AN EFFECTIVE SELECTION OF REGRESSION VARIABLES WHEN THE ERROR DISTRIBUTION IS INCORRECTLY SPECIFIED*

Wolfgang Härdle
FB Mathematik
Johann-Wolfgang-Goethe-Universität
D-6000 Frankfurt/M.
WEST GERMANY

and

Department of Statistics
University of North Carolina
Chapel Hill, North Carolina 27514

Abstract

In the situation where the statistician estimates regression parameters by maximum likelihood methods but fails to choose a likelihood function matching the true error distribution, an asymptotically efficient selection of regression variables is considered. The proposed procedure is especially useful when a robust regression is applied but the data in fact do not require that treatment. Examples are given and relationships to other selectors such as Mallows' C_p are investigated.

Keywords and Phrases: variable selection, regression analysis, robust regression, model choice

AMS Subject Classification: Primary 62J05, Secondary 62G99

*Research supported by Deutsche Forschungsgemeinschaft, Sonderforschungsbereich 123 "Stochastische Mathematische Modelle" and AFOSR Contract No. F49620 82 C 0009.

1. Introduction and Results

Suppose that $Y = (Y_1, \ldots, Y_n)'$ is a random vector of n observations with mean $\mu = (\mu_1, \ldots, \mu_n)'$ and assume that the i^{th} observation is associated with a covariate x_i, $i = 1, \ldots, n$. The dependence of μ_i on x_i is specified by an infinite parameter vector β; therefore at most n parameters can be estimated on the basis of the observations. Suppose that a certain likelihood function, not necessarily matching the true error distribution, has been chosen and that parameter estimates in a finite dimensional submodel p have been obtained by the maximum likelihood principle. The regression curve μ_i at x_i is estimated by $\hat{\mu}_i(p)$ and a loss $L_n(p) = \|\mu - \hat{\mu}(p)\|^2$ is suffered. We shall consider an efficient model selection procedure that asymptotically minimizes the loss $L_n(p)$ over a certain class of finite dimensional models of increasing dimension.

This paper completes earlier papers in various ways. Shibata (1981), Breiman and Freedman (1983) consider the problem of selecting regression variables when the true error distribution is known to be Gaussian and derive selectors that are equivalent to ours in this case. In the setting of least squares estimation Li (1984) recently gave conditions for asymptotic efficiency of model choice procedures based on cross validation, FPE and other means.

Schrader and Hettmansberger (1980) provided a robust analysis of variance based on Huber's M-estimates and propose a likelihood ratio type of test for testing between finite dimensional submodels. This approach was pursued by Ronchetti (1985) who derived a "robust model selection" procedure that is related to ours.

The mismatch of the chosen likelihood function and of the true error distribution can happen when a robust regression estimation (Huber, 1973) is applied but the data is in fact Gaussian. We may also think of the reverse

situation that a Gaussian maximum likelihood estimate (i.e. the least squares estimate) is computed but the true error distribution is different, possibly long tailed. This mismatch is mathematically reflected in the way that a larger constant in the dimensionality penalty term has to be chosen. For instance, Akaike's (1970) AIC has the penalty constant 2, whereas in case of mismatch this constant is changed depending on the type of mismatch. If there are outliers in the data and we apply AIC based on a Gaussian maximum likelihood estimate we will overfit the data since the model selection procedure tries to fit the outliers. On the other hand, if we apply the model selection procedure to be presented below, a high dimensional model is more penalized since the penalty constant is bigger than 2.

In the simple case that the data is Gaussian and the statistician chooses a Gaussian likelihood function, then our model selection procedure is equivalent to Mallow's C_p (1973) (see Section 4). This entails equivalence to many other selectors such as FPE, AIC, GCV, as was shown by Li (1984).

We will assume the control variables $x_i = (x_{i1}, x_{i2}, \ldots)'$, $i = 1, \ldots, n$, and the parameter vector $\beta = (\beta_1, \beta_2, \ldots)'$ are in ℓ_2.

The model can then be written as

$$Y = X\beta + e,$$

where $e = (e_1, \ldots, e_n)'$ is the vector of the independent observation errors having distribution F with density f and $X' = (x_1' x_2' \ldots x_n')$ is considered as a linear operator from ℓ_2 to $\mathbb{R}^n$. By $p = (p_1, p_2, \ldots, p_{k(p)})$ we denote a finite dimensional submodel with parameter

$$\beta'(p) = (0, \ldots, \beta_{p_1}, 0, \ldots, \beta_{p_2}, 0, \ldots, \beta_{p_{k(p)}}, 0, \ldots)$$

where $p_1 < p_2 < \ldots < p_{k(p)}$, $k(p) \geq 1$.

