Machine Learning, 48, 9–23, 2002

Model Selection for Small Sample Regression

OLIVIER CHAPELLE olivier.chapelle@liple.fr
LIP6, 15 rue du Capitaine Scott, 75015 Paris, France

VLADIMIR VAPNIK vlad@research.att.com
AT&T Research Labs, 200 Laurel Avenue, Middletown, NJ 07748, USA

YOSHUA BENGIO bengioy@IRO.UMontreal.CA
Dept. IRO, CP 6128, Université de Montréal, Succ. Centre-Ville, 2920 Chemin de la tour, Montréal, Québec, Canada, H3C 3J7

Editor: Dale Schuurmans

Abstract. Model selection is an important ingredient of many machine learning algorithms, in particular when the sample size in small, in order to strike the right trade-off between overfitting and underfitting. Previous classical results for linear regression are based on an asymptotic analysis. We present a new penalization method for performing model selection for regression that is appropriate even for small samples. Our penalization is based on an accurate estimator of the ratio of the expected training error and the expected generalization error, in terms of the expected eigenvalues of the input covariance matrix.

Keywords: model selection, parametric regression, uniform convergence bounds

1. Introduction

Consider the problem of estimating a regression function in the set of functions

$$f(x, \alpha) = \sum_{k=1}^{\infty} \alpha_k \varphi_k(x) \tag{1}$$

where $\{\varphi_k\}$ form a basis of $L_2(\mathbb{R}^p)$, e.g. a Fourier or wavelet basis.

Given a collection of data $(x_1, y_1), \ldots, (x_n, y_n)$, where $y_i = f(x_i, \alpha_0) + \xi_i$ and x_i, ξ_i are independently generated by unknown distributions $P(x)$ and $P(\xi)$, one wants to find the function $f(x, \alpha_*)$ that provides the smallest value of the expected loss

$$R(\alpha) = \int L(y, f(x, \alpha))\, dP(x)\, dP(\xi) \tag{2}$$

where $L(y, f(x, \alpha))$ is a given loss function, usually the quadratic loss $L(y, f(x, \alpha)) = (y - f(x, \alpha))^2$. To minimize the expected risk (2), one minimizes the empirical risk

functional

$$R_{emp}(\alpha) = \frac{1}{n}\sum_{i=1}^{n} L(y_i, f(x_i, \alpha))$$

However since the set (1) has an infinite expansion, this idea does not work: for any finite number of (different) examples there are functions which have zero empirical risk and a large value of the expected loss.

To guarantee a small expected risk, one can minimize the empirical functional over only the first $d = d(n)$ functions $\varphi_k(x)$. This is reasonable if the φ_k are ordered in such way that puts the "smoother" components first, introducing a preference for smooth functions. The problem of choosing an appropriate value $d = d(n)$ is called *model selection*.

For the case of quadratic loss and a large number of observations, several penalty-based methods were proposed in the mid-70's, and these are asymptotically optimal. All of these solutions, described in more detail below, minimize functionals of the form

$$R^*_{emp}(\hat{f}_d) = R_{emp}(\hat{f}_d)T(d, n) \tag{3}$$

where n is the sample size, $R_{emp}(f_d)$ is the minimum of the empirical risk when training with a model of size d (achieved by the function $\hat{f}_d$), and $T(d, n)$ is a correction factor for performing model selection.

In particular Akaike (Akaike, 1970) defined in the context of autoregressive models the "Future Prediction Error" (FPE) correction factor

$$T(d, n) = (1 + d/n)(1 - d/n)^{-1}, \tag{4}$$

For small ratios d/n this multiplicative factor has a linear approximation $(1 + 2\frac{d}{n})$. Generalized Cross-Validation (Wahba, Golub, & Heath, 1979) and Shibata's model selector (Shibata, 1981) have the same linear approximation. Some other criteria which provide a different asymptotic behavior have been proposed including RIC (Foster & George, 1994) BIC (Schwartz, 1978) as well as criteria derived from the Minimum Description Length (MDL) principle (Rissanen, 1986; Barron, Rissanen, & Yu, 1998).

During the same years, a general theory of minimizing the empirical risk (for any set of functions, any loss functions, and any number of samples) has been constructed (Vapnik, 1982). In the framework of this theory, the method of *Structural Risk Minimization* for model selection was proposed. In the case studied here, this yields the following multiplicative factor (Cherkassky, Mulier, & Vapnik, 1997), derived from *Uniform Convergence Bounds* (UCB):

$$T(d, n) = \left(1 - c\sqrt{\frac{d(\ln n/d + 1) - \ln \eta}{n}}\right)_+^{-1} \tag{5}$$

where $u_+ = \max(0, u)$ and c, η are some constants. In spite of the fact that in the asymptotic case, this factor is less accurate than classical ones, simulation experiments showed that this correction factor outperforms other classical ones (Cherkassky, Mulier, & Vapnik, 1997).

This article is the development of an idea described in Vapnik (1998). We first show that the expectation of the loss of the function minimizing the empirical risk depends both on the ratio d/n and the eigenvalues of a covariance matrix. It appears that by taking into account those eigenvalues we obtain a correction factor $T(d, n)$ which for small d/n coincides with Akaike's factor, but which is significantly different for larger d/n.

