The Annals of Statistics
1992, Vol. 20, No. 1, 414–427

SMOOTHING IN ADAPTIVE ESTIMATION[1]

BY JULIAN J. FARAWAY

University of Michigan

An adaptive maximum likelihood estimator based on the estimation of the log-density by B-splines is introduced. A data-driven method of selecting the smoothing parameter involved in the consequent density estimation is demonstrated. A Monte Carlo study is conducted to evaluate the small sample performance of the estimator in a location and a regression problem. The adaptive estimator is seen to compare favorably to some standard estimates. We show that the estimator is asymptotically efficient.

1. Introduction. The problem of adaptive estîmation was introduced by Stein (1956). One wishes to estimate a Euclidean parameter θ in the presence of an infinite-dimensional shape parameter G (usually the density). An adaptive estimate performs asymptotically as well (in the sense that the limiting distributions are the same) with G unknown as any estimate which utilizes knowledge of G. Note that the term adaptive estimation has been used elsewhere in the literature in the lesser sense of adapting to the data in some way. The estimates considered here are adaptive in a much stronger sense.

An adaptive estimator of the center of symmetry of an unknown distribution was constructed by Stone (1975). Bickel (1982) dealt with the multiple regression problem and simplified Stein's conditions for the circumstances under which adaptive estimation is possible. However, all the aforementioned work pertains to large sample behavior. Problems arise when one tries to apply these procedures in a practical small sample situation. The adaptive estimates proposed all depend on nonparametric density estimation and specifically the use of kernel density estimation. Essentially, one replaces the true density used in a one-step maximum likelihood estimate by an estimate of that density. A problem which pervades nonparametric density estimation is the choice of the smoothing parameters. Much work has been done on this subject and various schemes for the optimal choice of smoothing parameters have been proposed, mostly for the mean integrated square error criteria. Unfortunately, this is of little help in selecting the optimal smoothing parameters for the estimation of the parameters in the location or regression problem since these methods may give a good estimate of the density, but that is secondary to the problem at hand. We need to estimate the score, not the density, and our criterion is to minimize the MSE, say, of the estimate of θ. In fact, experimentation reveals that the optimal choice of smoothing parameters as regards the estimation of the location parameter tends to produce an under-

Received September 1988; revised April 1991.

[1] Research partially supported by ONR N00014-89-J-1563.

AMS 1980 *subject classifications*. Primary 62F35; secondary 62E25, 62F11, 62G20.

Key words and phrases. Adaptation, efficient estimation, B-splines.

smoothed density estimate. Theory in the aforementioned works gives required rates for the smoothing parameters, but is of no help given a specific sample size.

One might think that an optimal choice of smoothing parameters for a fixed sample size might be determined by empirical study. Unfortunately, experimentation reveals that the performance of adaptive estimates is highly sensitive to the choice of smoothing parameters, varying substantially from one distribution to another.

Hsieh and Manski (1987) tackle the problem in the simple regression case. Their approach is to bootstrap to estimate the MSE of the slope parameter for a given choice of smoothing parameters. Their estimate has the smoothing parameters that minimize the estimated MSE of the slope parameter. Their numerical results illustrate the virtue of their approach.

In this paper we take a somewhat different approach, using a plug-in method in which the theoretical MSE is replaced by its empirical value, rather than the bootstrap. The MSE of the parameter of interest can then be estimated directly. Furthermore, we use a different kind of density estimate. Previously, kernel density estimation has been used. However, when kernel estimates were used to estimate the MSE described above, the estimates were unstable and rough as a function of the smoothing parameter. This prompted the use of spline density estimation. The method is, however, different from maximum penalized likelihood density estimates, which also use splines. The reason that splines may work better in this situation is that we need estimates of the score and derivatives of the score, and splines are better suited to this purpose than kernel estimates. Of course, spline density estimation does not eliminate the problem of the choice of smoothing parameters; it just gives a different set of them to work with. Although spline density estimation is computationally more expensive than kernel density estimation for one evaluation, our method requires that the estimation be done only once, as opposed to many times for a bootstrap method, so that, overall, the method is probably faster. In Section 2, we consider the location problem and in Section 3 we consider the regression problem. We shall see that the adaptive estimator we propose produces estimates that compare favorably with the standard ones. In Section 4 we show that the adaptive estimator considered is indeed asymptotically efficient.

2. The location case. Suppose we observe $X_1, \dots, X_n$ i.i.d. random variables from a distribution $F(x - \theta)$, where F is symmetric and known. We may construct the one-step maximum likelihood estimate of θ

$$\hat{\theta} = \tilde{\theta} + \sum_{i=1}^{n} (\log f)'(X_i - \tilde{\theta}) \Big/ \sum_{i=1}^{n} (\log f)''(X_i - \tilde{\theta}), \tag{2.1}$$

where $\tilde{\theta}$ is an initial estimate of θ. However, suppose that F is not known. We propose to estimate it and use the resulting estimated density in (2.1).

