Statistica Sinica **6**(1996), 589-601

MODEL FITTING VIA TESTING

Prabir Burman

University of California, Davis

Abstract: Model selection methods are developed using the approach of hypothesis testing. The technique of cross-validation has been used to make the procedures fully automated. Simulation results are presented to show that the proposed methods work better than Akaike's FPE, Mallows' C_p or Schwartz's BIC.

Key words and phrases: Hypothesis testing, model selection, Akaike's FPE, Mallows' C_p, Schwartz's BIC.

1. Introduction

Model selection in regression and classification problems is one of the most important topics is statistics. This is so because the exact model is rarely known in practice. However, a class of reasonable models is often available and the statistical problem then is to select a model that is considered to be the most "appropriate". For instance, one of the most important issues in a multiple regression problem is to select a subset of variables that explains the response variable well. In fitting polynomials one faces the problem of deciding which degree polynomial describes the regression function best. In our discussion here we deal with standard Gauss-Markov type model fitting and we describe the columns as variables. It is worth noting that a large class of statistical problems including analysis of variance models, multiple regression, spline or polynomial (or any other finite element method) fitting can be viewed as Gauss-Markov models.

Suppose we have a Gauss-Markov model with p variables. Two of the most well known methods of model selection are Akaike's FPE (1970) and Mallows' C_p (1973). It is possible to view Akaike's method as a special case of Mallows'. The asymptotic theory of these model selection techniques have been studied in great detail by many authors including Shibata (1981), Li (1987), Burman and Chen (1989). All these theoretical studies assume that the number of competing models grows with the sample size n, but only algebraically as fast as n (i.e., the number of competing models is no larger than n^s for some $s > 0$). If the order of importance of the variables is known, then the number of competing models is p. A case like this is covered by the mathematical theories developed in the literature. However, in the case of stepwise or all subsets methods, the

number of competing models grows exponentially with p. In such cases it is no longer clear that Akaike's or Mallows' methods are appropriate, especially when p is not small. Let us discuss heuristically where these methods could have problems. The methods of Akaike and Mallows assume that when we fit a model with k variables, the loss in degrees of freedom is k. It is however not at all clear that only k degrees of freedom is lost in fitting a k-variables model obtained in a stepwise procedure. Intuitively, the loss of degrees of freedom should be higher than k especially when p is large. Our simulation results bring out these points clearly. The method proposed in this paper does not suffer from these difficulties.

We approach the problem of model selection here through the classical method of testing. Suppose that we have under consideration a linear model $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, where the columns are orthonormal. Note that any Gauss-Markov model can be reformulated to have an orthogonal design after a suitable orthogonalization. Also, there are a number of methods in the literature where orthogonal covariables arise rather naturally, e.g., principal components regression, regression with wavelets etc.. In the standard framework of hypothesis testing, variable $\boldsymbol{x}_j$ is kept in the model if $|\hat{\beta}_j / s_{\hat{\beta}_j}|$ is larger than a prespecified value, where $s_{\hat{\beta}_j}$ is the estimated standard error of $\hat{\beta}_j$. However, selection of an appropriate model would require simultaneous testing and the problems associated with simultaneous tests are well known. Our goal here is to use the classical procedure for model building except that we approach it from the point of view of model fitting rather than hypotheses testing. In our framework, we let the data choose the cutoff point associated with the tests and thus make it a fully automated procedure. We also present a more general method in which, instead of "acceptance" or "rejection", a smoothed version of the estimated β parameters are used. Our simulation results show that our methods work better than those of Akaike or Mallows (or their modifications due to Shibata (1984) and Zhang (1992)).

Various methods of model selection including the proposed method are discussed in Section 2. Some simulation results are presented in Section 3. Finally, in the appendix we obtain an approximation to the optimal smoothing function used in defining our hypothesis-type method for model selection.

2. Methods of Model Selection

The basic model we look at is

$$\boldsymbol{Y} = \boldsymbol{\mu} + \boldsymbol{\varepsilon},$$

where $\boldsymbol{Y}$ is the n dimensional vector of observations, $\boldsymbol{\mu}$ is the vector of means and $\boldsymbol{\varepsilon}$ is the error vector of i.i.d. errors with mean zero and variance σ^2.

Let us now assume that we have decided to fit the following linear model with an orthogonal design

$$\boldsymbol{Y} = \sum_{1 \le j \le p} \boldsymbol{x}_j \beta_j + \boldsymbol{\varepsilon}, \tag{1}$$

where $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_p$ are orthonormal. The form given in (1) arises naturally in methods like principal components regression, regression with wavelets etc.. As we have pointed out before, any Gauss-Markov model can be reexpressed in the form given in (1) after a suitable orthogonalization. Note that no claim is being made that the linear model we have decided to fit is correct. The statistical problem here is to select the most appropriate submodel. We are especially interested in the case when the number of variables p is not small compared to the number of observations. It is important to point out that any Gauss-Markov model can be rewritten in the form given above after a suitable orthogonalization.