This analysis aims at characterizing the relation between empirical risk and bias (residual of the approximation and noise) on one hand, and between bias and generalization error on the other hand. For this purpose we made an independence assumption which might not be satisfied in practice. However, in our experiments the obtained estimator has a very good accuracy which suggests that this assumption is reasonable.

In the last section of the article, we compare the estimation accuracy of our method with classical ones and show that one can use it to perform state-of-the-art model selection.

2. Risk of the mean square error estimator

We consider a linear model of dimension d,

$$\mathcal{F}_d = \left\{ x \to \sum_{i=1}^{d} \alpha_i \varphi_i(x) \right\} \tag{6}$$

with $\alpha_i \in \mathbb{R}$ and the family $\{\varphi_i(x)\}_{i\in\mathbb{N}}$ is orthonormal with respect to the probability measure $\mu(x)$, which means $E\varphi_p(x)\varphi_q(x) = \delta_{pq}$.[1] We assume without any loss of generality that this family is also a basis of $L_2(\mathbb{R}^p)$ (if it is not, it is always possible to extend it). Let $\hat{f}_d$ be the function minimizing the empirical mean square error over the set of functions $\mathcal{F}_d$, i.e.

$$\hat{f}_d = \arg\min_{f\in\mathcal{F}_d} R_{emp}(f),$$

The following section gives an estimator of the risk of $\hat{f}_d$. This risk estimator will lead directly to the choice of the correcting term in the model selection problem.

We suppose without loss of generality that the first function φ_1 is the constant function 1 and then by orthonormality we have for all $p > 1$,

$$E\varphi_p(x) = 0 \tag{7}$$

2.1. Derivation of the risk estimator

In the orthonormal basis $\{\varphi_i(x)\}_{i\in\mathbb{N}}$, the desired regression function can be written as

$$f(x) = \sum_{i=1}^{\infty} \alpha_i \varphi_i(x)$$

and the regression function minimizing the empirical risk is

$$\hat{f}_d(x) = \sum_{i=1}^{d} \hat{\alpha}_i \varphi_i(x)$$

Let the i.i.d. noise ξ have variance σ^2 and mean zero, then the risk of this function is

$$\begin{aligned} R(\hat{f}_d) &= \int (f(x) + \xi - \hat{f}_d(x))^2 \, d\mu(x) \, dP(\xi) \\ &= \sigma^2 + \int (f(x) - \hat{f}_d(x))^2 \, d\mu(x) \\ &= \sigma^2 + \sum_{i=1}^{d} (\alpha_i - \hat{\alpha}_i)^2 + \sum_{i=d+1}^{\infty} \alpha_i^2 \end{aligned} \tag{8}$$

The last equality comes from the orthonormality of the family $\{\varphi_i\}_{i \in \mathbb{N}}$.

The first term σ^2 corresponds to the risk of the true regression function, $R(f)$. The second term is the estimation error and the third term is the approximation error that we call r_d,

$$r_d = \sum_{i=d+1}^{\infty} \alpha_i^2 \tag{9}$$

To analyze Eq. (8), let us introduce the vector $\beta_i = \hat{\alpha}_i - \alpha_i$ of estimation errors and express the empirical risk in function of β,

$$\begin{aligned} R_{emp}(\beta) &= \frac{1}{n} \sum_{i=1}^{n} \left(y_i - \sum_{p=1}^{d} (\alpha_p + \beta_p) \varphi_p(x_i) \right)^2 \\ &= \frac{1}{n} \sum_{i=1}^{n} \tilde{y}_i{}^2 - \frac{2}{n} \sum_{p=1}^{d} \beta_p \sum_{i=1}^{n} \tilde{y}_i \varphi_p(x_i) + \sum_{p,q=1}^{d} \beta_p \beta_q \frac{1}{n} \sum_{i=1}^{n} \varphi_p(x_i) \varphi_q(x_i), \end{aligned} \tag{10}$$

where

$$\tilde{y}_i = \xi_i + \sum_{p=d+1}^{\infty} \alpha_p \varphi_p(x_i).$$

If we introduce the $n \times d$ matrix Φ, with $\Phi_{i,p} = \varphi_p(x_i)$, then the empirical risk is minimized for

$$\beta = (\Phi^T \Phi)^{-1} \Phi^T \tilde{Y}, \tag{11}$$

where $\tilde{Y} = (\tilde{y}_1, \ldots, \tilde{y}_n)^T$ and the minimum value of the empirical risk is

$$R_{emp}(\hat{f}_d) = \frac{1}{n}\tilde{Y}^T(I - \Phi(\Phi^T\Phi)^{-1}\Phi^T)\tilde{Y}. \tag{12}$$

The SVD decomposition of the matrix Φ writes $\Phi = USV^T$, where U and V are orthogonal matrices of size $n \times n$ and $d \times d$ respectively. S is a $n \times d$ diagonal matrix. Then

$$\Phi(\Phi^T\Phi)^{-1}\Phi^T = UI_dU^T,$$

with $S(S^TS)^{-1}S^T = I_d$ being a diagonal $n \times n$ matrix with its first d diagonal elements equal to 1 and the others zero. Thus Eq. (12) writes

$$\begin{aligned} R_{emp}(\hat{f}_d) &= \frac{1}{n}\tilde{Y}^T U(I_n - I_d)U^T\tilde{Y} \\ &= \frac{1}{n}\sum_{p=d+1}^{n}\left(\sum_{i=1}^{n}\tilde{y}_i U_{ip}\right)^2 \end{aligned} \tag{13}$$

