Multivariate Log-Spline Models

By

Charles J. Stone

Technical Report No. 319
August 1991

*This Research supported in part by NSF grant DMS-9100723.

Department of Statistics
University of California
Berkeley, California 94720

MULTIVARIATE LOG-SPLINE MODELS[1]

BY CHARLES J. STONE

University of California, Berkeley

August 21, 1991

Let f be a density function on $\mathcal{Y} = [0, 1]^N$ such that $\varphi = \log f$ is bounded on $\mathcal{Y}$. Consider the normalized approximation φ^* to φ having the form of a specified sum of functions of at most d of the variables and, subject to this form, chosen to maximize expected log-likelihood, and let p be a suitably defined lower bound to the smoothness of φ^*. Consider a random sample of size n from f. Maximum likelihood and sums of products of polynomial splines are used to construct estimates of φ^* and its components having the optimal L_2 rate of convergence $n^{-p/(2p+d)}$.

[1] This research was supported in part by National Science Foundation Grant DMS–9100723.

AMS 1991 *subject classifications*. Primary 62G07; secondary 62G20.

Key words and phrases. Log-linear model, interactions, polynomial splines, exponential family, maximum likelihood method, optimal rate of convergence.

1. Introduction. In this paper we will consider models for continuous data that are analogous to log-linear models for discrete (categorical) data. In order to motivate the form of our models, we first consider discrete random variables $Y_1, \ldots, Y_N$ that range over finite sets $\mathcal{Y}_1, \ldots, \mathcal{Y}_N$ respectively. Let f denote the joint probability function of these random variables or, equivalently, the probability function of $\mathbf{Y} = (Y_1, \ldots, Y_N)$, suppose that f is positive on $\mathcal{Y} = \mathcal{Y}_1 \times \cdots \times \mathcal{Y}_N$, and set $\varphi = \log f$. Then we can write

(1) $$\varphi(\mathbf{y}) = \varphi_0 + \sum_j \varphi_j(y_j) + \sum_{j<k}\sum \varphi_{jk}(y_j, y_k) + \sum_{j<k<l}\sum\sum \varphi_{jkl}(y_j, y_k, y_l) + \cdots.$$

The right side of (1) is referred to as the saturated log-linear model for θ or as its ANOVA decomposition. In order to obtain a unique such decomposition, suitable constraints have to be imposed on the main effects φ_j, the two-factor interactions φ_{jk}, the three factor interactions φ_{jkl}, and the other nonconstant components of φ.

In practice, unsaturated submodels of (1) are commonly employed. Let d be the maximum number of variables that are allowed in any one component of the model. When $d = 1$, we get that

(2) $$\varphi(\mathbf{y}) = \varphi_0 + \sum_j \varphi_j(y_j),$$

which corresponds to the assumption that $Y_1, \ldots, Y_N$ are independent; when $d = 2$, we get that

(3) $$\varphi(\mathbf{y}) = \varphi_0 + \sum_j \varphi_j(y_j) + \sum_{j<k}\sum \varphi_{jk}(y_j, y_k).$$

Given a random sample of size n from the distribution of $\mathbf{Y}$, we can use finite parameter maximum likelihood to come up with the estimate $\hat{\varphi}$ of φ given by

(4) $$\hat{\varphi}(\mathbf{y}) = \hat{\varphi}_0 + \sum_j \hat{\varphi}_j(y_j) + \sum_{j<k}\sum \hat{\varphi}_{jk}(y_j, y_k) + \sum_{j<k<l}\sum\sum \hat{\varphi}_{jkl}(y_j, y_k, y_l) + \cdots.$$

Again, in order to obtain a unique such ANOVA decomposition, we need to impose constraints on the various nonconstant components. Examination of the main effect components $\hat{\varphi}_j$, the two-factor interactions $\hat{\varphi}_{jk}$, and so forth can give insight into the shape of $\hat{\varphi}$ and hopefully of φ as well.

An example of a hierarchical, unsaturated log-linear submodel with $d = 2$ when $N = 3$ is given by

(5) $$\varphi(y_1, y_2, y_3) = \varphi_0 + \varphi_1(y_1) + \varphi_2(y_2) + \varphi_3(y_3) + \varphi_{12}(y_1, y_2) + \varphi_{13}(y_1, y_3).$$

The joint probability function of Y_1, Y_2 and Y_3 has the form given by (5) if and only if Y_2 and Y_3 are conditionally independent given Y_1. The corresponding maximum likelihood estimate is given by

$$(6) \qquad \hat{\varphi}(y_1, y_2, y_3) = \hat{\varphi}_0 + \hat{\varphi}_1(y_1) + \hat{\varphi}_2(y_2) + \hat{\varphi}_3(y_3) + \hat{\varphi}_{12}(y_1, y_2) + \hat{\varphi}_{13}(y_1, y_3).$$

If we do not know that φ has the form given by (5), we can think of $\hat{\varphi}$ as an estimate of the corresponding best theoretical approximation

$$(7) \qquad \varphi^*(y_1, y_2, y_3) = \varphi_0^* + \varphi_1^*(y_1) + \varphi_2^*(y_2) + \varphi_3^*(y_3) + \varphi_{12}^*(y_1, y_2) + \varphi_{13}^*(y_1, y_3)$$

to φ, where best means having maximum expected log-likelihood subject to the indicated form.

Equations (1)–(7) are also applicable when $Y_1, \ldots, Y_N$ are a mixture of discrete and continuous random variables having joint probability-density function f. In order to employ finite parameter maximum likelihood estimation in this more general context, we can associate the continuous variables with polynomial splines. From a methodological viewpoint, an attractive approach would be to use adaptive model selection techniques as in MARS [Friedman (1990, 1991)]. In the interest of mathematical tractability, however, we will confine our attention to nonadaptively selected models, which have the form of a multiparameter exponential family. We will further restrict attention to continuous random variables $Y_1, \ldots, Y_N$ that each range over a compact interval. Without further loss of generality, we can assume that each of these variables ranges over $[0, 1]$.