We can approximate $f(x)$ by a density from an exponential family indexed by k of the form

$$\log f_k(x, \underline{a}) = \sum_{i=1}^{k} a_{ik} B_{ik}(x) + c(\underline{a}), \tag{2.2}$$

where $c(\underline{a})$ is the appropriate normalizing constant. The parameters $a_{1k}, \dots, a_{kk}$ may be estimated by maximum likelihood. B_{ik} is a symmetric quadratic B-spline, which is the sum of the B-spline defend on the positive knotpoints and the B-spline defined on the corresponding negative knotpoints. Let M be the largest knotpoint and let $B_{kk}(x)$ be linear if $|x| > M$, defined so that there is continuity and continuity of the first derivative at M. So f_k will have exponential linear tails and exponential piecewise quadratic midsection. Other bases will suffice, but B-splines were chosen because of the availability of theory and software and specifically because their nonnegative and local support makes proofs easier and computations more stable; see de Boor (1978) for a detailed discussion of B-splines. In the special case of $k = 0$, we set f_k to be normal. For $k = 1$, we set $B_11(x) = I[|x| < M]$, so that f_1 has a Huber-type density. M may be chosen beforehand or by some data based rule. We aim to choose k to minimize the MSE of $\hat{\theta}$. The method has three steps.

STEP 1. $a_{1k}, \dots, a_{kk}$ are estimated by maximum likelihood. $\underline{\hat{a}}$ satisfies ML equations

$$E_{\underline{\hat{a}}} B_{ik} = n^{-1} \sum_{j=1}^{n} B_{ik}(T_j), \qquad i = 1, \dots, k, \tag{2.3}$$

where $E_{\underline{\hat{a}}}$ is with respect to $f_k(\cdot, \underline{\hat{a}}) \equiv f_{n,k}$ and where $T_j = X_j - \tilde{\theta}$.

The equations were solved numerically using a modification of the Powell method; see Press, Flannery, Teukolsky and Vetterling (1988). Some care was required for larger k to ensure convergence, it sometimes being necessary to restart the algorithm from different starting values if convergence failed to occur.

STEP 2. Estimate the MSE of $\hat{\theta}(k)$ for given k. Since the asymptotic variance of $\hat{\theta}(k)$ is

$$M(k) = \frac{E_F(\log f_k)'^2}{[E_F(\log f_k)'']^2}, \tag{2.4}$$

where $f_k \equiv f_k(\cdot, \underline{a})$ and where $\underline{a}$ solves $E_{\underline{a}} \underline{B} = E_F \underline{B}$. It is natural to estimate this by its empirical version:

$$\hat{M}(k) = n^{-1} \sum_{j=1}^{n} \left[\sum_{i=1}^{k} \hat{a}_{ik} B'_{ik}(T_j) \right]^2 \bigg/ \left[\sum_{j=1}^{n} \sum_{i=1}^{k} \hat{a}_{ik} B''_{ik}(T_j) \right]^2. \tag{2.5}$$

STEP 3. Choose $\hat{\theta} = \hat{\theta}(\hat{k})$, where $\hat{k}$ minimizes $\hat{M}(k)$ over $k = 0, 1, \dots, K$.

We can offer no automatic method for selecting K, the largest number of knotpoints we are prepared to consider. $K \approx n$ does not work in practice, since we find that $\hat{M}(\hat{k})$ tends to be an underestimate for large k. Furthermore, the computational expense would be considerable. However, based on the empirical evidence, we can recommend a simple fixed choice of $K = 2$ for small sample sizes ($n \leq 400$). Jin (1990) offers a data-dependent method of selecting K, which gives similar results to this fixed choice of K.

Various schemes for positioning the knots seem reasonable, but we have used knot points at iM/k, $i = -k, \dots, 0, \dots, k$, and $M = (T_{[0.95]} - T_{[0.05]})/2$. An alternative might be to have all the knots spaced according to the percentiles of the data. In fact this was tried, but the results were very similar to those below. Furthermore, this method requires more work in computing the splines and so it was decided to stay with the above formulation. We use $\tilde{\theta} = \bar{X}$.

Note that, in any case, the choice of k is crucial, while the choice of K or the position of the knots is not.

Monte Carlo results. $X_1, \dots, X_n$ were generated from these distributions: normal, contaminated normal $0.9N(0, 1/9) + 0.1N(0, 9)$, double exponential, bimodal mixture of normals $0.5N(3, 1) + 0.5N(-3, 1)$, beta $(2, 2)$ and t with three degrees of freedom, using standard methods. All the distributions were standardized to have mean 0 and variance 1 for the purpose of comparison.

We compared our estimate with the mean, the median, the midmean, the 10% trimmed mean, the Pitman estimate and the one-step m.l.e. The Pitman estimate is the minimal risk invariant estimate of the parameter and is a suitable benchmark. It may be computed by calculating the mean of the normalized likelihood. The one-step m.l.e. is based on the true density which we are assuming is not available in practice. Since our estimate seeks to emulate the one-step m.l.e., it is a fair measure of the performance of our estimate.