Let us now make the assumption that $\tilde{Y}$ and Φ are statistically independent. This assumption will be discussed at the end of the section. Then $\tilde{Y}$ and U are independent and $E\,\tilde{y}_i U_{ip} = E\,\tilde{y}_i E U_{ip} = 0$ from (7). From Eq. (13), we conclude

$$\begin{aligned} ER_{emp}(\hat{f}_d) &= \frac{1}{n}\sum_{p=d+1}^{n}\sum_{i=1}^{n} E\,\tilde{y}_i{}^2 EU_{ip}^2 \\ &= \left(1 - \frac{d}{n}\right)(r_d + \sigma^2) \end{aligned} \tag{14}$$

The second equality is derived using the independence of ξ_i and x_i, the orthonormality of the bases (yielding $E\tilde{y}_i{}^2 = (r_d + \sigma^2)$), and orthogonality of the matrix U (yielding $\sum_{i=1}^{n} EU_{ip}^2 = 1$).

In Eq. (8) we have to estimate $\sum_{p=1}^{d}(\alpha_p - \hat{\alpha}_p)^2 = \sum_{p=1}^{d}(\beta_p)^2$. To do this, let us write

$$\|\beta\|^2 = \tilde{Y}^T\Phi(\Phi^T\Phi)^{-2}\Phi^T\tilde{Y}.$$

and denote by $(\lambda_1, \ldots, \lambda_d)$ the eigenvalues of the covariance matrix $C = \frac{1}{n}\Phi^T\Phi$,

$$C_{pq} = \frac{1}{n}\sum_{i=1}^{n}\varphi_p(x_i)\varphi_q(x_i). \tag{15}$$

Then one can show using the same technique as above that

$$E\sum_{p=1}^{d}\beta_p^2 = \frac{\sum_{i=1}^{d} E(1/\lambda_i)}{n}(r_d + \sigma^2)$$

Finally combining this last equality with Eqs. (8) and (14), we obtain

$$ER(\hat{f}_d) = ER_{emp}(\hat{f}_d)\left(1-\frac{d}{n}\right)^{-1}\left(1+\frac{E\sum_{i=1}^{d}(1/\lambda_i)}{n}\right) \tag{16}$$

2.2. *Remarks*

1. We have made the assumption that $\tilde{Y}$ and Φ are independent. Actually, the matrix Φ depends only on the first d functions in the basis and $\tilde{Y}$ depends only on the functions beyond d and on the noise. Thus Φ and $\tilde{Y}$ are orthogonal but might not be statistically independent. However in practice this assumption seems to be reasonable (see figure 1). Also when the residual is small compared to the noise, $\tilde{y}_i \approx \xi_i$, and the independence of ξ_i and Φ motivates this assumption. Note that the assumption that there is no residual was also made in the derivation of the Akaike Information Criterion (Akaike, 1973). Finally,

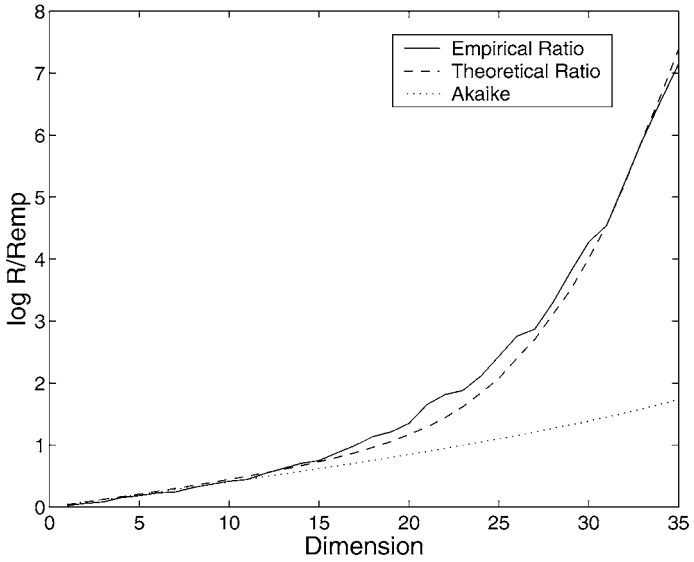

Figure 1. Comparison of the ratio (in log scale) of the median of the generalization error and training error (over 1000 trials) with the penalty term (17) and with Akaike's penalty. The latter is only accurate when d/n is small. The number of training examples n is 50, the target function is the step function, the noise level is 0.05, the training points are uniformly generated in $[-\pi, \pi]$ and the empirical risk minimization has been carried out in the Fourier basis.

the assumption would also be valid if $\varphi_i(x)$ is independent of $\varphi_j(x)$ (e.g. representing independent components of the vector x).

2. We computed the ratio of the expected generalization error and the expected empirical error. However, in practice, one would like to estimate the actual generalization error in function of the actual empirical error.

   To do this, in the previous derivation, one should replace equalities of the type

   $$E\frac{1}{k}\sum_{i=1}^{k}\tilde{y}_i{}^2 = r_d + \sigma^2$$

   by statements of the following type: With high probability,

   $$\left|\frac{1}{k}\sum_{i=1}^{k}\tilde{y}_i{}^2 - (r_d + \sigma^2)\right| \leq \frac{c}{\sqrt{k}}$$

   This kind of statement can be done if we have assumptions on the probability distribution of $\tilde{y}_i$ and would lead to risk bounds for the model selection strategy, as shown in Bartlett, Boucheron, and Lugosi (2000).