It is then natural to conjecture that (under suitable conditions) the integrated squared error of $\hat{\varphi}$ as an estimate of the corresponding best approximation φ^* and the integrated squared error of each component of $\hat{\varphi}$ as an estimate of the corresponding component of φ^* should approach zero as $n \to \infty$. Suppose the components of φ^* all have p derivatives. In light of Stone (1982, 1985, 1986, 1991a, 1991b) and Hasminskii and Ibragimov (1990), it is natural to conjecture that these integrated squared errors should converge to zero at the optimal rate $n^{-2p/(2p+d)}$ and hence that choosing $d < N$ should mitigate the "curse of dimensionality." The main purpose of the present paper is to verify the latter conjecture and thereby to provide theoretical motivation for the use of polynomial spline estimation as a building block in modelling the joint distribution of random variables some or all of which are continuous.

2. **Statement of Results.** Given a function h on $\mathcal{Y} = [0,1]^N$, set $c(h) = \log \int_{\mathcal{Y}} \exp(h(\mathrm{y}))dy$; if $c(h) < \infty$, then $\exp(h - c(h))$ is a density function on $\mathcal{Y}$. Given a subset s of $\{1, \ldots, N\}$, let $\mathcal{H}_s$ denote the space of functions on $\mathcal{Y}$ that only depend on the variables y_l, $l \in s$. Let $\mathcal{S}_0$ be a nonempty collection of subsets of $\{1, \ldots, N\}$. It is assumed that $\mathcal{S}_0$ is *hierarchical*; that is, that if s is a member of $\mathcal{S}_0$ and r is a subset of s then r is a member of $\mathcal{S}_0$. Let $\mathcal{H}_0$ be the collection of functions of the form $h = \sum_{s \in \mathcal{S}_0} h_s$ with $h_s \in \mathcal{H}_s$ for $s \in \mathcal{S}_0$ and such that $c(h) < \infty$.

Let f be a density function on $\mathcal{Y}$.

CONDITION 1. The function $\log f$ is bounded on $\mathcal{Y}$.

The expected log-likelihood function $\Lambda(h)$, $h \in \mathcal{H}_0$, is defined by

$$\Lambda(h) = \int_{\mathcal{Y}} [h(\mathrm{y}) - c(h)] f(\mathrm{y})dy = \int_{\mathcal{Y}} h(\mathrm{y}) f(\mathrm{y})dy - c(h).$$

The first two parts of the following theorem will be proven in Section 3; the third part, which is contained in the information inequality, is a consequence of Jensen's inequality.

THEOREM 1. *Suppose Condition* 1 *holds. Then there is a* $h^* \in \mathcal{H}_0$ *such that* $\Lambda(h^*) = \max_{h \in \mathcal{H}_0} \Lambda(h)$. *The function* $\varphi^* = h^* - c(h^*)$ *is essentially uniquely determined. If* $\varphi = h - c(h)$ *for some* $h \in \mathcal{H}_0$, *then* $\varphi^* = \varphi$ *almost everywhere.*

Set $\langle h_1, h_2 \rangle = \int_{\mathcal{Y}} h_1(\mathrm{y}) h_2(\mathrm{y}) f(\mathrm{y})dy$ and $\|h\|^2 = \langle h, h \rangle = \int_{\mathcal{Y}} h^2(\mathrm{y}) f(\mathrm{y})dy$ for square integrable functions h_1, h_2, h on $\mathcal{Y}$. For $s \in \mathcal{S}_0$, let $\mathcal{H}_s^2$ be the space of square integrable functions in $\mathcal{H}_s$ and set

$$\mathcal{H}_s^0 = \{h \in \mathcal{H}_s^2 \colon h \perp \mathcal{H}_r^2 \text{ for } r \subset s \text{ with } r \neq s\}, \quad s \in \mathcal{S}.$$

(Here $h \perp \mathcal{H}_r^2$ means that $\langle h, k \rangle = 0$ for $k \in \mathcal{H}_r^2$.)

Set $\mathcal{S} = \mathcal{S}_0 \setminus \{\emptyset\}$ and $d = \max_{s \in \mathcal{S}} \#(s)$. It is assumed that $d \geq 1$. Let $\mathcal{H}^2$ denote the direct sum of $\mathcal{H}_s^0$, $s \in \mathcal{S}$. Then each $h \in \mathcal{H}^2$ can be written in an essentially unique manner in the form $h = \sum_s h_s = \sum_{s \in \mathcal{S}} h_s$ with $h_s \in \mathcal{H}_s^0$ for $s \in \mathcal{S}$ [see Lemma 1 of Stone (1991a)].

Suppose the function φ^* in Theorem 1 is square integrable. Then it can be written in an essentially unique manner as $\varphi^* = \Sigma_s \varphi_s^* - c(\Sigma_s \varphi_s^*)$ with $\varphi_s^* \in \mathcal{H}_s^0$ for $s \in \mathcal{S}$.

Let $0 < \beta \leq 1$. A function h on $\mathcal{Y}$ is said to satisfy a Hölder condition with exponent β if there is a positive number B such that $|h(\mathbf{y}) - h(\mathbf{y}_0)| \leq B|\mathbf{y} - \mathbf{y}_0|^\beta$ for $\mathbf{y}_0, \mathbf{y} \in \mathcal{Y}$; here $|\mathbf{y}|$ is the Euclidean norm $(y_1^2 + \cdots + y_N^2)^{1/2}$ of $\mathbf{y} = (y_1, \ldots, y_N)$. Given an N-tuple $\alpha = (\alpha_1, \ldots, \alpha_N)$ of nonnegative integers, set $[\alpha] = \alpha_1 + \cdots + \alpha_N$ and let D^α denote the differentiable operator defined by

$$D^\alpha = \frac{\partial^{[\alpha]}}{\partial y_1^{\alpha_1} \cdots \partial y_N^{\alpha_N}}.$$

Let m be a nonnegative integer and set $p = m + \beta$. It is assumed that $p > d/2$.