The statistician chooses a likelihood function h of which he believes to represent the true error distribution, and estimates the parameters in a submodel p by maximizing

$$\prod_{i=1}^{n} h(Y_i - x_i'(p)\beta(p))$$

where $x_i'(p) = (0,\ldots, x_{ip_1},0,\ldots, x_{ip_2},0,\ldots, x_{ip_{k(p)}},0,\ldots)$.

Call this maximum likelihood estimate $\hat{\beta}(p)$. The regression surface point μ_i is estimated by

$$\hat{\mu}_i(p) = \langle x_i, \hat{\beta}(p)\rangle.$$

Define $\psi(u) = -\frac{d}{du}\log h(u)$, $\gamma = E_F\psi^2(e)/(E_F\psi'(e))^2$, $B_n = X'X$ and let P_n be a family of models p. A possible selection rule for choosing a model $p \in P_n$ could be defined by $W_n'(p) = -\|\hat{\mu}(p)\|^2 + 2\gamma k(p) + \|\mu\|^2$, since

$$W_n'(p) - L_n(p) = -\{\|\hat{\mu}(p) - \mu\|^2 + 2\langle\hat{\beta}(p) - \beta, \beta\rangle_{B_n} + \|\mu\|^2\}$$

$$+ 2\gamma k(p) + \|\mu\|^2 - \|\hat{\mu}(p) - \mu\|^2$$

$$= -2\|\hat{\mu}(p) - \mu\|^2 - 2\{-\|\hat{\mu}(p) - \mu\|^2 + \langle\hat{\beta}(p) - \beta, \hat{\beta}(p)\rangle_{B_n}\}$$

$$+ 2\gamma k(p)$$

$$= 2\{\gamma k(p) - \langle\hat{\beta}(p) - \beta, \hat{\beta}(p)\rangle_{B_n}\} \qquad (1)$$

where $\langle u,v\rangle_{B_n}$ denotes the bilinear form $u'B_nv$ for vectors $u,v \in \ell_2$. Suppose that it can be shown that the last term in (1) is tending to a constant uniformly over the model class P_n. Then minimizing $W_n'(p)$ over P_n will be the same task, at least asymptotically, as minimizing $L_n(p)$. However, W_n' cannot be computed directly from the data since it depends on the unknown regression curve μ. But note that the last term in $W_n'(p)$ is independent of the model p. We will therefore define

$$W_n(p) = -\|\hat{\mu}(p)\|^2 + 2\gamma k(p)$$

as the score function that is to be minimized over P_n. The problem of simultaneously estimating γ from the data, in order to make W_n completely data driven is considered in Section 4.

Remark 1

If the statistician is in the happy situation of knowing f, then he will choose $h \equiv f$. If f is symmetric, then by partial integration

$$I(F) = E_F\psi^2 = \int\psi^2 f = \int(f'/f)^2 f = \int f'\psi = \int\psi' f = E_F\psi'$$

and therefore the constant γ reduces to $(E_F\psi')^{-1} = I(F)^{-1}$, the Fisher-information number in a location family with density f.

We will use the concept of asymptotic efficiency as in Shibata (1981), Li (1984), Stone (1984): A selected $\hat{p}$ is called *asymptotically optimal* if, as $n \to \infty$

$$\frac{L_n(\hat{p})}{\inf_{p\in P_n} L_n(p)} \xrightarrow{p} 1. \qquad (2)$$

The following condition on ψ will be needed.

Condition 1

The function ψ is centered i.e., $E_F\psi(e) = 0$ and twice differentiable with bounded second derivative. We furthermore assume that $E_F[q^{-1}(\psi'(e) - q]^{2N} < \infty$ for some positive integer N and $q = E_F\psi'(e) > 0$.

