. A linear estimator (using algebraic polynomials) was used. Experiments were performed using a small training sample ($n = 30$) and a large sample size ($n = 100$). Figure 1 shows comparison results for AIC, BIC, and SRM for noise level $\sigma = 0.2$.

3.2.2 Experiment 2. The training data are generated using discontinuous piecewise polynomial target function corrupted by gaussian noise. A

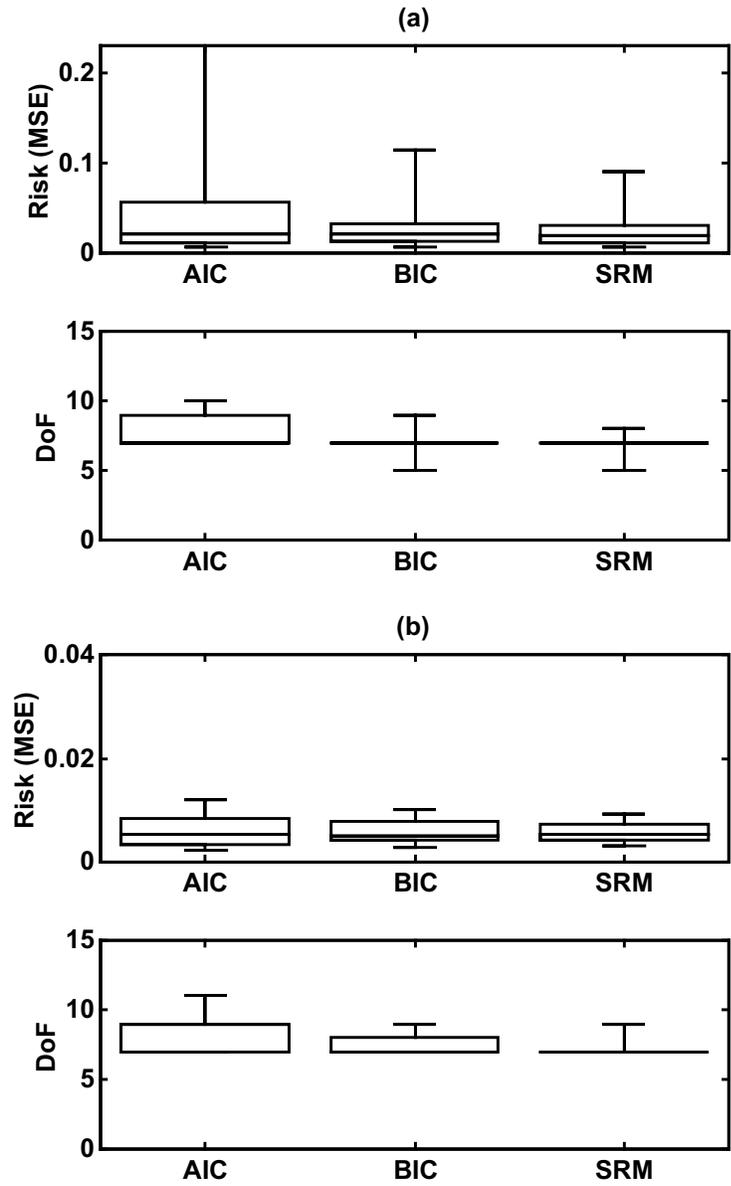


Figure 1: Comparison results for sine-squared target function estimated using polynomial regression, noise level $\sigma = 0.2$. (a) Small size, $n = 30$. (b) Large size, $n = 100$.

linear estimator (using trigonometric expansion) was used. Experiments were performed using a small training sample ($n = 30$) and different noise levels. Figure 2 shows comparison results for AIC, BIC, and SRM.

Comparison results presented in Figures 1 and 2 indicate that SRM and BIC methods work better than AIC for small sample sizes ($n = 30$). For large samples (see Figure 1b), all methods show comparable (similar) prediction accuracy; however, SRM is still preferable to other methods since it selects lower model complexity (i.e., lower-order polynomials). These findings are consistent with model selection comparisons for linear estimators presented in Cherkassky et al. (1999).

3.2.3 Experiment 3. Here we compare model selection methods using k -nearest neighbors regression. The training and test data are generated using the same target functions (sine-squared and piecewise polynomial) as in previous experiments. The training sample size is 30 and the noise level $\sigma = 0.2$. Comparison results are shown in Figure 3. We can see that all methods provide similar prediction accuracy (with SRM and BIC having a slight edge over AIC). However, SRM is more robust since it selects models of lower complexity (larger k) than BIC and AIC. Note that all comparisons use the proposed (estimated) model complexity (see equation 3.4) for the k -nearest neighbors method. Using (arguably incorrect) estimates of model complexity n/k as in Hastie et al. (2001) can obscure model selection comparisons. In this case, model selection comparisons for estimating sine-squared target function using k -nearest neighbors are shown in Figure 4. According to Figure 4, AIC provides the best model selection; however, its prediction accuracy is lower than with “more accurate” DoF estimates (see equation 3.4) shown in Figure 3a. By comparing results in Figure 3a and Figure 4, we conclude that the proposed model complexity estimate, equation 3.4, actually improves prediction performance of all three methods (AIC, BIC, and SRM) relative to using $\text{DoF} = n/k$. In addition, prediction performance of SRM relative to AIC and BIC degrades significantly when using estimates of $\text{DoF} = n/k$. This can be explained by the multiplicative form of VC-bound (see equation 2.5) that is affected more significantly by incorrect estimates of the model complexity (VC-dimension) than AIC and BIC, which have an additive form.