Confidence bands for θ may be constructed. A $100(1 - \alpha)\%$ confidence interval is proposed

$$\hat{\theta}(\hat{k}) \pm \Phi^{-1}(\alpha/2)\left[\hat{M}(\hat{k})\right]^{1/2}, \tag{2.6}$$

where $\Phi(\cdot)$ is the standard normal distribution. Note that the method might also be used to make hypothesis tests concerning θ.

We calculate our adaptive estimate for $K = 2, 3, 4$ and $n = 20, 40, 400$. See Table 1.

Discussion of results. K is the largest number of knots considered in the selection of the adaptive estimate. No one choice seems uniformly superior for any of the sample sizes considered here. RMSE performance does not appear

TABLE 1
Location case[a]

			10%	Mid-	1-Step	Adaptive estimates			Confidence intervals		
Distribution	**Mean**	**Median**	**Trim**	**mean**	**m.l.e.**	$K = 2$	$K = 3$	$K = 4$	$K = 2$	$K = 3$	$K = 4$
Normal	1.00	1.33	1.03	1.11	1.00	1.03	1.05	1.05	92.2	88.5	82.0
	1.00	1.29	1.02	1.10	1.00	1.03	1.04	1.05	91.9	87.1	79.7
	1.00	1.27	1.02	1.10	1.00	1.01	1.03	1.03	95.3	94.6	93.7
Contaminated	2.66	1.32	1.21	1.11	1.63	1.20	1.36	1.45	92.5	83.9	78.6
Normal	2.72	1.28	1.24	1.11	1.29	1.17	1.27	1.47	94.1	87.7	77.0
	2.68	1.24	1.16	1.09	1.00	1.33	1.33	1.35	94.5	93.6	92.0
Double	1.25	1.15	1.10	1.01	1.03	1.18	1.16	1.16	90.2	84.1	73.9
Exponential	1.31	1.09	1.17	1.04	1.06	1.18	1.15	1.15	92.4	88.2	81.4
	1.39	1.01	1.26	1.08	1.02	1.16	1.08	1.07	95.1	94.6	93.9
Bimodal	3.06	8.53	3.63	5.69	1.72	2.06	2.03	2.11	92.5	89.9	83.5
Normal	3.07	10.8	3.48	5.51	1.34	1.70	1.60	1.52	93.9	92.6	90.1
	3.12	23.4	3.41	5.32	1.00	1.60	1.57	1.55	94.1	93.8	93.7
Beta(2, 2)	1.22	1.90	1.35	1.56	1.19	1.22	1.22	1.22	88.9	82.9	74.1
	1.30	1.98	1.42	1.66	1.29	1.32	1.31	1.31	91.1	87.5	81.9
	1.59	2.40	1.71	2.02	1.55	1.60	1.62	1.62	94.9	94.2	93.2
t with	1.39	1.19	1.04	1.01	1.20	1.15	1.17	1.20	90.1	83.6	73.4
3 df's	1.29	1.06	1.01	1.00	1.10	1.02	1.03	1.06	92.7	87.9	80.0
	1.39	1.10	1.07	1.00	1.07	1.01	1.00	1.01	95.2	94.7	93.7

[a] For each estimator the ratio of its estimated RMSE to that of the Pitman estimate is given. Sample sizes are 20, 40 and 400, respectively. The actual level of the nominally 95% confidence interval is given in the last three columns. 10,000 trials were made. Estimated error is 2–3% of the given values.

to be very sensitive to this choice, but the confidence intervals are more accurate with smaller K. Overall, $K = 2$ seems to be the best choice.

The adaptive estimates compare favorably with the mean, median and the trimmed means. In the case of the Beta(2, 2) and the bimodal normal distributions, the adaptive estimator does well where the usually robust kind of estimators fail. Although, of course, no uniform claims may be made, it does seem that the adaptive estimator has the best overall performance.

Our estimator was conceived as an approximation to the one-step m.l.e. Generally, the adaptive estimates (for appropriate K) are close to the one-step m.l.e.

For sample sizes $n = 20$ and 40, the actual confidence levels fall short of the nominal 95%, especially for larger K. Probability plots of the adaptive estimates show some deviation from normality in the tails which is a partial explanation. Also, the method used picks the smallest of the estimated RMSE's and this will cause some downward bias and hence the undersized confidence intervals. For sample size $n = 400$, performance is a lot better. For suitable sample sizes, the method may provide a powerful alternative to the standard tests.

The selection method was evaluated by ascertaining the percent of the samples in which $\hat{\theta}(\hat{k})$ actually minimized $|\theta(\hat{k}) - \theta|$ over k. In cases such as the contaminated or bimodal normal where choice of k is important to performance, the selection method exhibits some discriminatory power. In other cases such as the normal where choice of k makes little difference, the selection method shows no distinction from random choice.

3. The regression case. We follow an approach similar to before. We seek to estimate the slope parameter β of the model

$$Y_i = \alpha + \beta X_i + \varepsilon_i, \qquad i = 1, \dots, n,$$

where the ε_i are i.i.d. with mean zero with unknown density f. The X_i are generated according to $P[X_i = 1] = P[X_i = 0] = 1/2$.

We aim to minimize the MSE of $\hat{\beta}$. The method used is the obvious generalization of the method used in the location case.

