A NONPARAMETRIC FRAMEWORK FOR STATISTICAL MODELLING[1]

by Charles J. Stone

Technical Report No. 73
August 1986

Research supported in part by National Science Foundation Grant DMS-8600409.

Department of Statistics
University of California
Berkeley, California

A NONPARAMETRIC FRAMEWORK FOR STATISTICAL MODELLING[1]

by Charles J. Stone

University of California, Berkeley

1. Introduction. Much of mathematical statistics deals with inference concerning the unknowns in a stochastic model for a random phenomenon. In the parametric approach the unknowns are a specific finite number of real parameters. In the nonparametric approach they are functions, perhaps subject to smoothness or other regularity conditions. These functions can be approximated by means of a flexible finite-dimensional function space. To some extent, this reduces the nonparametric approach to the parametric approach. But the asymptotic theory is different when the error of approximation is taken into account and the dimension of the approximating function space is allowed to tend to infinity with the sample size. We will illustrate this theory by means of three examples: density estimation, logistic regression, and additive logistic regression.

[1] This research was supported in part by National Science Foundation Grant DMS-8600409.

2. Density estimation. Let Y be a d-dimensional random variable taking on values from a known compact subset C of $\mathbb{R}^d$. It is assumed that the distribution of Y has a density function f which is continuous and positive on C. By definition $\int f = 1$. Set $g = \log f$. Let $\mathcal{G}_n$ denote a p_n-dimensional vector space of functions on C having basis B_{nj}, $1 \le j \le p_n$. It is assumed that $\Sigma_j B_{nj} = 1$ on C and that no nontrivial linear combination of B_{nj}, $1 \le j \le p_n$, is almost everywhere equal to zero on C. Given $\boldsymbol{\theta} \in \Theta_n$, the space of p_n-dimensional vectors, set

$$C_n(\boldsymbol{\theta}) = \log(\int \exp(\Sigma_j \theta_j B_{nj}))$$

and

$$f_n(\cdot;\ \boldsymbol{\theta}) = \exp(\Sigma_j \theta_j B_{nj} - C_n(\boldsymbol{\theta})).$$

Then $\int f_n(\cdot;\ \boldsymbol{\theta}) = 1$ for $\boldsymbol{\theta} \in \Theta_n$. Observe that $f_n(\cdot;\ \boldsymbol{\theta})$, $\boldsymbol{\theta} \in \Theta_n$, is an exponential family in canonical form. Let Θ_{n0} denote the (p_n-1)-dimensional space consisting of those $\boldsymbol{\theta} \in \Theta_n$ the sum of whose elements is zero. Let $\boldsymbol{\theta}_n$ denote the unique value of $\boldsymbol{\theta} \in \Theta_{n0}$ that maximizes the expected log-likelihood function $\Lambda_n(\cdot)$, defined by

$$\Lambda_n(\boldsymbol{\theta}) = E[\Sigma_j \theta_j B_{nj}(Y) - C_n(\boldsymbol{\theta})] = \Sigma_j \theta_j \int B_{nj} f - C_n(\boldsymbol{\theta}), \quad \boldsymbol{\theta} \in \Theta_{n0}.$$

Consider the *loglinear density approximation* $f_n = f_n(\cdot;\ \boldsymbol{\theta}_n)$ to f.

Let $Y_1, \cdots, Y_n$ be independent random variables each having density f. The log-likelihood function for the parametric model is given by

$$\ell_n(\boldsymbol{\theta}) = \Sigma_j \theta_j \Sigma_i B_{nj}(Y_i) - nC_n(\boldsymbol{\theta}).$$

The maximum likelihood estimator (MLE) $\hat{\boldsymbol{\theta}}_n$ is the value of $\boldsymbol{\theta} \in \Theta_{n0}$ that maximizes $\ell_n(\cdot)$. Since $\ell_n(\cdot)$ is a strictly concave function on Θ_{n0}, the MLE is unique if it exists. The corresponding estimate $\hat{f}_n = f_n(\cdot;\ \hat{\boldsymbol{\theta}}_n)$ of f is called a *loglinear density estimate*, since $\log \hat{f}_n \in \mathcal{G}_n$.