3.2.4 Experiment 4. Here we compare model selection using k -nearest neighbors regression to estimate two-dimensional *sinc* target function corrupted by gaussian noise (with $\sigma = 0.2$ and 0.4). The training size is 50, and the test size is 300. Comparison results in Figure 5 indicate that the VC-based approach yields better performance than AIC, and its performance is similar to BIC.

All experiments with low-dimensional data sets indicate that SRM model selection yields better prediction accuracy than AIC, and its performance is similar to BIC. This conclusion is obtained in spite of the fact that compar-

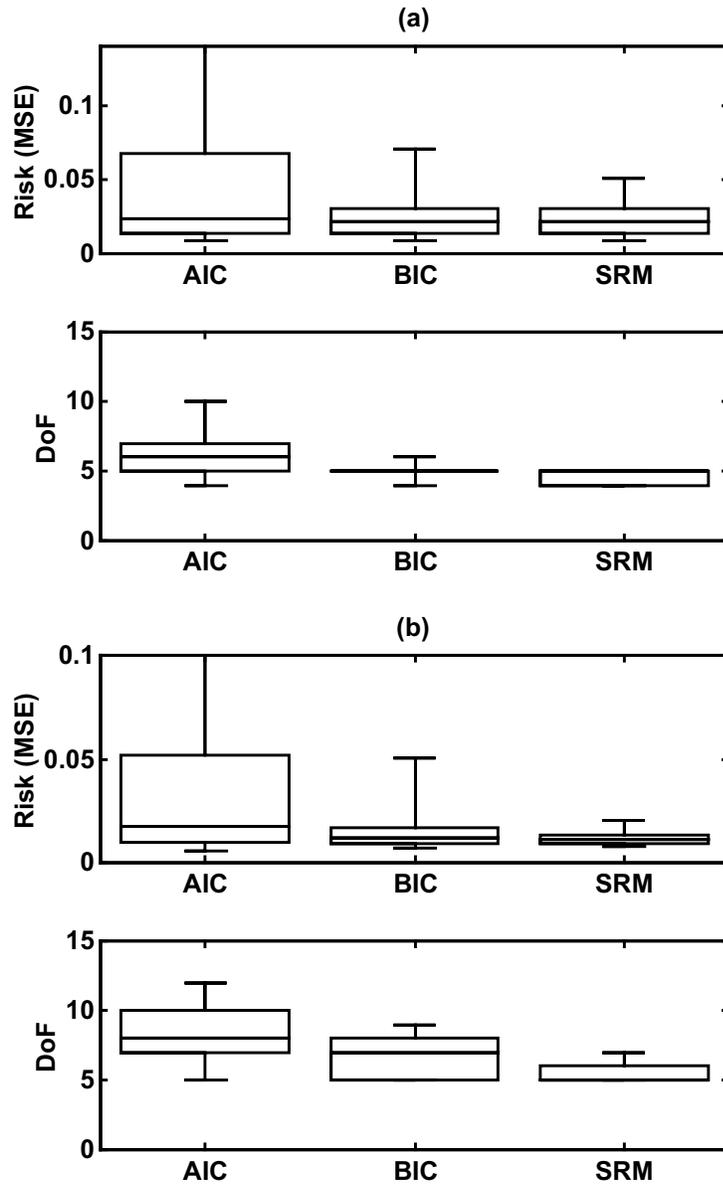


Figure 2: Comparison results for piecewise polynomial target function estimated using trigonometric regression. Sample size, $n = 30$. (a) $\sigma = 0.2$; SNR = 1.5. (b) $\sigma = 0.1$, SNR = 3.