Condition 2. The function φ^* is bounded and, for $s \in \mathcal{S}$ and $[\alpha] = m$, the function φ_s^* on $\mathcal{Y}$ is m-times continuously differentiable and $D^\alpha \varphi_s^*$ satisfies a Hölder condition with exponent β.

Let $\mathbf{Y}_1, \ldots, \mathbf{Y}_n$ be a random sample of size n from the distribution having density function f, and let $\langle \cdot , \cdot \rangle_n$ denote the semi-inner product defined by

$$\langle h_1, h_2 \rangle_n = n^{-1} \Sigma_i h_1(\mathbf{Y}_i) h_2(\mathbf{Y}_i).$$

The corresponding seminorm is given by $\|h\|_n^2 = \langle h, h \rangle$.

Let $K = K_n$ be a positive integer and let I_k, $1 \leq k \leq K$, denote the subintervals of $[0, 1]$ defined by $I_k = [(k-1)/K, k/K)$ for $1 \leq k < K$ and $I_k = [1 - 1/K, 1]$ for $k = K$. Let m and q be fixed integers such that $m \geq 0$ and $m > q$. Let $\mathcal{B} = \mathcal{B}_n$ denote the space of spline functions g on $[0, 1]$ such that

(i) the restriction of g to I_k is a polynomial of degree m (or less) for $1 \leq k \leq K$;

and, if $q \geq 0$,

(ii) g is q-times continuously differentiable on $[0, 1]$.

Let B_j, $1 \leq j \leq J$, denote the usual basis of $\mathcal{B}$ consisting of B-splines [see de Boor (1978)]. Then, in particular, $B_j \geq 0$ on $[0, 1]$ for $1 \leq j \leq J$ and $\Sigma_j B_j = 1$ on $[0,1]$. Observe that $K \leq J \leq (m + 1)K$. It is assumed that $J \geq 2$.

Given a subset s of $\{1,\ldots,N\}$, let $\mathcal{G}_s$ denote the space spanned by the functions g on $\mathcal{Y}$ of the form $g(\mathbf{y}) = \prod_{l \in s} g_l(y_l)$, where $\mathbf{y} = (y_1,\ldots,y_N)$ and $g_l \in \mathcal{B}$ for $l \in s$. Then $\mathcal{G}_s$ has dimension $J^{\#(s)}$. Set $\mathcal{G} = \{\sum_s g_s: g_s \in \mathcal{G}_s \text{ for } s \in \mathcal{S}\}$ and

$$\mathcal{G}_s^0 = \{g \in \mathcal{G}_s: g \perp_n \mathcal{G}_r \text{ for every proper subset } r \text{ of } s\}, \quad s \in \mathcal{S}.$$

(Here $g \perp_n \mathcal{G}_r$ means that $\langle g, h\rangle_n = 0$ for $h \in \mathcal{G}_r$.) Then $\mathcal{G} = \sum_s \mathcal{G}_s^0$.

The space $\mathcal{G}$ is said to be *identifiable* (relative to the random sample of size n) if the only function $g \in \mathcal{G}$ such that $g(\mathbf{Y}_i) = 0$ for $1 \le i \le n$ is the zero function; otherwise, $\mathcal{G}$ is said to be *nonidentifiable*. Suppose $\mathcal{G}$ is identifiable. Then $\langle \cdot, \cdot \rangle_n$ is an inner product on $\mathcal{G}$ and $\|\cdot\|_n$ is a norm on $\mathcal{G}$; that is, $\|g\|_n > 0$ for every nonzero function $g \in \mathcal{G}$. Moreover [see Lemma 2 of Stone (1991a)], $\mathcal{G}$ is the direct sum of $\mathcal{G}_s^0$, $s \in \mathcal{S}$; that is, each $g \in \mathcal{G}$ can be written uniquely in the form $g = \sum_s g_s$, where $g_s \in \mathcal{G}_s^0$ for $s \in \mathcal{S}$.

CONDITION 3. *$J^{2d} = o(n^{1-\delta})$ for some $\delta > 0$.*

It follows from Theorem 1 of Stone (1991a) that if Conditions 1 and 3 hold, then

$$P(\mathcal{G} \text{ is nonidentifiable}) = o(1).$$

We refer to the model corresponding to the assumption that $f = \exp(g - c(g))$ for some $g \in \mathcal{G}$ as a *multivariate log-spline model.* The corresponding log-likelihood function $l(g)$, $g \in \mathcal{G}$, is defined by $l(g) = \sum_i [g(\mathbf{Y}_i) - c(g)]$. If $\hat{g} \in \mathcal{G}$ and $l(\hat{g}) = \max_{g \in \mathcal{G}} l(g)$, then $\hat{\varphi} = \hat{g} - c(\hat{g})$ is referred to as the maximum likelihood estimate of φ^* and $\hat{f} = \exp(\hat{\varphi})$ is referred to as the maximum likelihood estimate of $f^* = \exp(\varphi^*)$. If $\mathcal{G}$ is identifiable and $\hat{\varphi}$ exists, then $\hat{\varphi} = \sum_s \hat{\varphi}_s - c(\sum_s \hat{\varphi}_s)$, where $\hat{\varphi}_s \in \mathcal{G}_s^0$ is uniquely determined for $s \in \mathcal{S}$. According to Lemma 8 in Section 4, if Conditions 1 and 3 hold, then $\hat{\varphi}$ exists except on an event whose probability tends to zero with n.