The estimates $\hat{\beta}(p)$ will be compared with the Gauss-Markov estimates in the model p based on the (unobservable) pseudodata $\tilde{Y}_i = \mu_i + \tilde{e}_i$, $\tilde{e}_i = \psi(e_i)/q$. Define X(p) as the (n,p) matrix containing the nonzero control variables in model p and assume that $B_n(p) = X'(p)X(p)$ has full rank k(p). Then the Gauss-Markov estimate of μ based on the pseudodata $\tilde{Y} = (\tilde{Y}_1, \ldots, \tilde{Y}_n)'$ is defined as $\tilde{\mu}(p) = M_n(p)\tilde{Y}$, where $M_n(p) = X(p)B_n^{-1}(p)X(p)$ denotes the hat matrix in model p. The loss for the Gauss-Markov estimate is $\tilde{L}_n(p) = \|\tilde{\mu}(p) - \mu\|^2$ which will be approximately $L_n(p)$ as will be seen later on. A well behaved design is guarenteed by

Condition 2

There exists a positive integer N such that with $\tilde{R}_n(p) = E_F\tilde{L}_n(p)$

$$\sum_{p \in P_n} \tilde{R}_n(p)^{-N} \to 0, \quad \text{as } n \to \infty \ .$$

Let h(p) be the largest diagonal element of the hat matrix $M_n(p)$. The speed of h(p) is controlled by

Condition 3

$$\sup_{p \in P_n} h(p)\tilde{R}_n(p) \to 0, \text{ as } n \to \infty \ .$$

Remark 2

It follows from Condition 3 that

$$k^2(p)/n \to 0, \text{ as } n \to \infty,$$

since $\tilde{R}_n(p) = \gamma k(p) + \|\mu - \mu(p)\|^2$, $\mu(p) = M_n(p)\mu$. This should be seen as an analogue of the necessary condition, $p^2/n \to 0$, that can be found in Huber (1981, p. 166). Conditions 2,3 imply also

$$\sum_{p \in P_n} h(p)^N \to 0. \tag{3}$$

Remark 3

If ψ is bounded, as is assumed in robust regression analysis, Condition 2 can be weakened. It is seen from the proofs that in this case Bernstein's inequality could be used instead of Whittle's (1960, Theorem 2). Condition 2 could be weakened to $\Sigma_{p \in P_n} \exp(-C\tilde{R}_n(p)) \to 0$, for some $C > 0$.

Denote by $\hat{p}$ a model $p \in P_n$ that minimizes $W_n(p)$ over P_n. The main result is as follows.

Theorem

Under Conditions 1-3, $\hat{p}$ is asymptotically optimal.

The rest of the paper is organized in four sections. In Section 2 the theorem above is shown, in Section 3 we give a variety of examples that satisfy our conditions, and in Section 4 the estimation of γ and the relation to other model selection procedures is investigated. Section 5 is devoted to detailed proofs of the lemmata that are needed in showing the asymptotic optimality.

2. Proof of the Theorem

In the proof of the Theorem, the following Lemmata will be used.

Lemma 1

Under the conditions of the Theorem, for all $\varepsilon > 0$

$$P\{\sup_{p\in P_n} \|\hat{\tilde{\mu}}(p) - \tilde{\mu}(p)\|^2 / \tilde{R}_n(p) > \varepsilon\} \to 0, \text{ as } n \to \infty.$$

Lemma 2

Under the conditions of the Theorem, for all $\varepsilon > 0$

$$P\{\sup_{p\in P_n} |\tilde{L}_n(p) - \tilde{R}_n(p)|/\tilde{R}_n(p) > \varepsilon\} \to 0, \text{ as } n \to \infty$$

Lemma 3

Under the conditions of the Theorem, for all $\varepsilon > 0$

$$P\{\sup_{p\in P_n} |\gamma k(p) - (\tilde{\mu}(p) - \mu)'\tilde{\mu}(p) + \mu'\tilde{e}|/\tilde{R}_n(p) > \varepsilon\} \to 0, \text{ as } n \to \infty.$$

Recall that the Gauss-Markov estimate based on the pseudodata $\tilde{Y}$ is $\tilde{\beta}(p) = B_n^{-1}(p)X'(p)\tilde{Y}$. The crossterm in (1) will be approximated by a corresponding crossterm based on the linearized estimates $\tilde{\beta}(p)$.

$$\begin{aligned} &\langle\hat{\beta}(p) - \beta, \hat{\beta}(p)\rangle_{B_n} - \langle\tilde{\beta}(p) - \beta, \tilde{\beta}(p)\rangle_{B_n} \\ &= \|\tilde{\mu}(p) - \hat{\mu}(p)\|^2 + \langle\hat{\beta}(p) - \tilde{\beta}(p), \tilde{\beta}(p) - \beta(p)\rangle_{B_n} \\ &+ \langle\hat{\beta}(p) - \tilde{\beta}(p), \beta(p)\rangle_{B_n} + \langle\tilde{\beta}(p) - \beta, \hat{\beta}(p) - \tilde{\beta}(p)\rangle_{B_n}. \end{aligned} \tag{4}$$