3. This derivation is based on the assumption that the set of basis functions is orthonormal with respect to the probability measure $\mu(x)$. However in the learning problem this probability distribution is usually unknown and therefore it is impossible to get an explicit orthonormal basis. Nevertheless, for any given independent set of basis functions $\{\Psi_i(x)\}$ and any probability distribution, using Gram-Schmidt orthonormalization, one can theoretically get a unique orthonormal family $\{\Phi_i(x)\}$ that describes the same set of functions $\mathcal{F}_d$.

   From the previous argument, one can still use (16) for a non orthonormal family, keeping in mind however that the eigenvalues appearing in this estimator are the ones corresponding to the covariance matrix constructed from the Gram-Schmidt orthonormalized basis.

   In practice this orthogonalization can be made using unlabeled data (more details are provided in the next section).

3. Application to model selection

As the goal in model selection is to choose the model with the smallest expected risk, the previous analysis (see Eq. (16)) suggests to take the correcting term $T(d, n)$ as

$$T(d,n) = \left(1 - \frac{d}{n}\right)^{-1}\left(1 + \frac{E\sum_{i=1}^{d}(1/\lambda_i)}{n}\right) \tag{17}$$

where λ_i is the i-th eigenvalue of the covariance matrix (15).

Note that in the asymptotic case, since the covariance matrix is almost the identity matrix (from the orthonormality assumption), $E(1/\lambda_i) \approx 1$ and we obtain Akaike's term (4).

However, in the non-asymptotic case the covariance matrix is not well-conditioned and it can happen that $E(1/\lambda_i) \gg 1$ (see figure 1).

Direct eigenvalue estimator method (DEE). In the case when along with training data, "unlabeled" data are available (x without y), one can compute two covariance matrices: one from unlabeled data $\tilde{C}$ and another from the training data C_{emp}.

There is a unique matrix P (Horn & Johnson, 1985; Corollary 7.6.5) such that

$$P^T \tilde{C} P = I \quad \text{and} \quad P^T C_{emp} P = \Lambda,$$

where Λ is a diagonal matrix with diagonal elements $\lambda_1, \ldots, \lambda_n$. To perform model selection, we used the correcting term (17) where we replace $E \sum_{i=1}^{d} 1/\lambda_i$ with its empirical value,

$$\begin{aligned} \sum_{i=1}^{d} 1/\lambda_i &= \operatorname{trace}\left(P^{-1} C_{emp}^{-1} (P^T)^{-1} P^T \tilde{C} P\right) \\ &= \operatorname{trace}\left(C_{emp}^{-1} \tilde{C}\right). \end{aligned}$$

This enables us to deal with a non orthonormal family. As before the quantity $\operatorname{trace}(C_{emp}^{-1}\tilde{C})$ is an indicator of the discrepancy between the empirical covariance matrix C_{emp} and its "expected" value $\tilde{C}$.

Smallest eigenvalue bound (SEB). To estimate $E \sum 1/\lambda_i$ appearing in Eq. (16), one can use a lower bound on the smallest eigenvalue of the covariance matrix.

Lemma 1. *With probability at least* $1 - \eta$

$$\lambda_{\min} > 1 - \sqrt{V_d \Lambda_d(n)}, \tag{18}$$

where

$$V_d = \sup_x \sup_{\|\alpha\|_2 = 1} \left(\sum_{i=1}^{d} \alpha_i \varphi_i(x) \right)^2 \quad \textit{and} \quad \Lambda_d(n) = \frac{d\left(\ln \frac{2n}{d} + 1\right) - \ln(\eta/4)}{n} \tag{19}$$

The proof is in appendix.

In practice, we take $\eta = 0.1$ and $V_d = 1^2$ and we get the following bound,

$$ER(\hat{f}_d) \leq ER_{emp}(\hat{f}_d) \left(1 - \frac{d}{n}\right)^{-1} \left(1 + \frac{d}{nk}\right) \tag{20}$$

where

$$k = \left(1 - \sqrt{\frac{d\left(\ln \frac{2n}{d} + 1\right) + 4}{n}}\right)_+ \tag{21}$$

Remark: Expected risk minimization and model selection. In Section 2, we derived an unbiased estimator of the risk of the function minimizing the mean square error on a linear model of dimension d. The model selection procedure we proposed is to choose the model minimizing this unbiased estimator.

However a more detailed analysis should be carried out. Indeed, if the variance of our estimator is large and the number of models tested is also large, then some "overfitting" problems might occur. To avoid this, one needs to increase the penalty in order to capture the variance of the risk estimator and the number of models. A related explanation can also be found in Remark 2.

We do not consider here the case where of lot of models are available, but just the standard case of nested regression (in which the number of models is less than the number of training points) and choosing the model which minimizes an unbiased estimator of the test error should give good results.

As explained before, the case of non-nested regression (choice of wavelet coefficients for example) needs some additional analysis and is left for future work.