The rate of convergence of $\hat{\varphi}$ to φ^* is given in the next result, which will be proven in Section 4.

THEOREM 2. *Suppose Conditions* 1–3 *hold. Then*

$$\|\hat{\varphi}_s - \varphi_s^*\| = O_P\left[J^{-p} + \sqrt{J^d/n}\right], \quad s \in \mathscr{S},$$

so

$$\|\hat{\varphi} - \varphi^*\| = O_P\left[J^{-p} + \sqrt{J^d/n}\right].$$

Observe that if Condition 3 holds with $J \sim n^{1/(2p+d)}$, then $p > d/2$.

COROLLARY 1. *Suppose Conditions* 1 *and* 2 *hold and that* $J \sim n^{1/(2p+d)}$. *Then*

$$\|\hat{\varphi}_s - \varphi_s^*\| = O_P(n^{-p/(2p+d)}), \quad s \in \mathscr{S},$$

so

$$\|\hat{\varphi} - \varphi^*\| = O_P(n^{-p/(2p+d)}).$$

The L_2 rate of convergence in Corollary 1 does not depend on N. It is clear [see Stone (1982) and Hasminskii and Ibragimov (1990)] with $d = N$ that this rate is optimal. When $d = N$, it is possible to use the tensor product extension of de Boor (1976) to obtain the pointwise and L_∞ rates of convergence of $\hat{\varphi}$ to φ^* [see Koo (1988)]. Stone (1990) contains a more extensive theory of univariate ($N = 1$) log-spline modelling, and the corresponding methodological issues are discussed in Stone and Koo (1986), Kooperberg and Stone (1991) and Kooperberg (1991). Koo (1991) uses AIC to select K adaptively in an asymptotically optimal manner in the context of univariate log-spline modelling. Presumably his techniques are applicable to multivariate log-spline modelling. The analog of Theorem 2 for interactive spline regression was obtained in Stone (1991a) and the analog for generalized interactive models was obtained in Stone (1991b).

3. Proof of Theorem 1. Let h_1 and h_2 be in $\mathscr{H}_0$. Set

$$h^{(t)} = (1-t)h_1 + th_2 \in \mathscr{H}_0, \quad C(t) = c(h^{(t)}) \quad \text{and} \quad f^{(t)} = \exp(h^{(t)} - C(t)), \quad t \in [0,1].$$

Then C is a continuous function on [0,1] and

$$C''(t) = \int_{\mathscr{Y}} [h_2(y) - h_1(y)]^2 f^{(t)}(y)dy - \left[\int_{\mathscr{Y}} [h_2(y) - h_1(y)] f^{(t)}(y)dy\right]^2 \tag{8}$$

for $0 < t < 1$. (It follows by a standard argument in the context of one parameter exponential families or that of moment generating functions that the various integrals appearing in (8) are finite.) We conclude from (8) that C is convex on [0, 1] and that it is strictly convex unless $h_2 - h_1$ is essentially constant on $\mathcal{Y}$. Moreover,

(9) $$\Lambda(h^{(t)}) = (1-t)\Lambda(h_1) + t\Lambda(h_2) - [(1-t)c(h_1) + tc(h_2) - C(t), \quad 0 \le t \le 1.$$

The first part of Theorem 1 will now be verified. It follows from Condition 1 and the information inequality that $\Lambda(h) = \int_{\mathcal{Y}} h(\mathrm{y})f(\mathrm{y})d\mathrm{y} - c(h) \le \int_{\mathcal{Y}} [\log f(\mathrm{y})]f(\mathrm{y})d\mathrm{y} < \infty$ for $h \in \mathcal{H}_0$ and hence that the numbers $\Lambda(h)$, $h \in \mathcal{H}_0$, have a finite least upper bound L. Let $|A|$ denote the Lebesgue measure of a subset A of $\mathcal{Y}$. Choose $h_k \in \mathcal{H}_0$ for $k \ge 1$ such that $\Lambda(h_k) \to L$ as $k \to \infty$. Since $f_k = \exp(h_k - c(h_k))$ is a density function on $\mathcal{Y}$,

$$|\{\mathrm{y} \in \mathcal{Y}: h_k(\mathrm{y}) - c(h_k) \ge M\}| \le \exp(-M), \quad M \in \mathbb{R}.$$

It now follows easily from the inequality

$$\log \frac{f_k}{f} \le \frac{f_k}{f} - 1,$$

that

(10) $$\lim_{M \to \infty} \limsup_{k \to \infty} |\{\mathrm{y} \in \mathcal{Y}: |h_k(\mathrm{y}) - c(h_k)| \ge M\}| = 0.$$

It is a straightforward consequence of (8)–(10), Lemma 1 of Stone (1991b), and the definition of L that there is a function $h^* \in \mathcal{H}_0$ such that $h_k - c(h_k) \to h^* - c(h^*)$ in measure as $k \to \infty$. Necessarily, $\Lambda(h^*) = L = \max_{h \in \mathcal{H}_0} \Lambda(h)$.

In order to verify that $h^* - c(h^*)$ is essentially uniquely determined, suppose that h_1^* and h_2^* are in $\mathcal{H}_0$ and that $\Lambda(h_1^*) = L$ and $\Lambda(h_2^*) = L$. It then follows from (8) and (9) that $h_2^* - h_1^*$ is essentially constant on $\mathcal{Y}$ and hence that $[h_2^* - c(h_2^*)] - [h_1^* - c(h_1^*)]$ is essentially constant. Since $\int_{\mathcal{Y}} \exp(h_1^*(\mathrm{y}) - c(h_1^*))d\mathrm{y} = 1$ and $\int_{\mathcal{Y}} \exp(h_2^*(\mathrm{y}) - c(h_2^*))d\mathrm{y} = 1$, the constant difference must equal zero. Therefore $h_1^* - c(h_1^*) = h_2^* - c(h_2^*)$ almost everywhere on $\mathcal{Y}$.