We fit here a model of the form $\hat{\boldsymbol{\mu}}(c) = \sum_{1 \le j \le p} \boldsymbol{x}_j \hat{\beta}_j I(|\hat{\beta}_j| > c)$ and our goal is to choose that cutoff point c for which this fitted model is closest to the true mean vector $\boldsymbol{\mu}$. Here I denotes the usual indicator function. Note that the value of c determines which variables are kept in the model and which variables are deleted. We also generalize this by fitting a smooth model of the form $\hat{\boldsymbol{\mu}}(c) = \sum_{1 \le j \le p} \boldsymbol{x}_j \hat{\beta}_j G(|\hat{\beta}_j|/c)$, where G is a known distribution function on $[0, \infty)$. It may also be worth noting here that the cutoff point c can be viewed as a smoothing parameter. Our simulation results show that models with a smooth G can often lead to better model fitting. Whether variable deletion is allowed depends on the form of G. We have used two different smooth distribution functions (one of which allows variable deletion) in our simulations and both of them provide quite satisfactory model fitting (see Section 3). Since simulation results show a smooth G leads to a better estimate for $\boldsymbol{\mu}$ than an unsmooth one, it leaves open the possibility that some function G is better than others. If we restrict ourselves to distribution functions with support on a compact set, then, under appropriate conditions, it can be shown that there is a distribution function G^* for which $E[\sum_{1 \le t \le n} \{\hat{\mu}_t(c) - \mu_t\}^2]$ is minimum. In the Appendix, we have given an approximation to the optimum G^*. However, it is difficult to find an exact expression for G^*. This issue needs further research.

Let $L(c)$ denote the squared error distance between the fitted vector $\hat{\boldsymbol{\mu}}(c)$ and $\boldsymbol{\mu}$, i.e.,

$$L(c) = \sum_{1 \le t \le n} \{\hat{\mu}_t(c) - \mu_t\}^2. \tag{2}$$

Clearly, the best value of the cutoff point c^* is obtained when L attains its minimum. However, we do not know c^* since L involves the unknown mean vector $\boldsymbol{\mu}$ and hence we need to estimate L. So if $\hat{L}$ is an estimate of L and $\hat{L}$

attains its minimum at $\hat{c}$, then $\hat{\boldsymbol{\mu}}(\hat{c})$ is declared to be the fitted mean vector. In order to estimate L we use the method of cross validation developed by Stone (1974). Note that

$$L(c) = \sum_{1\le t\le n} \{\hat{\mu}_t(c)\}^2 - 2\sum_{1\le t\le n} \mu_t\hat{\mu}_t(c) + \sum_{1\le t\le n} \mu_t^2.$$

The first term on the right hand side of the above expression is known. Since the last term does not involve c it does not play any role in the minimization. The second term is unknown and has to be estimated. Following Stone (1974) we estimate $\sum_{1\le t\le n} \hat{\mu}_t\mu_t(c)$ by $\sum_{1\le t\le n} Y_t\hat{\mu}_t^{(-t)}(c)$, where $\hat{\mu}_t^{(-t)}(c)$ is the estimate of μ_t obtained by deleting the tth case. So an estimate of $L(c)$ is given by

$$\hat{L}_1(c) = \sum_{1\le t\le n} \{\hat{\mu}_t(c)\}^2 - 2\sum_{1\le t\le n} Y_t\hat{\mu}_t^{(-t)}(c) + \sum_{1\le t\le n} \mu_t^2. \tag{3}$$

The second estimate of L is rather straightforward. Recall that the errors ε_t in model (1) have variance σ^2. Another estimate of $L(c)$ is given by

$$\hat{L}_2(c) = \sum_{1\le t\le n} \{Y_t - \hat{\mu}_t^{(-t)}(c)\}^2 - n\sigma^2. \tag{4}$$

The approximations $\hat{L}_1$ and $\hat{L}_2$ are quite reasonable and they are almost unbiased for $L(c)$ (see Burman (1994)). In a forthcoming paper we will discuss mathematically involved arguments about Shibata-type optimality of these estimates.

In order to develop the next two estimates we need to explain a little background. Let us first note that $\hat{\beta}_j = \sum_{1\le t\le n} Y_t x_{tj}$. It is easy to see that the mean and variance of $\hat{\beta}_j$ are β_j and σ^2 respectively. If we rank the absolute values of $\hat{\beta}_j$ in increasing order of magnitude, then we naturally obtain a sequence of p models. The smallest model will include that variable which has the highest absolute value of $\hat{\beta}$'s. The next model will include those two variables with the two highest absolute values of $\hat{\beta}$'s and so on. Let us denote the kth fitted model by $\hat{\boldsymbol{\mu}}_k$, $k = 1, \ldots, p$. Stretching the notation a little, we will denote by $L(k)$ the squared error distance between the kth fitted model and the true mean vector $\boldsymbol{\mu}$. So

$$L(k) = \sum_{1\le t\le n} \{\hat{\mu}_{kt} - \mu_t\}^2. \tag{5}$$

Once again the best model is obtained by minimizing $L(k)$ over $1 \le k \le p$. Since L involves the unknown mean vector $\boldsymbol{\mu}$, we need to estimate it. Following Mallows (1973) we get an estimate of $L(k)$ which is given by

$$\hat{L}(k) = \sum_{1\le t\le n} \{Y_t - \hat{\mu}_{kt}\}^2 + 2k\hat{\sigma}^2,$$