4. Experimental results

We performed toy experiments in order to compare model selection algorithms. The input distribution is the uniform distribution on $[-\pi, \pi]$ and the set of basis functions is the Fourier basis,

$$\begin{aligned} \varphi_1(x) &= 1 \\ \varphi_{2p} &= \sqrt{2}\,\cos(px) \\ \varphi_{2p+1} &= \sqrt{2}\,\sin(px) \end{aligned}$$

We compared our model selection methods, SEB (Smallest Empirical Bound), and DEE (Direct Eigenvalue Estimator), to eight other methods. Six of them are penalty-based: FPE (Akaike, Eq. (4)), Uniform Convergence Bound (UCB) (5), GCV (Wahba, Golub, & Heath, 1979), RIC (Foster & George, 1994), BIC (Schwartz, 1978), Mallow's C_p (CPM) (Mallows, 1973). For the UCB method, we took $c = 1$ and $\ln \eta = -3$ in Eq. (5).

The two other model selection algorithms we considered are ADJ (a state-of-the-art heuristic method (Schuurmans, 1997)), and CV_5 (5-fold cross-validation).

Note that both ADJ and DEE need some information about the distribution of input data $\mu(x)$ which can be provided by unlabeled data. In the experiments we used 1000 unlabeled training points.

We first compared the accuracy of some of these methods in the prediction of the generalization error. For this purpose, we considered the regression function

$$f(x) = \frac{1}{10}\left(x + \frac{3}{2}\right)^2,$$

a gaussian noise with standard deviation $\sigma = 0.05$ and a training set of 40 examples. For each $d \leq 23$, we computed the empirical risk minimizer $\hat{f}_d$ and tried to predict the generalization error $R(\hat{f}_d)$. The results are shown in figure 2 and are averaged over 1000 trials.

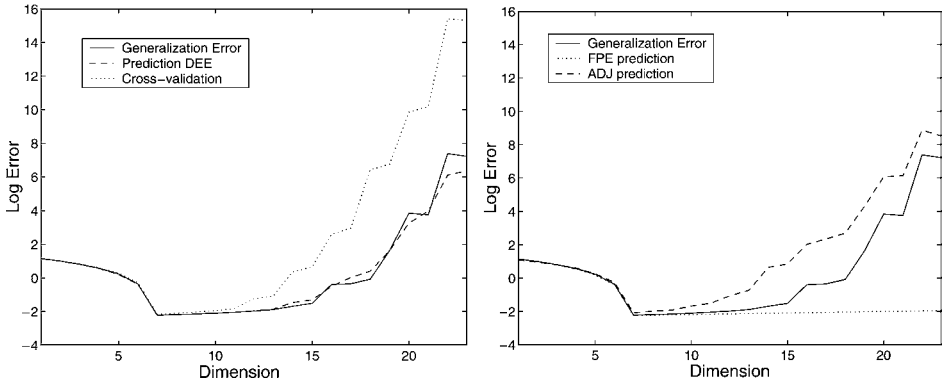

Figure 2. Prediction of the generalization error for the following methods: DEE, CV5 (left), FPE, ADJ (right).

Both DEE and ADJ predict accurately the test error, ADJ being a little bit over pessimistic. When the number of dimension becomes large, FPE underestimates the generalization error while CV5 overestimates (this is explained by the fact that during cross-validation a smaller training set is used).

For the model selection itself we are interested in the generalization error of the function chosen by the model selection procedure. Indeed, as explained at the end of Section 3, an unbiased estimator of the generalization error with a large variance might give a poor criterion for model selection.

Different experiments have been carried out by changing the variance of the gaussian noise, the number of training points or the target function. For each model selection procedure, if the model $\hat{d}$ is chosen, we compute the log of the approximation ratio,

$$\log \frac{R(\hat{f}_{\hat{d}})}{\min_d R(\hat{f}_d)}. \tag{22}$$

The results are shown in boxplot style in figures 3 and 4, each one corresponding to a different target function: sinc $(\sin(4x)/4x)$ and step $(1_{x>0})$ functions. All the experiments have been repeated 1000 times.

The plots for the sinc function (figure 3) show that the model selection procedures have a similar performance when the function is easy to estimate (the Fourier coefficients of this function decrease very rapidly). Only FPE is far from the optimal solution for 50 training points.

The second example is the step function (figure 4), which is difficult to approximate in the Fourier basis. In this case, traditional penalty based method (RIC, BIC, CPM, GCV, FPE) fail whereas DEE, SEB, UCB, ADJ and CV_5 are able to select a good model.

For each experiment, Tables 1 and 2 indicate the median and the mean (over 1000 trials) of the approximation ratio (22).

Judging from these experiments, both proposed methods DEE and SEB perform as well as the state-of-the-art methods, such as the ADJ heuristic or cross-validation and UCB,