Note that our interest here rests solely on estimation β. We could equally well center our interest on α. In a multiple regression, where one might be interested in estimating several parameters, it might be appropriate to minimize some function of the estimated covariance matrix.

The Monte Carlo results. Certain aspects of this setup follow Hsieh and Manski (1987) for the purposes of comparison. The distribution of the error F was chosen as for the location problem, except that we replace the double exponential by the lognormal.

We compare our estimate with least squares, least absolute deviations, a Huber-type estimate, Pitman estimate, one-step m.l.e. and Hsieh and Manski's estimate where possible ($n = 25, 50$). The Pitman estimate is gotten by computing the mean of the normalized likelihood function. The one-step m.l.e. uses the true density which our method assumes to be unknown. Some problems were encountered in computing this for the lognormal distribution because some of the residuals fall outside the support and it is not clear how this should be handled. For this reason we give the m.l.e. itself which is computed as a by-product of the Pitman estimate calculation. The initial estimate was least squares.

Confidence bands for β may be constructed. A $100(1 - \alpha)\%$ confidence interval is proposed:

$$\hat{\beta}(\hat{k}) \pm \Phi^{-1}(\alpha/2)\left[\hat{M}(\hat{k})\right]^{1/2}.$$

See Table 2.

Discussion of results. The adaptive estimates compare favorably with least squares, least absolute deviations and the Huber estimate. Performance varies, but $K = 2$ gives satisfactory overall performance and is cheapest to compute.

Again the adaptive estimates are comparable with the one-step m.l.e. Note that we had to give m.l.e. itself in the lognormal case so comparison is not reasonable here.

TABLE 2
Regression case[a]

				1-Step	Hsieh-	Adaptive estimates			Confidence intervals		
Distribution	$L2$	$L1$	**Huber**	**m.l.e.**	**Manski**	$K = 2$	$K = 3$	$K = 4$	$K = 2$	$K = 3$	$K = 4$
Normal	1.00	1.17	1.06	1.00	1.04	1.05	1.08	1.10	89.5	85.5	79.9
	1.00	1.16	1.05	1.00	1.09	1.03	1.05	1.06	92.1	89.6	86.1
	1.00	1.13	1.05	1.00	—	1.01	1.02	1.02	94.9	94.6	94.0
Contaminated	2.60	1.21	1.26	1.80	1.59	2.05	1.52	1.47	94.5	93.2	89.8
Normal	2.74	1.28	1.28	1.54	1.62	1.29	1.15	1.16	95.7	94.7	92.9
	2.74	1.39	1.29	1.10	—	1.27	1.27	1.28	94.5	93.7	93.6
t with	1.37	1.07	1.03	1.05	1.15	1.22	1.20	1.22	92.1	88.5	83.5
3 df's	1.37	1.09	1.04	1.10	1.13	1.15	1.13	1.14	94.3	92.1	88.9
	1.45	1.10	1.05	1.15	—	1.06	1.02	1.02	95.5	95.2	94.7
Bimodal	2.67	5.73	3.52	1.69	1.99	2.16	2.01	1.84	90.9	89.4	86.6
Normal	3.09	8.44	4.34	1.65	1.73	2.09	2.01	1.83	93.5	92.2	90.9
	3.08	14.8	4.59	1.20	—	1.58	1.55	1.53	94.6	94.0	93.6
Beta(2, 2)	1.15	1.52	1.34	1.17	1.19	1.14	1.14	1.12	88.3	85.0	80.8
	1.22	1.63	1.45	1.21	1.23	1.20	1.18	1.17	91.6	89.0	86.5
	1.51	1.95	1.80	1.50	—	1.50	1.48	1.45	94.6	94.0	93.3
Lognormal	3.09	2.03	1.70	1.06	2.02	2.17	1.96	1.96	93.5	91.3	86.4
	2.77	1.83	1.54	1.01	1.66	1.75	1.66	1.72	96.0	93.9	89.5
	2.12	1.47	1.20	1.01	—	1.04	1.03	1.04	97.3	94.0	90.5

[a] For each estimator the ratio of its estimated RMSE to that of the Pitman estimate is given. Sample sizes are 25, 50 and 400, respectively. The actual level of the nominally 95% confidence internal is given in the last three columns. 10,000 trials were made. Estimated error is 2–3% of the given values.

Results are roughly comparable to those of Hsieh and Manski. Their estimator is subject to a somewhat arbitrary choice of trimming parameters and ours to the choice of K. Since it is not clear how much the estimators have been optimized with respect to these choices (and to what extent it would be fair to do so), no closer comparison is reasonable. Comparison for larger sample sizes would be interesting. Our method probably requires less computation which might be important when dealing with larger samples.

4. Adaptiveness of the estimator.

We will show that the sequence of estimates $\hat{\theta}(k)$ as $k \to \infty$ are adaptive.

Let $X_1, X_2, \dots, X_n$ be i.i.d. $F(x - \theta)$ and have density f which is symmetric, but otherwise unknown. Without loss of generality we take θ to be zero. Let $k \equiv k_n$ be the number of knots and is a function of n and $M \equiv M_n$ be the largest knot point. Henceforth, we drop the subscript n on k and M. We state the following conditions.