where $\hat{\sigma}^2$ is a reasonable estimate of σ^2. If $\hat{\sigma}^2$ is taken to be the mean squared error for the kth model, then one obtains Akaike's final predictor error (FPE) estimate of $L(k)$. Shibata (1984) suggested using estimates of the form

$$\hat{L}(k) = \sum_{1 \le t \le n} \{Y_t - \hat{\mu}_{kt}\}^2 + \lambda k \hat{\sigma}^2,$$

where λ is between 2 and 5, and $\hat{\sigma}^2$ is an estimate of σ^2. We have obtained two Shibata type criteria by using two different estimates of σ^2. The first criterion $\hat{L}_3(k)$ is obtained by estimating σ^2 by $\hat{\sigma}_p^2$, the mean square error for the largest model involving all the variables. We get the second Shibata type criterion $\hat{L}_4(k)$ by employing Akaike's idea of estimating σ^2 by the mean square error of the kth fitted model. These estimates are given by

$$\hat{L}_3(k) = \sum_{1 \le t \le n} \{Y_t - \hat{\mu}_{kt}\}^2 + \lambda k \hat{\sigma}_p^2 \tag{6}$$

$$\hat{L}_4(k) = [(n + (\lambda - 1)k)/(n - k)] \sum_{1 \le t \le n} \{Y_t - \hat{\mu}_{kt}\}^2. \tag{7}$$

Our simulation results show that Shibata-type estimates work well, but different values of λ are needed for different cases. For instance $\lambda = 2$ is good in some examples, but it can be quite inadequate for others. Similarly, $\lambda = 8$ seem to be quite good in some cases, but can lead to very unsatisfactory results in others. Typically, the larger the value of p the higher the value of λ is needed. But an appropriate value of λ also depends on the coefficients β_j. This makes it difficult to use Shibata type estimates for model selection.

Finally, we write down the "Bayesian Information Criterion" (BIC) for model selection due to Schwartz (1978). Let $\hat{\sigma}_k = n^{-1} \sum_{1 \le t \le n} \{Y_t - \hat{\mu}_{kt}\}^2$. Then the BIC criterion for normal error model is

$$\hat{L}(k) = n \log(\hat{\sigma}_k) + (1/2) \log(n) k.$$

A first order approximation will reveal that BIC is roughly equivalent to a Shibata type criterion with $\lambda = \log(n)$. Our simulation results show that BIC sometimes does well, but other times its performance is not satisfactory.

3. Numerical Results

3.1. An example

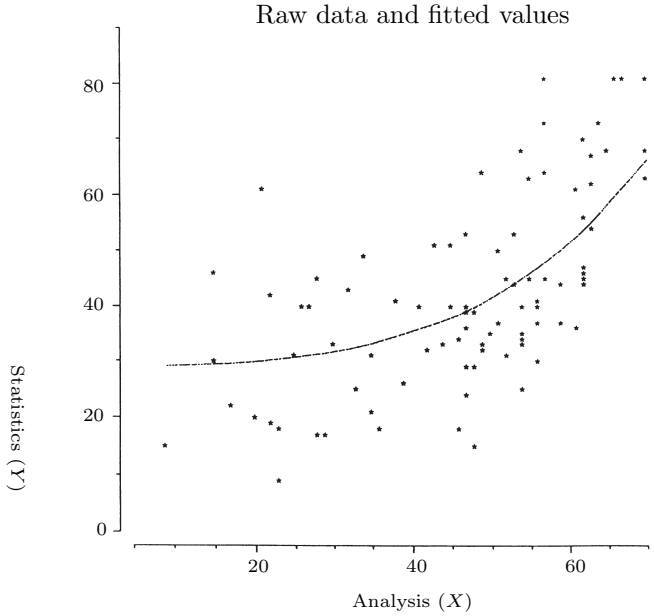
We first consider the problem of estimating a regression function from a data set given in Chapter 1 of the book by Mardia, Kent and Bibby (1979). The data consists of analysis(X) and statistics(Y) scores of 88 students. This data was analyzed in Burman and Chen (1989) using a kernel method. Here we consider

estimating the regression function using the method outlined in Section 2. All the analysis scores are between 9 and 70. We will consider a cubic spline with 10 equispaced knots between 5 and 75. Define the design matrix $\boldsymbol{X}$ by

$$\begin{aligned} x_{ij} &= (X_i - 5)^{j-1},\ j = 1, \ldots, 3, \quad \text{and} \\ x_{ij} &= (X_i - 5 - 70t_j)_+^3,\ j = 4, \ldots, 13,\ i = 1, \ldots, 88, \end{aligned}$$

where $t_j = (j-4)/10$, and $u_+ = \max(u, 0)$ for any real number u. The columns of $\boldsymbol{X}$ can be orthogonalized by many methods including Gram-Schmidt, QR, singular value decomposition. Here we obtain an orthogonal basis by using the singular value decomposition given in LINPACK subroutine. Note that since the design matrix is of order 88 by 13, when we use our method given in Section 2, we are in essence considering 2^{13} models. This is in contrast to the method in which one considers 13 models; where the kth model has k knots (either equispaced or at the sample quantiles), $k = 1, \ldots, 13$.