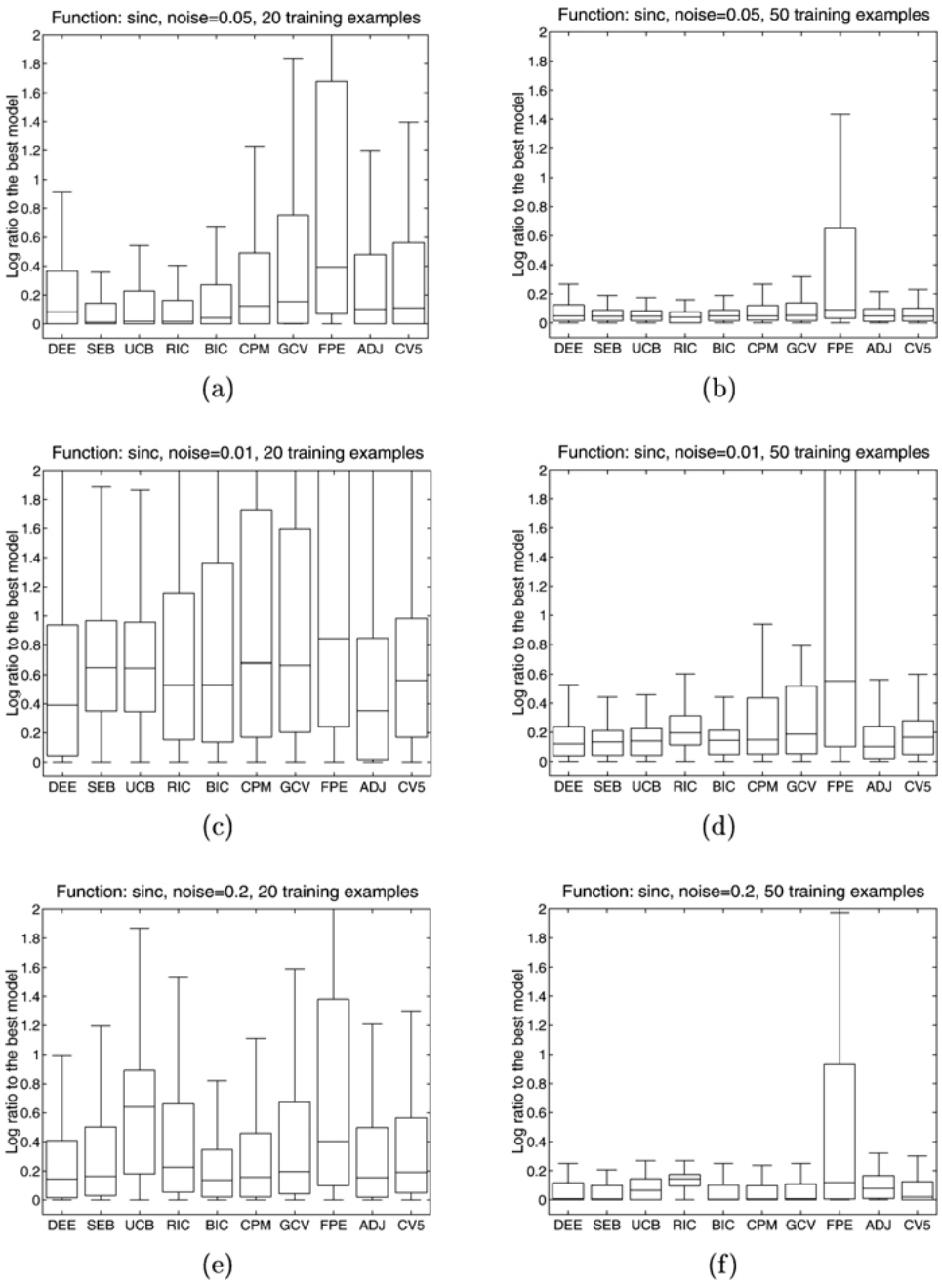

Figure 3. Approximation ratios for the sinc function. Numerical results can be found in Tables 1 and 2, each letter corresponding to the same experiment.