Let $\| \|_2$ and $\| \|_\infty$ denote the usual L_2 and L_∞ norms of functions on C. Let $\| \|_{\infty,A}$ denote the L_∞ norm of functions on A. It is assumed that $p_n \to \infty$ as $n \to \infty$, that $p_n = o(n^{.5-\epsilon})$ for some $\epsilon > 0$, and that

$$\lim_{n\to\infty} \inf_{s\in\mathcal{S}_n} \|s-h\|_\infty = 0 \quad \text{if } h \text{ is continuous on } C.$$

Let $\Pi_n h$ denote the orthogonal projection of h onto $\mathcal{S}_n$ with respect to $L_2(C)$. It is assumed that there is a positive constant M such that

$\|\Pi_n h\|_\infty \le M\|h\|_\infty$ for $n \ge 1$ and $h \in L_\infty(C)$;

$|B_{nj}| \le M$ on C for $n \ge 1$ and $1 \le j \le p_n$;

for $1 \le j \le p_n$, $B_{nj} = 0$ outside a set C_{nj} having diameter at most $Mp_n^{-1/d}$ and $C_{nj} \cap C_{nk}$ nonempty for at most M values of k;

and

$$M^{-1}|\theta_j|^2 \le \|\Sigma_k\theta_k B_{nk}\|^2_{\infty,C_{nj}} \le Mp_n\int_{C_{nj}}(\Sigma_k\theta_k B_{nk})^2 \quad \text{for } n \ge 1,\ \theta \in \Theta_n, \text{ and } 1 \le j \le p_n.$$

These properties can be satisfied with C_{nj}, $1 \le j \le p_n$, a partition of C and B_{nj} the indicator function for C_{nj}. Here $\hat{f}_n$ is the corresponding histogram density estimate. They can also be satisfied with $d = 1$, $\mathcal{S}_n$ a space of splines and B_{nj}, $1 \le j \le p_n$, a basis consisting of B-splines; see de Boor (1976, 1978) or Stone (1985, 1986a, 1986b) for details. Presumably, they can also be satisfied with tensor product spaces of splines and with spaces of the type that arise in the use of the finite element method (see arguments in Descloux, 1972, and de Boor, 1976).

Some asymptotic properties of loglinear density estimation will now be summarized. The proofs follow from arguments in Stone (1986b). Set

$$\delta_n = \inf_{s\in\mathcal{S}_n} \|s - g\|_\infty.$$

THEOREM 1. (i) $\|f_n - f\|_\infty = O(\delta_n)$;

(ii) $\hat{f}_n$ *exists except on an event whose probability tends to zero with* n;

(iii) $\|\hat{f}_n - f_n\|_2 = O_{pr}((n^{-1}p_n)^{1/2})$;

and

(iv) $\|\hat{f}_n - f_n\|_\infty = O_{pr}((n^{-1}p_n\log(p_n))^{1/2})$.

Write

$$\hat{f}_n - f = f_n - f + \hat{f}_n - f_n.$$

The quantity $f_n - f$ is a "bias" term, while $\hat{f}_n - f_n$ is a "noise" term whose magnitude is indicated by its asymptotic variance. Under typical smoothness assumptions on g, $\delta_n = O(p_n^{-q/d})$ for some positive number q (this holds with $q = m$ if g has a bounded mth derivative). Set $\gamma = 1/(2q+d)$ and $r = q\gamma$. Suppose that $\gamma d < 1/2$ or, equivalently, that $q > d/2$. To get the optimal rate of convergence of $\|\hat{f}_n - f\|_2$ to zero, choose $p_n \sim n^{\gamma d}$. Then $\delta_n^2 \sim n^{-1}p_n \sim n^{-2r}$ and hence

$$\|\hat{f}_n - f\|_2 = O_{pr}(n^{-r}).$$

To get the optimal rate of convergence of $\|\hat{f}_n - f\|_\infty$ to zero choose $p_n \sim (n/\log(n))^{\gamma d}$. Then $\delta_n^2 \sim n^{-1}p_n\log(p_n) \sim (n^{-1}\log(n))^{2r}$ and hence

$$\|\hat{f}_n - f\|_\infty = O_{pr}((n^{-1}\log(n))^{r}).$$

(See Stone, 1982, for a precise definition of *optimal rate of convergence*.)