CONDITIONS A. (i) There exist $n^{1/2}$-consistent estimates of θ such that $n^{1/2}(\tilde{\theta}_n - \theta) = O_p(1)$.

(ii) f has bounded Fisher information, $I(F) = \int f'(x)^2/f(x)\,dx < \infty$.
(iii) $\log f$ is three times continuous and differentiable and

$$(M/k)\|(\log f)'''\|_{L_\infty[0,M]} \to 0 \text{ as } n \to \infty.$$

(iv) $k/Mn^{1/2} \to 0$, $M/k \to 0$ and $M \to \infty$ as $n \to \infty$.

Condition A(iii) is somewhat artificial but not particularly restrictive. In order to carry through our proof, we find it convenient to use discretized estimates of θ. The construction is due to Le Cam (1969) and was employed by Bickel (1982). We take $\bar{\theta}_n$, a $n^{1/2}$-consistent estimate of θ. Let $R_n = \{n^{-1/2}i$, where i is an arbitrary integer$\}$ and then we define the discretized estimate $\tilde{\theta}_n$ as $\tilde{\theta}_n$ equals the point of in R_n closest to $\bar{\theta}_n$.

The estimator we will consider is $\hat{\theta}(k)$, where

$$\hat{\theta}(k) = \tilde{\theta} + \sum_{i=1}^{n} \log f'_{n,k}\left(X_i - \tilde{\theta}, \underline{\hat{a}}\right) \Big/ \sum_{i=1}^{n} \log f''_{n,k}\left(X_i - \tilde{\theta}, \underline{\hat{a}}\right). \tag{4.1}$$

THEOREM 4.1. *Under Conditions A, $\hat{\theta}(k)$ is adaptive, that is, for every regular θ, F,*

$$L_{\theta_n}\left\{n^{1/2}\left(\hat{\theta}(k_n) - \theta_n\right)\right\} \to N(0, I^{-1}(\theta, F)) \tag{4.2}$$

whenever $n^{1/2}|\theta_n - \theta|$ stays bounded.

This result has been obtained by previous authors for kernel density estimation, see Bickel (1982) for general results. Here we use a different method of density estimation. We are also able to avoid the artifice of sample splitting (half the data is used for density estimation, the other half is used for the estimate itself) thanks to a lemma due to Schick (1987). The proof of this theorem requires verification of the following two statements:

$$n^{-1/2}\left[\sum_{j=1}^{n} (\log f_{n,k})'(X_j - \theta_n) - \sum_{j=1}^{n} (\log f)'(X_j - \theta_n)\right] \to 0 \tag{S1}$$

whenever $n^{1/2}|\theta_n - \theta|$ is bounded and

$$n^{-1} \sum_{j=1}^{n} \log f''_{n,k}\left(X_j - \tilde{\theta}\right) \to I(\theta, F). \tag{S2}$$

We use a generalization of a lemma due to Schick to show (S1). See the Appendix for the proofs.

There exist some differences between the estimator considered in the theorem and the one actually used in the simulation studies.

1. M (the largest knot point) does not depend on the data, just on n.
2. k is a fixed sequence which depends only on n. Jin (1990) shows efficiency using $\hat{k}$. We have only a first order efficiency result here. It would be nice to show that the method of selecting k used in the simulation studies was optimal.
3. We set $B_{kk}(t) = 0$ for $|t| > M$ to simplify some proofs.

The result extends to regression in a straightforward manner.

5. Conclusion. Where the assumption of normality is considered inappropriate, the use of robust estimators is often proposed. However, these estimators are generally designed to tackle long-tailed deviations from normality. If the distribution is short-tailed or bimodal, say, they do not do so well. The adaptive estimator described here deserves consideration when the form of deviation from normality may not be known. More work on the implementation of this estimator might be desirable. Order of splines, knot placement and C.I.'s all deserve further investigation. On a theoretical level, results concerning the selection of k and K may be helpful.

APPENDIX

The proof of Theorem 4.1 depends on (S1) and (S2). (S1) may be verified using Schick's lemma which holds that the conditions sufficient to show that

$$n^{-1/2}\sum_{j=1}^{n}\left(\hat{h}_n(X_j) - h_n(X_j)\right) \to 0,$$

where $X_1, \dots, X_n$ are i.i.d. r.v.'s with distribution F, h_n is a measurable function satisfying $\int h_n\, dF = 0$ and $\int |h_n|^2\, dF < \infty$ and $\hat{h}_n$ is an estimate of h_n, that is, $\hat{h}_n = H_n(\cdot, X_1, \dots, X_n)$ for some measurable function H_n, are (dropping subscript n henceforth):