In order to get an idea about the range in which the smoothing parameter c should be looked at, it is useful to have a rough estimate of σ (the standard deviation of the errors ε_i's). A preliminary estimate of σ is obtained by fitting the full model and it turned out to be about 13. We calculate the criterion function $\sum_{1 \le t \le n} \{Y_t - \hat{\mu}_t^{(-t)}(c)\}^2 = \hat{L}_2(c) + n\sigma^2$ for the values of c between 1 and 150 over a grid of 150 equispaced values. The criterion function attains its minimum at $\hat{c} = 22$. Here we have used $G_2(u) = (u - .5)_+^2/[1 + (u - .5)_+^2]$, $u > 0$, as our smoothing function. The raw data and the estimated regression function are given in the graph below.



3.2. Simulation results

In this section we present a number of results based on simulation in support of our proposed estimates $\hat{L}_1$ and $\hat{L}_2$. The model we use is

$$Y_t = \sum_{1 \le j \le p} x_{tj}\beta_j + \varepsilon_t, \quad t = 1, \ldots, 100,$$

where the vectors $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_p$ are orthonormalized from p vectors on R^{100} with entries as i.i.d. Uniform$[0,1]$ variates and ε_t' are i.i.d. Normal$(0,1)$. Four different values of p (5, 25, 50 and 75) have been considered here. For each value of p, we have considered two different sets of values for the β's:

$$\text{(i) } \beta_j = 10/j,\ j = 1, \ldots, p \quad \text{and} \quad \text{(ii) } \beta_j = 10/j^2,\ j = 1, \ldots, p.$$

For each value of p and $\{\beta_j\}$, we consider three estimates of the form

$$\text{(I)} \sum_{1 \le j \le p} x_{tj}\hat{\beta}_j I(|\hat{\beta}_j| > c), \quad \text{(II)} \sum_{1 \le j \le p} x_{tj}\hat{\beta}_j G_1(|\hat{\beta}_j|/c), \quad \text{and}$$
$$\text{(III)} \sum_{1 \le j \le p} x_{tj}\hat{\beta}_j G_2(|\hat{\beta}_j|/c),$$

where $G_1(u) = u^2/(1+u^2)$, and $G_2(u) = 0$, $0 \le u \le .5$, and $G_2(u) = (u - .5)^2/[1 + (u - .5)^2]$ for $u > .5$. Note that a model fitting method with G_1 retains all the variables, whereas the use of G_2 allows for variable deletion in a smooth manner.

It is important to point out that all the models have been fitted without any intercept term. For each p and $\{\beta_j\}$, we calculate c^* and k^* which are minimizers of $L(c)$ and $L(k)$ respectively (defined in (2) and (5) respectively). Note that $L(c^*)$ and $L(k^*)$ should be the same. However, since we evaluate $L(c)$ at 51 values of c between 0 and 5, they turn out to be slighty different. We have calculated the mean, median and standard deviations of $L(c^*)$ and $L(k^*)$. But we report only the summary statistics for $L(k^*)$.

Note that $\hat{c}_1$ and $\hat{c}_2$ are the minimizers of $\hat{L}_1$ and $\hat{L}_2$ respectively (defined in (4) and (5)). Also $\hat{c}_{2G_1}$ and $\hat{c}_{2G_2}$ are the minimizers of $\hat{L}_2$ when smoothed estimated (II) and (III) above are used. In the same way $\hat{k}_{BIC}$ is defined to be the minimizer of the BIC criterion. Performances of various criteria are judged by the behaviour of $L(\hat{c}_1)$, $L(\hat{c}_2)$, $L(\hat{c}_{2G_1})$, $L(\hat{c}_{2G_2})$ and $L(\hat{k}_{BIC})$. Similarly, $\hat{k}_3$ and $\hat{k}_4$ are the minimizers of Shibata-type estimates $\hat{L}_3$ and $\hat{L}_4$ (defined in (6) and (7)). We have calculated the summary statistics of $L(\hat{k}_3)$ and $L(\hat{k}_4)$ for various values λ between 2 and 12. All the estimates given in the tables have been calculated on the basis of 2500 repeats.

Table 1. $p = 5$
Mean, standard deviation and median of the true error L for various model selection methods