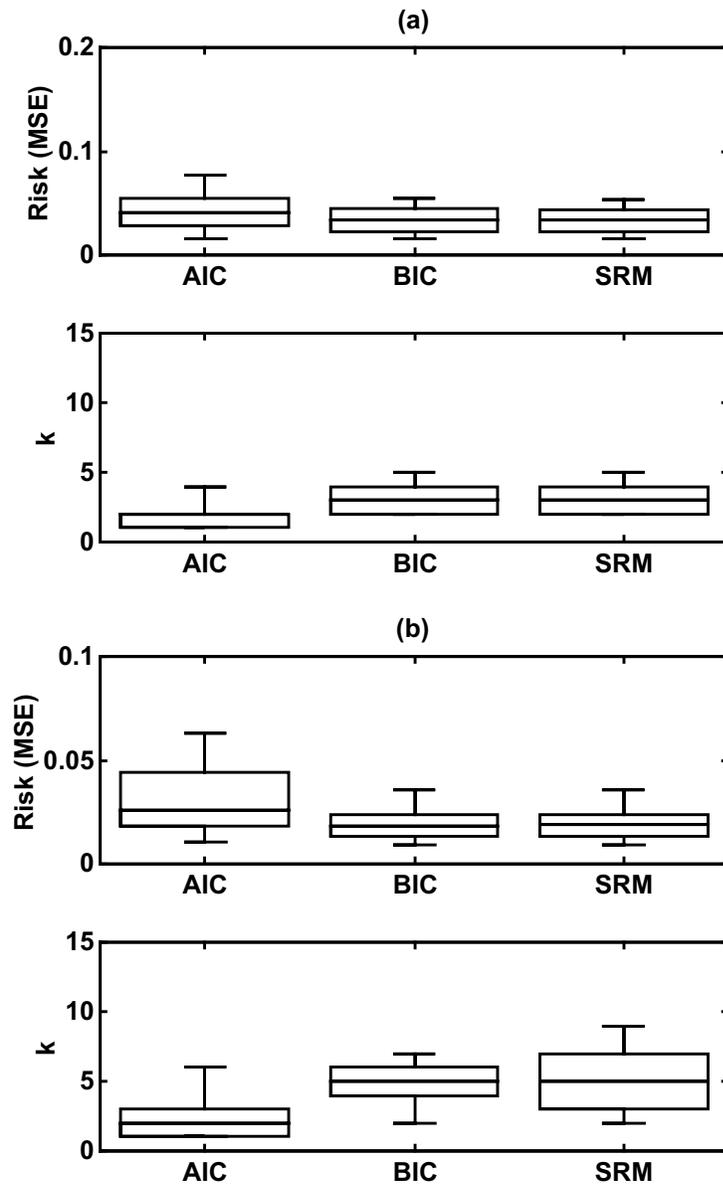


Figure 3: Comparison results for univariate regression using k -nearest neighbors. Training data: $n = 30$; noise level $\sigma = 0.2$. (a) Sine squared target function. (b) Piecewise polynomial target function.

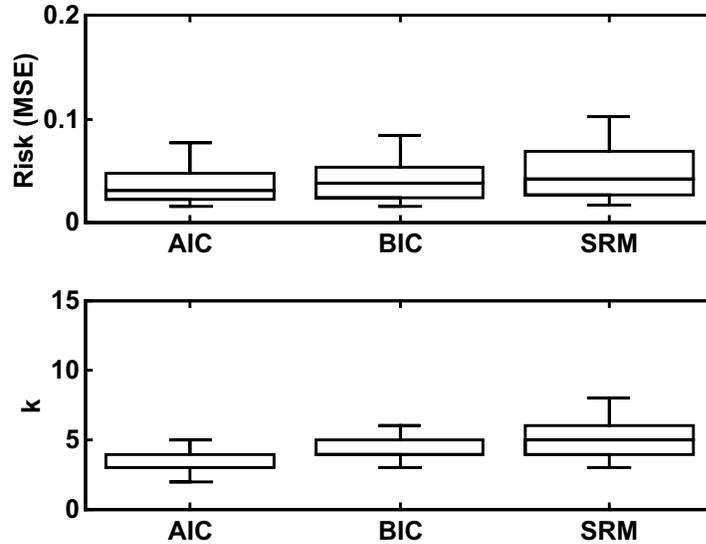


Figure 4: Comparison results for sine-squared target function using k -nearest neighbors when $\text{DoF} = n/k$ for all methods. Training data: $n = 30$; noise level $\sigma = 0.2$.

isons were biased in favor of AIC/BIC since we used the true noise variance for AIC and BIC. One can also see from our experimental results that SRM is more robust than BIC, since it consistently selects lower model complexity, when both methods provide similar prediction performance. Hence, we conclude that SRM is a better choice for practical model selection than AIC or BIC. Our conclusions strongly disagree with conclusions presented in Hastie et al. (2001), who used high-dimensional data sets for model selection comparisons. The reasons for such disagreement are discussed in the next section.

4 Comparisons for High-Dimensional Data Sets

This section presents comparisons for higher-dimensional data sets taken from (or similar to) Hastie et al. (2001). These data sets are used in Hastie et al. (2001) to compare model selection methods for two types of estimators: k -nearest neighbors (described in section 3) and linear subset selection (discussed later in this section). Next, we present comparisons using the same or similar data sets, but our results show an overall superiority of SRM method.