1. $n^{1/2}\int \hat{h}(x)\, dF(x) \to 0$ in probability,
2. $E\int |\hat{h}(x) - h(x)|^2\, dF(x) \to 0$,
3. $\Sigma_{j=1}^{n}(\hat{h}(X_j) - \bar{h}_j(X_j))^2 \to 0$ in probability,
4. $\Sigma_{j=1}^{n} E\int |\hat{h}(x) - \bar{h}_j(x)|^2\, dF(x) \to 0$,

where $\bar{h}_j = H(\cdot, X_1, \dots, X_{j-1}, X_{j+1}, \dots, X_n)$; see Schick (1987) for further details. Set

$$\hat{h}(x) = \sum_{i=1}^{k} \hat{a}_{ik} B'_{ik}(x - \theta_n),$$

$$h(x) = (\log f)'(x - \theta_n), \qquad \bar{h}_j(x) = \sum_{i=1}^{k} \hat{a}_{ij}^{-j} B'_{ik}(x - \theta_n).$$

Lemmas 6.1–6.5 are sufficient to verify these conditions. Write

$$\log f_n = -\sum_{i=1}^{k} a_{ik} B_{ik}(T_j) - c(\underline{a}), \qquad \log \hat{f}_n = -\sum_{i=1}^{k} \hat{a}_{ik} B_{ik}(T_j) - c(\hat{\underline{a}}),$$

$$T_j = X_j - \tilde{\theta}.$$

ML equations give

$$E_{\underline{a}}\underline{B} - E_{\hat{\underline{a}}}\underline{B} = E_F\underline{B} - n^{-1}\sum_{j=1}^{n}\underline{B}(T_j),$$

which may be expressed as

$$C(\underline{a} - \hat{\underline{a}}) + \text{rem} = E_F\underline{B} - n^{-1}\sum_{j=1}^{n}\underline{B}(T_j)$$

for matrix C. Lemma 6.1 investigates the LHS of this equation and Lemma 6.2 the RHS. From this we determine the closeness of $\hat{\underline{a}}$ to $\underline{a}$ and hence of $\hat{f}_n$ to f_n. All the following lemmas assume Conditions A.

LEMMA 6.1.

$$\sup_{i,k}\left|n^{-1}\sum_{j=1}^{n}B_{ik}(T_j) - E_F B_{ik}\right| = O_p(n^{-1/2}).$$

PROOF. Consider the stochastic processes

$$\Xi_{n\alpha}(s) = n^{1/2}\left(n^{-1}\sum_{j=1}^{n}T_j^{\alpha}I(T_i \le s) - \int_0^s t^{\alpha}f(t)\,dt\right)$$

for $\alpha = 0, 1, 2$. So $\Xi_{n0}(s) \Rightarrow_{\mathscr{D}} \Xi_0(s) = U(F(s))$, where $U(\cdot)$ is the Brownian bridge. $n^{-1}\sum_{j=1}^{n}B_{ik}(T_j) - E_F B_{ik}$ may be expressed in terms of $\Xi_{n0}, \Xi_{n1}, \Xi_{n2}$. Now given $F(0) = 0$,

$$\begin{aligned}\int_0^s \Xi_{n0}(u)\,du &= n^{1/2}\left(n^{-1}\sum_{j=1}^{n}(s - T_j)I(T_j \le s) - \int_0^s\left[\int_0^u f(t)\,dt\right]du\right)\\ &= n^{1/2}\left(s\left(n^{-1}\sum_{j=1}^{n}I(T_j \le s) - \int_0^s f(t)\,dt\right) - n^{-1}\sum_{j=1}^{n}T_jI(T_j \le s)\right.\\ &\qquad\left. + \int_0^s tf(t)\,dt + s\int_0^s f(t)\,dt - \int_0^s tf(t)\,dt - \int_0^s\left[\int_0^t f(u)\,du\right]dt\right)\end{aligned}$$

but by integrating $\int_0^s[\int_0^t f(u)\,du]\,dt$ by parts we see that the last three terms add up to zero.

Hence

$$\int_0^s \Xi_{n0}(u)\,du = s\Xi_{n0}(s) - \Xi_{n1}(s).$$

So

$$\begin{aligned}\sup_s|\Xi_{1n}(s)| &\le \sup_s\left|\int_0^s \Xi_{n0}(u)\,du\right| + \sup_s|s\Xi_{n0}(s)|\\ &\le 2s\sup_s|\Xi_{n0}(s)| = sO_p(1).\end{aligned}$$

Similarly, $\int_0^s \Xi_{n1}(u)\,du = s\Xi_{n1}(s) - \Xi_{n0}(s)$, hence $\sup_s|\Xi_{2n}(s)| = s^2O_p(1)$.