Let $\mathcal{I}_n(\cdot)$ denote the information function based on the random sample of size n. Then $\mathcal{I}_n(\boldsymbol{\theta})$ is the Hessian matrix of $nC_n(\cdot)$ at $\boldsymbol{\theta}$; that is, the $p_n \times p_n$ matrix whose (j, k)th element is

$$n\frac{\partial^2 C_n(\boldsymbol{\theta})}{\partial\theta_j\partial\theta_k}.$$

Let $\mathcal{I}_n^{-1}(\boldsymbol{\theta})$ denote the inverse to $\mathcal{I}_n(\boldsymbol{\theta})$ viewed as a linear transformation of Θ_{n0}. Set $\mathcal{I}_n^{-1} = \mathcal{I}_n^{-1}(\boldsymbol{\theta}_n)$ and $\hat{\mathcal{I}}_n^{-1} = \mathcal{I}_n^{-1}(\hat{\boldsymbol{\theta}}_n)$. Let $G_n(y), \hat{G}_n(y) \in \Theta_{n0}$

denote the p_n-dimensional vectors having elements

$$G_{nj}(y) = B_{nj}(y) - \frac{\partial C_n}{\partial \theta_j}(\boldsymbol{\theta}_n)$$

and

$$\hat{G}_{nj}(y) = B_{nj}(y) - \frac{\partial C_n}{\partial \theta_j}(\hat{\boldsymbol{\theta}}_n).$$

respectively. Set

$$SE(\hat{f}_n(y)) = f_n(y)(\mathbf{G}_n(y)'\boldsymbol{\mathcal{I}}_n^{-1}\mathbf{G}_n(y))^{1/2}$$

and

$$\hat{S}E(\hat{f}_n(y)) = \hat{f}_n(y)(\hat{\mathbf{G}}_n(y)'\hat{\boldsymbol{\mathcal{I}}}_n^{-1}\hat{\mathbf{G}}_n(y))^{1/2}.$$

THEOREM 2. *Uniformly in* $y \in I$,

$$SE(\hat{f}_n(y)) \sim (n^{-1}p_n)^{1/2},$$

$$\frac{\hat{S}E(\hat{f}_n(y))}{SE(\hat{f}_n(y))} = 1 + o_{pr}(1),$$

and

$$\mathcal{L}\left\{\frac{\hat{f}_n(y) - f_n(y)}{SE(\hat{f}_n(y))}\right\} \to \mathcal{N}(0, 1).$$

It follows from Theorem 2 that $\hat{f}_n(y) \pm z_{1-.5\alpha}\hat{S}E(\hat{f}_n(y))$ is an asymptotic $(1-\alpha)$-level confidence interval for $f_n(y)$; if $\delta_n = o((n^{-1}p_n)^{1/2})$, it is also an asymptotic $(1-\alpha)$-level confidence interval for $f(y)$. Here $\Phi(z_q) = q$, Φ being the standard normal distribution function.

Let P denote the distribution corresponding to f, defined by $P(A) = \int_A f$, and let P_n and $\hat{P}_n$ be defined similarly in terms of f_n and $\hat{f}_n$. Let $\mathcal{A}$ denote a class of subsets of C. Given distributions Q_1 and Q_2 on C set

$$\|Q_1 - Q_2\|_\infty = \sup_{A \in \mathcal{A}} |Q_1(A) - Q_2(A)|.$$

Under reasonable conditions on $\mathcal{A}$, $\mathcal{S}_n$ and f

$$\|P_n - P\|_\infty = O(p_n^{-1/d}\delta_n)$$

and

$$\mathcal{L}\left[\frac{\hat{P}_n(A)-P_n(A)}{SE(\hat{P}_n(A))}\right] \to \mathcal{N}(0, 1) \qquad \text{with} \quad SE(\hat{P}_n(A)) = (P(A)(1-P(A))/n)^{1/2}.$$

It was shown in Stone (1986b) for the special case of $d = 1$, bases consisting of B-splines, and $\mathcal{A}$ the collection of subintervals of a compact interval C that

$$\|\hat{P}_n - P_n\|_\infty = 0_{pr}(n^{-1/2}).$$

What is a corresponding result in the more general context of the present paper?

3. Logistic regression. Let X, Y be a pair of random variables such that X ranges over a known compact subset C of $\mathbb{R}^d$ and Y takes on only two values, 0 and 1. It is assumed that the distribution of X is absolutely continuous and that its density is bounded away from zero and infinity on C. Let f be the regression function, defined on C by

$$f(x) = \Pr(Y = 1|X = x).$$

It is assumed that f is continuous and that $0 < f < 1$. Let g denote the corresponding logistic regression function, defined by $g = \operatorname{logit}(f) = \log(f/(1-f))$; so that $f = \exp(g)/(1+\exp(g))$.