4. Proof of Theorem 2. Throughout this section it is assumed that Conditions 1–3 hold. Let $\|h\|_\infty = \sup_{\mathrm{y} \in \mathcal{Y}} |h(\mathrm{y})|$ denote the L_∞ norm of a function h on $\mathcal{Y}$.

LEMMA 1. *Let T be a positive constant. Then there are positive numbers* M_1 and M_2 *such that*

$$-M_1\|h-c(h)-\varphi^*\|^2 \le \Lambda(h)-\Lambda(\varphi^*) \le -M_2\|h-c(h)-\varphi^*\|^2$$

for all $h \in \mathscr{H}_0$ such that $\|h-c(h)\|_\infty \le T$.

PROOF. Given $h \in \mathscr{H}_0$ with $\|h-c(h)\|_\infty \le T$ and given $t \in [0, 1]$, set

$$h^{(t)} = (1-t)\varphi^* + th \quad \text{and} \quad C(t) = c(h^{(t)}).$$

Then

$$\frac{d}{dt}\Lambda(h^{(t)})\Big|_{t=0} = 0$$

and hence, by (9),

$$\Lambda(h)-\Lambda(\varphi^*) = \int_0^1 (1-t)\frac{d^2}{dt^2}\Lambda(h^{(t)})\,dt = -\int_0^1 (1-t)C''(t)dt.$$

Thus, by (8), there is a positive number M_1 such that

$$\Lambda(h)-\Lambda(\varphi^*) \ge -M_1\|h-c(h)-\varphi^*\|^2, \quad h \in \mathscr{H}_0 \text{ with } \|h-c(h)\|_\infty \le T.$$

By another application of (8), in order to complete the proof of the lemma, it suffices to show that if $h_k \in \mathscr{H}_0$ and $\|h_k - c(h_k)\|_\infty \le T$ for $k \ge 1$, then there is an $\varepsilon > 0$ such that

$$\left[\int_{\mathscr{Y}} [h_k(y) - c(h_k) - \varphi^*(y)]f^*(y)dy\right]^2 \le (1-\varepsilon)\int_{\mathscr{Y}} [h_k(y) - c(h_k) - \varphi^*(y)]^2 f^*(y)dy, \quad k \gg 1.$$

This result is easily established under the additional assumption that

$$\liminf_{k\to\infty} \int_{\mathscr{Y}} [h_k(y) - c(h_k) - \varphi^*(y)]^2 dy > 0.$$

(Note that if $h \in \mathscr{H}_0$ and $h - c(h) - \varphi^*$ is essentially constant on $\mathscr{Y}$, then the constant equals zero.) Otherwise, we can assume that

$$\lim_{k\to\infty} \int_{\mathscr{Y}} [h_k(y) - c(h_k) - \varphi^*(y)]^2 dy = 0.$$

Then there is a bounded function R such that

$$\begin{aligned}
1 &= \int_{\mathscr{Y}} \exp(h_k(y) - c(h_k))dy \\
&= \int_{\mathscr{Y}} \exp(h_k(y) - c(h_k) - \varphi^*(y))f^*(y)dy \\
&= 1 + \int_{\mathscr{Y}} [h_k - c(h_k) - \varphi^*(y)]f^*(y)dy + \int_{\mathscr{Y}} R(y)[h_k(y) - c(h_k) - \varphi^*(y)]^2 f^*(y)dy,
\end{aligned}$$

which yields the desired result. □

The next result is Lemma 3 in Stone (1991b).

LEMMA 2. *There is a positive number M_3 such that $\|g\|_\infty \le M_3 J^{d/2}\|g\|$ for $g \in \mathcal{G}$.*

According to a simplification of the argument used in Section 3 to prove Theorem 1, there is a unique function $g_n^* \in \mathcal{G}$ such that $\Lambda(g_n^*) = \min_{g\in\mathcal{G}}\Lambda(g)$. Set $\varphi_n^* = g_n^* - c(g_n^*)$. (Actually, g_n^* and φ_n^* depend on J rather than n, but we are mainly thinking of J as depending on n.) If $\mathcal{G}$ is identifiable, then $g_n^* = \sum_s \varphi_{ns}^*$, where $\varphi_s^* \in \mathcal{G}_s^0$ is uniquely determined for $s \in \mathcal{S}$.

LEMMA 3. $\|\varphi_n^* - \varphi^*\|^2 = O(J^{-2p})$ *and* $\|\varphi_n^* - \varphi^*\|_\infty = O(J^{d/2-p})$.

PROOF. By Condition 2 [see Theorem 12.8 of Schumaker (1981)], there is a function $g_n \in \mathcal{G}$ such that $\|g_n - \varphi^*\|_\infty \le M_4 J^{-p}$; here M_4 is a positive constant. Set $\varphi_n = g_n - c(g_n)$. Then $\|\varphi_n - \varphi^*\|_\infty \le M_5 J^{-p}$, where $M_5 = 2M_4$. Consequently, $\|\varphi_n - \varphi^*\|^2 \le M_5^2 J^{-2p}$. Thus by Lemma 1 there is a positive constant M_6 such that

$$\Lambda(\varphi_n) - \Lambda(\varphi^*) \ge -M_6 J^{-2p}. \tag{11}$$

Let a denote a large positive constant. Choose $g \in \mathcal{G}$ with $\|g - c(g) - \varphi^*\|^2 = aJ^{-2p}$. Then $\|g - c(g) - \varphi_n\|^2 \le 2(a + M_5^2)J^{-2p}$. Since $p > d/2$, it follows from Lemma 2 that, for J sufficiently large, $\|g - c(g)\|_\infty \le \|\varphi^*\|_\infty + 1$ for all such functions g. Thus by Lemma 1 there is a positive constant M_7 such that, for J sufficiently large,

$$\Lambda(g) - \Lambda(\varphi^*) \le -M_7 aJ^{-2p} \quad \text{for all } g \in \mathcal{G} \text{ with } \|g - c(g) - \varphi^*\|^2 = aJ^{-2p}. \tag{12}$$