By Lemma 1, the first term is of lower order than $\tilde{R}_n(p)$ uniformly over P_n, the second term is bounded by the Cauchy-Schwarz inequality and then Lemma 1 and Lemma 2 are applied. The third term is handled by formula (8), given in the proof of Lemma 1, by setting $a = \beta(p)$, $\eta = \hat{\beta}(p)$. The fourth term is handled as the second term. Suppose that

$$\sup_{p,p' \epsilon P_n} \left| \frac{(W_n(p) - W_n(p')) - (L_n(p) - L_n(p'))}{L_n(p) + L_n(p')} \right| = o_p(1), \qquad (5)$$

and let p^* denote a minimizer of $L_n(p)$ over P_n. Then by (5) with probability greater than $1 - \epsilon$,

$$\frac{W_n(\hat{p}) - W_n(p^*) - (L_n(\hat{p}) - L_n(p^*))}{L_n(\hat{p}) + L_n(p^*)} \geq -\epsilon.$$

By the definition of $\hat{p}$, $W_n(\hat{p}) - W_n(p^*) \leq 0$; therefore,

$$-(L_n(\hat{p}) - L_n(p^*)) \geq -\epsilon(L_n(\hat{p}) + L_n(p^*))$$

$$L_n(p^*)(1 + \epsilon) \geq L_n(\hat{p})(1 - \epsilon)$$

$$1 \geq \frac{L_n(p^*)}{L_n(\hat{p})} \geq \frac{1 - \epsilon}{1 + \epsilon}$$

which shows that (2) holds, i.e. $\hat{p}$ is asymptotically optimal. Formula (5) follows by observing that

$$\frac{(W_n'(p) + \mu'\tilde{e} - L_n(p))}{L_n(p)}$$

$$= \frac{2(\gamma k(p) - \langle \hat{\beta}(p) - \beta, \hat{\beta}(p) \rangle_{B_n} + \mu'\tilde{e})}{\tilde{R}_n(p)} \cdot \frac{\tilde{L}_n(p)}{L_n(p)} \cdot \frac{\tilde{R}_n(p)}{\tilde{L}_n(p)}$$

The first factor is tending to zero in probability, uniformly over P_n by Lemma 3 and formula (4). The two other factors tend to one in probability, uniformly over P_n, by Lemma 1 and Lemma 2.

3. Examples

We start with a reformulation of Condition 2 in the case of hierarchical model sequences, i.e. $P_n = \{(1),(1,2),\ldots, (1,2,\ldots, p_n)\}$ with p_n tending to infinity. In this case Condition 2 follows for $N = 2$ from

Condition 2'

$$\inf_{p\in P_n} \tilde{R}_n(p) \to \infty, \text{ as } n \to \infty.$$

We slightly abuse notation by writing j for $(1,\ldots, j)$. Then Condition 2 follows from $\tilde{R}_n(j) = \gamma j + \|\mu(p) - \mu\|^2$ and

$$\sum_{p\in P_n} \tilde{R}_n(p)^{-2} = \sum_{j=1}^{J_n} \tilde{R}_n(j)^{-2} + \sum_{j=J_n+1}^{P_n} \tilde{R}_n(j)^{-2}$$

$$\leq J_n\{\inf_{p\in P_n} \tilde{R}_n(p)\}^{-2} + \gamma^{-2} \sum_{j=J_n+1}^{\infty} j^{-2} \to 0,$$

if J_n tends to infinity slowly enough. In the following examples we assume that P_n represents a hierarchical model sequence. The following lemma, which is due to Shibata (1981), is useful in checking Condition 2'.

Lemma 4

Assume that with a positive divergent sequence $\{c_n\}$ the linear operator $c_n^{-1}B_n$ converges weakly to a nonsingular operator $B:\ell_2 \to \ell_2$, such that every $p\times p$ principal submatrix $B(p)$ has full rank p for all $p > 0$. If β has infinitely many nonzero coordinates, then Condition 2' holds and p^* diverges to infinity, as $n \to \infty$.