We can approximate g by a member of a p_n-dimensional vector space $\mathcal{G}_n$ of functions on C. Let B_{nj}, $1 \le j \le p_n$, denote a basis of $\mathcal{G}_n$. Then the expected log-likelihood function $\Lambda_n(\cdot)$ is defined by

$$\Lambda_n(\boldsymbol{\theta}) = E[\Sigma_j \theta_j B_{nj}(X)Y - \log(1+\exp(\Sigma_j \theta_j B_{nj}(X)))], \quad \boldsymbol{\theta} \in \Theta_n.$$

Let $\boldsymbol{\theta}_n$ be the unique $\boldsymbol{\theta} \in \Theta_n$ that maximizes $\Lambda_n(\boldsymbol{\theta})$ and set

$$g_n = \Sigma_j \theta_{nj} B_{nj} \qquad \text{and} \qquad f_n = \exp(g_n)/(1+\exp(g_n)).$$

Let $(X_1, Y_1), \cdots, (X_n, Y_n)$ be independent random pairs, each having the same distribution as (X, Y). The log-likelihood function for the parametric model is given by

$$\ell_n(\boldsymbol{\theta}) = \Sigma_j \theta_j \Sigma_i Y_i B_{nj}(X_i) - \Sigma_i \log(1+\exp(\Sigma_j \theta_j B_{nj}(X_i)),$$

which corresponds to an exponential family in canonical form. Let $\hat{\boldsymbol{\theta}}_n$ denote the MLE of $\boldsymbol{\theta}$ and set $\hat{g}_n = \Sigma_j \hat{\theta}_{nj} B_{nj}$ and $\hat{f}_n = \exp(\hat{g}_n)/(1+\exp(\hat{g}_n))$. Under appropriate regularity conditions, analogs of Theorems 1 and 2 of Section 2 should hold.

4. Additive logistic regression. Let C be a rectangle, say, $C = [0, 1]^d$. It is then useful in practice to assume that g is additive or, more generally, to replace g by its best additive approximation g^*; this is defined to be the unique additive function h on C that maximizes the expected log-likelihood

$$E[h(X)Y-\log(1+e^{h(X)})].$$

Set $f^* = \exp(g^*)/(1+\exp(g^*))$. If g itself is additive, then $g^* = g$ and $f^* = f$.

To obtain a p_n-dimensional space of additive approximations to g^*, we consider p_{nk}-dimensional vector spaces $\mathcal{S}_{nk}$ of functions on $[0, 1]$ for $1 \le k \le d$, each containing the constant functions, and let $\mathcal{S}_n$ be the collection of all functions of the form

$$s(x_1,\cdots,x_d) = \Sigma_k s_k(x_k), \quad \text{where} \quad s_k \in \mathcal{S}_{nk} \quad \text{for} \quad 1 \le k \le d.$$

Then

$$p_n = 1 + \Sigma_k(p_{nk}-1).$$

Analogs of Theorem 1 of Section 2 and its consequences for optimal rates of convergence should hold with f replaced by f^*, g replaced by g^*, and d replaced by 1; see Stone (1985, 1986a) for what has been rigorously verified to date. An analog to Theorem 2 should also hold if g itself is additive. Otherwise, a more complicated standard error formula would be required since $\Pr(Y = 1| X = x)$ would not be exactly equal to $f^*(x)$.

REFERENCES

de BOOR, C. (1976). A bound on the L_∞-norm of the L_2-approximation by splines in terms of a global mesh ratio. *Math. Comp.* **30** 765-771.

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

DESCLOUX, J. (1972). On finite element matrices. *SIAM J. Numer. Anal.* **9** 260-265.

STONE, C. J. (1980). Optimal rates of convergence for nonparametric estimators. *Ann. Statist.* **8** 1348-1360.

STONE, C. J. (1982). Optimal global rates of convergence for nonparametric regression. *Ann. Statist.* **10** 1040-1053.

STONE, C. J. (1983). Optimal uniform rate of convergence for nonparametric estimators of a density function or its derivatives. Pp 393-406 in *Recent Advances in Statistics: Papers in Honor of Herman Chernoff on his Sixtieth Birthday*, edited by M. H. Rezvi, J. S. Rustagi, and D. Siegmund. Academic Press, New York.

STONE, C. J. (1985). Additive regression and other nonparametric models. *Ann. Statist.* **13** 689-705.

STONE, C. J. (1986a). The dimensionality reduction principle for generalized additive models. *Ann. Statist.* **14** 590-606.

STONE, C. J. (1986b). Asymptotic properties of logspline density estimation, Technical Report No. 69, Department of Statistics, Univ. of California, Berkeley.