Applying this to the quadratic B-splines we get

$$\sup_{i,k}\left|n^{-1}\sum_{j=1}^{n}B_{ik}(T_j)-E_F B_{ik}\right|=O_p(n^{-1/2}).$$ □

LEMMA 6.2.

$$\|\underline{\hat{a}}-\underline{a}\|_2=O_p(k/Mn^{1/2}).$$

PROOF.

$$\begin{aligned}E_{\underline{a}}B_{ik}&-E_{\underline{\hat{a}}}B_{ik}\\&=\int B_{ik}\left[\exp\left[-\sum_{j=1}^{k}a_{jk}B_{jk}(x)-c(\underline{a})\right]\right.\\&\qquad\left.-\exp\left[-\sum_{j=1}^{k}\hat{a}_{jk}B_{jk}(x)-c(\underline{\hat{a}})\right]\right]dx\\&=\int B_{ik}\left[-\sum_{j=1}^{k}(a_{jk}-\hat{a}_{jk})B_{jk}(x)-[c(\underline{a})-c(\underline{\hat{a}})]\right]dx+R_k\\&=-\sum_{j=1}^{k}(a_{jk}-\hat{a}_{jk})\int B_{ik}(x)B_{jk}(x)\,dx-(M/k)[c(\underline{a})-c(\underline{\hat{a}})]+R_k.\end{aligned}$$

Now both R_k and $(M/k)[c(\underline{a})-c(\underline{\hat{a}})]$ may be shown to be $o_p(n^{-1/2})$ and $(k/M)\int B_{ik}(x)B_{jk}(x)\,dx$ may be evaluated explicitly. It is a banded matrix, C, where the entries are zero except close to the diagonal and its eigenvalues are greater than or equal to 0.1 by Gerčgorin's theorem. [Gerčgorin's theorem states that the eigenvalues of a symmetric matrix are contained within discs centered on the diagonal entries with radii the sum of the absolute values of the off-diagonal elements of the corresponding rows or columns; see Wilkinson (1965).]

So by this and Lemma 6.1, the lemma is proven. □

LEMMA 6.3. *Let $\hat{a}_{ik}^{-l}$ be the estimate of a_{ik} when the lth observation is omitted.*

$$\sup_{l}\left\|\hat{a}_{ik}^{-l}-\hat{a}_{ik}\right\|_2=O_p(k/Mn).$$

PROOF. $\underline{\hat{a}}^{-l}$ is the solution to

$$(n-1)^{-1}\sum_{i\neq l}^{n}\underline{B}(t_i)=E_{\underline{a}}\underline{B},\qquad l=1,\dots,k.$$

Now

$$\begin{aligned}|E_{\underline{\hat{a}}}\underline{B} - E_{\underline{\hat{a}}^{-l}}\underline{B}| &= \left| n^{-1}\sum_{i=1}^{n}\underline{B}(T_j) - (n-1)^{-1}\sum_{i\neq l}^{n}\underline{B}(t_i)\right| \\ &= \left| -(n(n-1))^{-1}\sum_{i\neq l}^{n}\underline{B}(t_i) + \underline{B}\big(X_l - \tilde{\theta}\big)\Big/ n\right| \\ &\leq (n-1)^{-1} + n^{-1}.\end{aligned}$$

The result follows as in Lemma 6.2. □

Lemma 6.4. *Under Conditions A*

$$\|(\log f_n)'' - (\log f)''\|_{L_\infty[0,M]} \to 0 \quad \textit{as } n \to \infty.$$

Proof. Let $\log f_c(t) = \sum_{i=1}^{k}\log f(iM/k)B_{ik}(t)$. We show that $(\log f_c)''$ is close to $(\log f)''$ and then that $(\log f_c)''$ is close to $(\log f_n)''$.

First we show that

$$\|(\log f_c)'' - (\log f)''\|_{L_\infty[0,M]} \to 0 \quad \text{as } n \to \infty.$$

Suppose $t^* \in [(i-1)M/k, iM/k]$, then

$$\begin{aligned}(\log f_c)''(t^*) = [M/k]^{-2}(\log f((i-2)M/k)& \\ -2\log f((i-1)M/k) + \log f(iM/k)&).\end{aligned}$$

We expand twice to get

$$\begin{aligned}(\log f_c)''(t^*) = (\log f)''(iM/k) &+ [M/2k](\log f)'''(\xi_4) \\ &+ [M/k](\log f)'''(\xi_3) + [M/6k](\log f)'''(\xi_1) \\ &- [M/6k](\log f)'''(\xi_2),\end{aligned}$$

where $\xi_1 \in [(i-2)M/k, (i-1)M/k]$, $\xi_2 \in [(i-1)M/k, iM/k]$, $\xi_3, \xi_4 \in [(i-1)M/k, iM/k]$.

So

$$|(\log f_c)''(t^*) - (\log f)''(iM/k)| = O(M/k)\|(\log f)'''\|_{L_\infty[0,M]}$$

which by assumption approaches 0.

Now we show that

$$\|(\log f_c)'' - (\log f_n)''\|_{L_\infty[0,M]} \to 0 \quad \text{as } n \to \infty.$$

This will be true provided that $\max_{i,k}|a_{ik} - (\log f)(iM/k)| \to 0$.

From the maximum likelihood equations we know that

$$E_F\underline{B} - E_{f_c}\underline{B} = E_a\underline{B} + E_{f_c}\underline{B} = \int \underline{B}(t)(f_c - f)\,dt.$$

From de Boor [(1978), pages 167–170] we know that

$$\|(\log f_c) - (\log f)\|_{L_\infty[0,M]} \leq (M/k)^3\|(\log f)'''\|_{L_\infty[0,M]} = O(M/k)^2$$

and so $\|f_c - f\|$ is $O(M/k)^2$.