Are the conditions of the Theorem fulfilled for typical examples? We check conditions in examples given by Shibata (1981).

Example 1

Consider the polynomial regression on the interval [0,1). Here

$$X_{ij} = \left(\frac{i-1}{n}\right)^{j-1},\ i = 1,\ldots,\ n,\quad j = 1,2,\ldots$$

and

$$Y_i = \sum_{j=1}^{\infty} \left(\frac{i-1}{n}\right)^{j-1}\beta_j + e_i,\ i = 1,\ldots,\ n$$

is observed.

Condition 1 is model independent and is an assumption about the error distribution. Condition 2' is satisfied via Lemma 4 (set $c_n = n$). It remains to check Condition 3. The symmetric matrix $B_n^{-1}(p)$ has a spectral decomposition (Mardia, Kent, Bibby, 1979, Theorem A.6.4)

$$B_n^{-1}(p) = \Gamma_n \Lambda_n \Gamma_n',$$

where $\Lambda_n = \mathrm{diag}(\lambda_1(p),\ldots,\ \lambda_p(p))$ and $\Gamma_n = (\gamma_1,\ldots,\ \gamma_p)$, $\gamma_j = (\gamma_{1j},\ldots,\ \gamma_{pj})'$ the i^{th} normalized eigenvector of $B_n^{-1}(p)$. Lemma 4 insures that $\lambda_{min}(p)$ the smallest eigenvalue of $n^{-1}B_n(p)$ is bounded above zero by a constant C. Therefore each diagonal element h_i of $M_n(p)$ can be estimated by

$$h_i = \sum_{j=1}^{p}\sum_{k=1}^{p} x_{ij}x_{ik}\left(\sum_{\ell=1}^{p}\gamma_{\ell j}\gamma_{\ell k}\lambda_\ell(p)\right)$$

$$\le \sum_{\ell=1}^{p}\lambda_\ell(p)\left(\sum_{j=1}^{p}x_{ij}^2\right)\left(\sum_{j=1}^{p}\gamma_{\ell j}^2\right)$$

$$\le p\lambda_{max}(p)\sum_{j=1}^{p}x_{ij}^2 \le p^2\lambda_{min}(p)^{-1} \le C^{-1}p^2/n.$$

So Condition 3 is fulfilled if we ask for

$$\sup_{1\le p\le p_n} p^2\tilde{R}_n(p)/n \to 0. \qquad (6)$$

A necessary condition is $p^3/n \to 0$ which is slightly stronger than Huber's (1981) conditions.

Example 2

Consider the following representation of the regression curve

$$\mu_i = \sum_{j=1}^{\infty} \beta_j \cos(\pi(j-1)(i-1)/nj).$$

Here the observations are taken at $x = 0, n^{-1}, \ldots, (\frac{n-1}{n})$. As in the example above Condition 2 is satisfied by Lemma 4, setting $c_n = n/2$. Condition 3 is satisfied by similar arguments as above if we assume that (6) holds.

Example 3

Consider the robust M-estimation of location at different units x_j. Observations are taken repeatedly at p_n different units and n/p_n observations are taken at the point x_j, $j = 1, \ldots, p_n$. Assume that $E_F\psi(e) = 0$, then Condition 1 is satisfied if ψ, ψ' are bounded. Shibata (1981, p. 51) shows that Condition 2 is satisfied if the vectors of the control-variables $(x_1, \ldots, x_{p_n})$ are linearly independent. Condition 3 can be checked as above.

4. Other methods and estimation of γ

There are a variety of other model selection methods, most of which were shown to be equivalent to Mallows' C_p. We therefore compare our method with C_p only. For simplicity, we work with the linearized estimate $\tilde{\mu}(p)$ based on the pseudodata $\tilde{Y}$. Mallows' (1973) score function reads

$$C_p(p) = \|\tilde{Y} - \tilde{\mu}(p)\|^2 + 2\gamma k(p)$$

$$= \|\tilde{e}\|^2 + \tilde{L}_n(p)$$

$$+ 2\tilde{e}'(I_n - M_n(p))\mu$$

$$+ 2\{\gamma k(p) - \tilde{e}' M_n(p)\tilde{e}\}.$$

The first term is independent of p, the third and the last term vanish uniformly over model classes P_n, as can be seen in the next section. This shows that a model selected by C_p is asymptotically optimal.