		$\beta_j = 10/j,\ j = 1, \ldots, 5$		$\beta_j = 10/j^2,\ j = 1, \ldots, 5$	
		Mean (SD)	Median	Mean (SD)	Median
	$L(k^*)$	4.91 (3.08)	4.31	3.72 (2.44)	3.03
	$L(\hat{c}_1)$	6.11 (4.12)	5.28	5.51 (3.38)	4.53
	$L(\hat{c}_2)$	6.20 (4.22)	5.33	5.51 (3.39)	4.50
	$L(\hat{c}_{2G_1})$	5.10 (3.32)	4.40	5.05 (3.05)	4.44
	$L(\hat{c}_{2G_2})$	5.14 (3.40)	4.41	5.34 (3.26)	4.50
	$L(\hat{c}_{BIC})$	9.46 (5.51)	8.52	6.25 (3.55)	5.97
$\lambda = 2$	$L(\hat{k}_3)$	6.16 (3.98)	5.47	5.36 (3.44)	4.30
	$L(\hat{k}_4)$	6.09 (3.94)	5.39	5.34 (3.42)	4.29
$\lambda = 4$	$L(\hat{k}_3)$	8.62 (5.14)	7.63	6.02 (3.57)	5.20
	$L(\hat{k}_4)$	8.12 (5.00)	7.18	5.97 (3.55)	5.12
$\lambda = 6$	$L(\hat{k}_3)$	11.61 (6.06)	10.97	6.75 (3.51)	8.04
	$L(\hat{k}_4)$	10.54 (6.04)	9.93	6.63 (3.51)	8.04
$\lambda = 8$	$L(\hat{k}_3)$	14.50 (6.97)	12.78	7.41 (3.24)	8.14
	$L(\hat{k}_4)$	12.95 (7.08)	11.56	7.25 (3.28)	8.12
$\lambda = 10$	$L(\hat{k}_3)$	17.31 (7.92)	16.08	7.98 (2.79)	8.24
	$L(\hat{k}_4)$	15.35 (8.17)	13.28	7.78 (2.91)	8.21
$\lambda = 12$	$L(\hat{k}_3)$	19.85 (8.81)	21.43	8.41 (2.38)	8.33
	$L(\hat{k}_4)$	17.68 (9.39)	15.88	8.22 (2.53)	8.28

Table 2. $p = 25$
Mean, standard deviation and median of the true error L for various model selection methods

		$\beta_j = 10/j,\ j = 1, \ldots, 25$		$\beta_j = 10/j^2,\ j = 1, \ldots, 25$	
		Mean (SD)	Median	Mean (SD)	Median
	$L(k^*)$	21.07 (5.56)	20.45	5.91 (3.06)	5.40
	$L(\hat{c}_1)$	27.22 (7.30)	26.29	12.15 (8.10)	9.39
	$L(\hat{c}_2)$	27.76 (7.35)	26.74	11.24 (7.40)	9.05
	$L(\hat{c}_{2G_1})$	19.62 (5.90)	18.82	7.75 (3.67)	7.34
	$L(\hat{c}_{2G_2})$	21.70 (6.55)	20.73	7.89 (4.32)	7.77
	$L(\hat{c}_{BIC})$	27.60 (7.19)	26.62	11.06 (6.63)	9.53
$\lambda = 2$	$L(\hat{k}_3)$	26.00 (7.03)	25.11	16.97 (7.68)	16.24
	$L(\hat{k}_4)$	25.76 (7.04)	24.85	18.23 (7.76)	17.44
$\lambda = 4$	$L(\hat{k}_3)$	27.17 (7.17)	26.16	11.74 (6.91)	10.68
	$L(\hat{k}_4)$	26.79 (7.05)	26.02	12.86 (7.35)	11.78
$\lambda = 6$	$L(\hat{k}_3)$	28.93 (7.46)	27.86	9.39 (5.57)	8.57
	$L(\hat{k}_4)$	28.15 (7.40)	27.05	10.00 (6.11)	8.74
$\lambda = 8$	$L(\hat{k}_3)$	30.74 (7.81)	29.20	8.64 (4.48)	8.43
	$L(\hat{k}_4)$	30.10 (7.87)	28.73	8.85 (4.81)	8.46
$\lambda = 10$	$L(\hat{k}_3)$	32.63 (8.27)	31.82	8.67 (3.60)	8.47
	$L(\hat{k}_4)$	32.24 (8.84)	31.18	8.61 (3.80)	8.46
$\lambda = 12$	$L(\hat{k}_3)$	34.85 (9.13)	35.86	8.82 (2.86)	8.55
	$L(\hat{k}_4)$	34.59 (10.07)	35.62	8.73 (3.14)	8.51

Table 3. $p = 50$
Mean, standard deviation and median of the true error L for various model selection methods

		$\beta_j = 10/j,\ j = 1, \ldots, 50$		$\beta_j = 10/j^2,\ j = 1, \ldots, 50$	
		Mean (SD)	Median	Mean (SD)	Median
	$L(k^*)$	27.65 (5.94)	26.91	6.63 (3.13)	8.09
	$L(\hat{c}_1)$	37.58 (10.72)	36.39	14.53 (11.77)	9.81
	$L(\hat{c}_2)$	35.38 (9.29)	34.80	11.86 (8.19)	9.13
	$L(\hat{c}_{2G_1})$	26.34 (6.87)	25.40	9.61 (4.37)	8.89
	$L(\hat{c}_{2G_2})$	27.17 (7.01)	26.08	8.55 (3.85)	8.31
	$L(\hat{c}_{BIC})$	36.03 (9.51)	35.27	18.32 (9.99)	17.03
$\lambda = 2$	$L(\hat{k}_3)$	42.09 (10.10)	41.07	31.74 (11.18)	31.12
	$L(\hat{k}_4)$	45.12 (10.20)	44.51	38.49 (11.15)	38.00
$\lambda = 4$	$L(\hat{k}_3)$	36.44 (9.59)	35.47	18.65 (9.88)	17.52
	$L(\hat{k}_4)$	40.31 (10.36)	39.56	25.41 (12.11)	24.25
$\lambda = 6$	$L(\hat{k}_3)$	34.20 (8.81)	34.03	12.65 (7.74)	10.60
	$L(\hat{k}_4)$	36.33 (9.86)	35.38	15.72 (10.00)	14.08
$\lambda = 8$	$L(\hat{k}_3)$	33.94 (8.37)	32.75	10.29 (5.93)	8.88
	$L(\hat{k}_4)$	35.00 (9.10)	33.76	11.24 (7.20)	9.16
$\lambda = 10$	$L(\hat{k}_3)$	35.10 (8.74)	34.63	9.43 (4.48)	8.64
	$L(\hat{k}_4)$	35.39 (9.21)	34.81	9.68 (5.11)	8.69
$\lambda = 12$	$L(\hat{k}_3)$	36.73 (9.24)	37.61	9.16 (3.44)	8.62
	$L(\hat{k}_4)$	36.87 (9.92)	37.62	9.18 (3.82)	8.61