Let a be chosen so that $a > M_5^2$ and $M_7 a > M_6$. It follows from (11) and (12) that, for J sufficiently large,

$$\Lambda(g) < \Lambda(\varphi_n) \quad \text{for all } g \in \mathcal{G} \text{ with } \|g - c(g) - \varphi^*\|^2 = aJ^{-2p}.$$

Therefore, by the concavity of $\Lambda(g)$ as a function g, $\|\varphi_n^* - \varphi^*\|^2 < aJ^{-2p}$ for J sufficiently large. This verifies the first conclusion of the lemma. Observe that $\|\varphi_n^* - \varphi_n\|^2 = O(J^{-2p})$ and hence by Lemma 2 that $\|\varphi_n^* - \varphi_n\|_\infty = O(J^{d/2-p})$. Consequently, $\|\varphi_n^* - \varphi^*\|_\infty = O(J^{d/2-p})$, so the second conclusion of the lemma is valid. □

LEMMA 4. *$\|\varphi^*_{ns} - \varphi^*_s\|^2 = O_P(J^{-2p} + J^d/n)$ for $s \in \mathscr{S}$.*

PROOF. Suppose $\mathscr{G}$ is identifiable, and let $\tilde{g}_n$ denote the orthogonal projection of φ^* onto $\mathscr{G}$ relative to $\perp_n$. Then $\tilde{g}_n = \sum_s \tilde{\varphi}_{ns}$, where $\tilde{\varphi}_{ns} \in \mathscr{G}^0_{ns}$ is uniquely determined for $s \in \mathscr{S}$. Set $\tilde{\varphi}_n = \tilde{g}_n - c(\tilde{g}_n)$. It follows from Theorem 3 in Stone (1991a) that

$$(13) \qquad \|\tilde{\varphi}_{ns} - \varphi^*_s\|^2 = O_P(J^{-2p} + J^d/n), \quad s \in \mathscr{S},$$

and hence from Lemma 2 that

$$\|\tilde{\varphi}_n - \varphi^*\|^2 = O_P(J^{-2p} + J^d/n).$$

Thus, by Lemma 3,

$$\|\tilde{\varphi}_n - \varphi^*_n\|^2 = O_P(J^{-2p} + J^d/n).$$

Consequently, by Lemma 6 of Stone (1991a),

$$(14) \qquad \|\tilde{\varphi}_{ns} - \varphi^*_{ns}\|^2 = O_P(J^{-2p} + J^d/n), \quad s \in \mathscr{S}.$$

The desired result follows from (13) and (14). □

Let τ_n, $n \ge 1$, be positive numbers such that $J^d\tau_n^2 = O(1)$ and $J^d \log n = o(n\tau_n^2)$. The next result follows from Lemma 2 and Bernstein's inequality (see the proof of Lemma 5 in Stone (1990)].

LEMMA 5. *Given $a > 0$ and $\varepsilon > 0$, there is a $\delta > 0$ such that, for n sufficiently large,*

$$P\left[\left|\frac{l(g) - l(\varphi^*_n)}{n} - [\Lambda(g) - \Lambda(\varphi^*_n)]\right| \ge \varepsilon\tau_n^2\right] \le 2\exp(-\delta n\tau_n^2)$$

for all $g \in \mathscr{G}$ with $\|g - c(g) - \varphi^*_n\| \le a\tau_n$.

We define the *diameter* of a set B of functions on $\mathscr{Y}$ as

$$\sup\{\|g_2 - g_1\|_\infty : g_1, g_2 \in B\}.$$

The proof of the next result is essentially the same as that of Lemma 8 of Stone (1991b).

LEMMA 6. *Given $a > 0$ and $\delta > 0$, there is a positive constant M_4 such that*

$$\{g - c(g) : g \in \mathscr{G} \text{ and } \|g\, c(g) - \varphi^*_n\| \le a\tau_n\}$$

can be covered by $O(\exp(M_4 J^d \log n))$ subsets each having diameter at most $\delta\tau_n^2$.

LEMMA 7. *Let $a > 0$. Then, except on an event whose probability tends to zero with n, $l(g) < l(\varphi_n^*)$ for all $g \in \mathscr{G}$ such that $\|g - c(g) - \varphi_n^*\| = a\tau_n$.*

PROOF. This result follows from Lemma 1, with φ^* replaced by φ_n^* and $\mathscr{H}_0$ replaced by $\mathscr{G}$, Lemmas 5 and 6, and the inequality

$$\left|\frac{l(g_2) - l(g_1)}{n}\right| \le \|g_2 - c(g_2) - [g_1 - c(g_1)]\|_\infty, \quad g_1, g_2 \in \mathscr{G}. \ \square$$

LEMMA 8. *The maximum likelihood estimate* of φ *of the form $\hat{\varphi} = \hat{g} - c(\hat{g})$ with $\hat{g} \in \mathscr{G}$ exists and is unique except on an event whose probability tends to zero with n. Moreover, $\|\hat{\varphi} - \varphi_n^*\|_\infty = o_P(1)$.*

PROOF. It follows from Lemma 7 and the concavity of $\Lambda(g)$ as a function of g that $\|\hat{\varphi} - \varphi_n^*\| = o_P(\tau_n)$ and hence from Lemma 2 that $\|\hat{\varphi} - \varphi_n^*\|_\infty = o_P(J^{d/2}\tau_n) = o_P(1)$. $\square$

For $s \in \mathscr{S}$, let $\mathscr{J}_s$ denote the collection of ordered $\#(s)$-tuples j_l, $l \in s$, with $j_l \in \{1, \ldots, J\}$ for $l \in s$. Then $\#(\mathscr{J}_s) = J^{\#(s)}$. For $\mathbf{j} \in \mathscr{J}_s$, let $B_{s\mathbf{j}}$ denote the function on $\mathscr{Y}$ given by

$$B_{s\mathbf{j}}(\mathbf{y}) = \prod_{l \in s} B_{j_l}(y_l), \quad \mathbf{y} = (y_1, \ldots, y_N).$$

Then the functions $B_{s\mathbf{j}}$, $\mathbf{j} \in \mathscr{J}_s$, which are nonnegative and have sum one, form a basis of $\mathscr{G}_s$.