It can now be seen that $W_n(p)$ has a similar structure.

$$W_n(p) = -\|\tilde{\mu}(p)\|^2 + 2\gamma k(p)$$

$$= -\|\tilde{\mu}(p) - \mu\|^2 - \|\mu\|^2 - 2(\tilde{\mu} - \mu)'(\mu - \tilde{\mu})$$

$$- 2(\tilde{\mu} - \mu)'\tilde{\mu} + 2\gamma k(p)$$

$$= \tilde{L}_n(p) + 2\gamma k(p) - 2\tilde{e}' M_n(p)\tilde{e} + 2\tilde{e}'(I - M_n(p))\mu$$

$$+ 2\mu'\tilde{e} - \|\mu\|^2.$$

Here the last two terms are independent of the model. The remaining terms are identical to those in Mallows C_p, which shows that $W_n(p)$ is equivalent to C_p.

It could be argued that the score function that is proposed here is not very reasonable in a practical application since the constant γ is unknown to the statistician. However, if the constant γ can be consistently estimated (independent of p) then the score function based on an estimated γ is also asymptotically optimal. A consistent estimate $\hat{\gamma}_n$ of γ is provided, for instance, by

$$n^{-1} \sum_{i=1}^{n} \psi^2(\hat{e}_i(p_n)) / (n^{-1} \sum_{i=1}^{n} \psi'(\hat{e}_i(p_n)))^2$$

where $\hat{\varepsilon}_i(p_n)$ denote residuals from a fit with a deterministic model p_n, increasing in magnitude as $n \to \infty$. A Taylor expansion and the Cauchy-Schwarz inequality show that $\hat{\gamma}_n \xrightarrow{p} \gamma$, as $n \to \infty$.

5. Proofs

In this section we give the proofs of Lemmas 1-3. The proof of Lemma 1 follows a related proof in Huber (1981), Section 7.4. Similar ideas were used by Cox (1983) who considered M-type smoothing splines. In order to simplify notation we will consider the hierarchical case only, i.e. the model "p" is identified with "(1,2,..., k(p)), k(p) = p." Furthermore we assume without loss of generality that the coordinate system in the p-dimensional subspace of the first p components has been chosen so that $X'(p)X(p) = I_p$. Consider the mapping $\Phi: \mathbb{R}^p \to \mathbb{R}^p$, $\Phi_k(\eta) = -q^{-1}\Sigma_{i=1}^n \psi(Y_i - \Sigma_{j=1}^p x_{ij}\eta_j)x_{ik}$, $k = 1,\ldots, p$ where $\eta = (\eta_1,\ldots, \eta_p)' \in \mathbb{R}^p$. A zero (with respect to η) of Φ will be compared with a zero of $\psi_k(\eta) = \eta_k - \Sigma_{i=1}^n (\mu_i + \tilde{e}_i)x_{ik}$, where $\tilde{e}_i = \psi(e_i)/q$, $q = E_F\psi'(e)$. The zero of $\psi_k(\eta)$ is the least squares estimate $\tilde{\beta}(p) = X(p)\tilde{Y}$ based on the pseudodata $\tilde{Y}$. Consider an arbitrary normalized vector $a \in \mathbb{R}^p$, $\|a\| = 1$. Taylor expansion of Φ, using Condition 1, leads to

$$\sum_{k=1}^p a_k(\Phi_k(\eta) - \psi_k(\eta))$$

$$= -q^{-1}\sum_{k=1}^p a_k \sum_{i=1}^n (\psi'(e_i) - q)\Big(\sum_{j=p+1}^{\infty} x_{ij}\beta_j\Big)x_{ik}$$

$$- q^{-1}\sum_{k=1}^p a_k \sum_{i=1}^n (\psi'(e_i) - q)\sum_{j=1}^p x_{ij}x_{ik}(\beta_j - \eta_j)$$

$$- \frac{1}{2}q^{-1}\sum_{k=1}^p a_k \sum_{i=1}^n \psi''(e_i + \nu(\sum_{j=1}^{\infty} x_{ij}\beta_j - \sum_{j=1}^p x_{ij}\eta_j))(\sum_{j=1}^{\infty} x_{ij}\beta_j - \sum_{j=1}^p x_{ij}\eta_j)^2 x_{ik}$$

$$= T_{1,n}(p) + T_{2,n}(p) + T_{3,n}(p),\ \nu \in (-1,1).$$

We will now show that each of these terms uniformly vanishes over P_n, in the sense that

$$\sup_{p\epsilon P_n} T_{\alpha,n}(p)/\tilde{R}_n^{1/2}(p) \xrightarrow{P} 0, \quad \alpha = 1,2,3 \tag{7}$$

for all (η,a) in the set

$$F_n = \bigcap_{p\epsilon P_n} \left\{ (\eta,a) : \sum_{i=1}^{n} \left(\sum_{j=1}^{p} x_{ij}(\beta_j - \eta_j) \right)^2 \le K\tilde{R}_n(p), \right.$$
$$\left. \|a\| = 1 \right\}$$