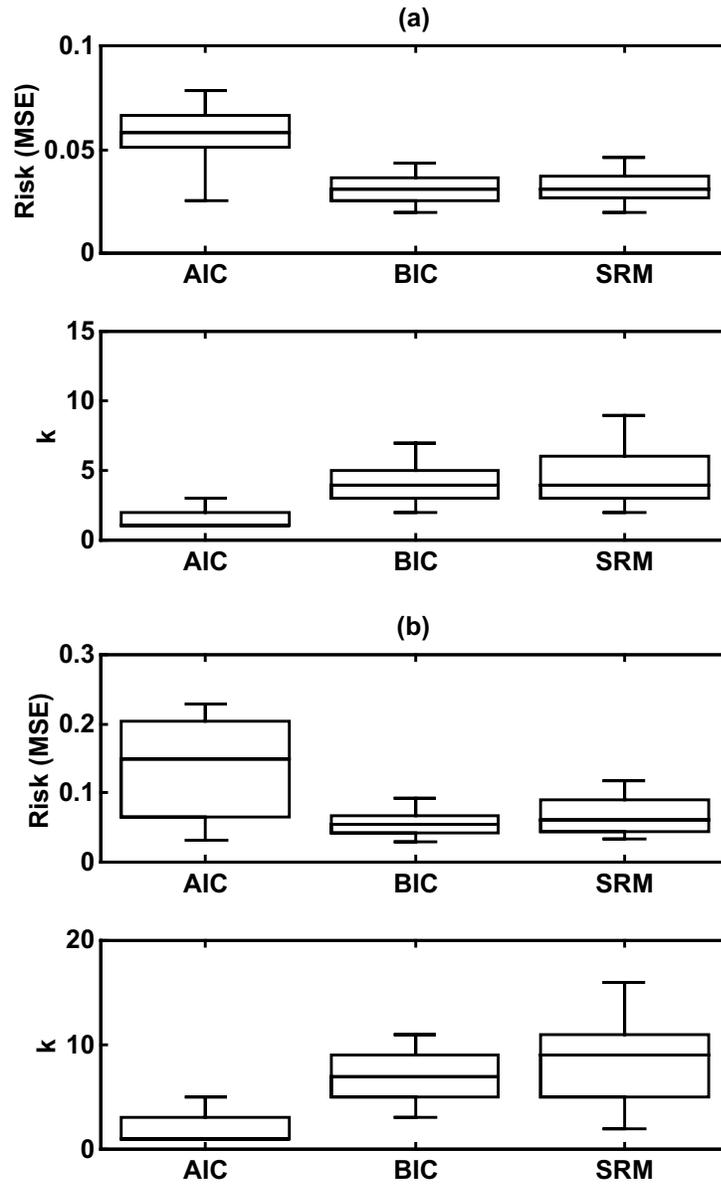


Figure 5: Comparison results for two-dimensional *sinc* target function using k -nearest neighbors; sample size $n = 50$. (a) $\sigma = 0.2$. (b) $\sigma = 0.4$.

4.1 Comparisons for k -Nearest Neighbors. Let us consider a multivariate target function of 20 input variables $\mathbf{x} \in R^{20}$ uniformly distributed in $[0, 1]^{20}$:

$$g_4(\mathbf{x}) = \begin{cases} 1 & \text{if } x_1 > 0.5 \\ 0 & \text{if } x_1 \leq 0.5 \end{cases} \quad (4.1)$$

4.1.1 Experiment 5. We estimate target function 4.1 from $n = 50$ training samples, using k -nearest neighbor regression. The quality of estimated models is evaluated using the test set of 500 samples. We use the true noise variance for AIC and BIC methods. All methods use the proposed DoF estimate, equation 3.4. Results in Figure 6a correspond to the example in Hastie et al. (2001) when training samples are not corrupted by noise. Results in Figure 6b show comparisons for noisy training samples. In both cases, SRM achieves better prediction performance than other methods. Results presented in Figure 6 are in complete disagreement with empirical results and conclusions presented in Hastie et al. (2001). There are several reasons for this disagreement. First, Hastie et al. (2001) use their own (incorrect) interpretation of theoretical VC-bound that is completely different from the VC-bound equation 2.5. Second, discontinuous target function (see equation 4.1) used by Hastie et al. (2001) prevents any meaningful application of AIC/BIC model selection with k -nearest neighbors. That is, kernel estimation methods assume sufficiently smooth (or at least, continuous) target functions. When this assumption is violated (as in the case of discontinuous step function), it becomes impossible to estimate the noise variance (for AIC/BIC method) with any reasonable accuracy. For example, using 5-nearest neighbor regression to estimate noise for this data set, as recommended by Hastie et al. (2001), gives the noise variance 0.15, which is completely wrong (since the data has no additive noise). Third, application of AIC/BIC to the data set with no additive noise (as suggested by Hastie et al., 2001) makes little sense as well, because the additive form 2.1 and 2.2 for AIC and BIC implies that with zero noise, these methods should select the highest model complexity. That is why we used a small noise variance 0.001 (rather than zero) for AIC/BIC in our comparisons for this data set. Graphical illustration of overfitting by AIC/BIC for this data set can be clearly seen in Figure 7, showing the values of AIC, BIC and VC-bound as a function of k , along with the true prediction risk. In contrast, the SRM method performs very accurate model selection for this data set.

4.2 Comparisons for Linear Subset Selection. Here, the training and test data are generated using a five-dimensional target function $\mathbf{x} \in R^5$ and