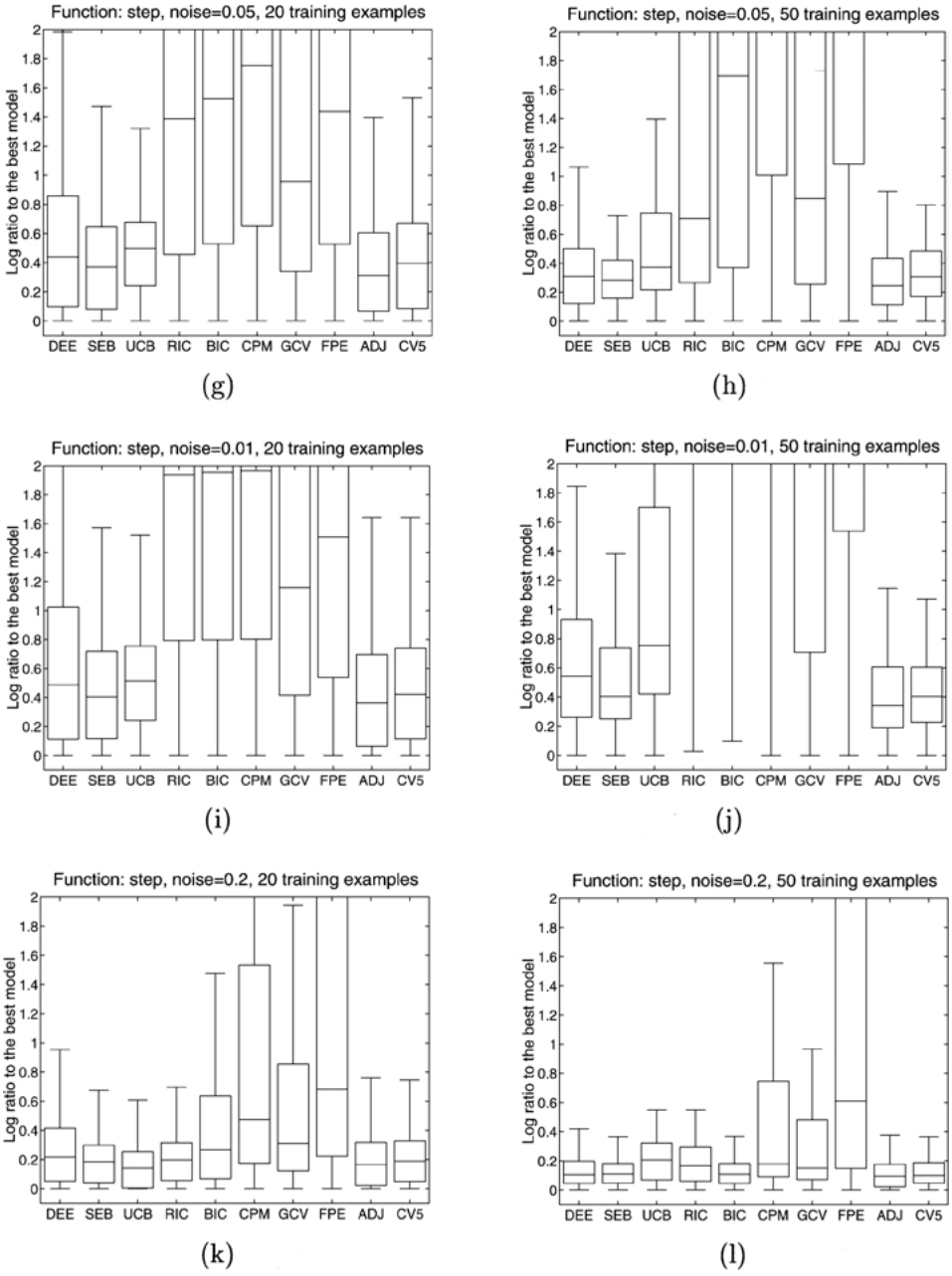

Figure 4. Approximation ratios for the step function. Numerical results can be found in Tables 1 and 2, each letter corresponding to the same experiment.

Table 1. Median of the ratio of the test error to the best model for the 12 experiments reported in figures 3 and 4. The last row is an average over the 12 experiments.

	ADJ	SEB	CV_5	DEE	UCB	GCV	RIC	BIC	FPE	CPM
a	1.11	1.01	1.12	1.09	1.02	1.16	1.01	1.04	1.48	1.13
b	1.05	1.05	1.05	1.05	1.05	1.05	1.04	1.05	1.10	1.05
c	1.42	1.91	1.75	1.48	1.91	1.94	1.70	1.70	2.32	1.97
d	1.11	1.14	1.18	1.13	1.15	1.21	1.22	1.15	1.74	1.16
e	1.17	1.18	1.21	1.15	1.90	1.21	1.25	1.15	1.50	1.17
f	1.08	1.00	1.02	1.01	1.07	1.01	1.15	1.00	1.13	1.01
g	1.37	1.45	1.49	1.55	1.64	2.60	4.00	4.60	4.21	5.77
h	1.28	1.33	1.36	1.36	1.45	2.34	2.03	5.45	12.33	13.14
i	1.44	1.50	1.52	1.63	1.67	3.19	6.94	7.06	4.52	7.14
j	1.41	1.49	1.49	1.72	2.12	18.11	36.01	36.05	30.22	42.93
k	1.18	1.20	1.21	1.24	1.15	1.37	1.22	1.31	1.98	1.61
l	1.10	1.12	1.11	1.11	1.23	1.16	1.18	1.11	1.84	1.20
	1.23	**1.28**	**1.29**	**1.29**	**1.45**	**3.03**	**4.90**	**5.22**	**5.36**	**6.61**

Table 2. Mean of the ratio of the test error to the best model for the 12 experiments reported in figures 3 and 4. The last row is an average over the 12 experiments.

	ADJ	DEE	SEB	CV_5	UCB	RIC	GCV	FPE	BIC	CPM
a	2.63	2.48	1.77	2.85	2.27	1.64	3.6e3	4.3e3	7	327
b	1.1	1.16	1.09	1.1	1.09	1.05	1.31	26	1.09	1.2
c	3.51	3.12	2.71	7.45	2.67	19.5	158	1.3e4	289	1.3e4
d	1.26	1.29	1.27	1.31	1.32	1.35	2.4e3	6.5e3	1.33	11.3
e	1.42	1.44	1.47	1.58	1.93	1.59	139	3.5e4	1.51	1.2e4
f	1.12	1.07	1.06	1.08	1.08	1.14	7.15	1e4	1.05	1.06
g	1.69	2.44	2.44	2.53	1.88	3.1e4	3.4e4	3.4e4	3.5e4	3.5e4
h	1.44	1.69	1.43	1.48	12.2	637	464	3.4e3	5.9e8	5.9e8
i	2.39	3.67	2.21	2.36	2.18	4e3	3.8e3	4e3	4e3	4e3
j	1.8	2.69	8.29	2.28	156	1.8e4	5.4e4	1e5	1.8e4	5.6e4
k	1.32	1.49	1.38	1.94	1.22	15.3	615	9.7e5	159	700
l	1.14	1.19	1.15	1.16	1.24	1.21	17.9	1.5e3	1.48	63.5
	1.73	**1.98**	**2.19**	**2.26**	**15.4**	**4.5e3**	**8.3e3**	**9.9e4**	**4.9e7**	**4.9e7**

while classical penalty-based methods fail. It is worth noting that the ADJ heuristic seems to be the best model selection procedure among all the ones we tested.