Define for $i = 1,\ldots, n$

$$V_i = q^{-1}(\psi'(e_i) - q),$$

$$B_{i,n}(p) = \sum_{j=p+1}^{\infty} x_{ij}\beta_j,$$

$$s_i = \sum_{k=1}^{p} a_k x_{ik},$$

$$\Delta_{i,n}(p) = \sum_{j=1}^{p} x_{ij}(\beta_j - \eta_j).$$

Note that

$$\|s\|^2 = \sum_{i=1}^{n} s_i^2 = \sum_{i=1}^{n} \left(\sum_{k=1}^{p} a_k x_{ik} \right)^2 = \|X(p)a\|^2 = \|a\|^2 = 1.$$

The first term $T_{1,n}(p)$ is estimated as follows.

$$P\{\sup_{p\epsilon P_n} |T_{1,n}(p)|/\tilde{R}_n^{1/2}(p) > \varepsilon\}$$

$$\le \sum_{p\epsilon P_n} \varepsilon^{-2N} E\{|T_{1,n}(p)|^{2N}/\tilde{R}_n^N(p)\}$$

$$= \sum_{p\in P_n} \varepsilon^{-2N} E\{ | \sum_{i=1}^{n} s_i B_{i,n}(p) V_i |^{2N} / \tilde{R}_n^N(p) \}.$$

Applying Condition 1 and Whittle's (1960), Th. 2 inequality, this term is bounded by

$$\sum_{p\in P_n} C_1 \varepsilon^{-2N} (\sum_{i=1}^{n} s_i^2 B_{i,n}(p))^N / \tilde{R}_n^N(p)$$

$$\leq C_1 \varepsilon^{-2N} \sum_{p\in P_n} (\max_{i=1}^{n} s_i^2)^N (\sum_{i=1}^{n} B_{i,n}^2(p))^N / \tilde{R}_n^N(p), \; C_1 > 0.$$

Recall that $\tilde{R}_n(p) = \gamma p + \|(I_n - M_n(p))\mu\|^2 \geq \Sigma_{i=1}^n B_{i,n}^2(p)$. Condition 2,3 (see Remark 2, Formula (3)) and the simple inequality $s_i^2 \leq \Sigma_{j=1}^p x_{ij}^2 \Sigma_{k=1}^p a_k^2 = h_i(p)$, where $h_i(p)$ denotes the i^{th} diagonal element of the hat matrix $M_n(p)$, imply that (7) holds for $\alpha = 1$.

The second term is estimated similarly. We omit some details. If $(\eta,a) \in F_n$ then as above

$$P\{\sup_{p\in P_n} |T_{2,n}(p)| / \tilde{R}_n^{1/2}(p) > \varepsilon\}$$

$$\leq C_1 \varepsilon^{-2N} \sum_{p\in P_n} h(p)^N (\sum_{i=1}^{n} \Delta_{i,n}^2(p))^N / \tilde{R}_n(p)^N.$$

$$\leq C_1 \gamma^{-N} K^N \varepsilon^{-2N} \sum_{p\in P_n} h(p)^N.$$

Now apply Condition 2,3 as described in Remark 2, Formula (3).

The third term, involving the second derivative of ψ is bounded by

$$\frac{1}{2}\, q^{-1} \sup|\psi''| \max_{i=1}^{n} |s_i| \sum_{i=1}^{n} (B_{i,n}(p) + \Delta_{i,n}(p))^2$$

If $(\eta,a) \in F_n$ we obtain that with a constant C_2

$$|T_{3,n}(p)| / \tilde{R}_n^{1/2}(p) \le C_2 h(p)^{1/2} \tilde{R}_n(p)^{1/2},$$

which tends to zero by Condition 3.