Table 4. $p = 75$
Mean, standard deviation and median of the true error L for various model selection methods

		$\beta_j = 10/j,\ j = 1, \ldots, 75$		$\beta_j = 10/j^2,\ j = 1, \ldots, 75$	
		Mean (SD)	Median	Mean (SD)	Median
	$L(k^*)$	29.83 (6.26)	28.80	6.88 (3.02)	8.25
	$L(\hat{c}_1)$	43.73 (15.89)	39.33	16.84 (16.61)	9.35
	$L(\hat{c}_2)$	38.09 (10.02)	38.32	10.78 (6.80)	8.71
	$L(\hat{c}_{2G_1})$	33.63 (8.24)	32.78	11.16 (4.84)	10.40
	$L(\hat{c}_{2G_2})$	32.66 (8.24)	31.48	9.12 (3.88)	8.71
	$L(\hat{c}_{BIC})$	46.50 (13.31)	45.15	30.28 (14.67)	28.58
$\lambda = 2$	$L(\hat{k}_3)$	57.23 (13.65)	56.56	46.35 (15.33)	45.83
	$L(\hat{k}_4)$	68.66 (13.24)	68.01	64.76 (13.85)	64.30
$\lambda = 4$	$L(\hat{k}_3)$	44.58 (12.65)	43.52	26.61 (13.93)	25.14
	$L(\hat{k}_4)$	62.43 (14.87)	62.19	51.15 (18.60)	51.15
$\lambda = 6$	$L(\hat{k}_3)$	39.05 (10.84)	37.90	16.97 (10.93)	14.86
	$L(\hat{k}_4)$	52.56 (17.17)	51.23	30.69 (21.02)	25.47
$\lambda = 8$	$L(\hat{k}_3)$	37.47 (9.71)	36.68	12.60 (8.22)	9.47
	$L(\hat{k}_4)$	43.71 (15.89)	39.71	16.37 (14.10)	10.79
$\lambda = 10$	$L(\hat{k}_3)$	37.45 (9.41)	37.89	10.66 (6.25)	8.75
	$L(\hat{k}_4)$	39.53 (12.65)	38.40	11.13 (8.11)	8.75
$\lambda = 12$	$L(\hat{k}_3)$	38.67 (10.31)	38.50	9.94 (5.88)	8.64
	$L(\hat{k}_4)$	39.41 (11.48)	38.60	9.84 (6.23)	8.62

It is clear that the best model is $\hat{\boldsymbol{\mu}}(c^*)$ [or $\hat{\boldsymbol{\mu}}(k^*)$ since $\hat{\boldsymbol{\mu}}(c^*) = \hat{\boldsymbol{\mu}}(k^*)$] and hence $L(c^*)$ is the smallest error we find between the true mean vector and any fitted model. Note that when $\lambda = 2$, $\hat{L}_3$ and $\hat{L}_4$ are the Mallows and Akaike's estimates. The simulation results clearly show that Akaike's or Mallows' methods do not work well when p is not small. The performances of Shibata-type estimates are somewhat spotty. They are clearly good for some λ, but that "right" value of λ varies from case to case. The performances of $L(\hat{c}_1)$ and $L(\hat{c}_2)$ are quite good, but we believe they could be improved. The functions $\hat{L}_1$ and $\hat{L}_2$ are quite rough. By using fairly simple smoothing methods, $\hat{L}_1$ and $\hat{L}_2$ could be smoothed and then the behaviour of $L(\hat{c}_1)$ and $L(\hat{c}_2)$ might improve. The behaviour of the BIC criterion is also spotty. In many cases it does well, but in other cases its performance is not satisfactory. It is worth noting here that the performance of $L(\hat{c}_{2G_1})$ and $L(\hat{c}_{2G_2})$ are perhaps the most satisfactory in all the cases. What is more important is that it can be substantially better than all others in some cases.

The simulation results clearly show that $\hat{\boldsymbol{\mu}}(\hat{c}_1)$, $\hat{\boldsymbol{\mu}}(\hat{c}_2)$ and $\hat{\boldsymbol{\mu}}(\hat{c}_{2G})$ are doing uniformly well, with $\hat{\boldsymbol{\mu}}(\hat{c}_{2G_1})$ and $\hat{\boldsymbol{\mu}}(\hat{c}_{2G_2})$ clearly the best. The estimates obtained using Akaike's or Mallows' methods can sometimes be very inadequate especially when p is not small. The Shibata-type estimates do work well, but different values of λ are needed for different cases. At this point of time there is no guide to tell us which value of λ is appropriate for a given data set, and hence it is difficult to recommend using Shibata-type criteria.