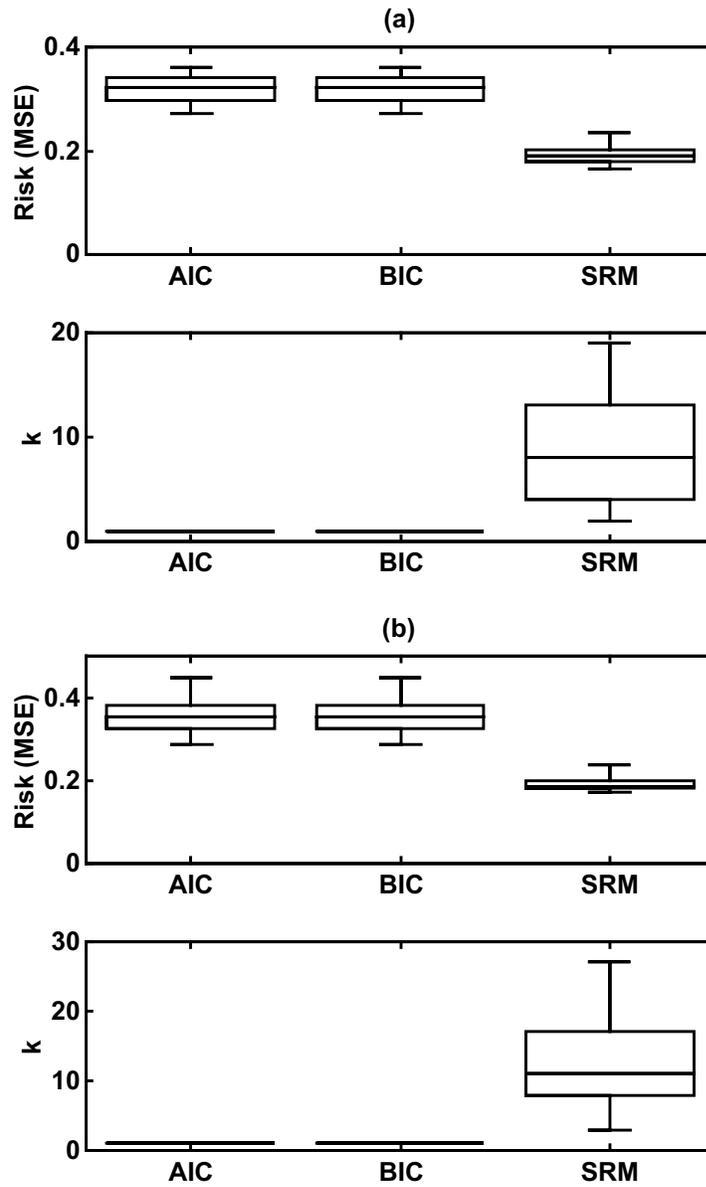


Figure 6: Comparisons for high-dimensional target function 4.1 using k -nearest neighbors method; sample size $n = 50$. (a) $\sigma = 0$. (b) $\sigma = 0.2$.

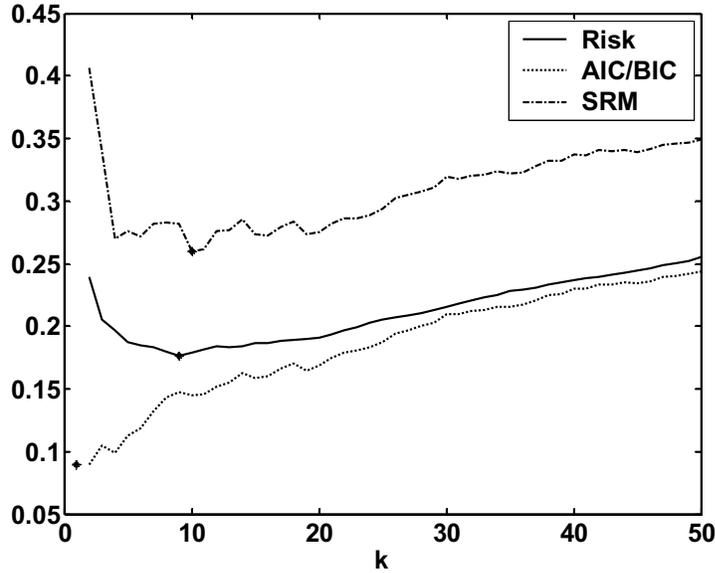


Figure 7: True prediction risk and its estimates provided by AIC/BIC and SRM (VC-bound) as a function of k (number of nearest neighbors) for a single realization of training data for high-dimensional target function 4.1 with sample size $n = 50$ and noise $\sigma = 0$. Risk estimates given by AIC and BIC are identical due to zero noise (assumed to be known). Minimal values of risk and estimated risk are marked with a star.

$y \in R$, defined as

$$g_5(\mathbf{x}) = \begin{cases} 1 & \text{if } \sum_{j=1}^3 x_j > 1.5 \\ 0 & \text{if } \sum_{j=1}^3 x_j \leq 1.5 \end{cases} \quad (4.2)$$

with random \mathbf{x} -values uniformly distributed in $[0, 1]^5$.

The linear subset selection method amounts to selecting the best subset of m input variables for a given training sample. Here the “best” subset of m variables yields the linear model with the lowest empirical risk (MSE fitting error) among all linear models with m variables, for a given training sample. Hence, for linear subset selection, model selection corresponds to selecting an optimal value of m (providing minimum prediction risk). Also note that linear subset selection is a nonlinear estimator, even though it produces models linear in parameters. Hence, there may be a problem of

estimating its model complexity when applying AIC, BIC, or SRM for model selection. In this article, we estimate the model complexity (DoF) as $m + 1$ (where m is the number of chosen input variables) for all methods, similar to Hastie et al. (2001). We do not know, however, whether this estimate is correct. Implementation of subset selection used in this article performs an exhaustive search over all possible subsets of m variables (out of total d input variables) for choosing the best subset (minimizing the empirical risk). Hence, we used a moderate number of input variables ($d = 5$) to avoid the combinatorial explosion of subset selection.

4.2.1 Experiment 6. Training data and test data are generated using target function 4.2. We used true noise variance for AIC and BIC methods. Our comparisons use 30 training samples and 200 test samples. Regression estimates are obtained from (noisy) training data using linear subset selection. Comparison results in Figure 8 show that all methods achieve similar prediction performance when the training data have no noise. For noisy data, BIC provides superior prediction performance, and both AIC and BIC perform better than SRM. A closer look at results in Figure 8 helps to explain the methods' performance for this data set. Results in Figure 8a clearly show that overfitting by AIC and BIC does not result in the degradation of prediction risk, whereas underfitting (by SRM) can adversely affect prediction performance (see Figure 8b). Also, from DoF box plots in Figure 8b, SRM selects optimal model complexity (DoF = 4) most of the time, unlike AIC/BIC, which tend to overfit. However, this does not yield better prediction performance for SRM in terms of prediction risk. This is further explained in Figure 9, which shows the dependence of prediction risk on degrees of freedom (the number of variables +1) for one (representative) realization of training data. From Figure 9, we can clearly see that there is no overfitting. This happens due to the choice of approximating functions (i.e., linear subset selection), which results in a large bias (mismatch) for the chosen target function, 4.2. Even using the most complex linear subset selection model, one cannot accurately fit the training data generated by this target function. In other words, there is no possibility of overfitting for this data set; hence, a more conservative model selection approach (such as SRM) tends to underperform relative to AIC and BIC.