Altogether we have shown that

$$\sup_{p \in P_n} \Big| \sum_{k=1}^{p} a_k (\Phi_k(\eta) - \Psi_k(\eta)) \Big| / \tilde{R}_n^{1/2}(p) \xrightarrow{P} 0 \tag{8}$$

for all $(\eta,a) \in F_n$. This entails that for all η in the set

$$G_n = \{\eta \in \mathbb{R}^p : \sup_{p \in P_n} \Sigma_{i=1}^{n} (\Sigma_{k=1}^{p} (\eta_k - \beta_k) X_{ik})^2 / \tilde{R}_n(p) \le K\},$$

$$\sup_{p \in P_n} \|\Phi(\eta) - \Psi(\eta)\| / \tilde{R}_n^{1/2}(p) \xrightarrow{P} 0. \tag{9}$$

Condition 2 and bounds on higher moments as above imply that with probability greater than $1 - \delta$,

$$\sup_{p \in P_n} \|\tilde{\mu}(p) - \mu(p)\|^2 / \tilde{R}_n(p) < \gamma + \varepsilon. \tag{10}$$

This shows that $\tilde{\beta}(p) \in G_n$ with high probability. Note that

$$\|\Phi(\eta) - \eta\| = \|\Phi(\eta) - \Psi(\eta) + (\beta(p) - \tilde{\beta}(p)) + \beta(p)\|$$

$$\le \|\Phi(\eta) - \Psi(\eta)\| + \|\beta(p) - \tilde{\beta}(p)\| + \|\beta(p)\| \tag{11}$$

From formula (9) we know that the first term vanishes asymptotically. From (10) we conclude that for K big enough

$$\sup_{p \in P_n} \|\beta(p) - \tilde{\beta}(p)\| / \tilde{R}_n^{1/2}(p) \leq \frac{1}{2} K^{1/2}.$$

Certainly the third term can be made less than $\frac{1}{2} K^{1/2} p^{1/2}$. Thus the function $\eta \to \eta - \Phi(\eta)$ has a fixed point η^* in the compact, convex set G_n. Since this fixed point is necessarily a zero of Φ, it is seen that $\hat{\beta}(p)$ is in G_n with probability greater than $1 - \delta$. Substituting $\hat{\beta}(p)$ into equation (9) shows that Lemma 1 holds.

Lemma 2 is seen by the following equation.

$$\tilde{L}_n(p) - \tilde{R}_n(p) = \tilde{e}' M_n(p) \tilde{e} - \gamma k(p).$$

Condition 2 implies that

$$\sup_{p \in P_n} \left| \|M_n(p)\tilde{e}\|^2 - \gamma k(p) \right| / \tilde{R}_n(p) \xrightarrow{p} 0$$

which shows Lemma 2.

Lemma 3 follows similarly observing that

$$\langle \beta - \tilde{\beta}(p), \tilde{\beta}(p) \rangle_{B_n}$$

$$= \tilde{e}' (I_n - M_n(p))\mu - \tilde{e}' M_n(p) \tilde{e} - \tilde{e}' \mu.$$

Acknowledgement

I would like to thank Charles J. Stone and Johanna Behrens for helpful discussions.

References

Akaike, H. (1970). Statistical Predictor Identification. Ann. Inst. Math. Stat. 22, 203-217.

Breiman, L. and Freedman, D. (1983). How many variables should be entered in a regression equation. J. Amer. Stat. Assoc. 78, 131-136.

Cox, D. (1983). Asymptotics for M-type smoothing splines. Ann. Statist. 11, 530-551.

Huber, P. (1973). Robust regression: Asymptotics, conjectures, and Monte Carlo. Ann. Statist. 1, 799-821.

Huber, P. (1981). Robust Statistics. Wiley, New York.

Mallows, C. (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, London.

Li, K.C. (1984). Asymptotic optimality for C_p,C_l, cross-validation and generalized cross-validation: Discrete index set. Manuscript.

Ronchetti, E. (1985). Robust model selection in regression. Stat. and Prob. Letters 3, 21-23.

Schrader, R.M. and Hettmansberger, T.P. (1980). Robust analysis of variance based upon a likelihood ratio criterion. Biometrika, 67, 93-101.

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

Stone, C.J. (1984). An asymptotically optimal window selection rule for kernel density estimates. Ann. Statist. 12, 1285-1297.

Whittle, P. (1960). Bounds for the moments of linear and quadratic forms in independent variables. Theor. Prob. Appl. 3, 302-305.