Acknowledgement

The author wishes to thank the Associate Editor and a referee for their suggestions which have been very useful in revising the paper. This research has been supported in part by NSF grant DMS-9108295.

Appendix. Optimal Choice of the Cutoff Point c and the Smoothing Function G

The section will be devoted to a discussion on the choice of the cutoff point c and the smoothing function G. There are a number of approaches to this problem and we present only one of them. At this point of time we are unable to provide a definitive answer on how to obtain the exact form of the optimum G. This issue needs further research. Here we obtain an approximate value of the optimal cutoff point c for a reasonably broad class of coefficients $\{\beta_j\}$ and then we derive a form for the optimal smoothing function G. In order to save space, we provide only an outline of the arguments and delete the mathematical details.

For the discussion here we change the notations a little. Assume here that $X'X = nI$ and consider estimates of the form

$$\hat{\mu}_t(c, G) = \sum_{1 \le j \le p} x_{tj} \hat{\beta}_j G(\sqrt{n}\hat{\beta}_j / c), \quad t = 1, \ldots, n.$$

We first find the value of the cutoff point c which minimizes $L(c, G) = n^{-1} \sum_{1 \le t \le n} \{\hat{\mu}_t(c, G) - \mu_t\}^2$. Without loss of generality assume that β_j's are nonnegative and $\sum \beta_j^2 < \infty$. Also assume that β_j is given by $\beta_j = \beta(1/j)$, where β is a nonnegative smooth function on $[0, 1]$ with $\beta(0) = 0$. Let s be such that $\beta^{(j)}(0) = 0$, $j = 1, \ldots, s-1$, $\beta^{(s)}(0) \neq 0$. Then, by Taylor series we get $\beta(u) = d_s u^s + O(u^{s+1})$, for some constant d_s, $0 \le u \le 1$. In order to keep the calculations simple, assume that $\beta(u) = u^s$. Note that there is no need for s to be an integer. It is enough to have $s > 1/2$ so that $\sum \beta_j^2 < \infty$.

Consider the problem of estimating β_j with an estimate of the form $d\hat{\beta}_j$, where d is a constant. Then the quantity $E[d\hat{\beta}_j - \beta_j]^2$ is minimized at $d = \beta_j^2/(\sigma^2/n + \beta_j^2)$. This suggests that a reasonable weight function is of the form $\tilde{G}(u) = u^2/(b + u^2)$, where b is a constant. We show that, under some restrictions, the optimal weight function is indeed of this form. In the subsequent discussion we consider a weight function of the form $G(u) = u^2 H(u)$, where H is square integrable and $H(0) \neq 0$. Assuming that the errors in model (1) are i.i.d. $N(0, \sigma^2)$ we get

$$L(c, G) = \sum E\{\hat{\beta}_j G(\sqrt{n}\hat{\beta}_j / c) - \beta_j\}^2 = n^{-1} \sum E\{(Z + \gamma_j) G(Z/c + \gamma_j/c) - \gamma_j\}^2, \tag{A.1}$$

where $\gamma_j = \sqrt{n}\beta_j$ and Z is a $N(0, \sigma^2)$ variable. Assuming that $c \to \infty$ but $c/\sqrt{n} \to 0$, we get

$$\begin{aligned}
& \sum \Big[E\{(Z + \gamma_j) G(Z/c + \gamma_j/c)\} - \gamma_j \Big]^2 \\
= \; & \sum \Big[E\{c^{-2}(Z + \gamma_j)^3 H(Z/c + \gamma_j/c)\} - \gamma_j \Big]^2 \\
\approx \; & \sum \Big[E\{c^{-2}(Z + \gamma_j)^3 H(\gamma_j/c)\} - \gamma_j \Big]^2 \\
= \; & \sum \Big[3c^{-2}\sigma^2 \gamma_j H(\gamma_j/c) + \gamma_j \{(\gamma_j/c)^2 H(\gamma_j/c) - \gamma_j\} \Big]^2 \\
= \; & 9c^{-4}\sigma^4 \sum \Big[\gamma_j H(\gamma_j/c) \Big]^2 + \sum \gamma_j^2 \Big[(\gamma_j/c)^2 H(\gamma_j/c) - \gamma_j \Big]^2 \\
& + 6c^{-2}\sigma^2 \sum \gamma_j^2 H(\gamma_j/c) \Big[(\gamma_j/c)^2 H(\gamma_j/c) - \gamma_j \Big] = I_1 + I_2 + I_3, \quad \text{say.}
\end{aligned}$$

Denoting the interval $[(\sqrt{n}/c)p^{-s}, \sqrt{n}/c]$ by T, we get

$$I_1 \approx 9c^{-4}\sigma^4 \int_1^p [(\sqrt{n}u^{-s}) H(\sqrt{n}u^{-s}/c)]^2 du = 9c^{-2}\sigma^4 s^{-1} (\sqrt{n}/c)^{1/s} \int_T t^{1-1/s} H^2(t) dt.$$