So

$$\int B_{ik}(t)(f_c - f)\,dt \le (M/k)\|f_c - f\| = O(M/k)^3.$$

Now by expanding $E_a\underline{B} - E_{f_c}\underline{B}$ as in Lemma 6.2, we obtain that $\max_{i,k}|a_{ik} - (\log f)(iM/k)| \to 0$.

Hence the lemma is proved. □

Lemma 6.5.

$$\|(\log f)' - (\log f_n)'\|_{L_\infty[0,M]} = O(M/k).$$

Proof. The ML equations are $\int B_{ik}(t)(f_n - f)\,dt = 0$ for $i = 1, \dots, k$. Since $B_{ik}(t)$ have bounded support we know that there is a zero of $f_n - f$ in each interval $[(i-3)M/k, iM/k]$. This is also a zero of $\log f_n - \log f$. By Rolle's theorem, we know that there is zero of $(\log f_n') - (\log f)'$ between each of these zeros so

$$\begin{aligned}
&\|(\log f)' - (\log f_n)'\|_{L_\infty[0,M]}\\
&\quad\le (\text{max. dist. between zeros}) \times (\text{max. deriv. of} (\log f') - (\log f_n)')\\
&\quad\le 6(M/k)\|(\log f)'' - (\log f_n)''\|_{L_\infty[0,M]} = O(M/k).
\end{aligned}$$

□

Now the conditions of Schick's lemma are easily verified. (i) is true by symmetry and (ii) follows by application of Lemmas 6.2 and 6.5. (iii) and (iv) hold by observing that

$$\begin{aligned}
\left(\hat{h}(x) - \bar{h}_j(x)\right)^2 &= \left[\sum_{i=1}^{k} \left(\hat{a}_{ik} - \hat{a}_{ik}^{-j}\right) B'_{ik}(x - \theta_n)\right]^2\\
&= \sum_{i=1}^{k} \left[\left(\hat{a}_{ik} - \hat{a}_{ik}^{-j}\right)^2 - 2\left(\hat{a}_{ik} - \hat{a}_{ik}^{-j}\right)\left(\hat{a}_{i,k-1} - \hat{a}_{i,k-1}^{-j}\right)\right]\\
&\qquad \times B'^2_{ik}(x - \theta_n)\\
&\le 2\sum_{i=1}^{k} \left(\hat{a}_{ik} - \hat{a}_{ik}^{-j}\right)^2 B'^2_{ik}(x - \theta_n).
\end{aligned}$$

The equality follows from $B'_{ik}B'_{i,k-1} = -B'^2_{ik}$ and the inequality by application of Cauchy–Schwarz. The result then follows by application of Lemma 6.3. □

Lemma 6.6.

$$n^{-1}\sum_{i=1}^{n} \log \hat{f}''_{n,k}\left(X_i - \tilde{\theta}\right) - E_F\left((\log f)'\right)^2 \to 0 \quad \textit{in probability}.$$

Proof. Result follows from Lemma 6.4. □

Proof of Theorem 4.1. The two conditions (S1 and S2) have been met. S1 by the verification of Schick's lemma and S2 by Lemma 6.4. □

Acknowledgments. This work is based on the author's Ph.D. thesis written under the supervision of Professor P. J. Bickel, whose help is gratefully acknowledged. Thanks are also due to an Associate Editor and two referees for some helpful comments.

REFERENCES

BICKEL, P. J. (1982). On adaptive estimation. *Ann. Statist.* **10** 647–671.

DE BOOR, C. (1978). *A Practical Guide to Splines*. Springer, New York.

HSIEH, D and MANSKI, C. (1987). Monte-Carlo evidence on adaptive maximum likelihood estimation of a regression. *Ann. Statist.* **15** 541–551.

JIN, K. (1990). Empirical smoothing parameter selection in adaptive estimation. Ph.D. dissertation, Dept. Statistics, Univ. California, Berkeley.

LE CAM, L. (1969). *Théorie Asymptotique de la Décision Statistique*. Les Presses de l'Université de Montréal.

PRESS, W. H., FLANNERY, B. P., TEUKOLSKY, S. A. and VETTERLING, W. T. (1988). *Numerical Recipes: The Art of Scientific Computing*. Cambridge Univ. Press.

SCHICK, A. (1987). A note on the construction of asymptotically linear estimates. *J. Statist. Plann. Inference* **16** 89–105.

STEIN, C. (1956). Efficient non-parametric testing and estimation. *Proc. Third Berkeley Symp. Math. Statist. Probab.* **1** 187–196. Univ. California Press, Berkeley.

STONE, C. J. (1975). Adaptive maximum likelihood estimators of a location parameter. *Ann. Statist.* **3** 267–284.

WILKINSON, J. (1965). *The Algebraic Eigenvalue Problem*. Oxford Univ. Press, New York.

DEPARTMENT OF STATISTICS
UNIVERSITY OF MICHIGAN
419 SOUTH STATE STREET
ANN ARBOR, MICHIGAN 48109-1027