Set $K = \sum_s \#(\mathscr{J}_s)$. Given a K-dimensional (column) vector $\boldsymbol{\theta}$ having entries $\theta_{s\mathbf{j}}$, $s \in \mathscr{S}$ and $\mathbf{j} \in \mathscr{J}_s$, set

$$g_s(\cdot\,; \boldsymbol{\theta}) = \sum_{\mathbf{j} \in \mathscr{J}_s} \theta_{s\mathbf{j}} B_{s\mathbf{j}}, \quad s \in \mathscr{S}, \quad \text{and} \quad g(\cdot\,; \boldsymbol{\theta}) = \sum_s g_s(\cdot\,; \boldsymbol{\theta}).$$

Also, set $C(\boldsymbol{\theta}) = c(g(\cdot\,; \boldsymbol{\theta})) = \log \int_{\mathscr{Y}} \exp(g(\mathbf{y}; \boldsymbol{\theta}) d\mathbf{y}$ and $f(\cdot\,; \boldsymbol{\theta}) = \exp(g(\cdot\,; \boldsymbol{\theta}) - C(\boldsymbol{\theta}))$. Then the log-likelihood function can be written as

$$l(\boldsymbol{\theta}) = \sum_i \log f(\mathbf{Y}_i; \boldsymbol{\theta}) = \sum_i [g(\mathbf{Y}_i; \boldsymbol{\theta}) - C(\boldsymbol{\theta})].$$

Let

$$\mathbf{S}(\boldsymbol{\theta}) = \frac{\partial}{\partial \boldsymbol{\theta}} l(\boldsymbol{\theta})$$

denote the score at $\boldsymbol{\theta}$; that is, the K-dimensional vector having entries

$$\frac{\partial}{\partial\theta_{s\mathbf{j}}}l(\boldsymbol{\theta}) = \sum_i B_{s\mathbf{j}}(\mathbf{Y}_i) - \int_{\mathcal{Y}} B_{s\mathbf{j}}(\mathbf{y})f(\mathbf{y};\boldsymbol{\theta})d\mathbf{y}\Bigr].$$

Let

$$\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^t}l(\boldsymbol{\theta})$$

be the $K \times K$ matrix having entries

(15) $$\frac{\partial^2}{\partial\theta_{s_1\mathbf{j}_1}\partial\theta_{s_1\mathbf{j}_2}}l(\boldsymbol{\theta}) = -n\Bigl[\int_{\mathcal{Y}} B_{s_1\mathbf{j}_1}(\mathbf{y})B_{s_1\mathbf{j}_2}(\mathbf{y})f(\mathbf{y};\boldsymbol{\theta})d\mathbf{y}$$
$$- \Bigl[\int_{\mathcal{Y}} B_{s_1\mathbf{j}_1}(\mathbf{y})f(\mathbf{y};\boldsymbol{\theta})d\mathbf{y}\Bigr]\Bigl[\int_{\mathcal{Y}} B_{s_1\mathbf{j}_2}(\mathbf{y})f(\mathbf{y};\boldsymbol{\theta})d\mathbf{y}\Bigr]\Bigr].$$

Set $\Theta = \{\boldsymbol{\theta} \in \mathbb{R}^K: g_s(\cdot\,;\boldsymbol{\theta}) \in \mathcal{G}_s^0 \text{ for } s \in \mathcal{S}\}$.

Let $\boldsymbol{\theta}^*$ be given by $\varphi_n^* = \sum_s \varphi_{ns}^* - C(\boldsymbol{\theta}^*)$, where $\varphi_{ns}^* = g_s(\cdot\,;\boldsymbol{\theta}^*) \in \mathcal{G}_s^0$ for $s \in \mathcal{S}$. Let $\hat{\boldsymbol{\theta}}$ denote the maximum likelihood estimate of $\boldsymbol{\theta}$, so that $\hat{\varphi} = \sum_s \hat{\varphi}_s - C(\hat{\boldsymbol{\theta}})$, where $\hat{\varphi}_s = g_s(\cdot\,;\hat{\boldsymbol{\theta}}) \in \mathcal{G}_s^0$ for $s \in \mathcal{S}$. Then $\boldsymbol{\theta}^*$ and $\hat{\boldsymbol{\theta}}$ are in Θ. The maximum likelihood equation $\mathbf{S}(\hat{\boldsymbol{\theta}}) = \mathbf{0}$ can be written as

$$\int_0^1 \frac{d}{dt}\mathbf{S}(\boldsymbol{\theta}^* + t(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*))dt = -\,\mathbf{S}(\boldsymbol{\theta}^*).$$

Thus it can be written as $\mathbf{D}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) = -\mathbf{S}(\boldsymbol{\theta}^*)$, where $\mathbf{D}$ is the $K \times K$ matrix given by

$$D = \int_0^1 \frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^t}l(\boldsymbol{\theta}^* + t(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)dt.$$

Let $|\ |$ denote the Euclidean norm on $\mathbb{R}^K$. It follows from the maximum likelihood equation that