The comparison between Table 1 (median of the approximation ratio) and Table 2 (mean of the approximation) gives a better insight of the behavior of some model selection algorithms. For example, UCB has a median of 1.45, but a mean of 15.4. This is due to the fact

that sometimes it selects a very large model (i.e. it overfits) incurring a catastrophic generalization error. The same explanation applies obviously to other penalty-based methods (which have terrible approximation ratios in mean) and to a certain extent to CV5 (see row c of Table 2) and SEB (see row j). Intuitively, cross-validation gives an almost unbiased estimator of the generalization error, but because of its variance, it might select sometimes a model which is far from the optimal one. This is also true for SEB and DEE, even though we expect these methods to have a smaller variance. A discussion on this topic can be found at the end of Section 3.

5. Conclusion

In this article we showed that to select models using small sample size the formulas obtained for asymptotic classical models are insufficient. In our analysis, we pointed out that the discrepancy between the empirical covariance matrix and its expectation is critical for small sample size regression. Taking this discrepancy into account, we obtain a model selection algorithm which behaves similarly to the state-of-the-art.

Further research includes improvement of the SEB method thanks to a deeper analysis of the distribution of the eigenvalues of a covariance matrix. The DEE method is very attractive since it provides an unbiased estimator of the generalization error of a given model. Unfortunately it requires unlabeled data. If such data is not available, we believe this method will still be efficient by generating unlabeled data from a Parzen window estimator of the input density. New experiments will be carried out to assess this supposition.

From a theoretical point of view, we will focus on the remark at the end of Section 3 and try to extend this method for non-nested regression. A typical application of this in machine learning would be to determine the number of centers in a RBF network.

Appendix

Proof of Lemma 1: Consider the quantity

$$Q(x,\alpha) = \left(\sum_{p=1}^{d} \alpha_p \varphi_p(x)\right)^2$$

For all α such that $\|\alpha\| = 1$, we have $EQ(x,\alpha) = 1$. On the other hand,

$$\frac{1}{n}\sum_{i=1}^{n} Q(x_i,\alpha) = \alpha^T C \alpha$$

where $C = \Phi^T\Phi/n$ is the covariance matrix and then

$$\min_{\|\alpha\|=1} \frac{1}{n}\sum_{i=1}^{n} Q(x_i,\alpha) = \lambda_{\min},$$

where $\lambda_{\min}$ is the smallest eigenvalue of C. In Vapnik (1982), it is shown that for any family of functions Q satisfying $0 \leq Q(x, \alpha) \leq B$ and of VC dimension d, the following inequality holds

$$P\left(\sup_{\alpha} EQ(x,\alpha) - \frac{1}{n}\sum_{i=1}^{n} Q(x_i,\alpha) > \epsilon\right) \leq \exp\left(\frac{d}{n}(\log(2n/d)+1) - \frac{\epsilon^2}{B^2}\right) n$$

Using this last inequality, we get that with probability $1-\eta$,

$$1 - \lambda_{\min} < \sqrt{V_d \Lambda_d(n)} \qquad \square$$

Acknowledgment

The authors would like to thank a anonymous referee for helpful and valuable comments.

Notes

1. Note that the choice of such a family requires knowledge about $\mu(x)$. See Remark 3 of Section 2.2 for more details.
2. V_d might be much larger than 1 for some basis φ, but in our experiments $V_d = 1$ seems to be a good choice.

References

Akaike, H. (1970). Statistical predictor identification. *Ann. Inst. Stat. Math., 22*, 202–217.

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. Petrov, & F. Csaki (Eds.), *2nd International Symposium on Information Theory*, Budapest (Vol. 22, pp. 267–281).

Barron, A., Rissanen, J., & Yu, B. (1998). The minimum description length principle in coding and modeling. *IEEE Trans. Inf. Theory, 44*, 2743–2760.

Bartlett, P., Boucheron, S., & Lugosi, G. (2000). Model selection and error estimation. In *COLT'00*.

Cherkassky, V., Mulier, F., & Vapnik, V. (1997). Comparison of VC method with classical methods for model selection. In *Proceedings of the World Congress on Neural Networks* (pp. 957–962).

Foster, D., & George, E. (1994). The risk inflation criterion for multiple regression. *Annals of Statistics, 22:4*, 1947–1975.

Horn, R. A., & Johnson, C. R. (1985). *Matrix analysis*. Cambridge: Cambridge University Press.

Mallows, C. L. (1973). Some comments on C_p. *Technometrics, 15:4*, 661–675.

Rissanen, J. (1986). Stochastic complexity and modeling. *Annals of Statistics, 14*, 1080–1100.

Schuurmans, D. (1997). A new metric-based approach to model selection. In *Proceedings of the Fourteenth National Conference on Artificial Intelligence*.

Schwartz, G. (1978). Estimating the dimension of a model. *Ann. Stat., 6*, 461–464.

Shibata, R. (1981). An optimal selection of regression variables. *Biometrica, 68*, 461–464.

Vapnik, V. (1982). *Estimation of dependencies based on empirical data*. Berlin: Springer.

Vapnik, V. (1998). *Statistical learning theory*. New York: Wiley.

Wahba, G., Golub, G., & Heath, M. (1979). Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics, 21*, 215–223.

Received August 23, 2000
Revised January 12, 2001
Accepted January 17, 2001
Final manuscript February 20, 2001