In order to perform more meaningful comparisons for the linear subset selection method, consider the data set where the target function belongs to a set of possible models (approximating functions): the training and test data are generated using five-dimensional target function $\mathbf{x} \in R^5$ and $y \in R$, defined as

$$g_6(\mathbf{x}) = x_1 + 2x_2 + x_3 \quad (4.3)$$

with \mathbf{x} -values with uniformly distributed in $[0, 1]^5$.

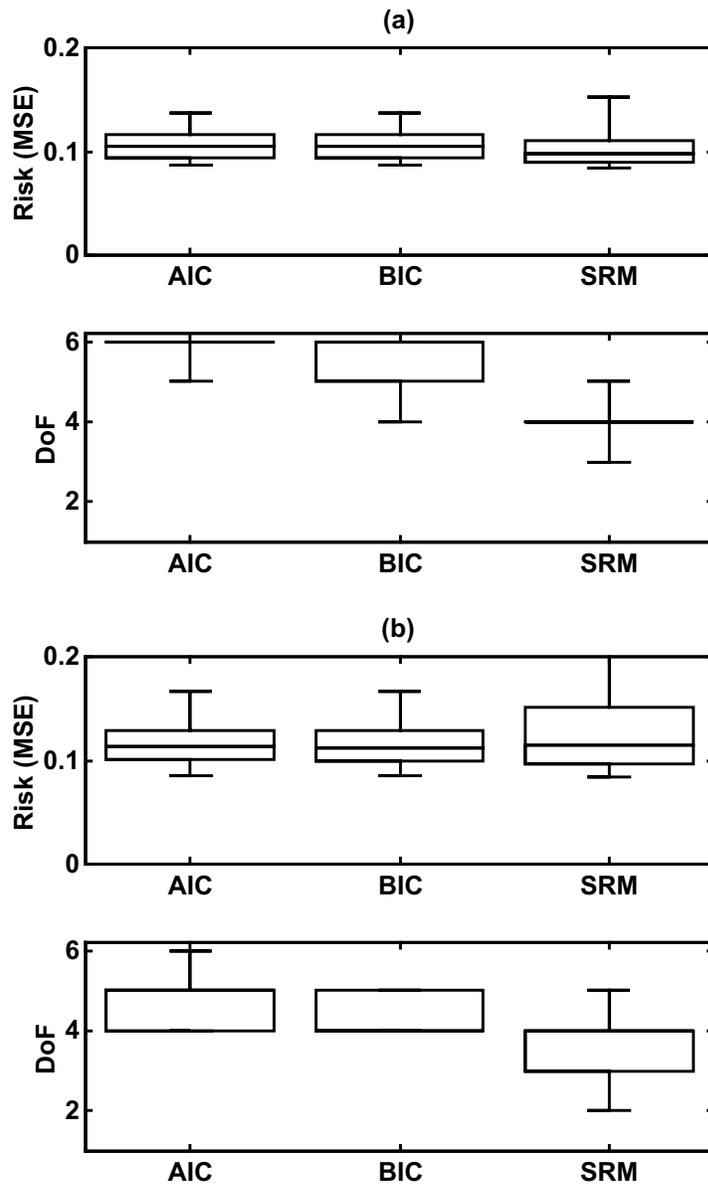


Figure 8: Comparison results for high-dimensional target function 4.2 using linear subset selection for $n = 30$ samples. (a) $\sigma = 0$. (b) $\sigma = 0.2$.

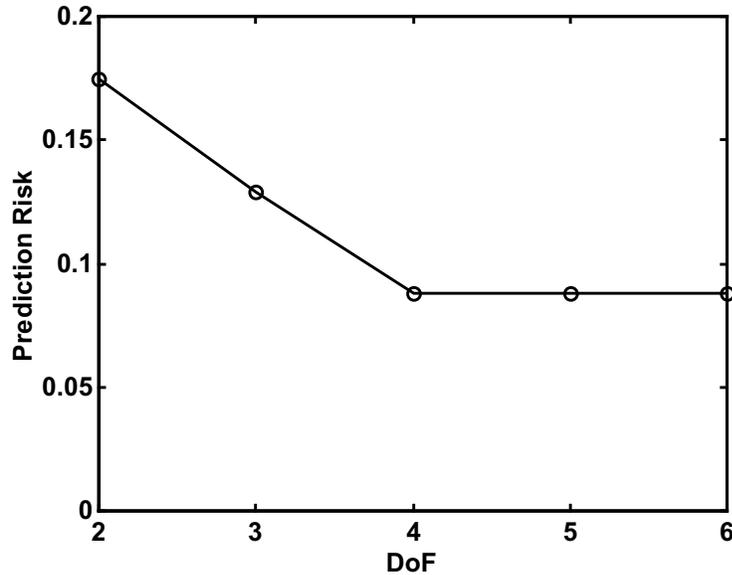


Figure 9: Prediction risk as a function of DoF with linear subset selection for a single realization of training data generated using high-dimensional target function 4.2; $n = 30$ samples; $\sigma = 0$.