(16) $$(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)^t\mathbf{D}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) = -\,(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)^t\mathbf{S}(\boldsymbol{\theta}^*).$$

We claim that

(17) $$|\mathbf{S}(\boldsymbol{\theta}^*)|^2 = O_P(n)$$

and that (for some positive constant M_5)

(18) $$(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)^t\mathbf{D}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) \le -M_5 nJ^{-d}|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*|^2$$

except on an event whose probability tends to zero with n. It follows from (16)–(18) that $|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*| = O_P(J^{2d}/n)$ and hence that

(19) $$\|\hat{\varphi}_s - \varphi_{ns}^*\|^2 = O_P(J^d/n), \quad s \in \mathcal{S},$$

and

(20) $$\|\hat{\varphi} - \varphi_n^*\|^2 = O_P(J^d/n).$$

Theorem 2 follows from (19), (20) and Lemmas 3 and 4.

To verify (17) note that

$$E[B_{sj}(Y) = \int_{\mathcal{Y}} B_{sj}(y) f(y;\theta^*)dy, \quad s \in \mathcal{S} \text{ and } j \in \mathcal{J}_s.$$

Consequently,

$$E\,|\,S(\theta^*)\,|^2 = n\sum_s \sum_{j\in\mathcal{J}_s} \mathrm{var}(B_{sj}(Y)) \le n\sum_s \sum_{j\in\mathcal{J}_s} E[B_{sj}^2(Y)] = O(n),$$

so (17) holds.

Finally, (18) will be verified. It follows from (15) that

$$(21)\quad \delta^t \frac{\partial^2 l}{\partial\theta\partial\theta^t}(\theta)\delta = -n\left[\int_{\mathcal{Y}} g^2(y;\delta)f(y;\theta)dy - \left[\int_{\mathcal{Y}} g(y;\delta)f(y;\theta)dy\right]^2\right], \quad \delta, \theta \in \mathbb{R}^K.$$

By Condition 2, the inequality $p > d/2$, and Lemmas 3 and 8, there is a positive constant T such that

$$(22)\qquad \lim_{n\to\infty} P(\|\varphi_n^*\|_\infty \le T \text{ and } \|\hat{\varphi}\|_\infty \le T) = 1.$$

It follows from (21), (22) and Lemma 7 of Stone (1991a) that there is an $\varepsilon > 0$ such that, except on an event whose probability tends to zero with n,

$$(23)\qquad \delta^t \mathbf{D}\delta \le -\varepsilon n \int_{\mathcal{Y}} g^2(y;\delta)dy, \quad \delta \in \Theta.$$

(Note that $\sum_i g(Y_i;\delta) = 0$ for $\delta \in \Theta$.) According to Conditions 1 and 3 and Lemma 6 of Stone (1991a), there is an $\varepsilon > 0$ such that, except on an event whose probability tends to zero with n,

$$(24)\qquad \int_{\mathcal{Y}} g^2(y;\delta)dy \ge \varepsilon \sum_s \int_{\mathcal{Y}} g_s^2(y;\delta)dy, \quad \delta \in \Theta.$$

It follows from the basic properties of B-splines and repeated use of (viii) on page 155 of de Boor (1978) that, for some $\varepsilon > 0$,

$$\int_{\mathcal{Y}} g_s^2(y;\delta)dy \ge \varepsilon J^{-\#(s)} \sum_j \delta_{sj}^2, \quad s \in \mathcal{S} \text{ and } \delta \in \mathbb{R}^K$$

and hence

$$(25)\qquad \sum_s \int_{\mathcal{Y}} g_s^2(y;\delta)dy \ge \varepsilon J^{-d}|\delta|^2, \quad \delta \in \mathbb{R}^K.$$

Equation (16) follows from (23)–(25) applied to $\delta = \hat{\theta} - \theta^*$. This completes the proof of Theorem 2.

REFERENCES

DE BOOR, C. (1976). A bound on the L_∞-norm of 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.

FRIEDMAN, J. H. (1991). Multivariate Adaptive Regression Splines (with discussion). *Ann. Statist.* 19 1–141.

HASMINSKII, R. and IBRAGIMOV, I. (1990). Kolmogorov's contributions to mathematical statistics. *Ann. Statist.* **18** 1011–1016.

KOO, C.-Y. (1991). A model selection rule for logspline density estimation. Manuscript.

KOO, J.-Y. (1988). Tensor product splines in the estimation of regression, exponential response functions and multivariate densities. Ph. D. Dissertation, Dept. Statist., Univ. California, Berkeley.

KOOPERBERG, C. (1991). Smoothing images, curves and densities. Ph. D. Dissertation. Dept. Statist., Univ. California, Berkeley.

KOOPERBERG, C. and STONE. C. J. (1991). A study of logspline density estimation. *Computational Statistics and Data Analysis*, to appear.

SCHUMAKER, L. L. (1981). *Spline Functions: Basic Theory*. Wiley, New York.

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

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

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

STONE, C. J. (1989). Uniform error bounds involving logspline models. *In Probability, Statistics and Mathematics: Papers in Honor of Samuel Karlin* (T. W. Anderson, K. B. Athreya, and D. L. Iglehart, eds.) 335–355. Academic Press, Boston.

STONE, C. J. (1990). Large-sample inference for log-spline models. *Ann. Statist.* 18 717–741.

STONE, C. J. (1991a). Multivariate regression splines. Technical Report No. 317, Dept. Statist., Univ. California, Berkeley.

STONE, C. J. (1991b). Generalized multivariate regression splines. Technical Report No. 318, Dept. Statist. Univ. California, Berkeley.

STONE, C. J. and KOO, C.–Y. (1986) Logspline density estimation. In *AMS Contemporary Math. Ser.* 29 1–15. Amer. Math. Soc., Providence.