Similar arguments show

$$I_2 \approx s^{-1}c^2(\sqrt{n}/c)^{1/s}\int_T t^{1-1/s}[t^2H(t)-1]^2dt$$
$$I_3 \approx 6\sigma^2 s^{-1}(\sqrt{n}/c)^{1/s}\int_T t^{1-1/s}H(t)[t^2H(t)-1]dt.$$

Note that the interval T converges to $(0,\infty)$ as $n\to\infty$. Assuming that $J_1(H)=\int_0^\infty t^{1-1/s}H^2(t)dt<\infty$ and $J_0(H)=\int_0^\infty t^{1-1/s}[t^2H(t)-1]^2dt<\infty$, and by keeping only the dominant term, which is I_2 here, we get

$$\sum\Big[E\{(Z+\gamma_j)G(Z/c+\gamma_j/c)-\gamma_j\}\Big]^2 \approx s^{-1}c^2(\sqrt{n}/c)^{1/s}J_0(H). \tag{A.2}$$

Similarly, if we also assume that $J_2(H)=\int_0^\infty t^{3-1/s}H^2(t)dt<\infty$, then it can be shown that by keeping only the dominant term the following is true

$$\sum \mathrm{Var}\left((Z+\gamma_j)G(Z/c+\gamma_j/c)-\gamma_j\right) \approx c^{-4}p15\sigma^6H^2(0). \tag{A.3}$$

Assume that $p/n\to\tau$, where $0<\tau\le 1$. Combining (A.1), (A.2) and (A.3) we get,

$$L(c,G)\approx n^{-1}(1/s)c^2(\sqrt{n}/c)^{1/s}J_0(H)+c^{-4}\tau 15\sigma^6H^2(0).$$

It is easy to verify that $L(c,G)$ is minimized at

$$c^*=n^{1-1/(2s)}[60\tau\sigma^6H^2(0)/\{(2-s^{-1})s^{-1}\}]^{s/(6s-1)}. \tag{A.4}$$

By denoting $r=(2s-1)/(6s-1)$, the value of the error $L(c,G)$ at $c=c^*$ is

$$L(c^*,G)\approx d_0n^{-2r}[J_0(H)]^{1-r}[H^2(0)]^r[60\tau\sigma^6/\{(2-s^{-1})s^{-1}\}]^r(1.5s^{-1}-.25s^{-2}).$$

The expression for $L(c^*,G)$ given above tells us that in order to obtain the optimal weight function we need to minimize the functional $J_0(H)$ with respect to the square integrable function H, with the side condition $H(0)=h_0\neq 0$. However, this minimum is zero and it will be obvious from the following example. Let $H_\delta(u)=u^{-2}$, $u>\delta$, but $H_\delta(u)=h_0$, $u\le\delta$. Then $J_0(H_\delta)\to 0$ as $\delta\to 0$, but $J_1(H_\delta)$ and $J_2(H_\delta)$ converge to ∞. However, the optimum value of c given in (A.4) was obtained on the basis of the assumption that $J_1(H)$ and $J_2(H)$ are finite. This suggests that we should minimize $J_0(H)$ subject to the constraints that $J_1(H)$ and $J_2(H)$ are bounded. This leads to the following minimization problem:
Minimize $J_0(H)+\lambda_1J_1(H)+\lambda_2J_2(H)$ subject to the constraint $H(0)=h_0$. A calculus of variations technique tells us that the minimum is attained at $H^*(u)=u^3/(\lambda_1u+\lambda_2u^3+u^5)$. The constraint $H(0)=h_0\neq 0$ forces $\lambda_1=0$, $\lambda_2=1/h_0$. Consequently, the optimal weight function is

$$G^*(u)=u^2H^*(u)=u^2/(1/h_0+u^2).$$

References

Akaike, H. (1970). Statistical predictor identification. *Ann. Inst. Statist. Math.* **22**, 203-217.

Burman, P. (1994). Model fitting via testing. Technical Report #295, Division of Statistics, University of California, Davis.

Burman, P. and Chen, K. W. (1989). Nonparametric estimation of a regression function. *Ann. Statist.* **17**, 1567-1596.

Li, K.-C. (1987). Asymptotic optimality for C_p, C_L, cross-validation and generalized cross validation: Discrete index set. *Ann. Statist.* **15**, 958-975.

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

Mardia, K. V., Kent, J. T. and Bibby, J. M. (1979). *Multivariate Analysis.* Academic Press, Orlando.

Schwarz, G. (1978). Estimating the dimension of a model. *Ann. Statist.* **6**, 461-464.

Shibata, R. (1981). An optimal selection of regression variables. *Biometrika* **68**, 45-54.

Shibata, R. (1984). Approximate efficiency of a selection procedure for the number of regression variables. *Biometrika* **71**, 43-49.

Stone, M. (1974). Cross-validatory choice and assessment of statistical prediction. *J. Roy. Statist. Soc. Ser. B* **36**, 111-147.

Zhang, P. (1992). On the distribution properties of model selection criteria. *J. Amer. Statist. Assoc.* **87**, 732-737.

Division of Statistics, 469 Kerr Hall, University of California, Davis, CA 95616, U.S.A.

(Received October 1993; accepted October 1995)