Experimental comparisons of model selection for this data set are shown in Figure 10. Note that experimental setup and the properties of training data (sample size, noise level $\sigma = 0.2$) are identical to the setting used to produce comparisons in Figure 8b, except that we use the target function, 4.3. Results shown in Figure 10 indicate that SRM and BIC have similar prediction performance (both better than AIC) for this data set.

Note that all comparisons assume known noise level for AIC/BIC; hence, they are biased in favor of AIC/BIC. Even with this bias, SRM performs better (or at least not worse) than AIC/BIC for linear subset selection. Our findings contradict comparisons presented by Hastie et al. (2001). This can be explained by the contrived data set used in their comparisons (similar to equation 4.2), such that no overfitting is possible with linear subset selection.

5 Discussion

Given the proliferation of ad hoc empirical comparisons and heuristic learning methods, the issue of what constitutes meaningful comparisons becomes of great practical importance. Hence, in this section, we discuss several methodological implications of our comparison study and suggest future research directions. The first issue to address is what the goal of empirical

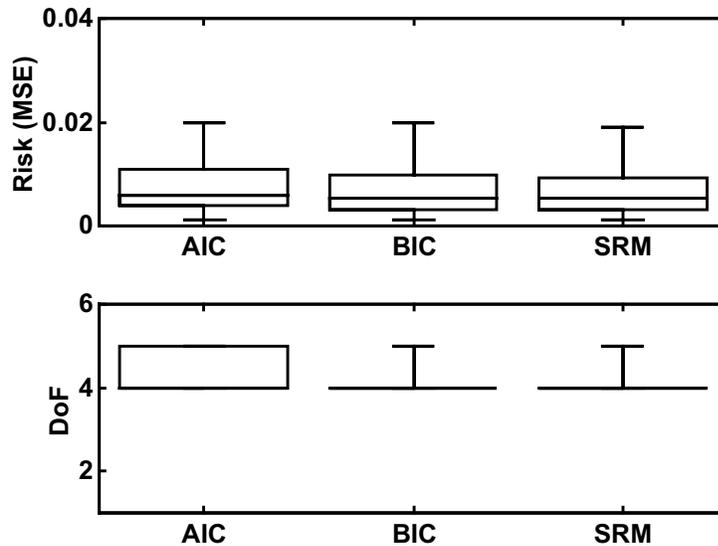


Figure 10: Comparison results for high-dimensional target function 4.3 using linear subset selection for $n = 30$ samples; noise level $\sigma = 0.2$.

comparisons is. Often, the goal seems to be demonstrating that one method (learning technique, model selection criterion) is better than the others, and this can easily be accomplished by using specially chosen (contrived) data sets that favor particular methods or tuning well one method (learning algorithm), while tuning poorly other (competing) methods, often done unintentionally since a person performing comparisons is usually an expert in one method (being proposed).

In fact, according to Vapnik (1998), generalization from finite data is possible only when an estimator has limited capacity (i.e., complexity). Therefore, a learning method cannot solve most practical problems with finite samples unless it uses a set of approximating functions (admissible models) appropriate for a problem at hand. Ignoring this commonsense observation leads to formal proofs that no learning method generalizes better than another for all data sets (i.e., results collectively known as *no-free-lunch theorems*; see Duda et al., 2001). In this sense, a mismatch between estimation methods and data sets used for model selection comparisons in Hastie et al. (2001) leads to particularly confusing and misleading conclusions.

In our opinion, the goal of empirical comparisons of methods for model selection should be improved understanding of their relative performance for finite-sample problems. Since the model complexity (VC-dimension) can be accurately estimated for linear methods, it is methodologically appropriate to perform such comparisons for linear regression first, before consid-

ering other types of estimators. Recent comparison studies (Cherkassky & Mulier, 1998; Cherkassky et al., 1999; Shao et al., 2000) indicate that VC-based model selection is superior to other analytic model selection approaches for linear and penalized linear regression with finite samples. Empirical comparisons between SRM, AIC, and BIC for linear estimators presented in this article confirm practical advantages of SRM model selection. In addition, we propose a new practical estimate of model complexity for k -nearest neighbor regression and use it for model selection comparisons. This article presents the first known practical application of VC-bound equation 2.5 with k -nearest neighbors regression. Our empirical results show the advantages of SRM and BIC (relative to AIC) model selection for k -nearest neighbors. Our empirical comparisons clearly show the practical advantages of SRM model selection for linear subset selection regression.

Future research may be directed toward meaningful comparisons of model selection methods for other (nonlinear) estimators. Here, the main challenge is estimating model complexity and avoiding potential methodological pitfalls of ad hoc comparisons. Empirical comparisons performed by Hastie et al. (2001) illustrate possible dangers of ad hoc comparisons, such as these:

- *Inaccurate estimation of model complexity.* It is impossible to draw any conclusions based on empirical comparisons unless one is confident that model selection criteria use accurate estimates of model complexity. There exist experimental methods for measuring the VC-dimension of an estimator (Vapnik et al., 1994; Shao et al., 2000); however, they may be difficult to apply for general practitioners. An alternative practical approach is to come up with empirical commonsense estimates of model complexity to be used in model selection criteria—for example, in this article, where we successfully used a new complexity measure for k -nearest neighbor regression. Essentially, under this approach, we combine the known analytical form of a model selection criterion with an appropriately tuned measure of model complexity taken as a function of (some) complexity parameter (i.e., the value of k in the k -nearest neighbor method).
- *Using contrived data sets.* For both k -nearest neighbors and linear subset selection comparisons, Hastie et al. (2001) used specially chosen data sets so that meaningful model selection comparisons become difficult or impossible. For example, for subset selection comparisons, they use a data set such that regression estimates do not exhibit any overfitting. This (contrived) setting obviously favors model selection methods (such as AIC) that tend to overfit.
- *Using small number of data sets.* Comparisons presented in Hastie et al. (2001) use just two data sets of fixed size with no noise (added to response variable). Such comparisons can easily be misleading, since

model selection methods typically exhibit different relative performance for different noise levels, sample size, and other factors. Reasonable comparisons (leading to meaningful conclusions) should apply model selection criteria to a broad cross-section of data sets with different statistical characteristics (Cherkassky et al., 1999).

- *Using an inappropriate form of VC-bounds.* The VC-theory provides an analytical form of VC-bounds; however, it does not give specific (practical) values of theoretical parameters and constants. Such (practical) parameter values lead to the practical form of VC-bounds for regression problems used in this article. It appears that Hastie et al. (2001) used the original theoretical VC-bound with poorly chosen parameter values instead of the practical form of VC-bound 2.5 described in Cherkassky and Mulier (1998), Vapnik (1998), and Cherkassky et al. (1999).

Finally, we briefly comment on empirical comparisons of model selection methods for classification, presented in Hastie et al. (2001) using an experimental setting for high-dimensional data sets described in section 4. Whereas Hastie et al. (2001) acknowledge that they do not know how to set practical values of constants in VC-bounds for classification, they proceed with empirical comparisons anyway. We remain highly skeptical about the scientific value of such comparisons. Selecting appropriate “practical” values of theoretical constants in VC bounds for classification is an important and interesting research area.

It may be worthwhile to point out that Hastie et al. (2001) use identical data sets for regression and classification comparisons; the only difference is in the choice of the loss function. Likewise, classical model selection criteria (AIC and BIC) have an identical form (for regression and classification). This observation underscores an important distinction between classical statistics and a VC-based approach to learning and estimating dependencies from data—namely, the classical approach to learning (estimation) problems is to apply maximum likelihood (ML) method for estimating parameters of a density function. Hence, the same approach (density estimation via ML) can be used for both classification and regression, and AIC/BIC prescriptions have an identical form for both types of learning problems. In contrast, the VC approach clearly differentiates between regression and classification settings, since the goal of learning is not density estimation, but rather estimation of certain properties of unknown densities via empirical risk minimization or structural risk minimization. This results in completely different forms of VC-bounds for classification (additive form) and regression (multiplicative form). Based on the above discussion, it becomes clear that Hastie et al. (2001) use identical data sets for both classification and regression comparisons following the classical statistical approach. In contrast, under VC-based methodology, it makes little sense to use the same data sets for classification and regression comparisons, since we are faced with two completely different learning problems.

Acknowledgments

This work was supported by NSF grant ECS-0099906. We also acknowledge two anonymous referees for their helpful comments.

References

- Akaike, H. (1970). Statistical prediction information. *Ann. Inst. Statist. Math*, 22, 203–217.
- Bishop, C. (1995). *Neural networks for pattern recognition*. New York: Oxford University Press.
- Chapelle, O., Vapnik, V., & Bengio, Y. (2002). Model selection for small sample regression. *Machine Learning*, 48(1), 9–23.
- Cherkassky, V., & Mulier, F. (1998). *Learning from data: Concepts, theory and methods*. New York: Wiley.
- Cherkassky, V., Shao, X., Mulier, F., & Vapnik, V. (1999). Model complexity control for regression using VC generalization bounds. *IEEE Transactions on Neural Networks*, 10(5), 1075–1089.
- Cherkassky, V., & Shao, X. (2001). Signal estimation and denoising using VC-theory. *Neural Networks*, 14, 37–52.
- Cherkassky, V., & Kilts, S. (2001). Myopotential denoising of ECG signals using wavelet thresholding methods. *Neural Networks*, 14, 1129–1137.
- Duda, R., Hart, P., & Stork, D. (2001). *Pattern classification* (2nd ed.). New York: Wiley.
- Hardle, W. (1995). *Applied nonparametric regression*. Cambridge: Cambridge University Press.
- Hastie, T., Tibshirani, R., & Friedman, J. (2001). *The elements of statistical learning: Data mining, inference and prediction*. New York: Springer-Verlag.
- Ripley, B. D. (1996). *Pattern recognition and neural networks*. Cambridge: Cambridge University Press.
- Shao, X., Cherkassky, V., & Li, W. (2000). Measuring the VC-dimension using optimized experimental design. *Neural Computation*, 12, 1969–1986.
- Vapnik, V. (1995). *The nature of statistical learning theory*. New York: Springer-Verlag.
- Vapnik, V. (1998). *Statistical learning theory*. New York: Wiley.
- Vapnik, V., Levin, E., & Cun, Y. (1994). Measuring the VC-dimension of a learning machine. *Neural Computation*, 6, 851–876.