The Annals of Statistics
1991, Vol. 19, No. 4, 1832–1854

ASYMPTOTICS FOR DOUBLY FLEXIBLE LOGSPLINE RESPONSE MODELS[1]

BY CHARLES J. STONE

University of California, Berkeley

Consider a $\mathscr{Y}$-valued response variable having a density function $f(\cdot|x)$ that depends on an $\mathscr{X}$-valued input variable x. It is assumed that $\mathscr{X}$ and $\mathscr{Y}$ are compact intervals and that $f(\cdot|\cdot)$ is continuous and positive on $\mathscr{X} \times \mathscr{Y}$. Let $F(\cdot|x)$ denote the distribution function of $f(\cdot|x)$ and let $Q(\cdot|x)$ denote its quantile function. A finite-parameter exponential family model based on tensor-product B-splines is constructed. Maximum likelihood estimation of the parameters of the model based on independent observations of the response variable at fixed settings of the input variable yields estimates of $f(\cdot|\cdot)$, $F(\cdot|\cdot)$ and $Q(\cdot|\cdot)$. Under mild conditions, if the number of parameters suitably tends to infinity as $n \to \infty$, these estimates have optimal rates of convergence. The asymptotic behavior of the corresponding confidence bounds is also investigated.

1. Discussion of results. Consider a $\mathscr{Y}$-valued random response variable Y having an unknown density function $f(\cdot|x)$ that depends on an $\mathscr{X}$-valued input variable x; here $\mathscr{X}$ and $\mathscr{Y}$ are intervals in $\mathbb{R}$ having positive length. It is assumed that $f(\cdot|\cdot)$ is continuous and positive on $\mathscr{X} \times \mathscr{Y}$. Let $F(\cdot|x)$ and $Q(\cdot|x)$ denote the distribution function and quantile function, respectively, corresponding to $f(\cdot|x)$. Let fixed inputs (design points) $x_1, \ldots, x_n \in \mathscr{X}$ be given and let $Y_1, \ldots, Y_n$ be independent random variables such that Y_i has density function $f(\cdot|x_i)$ for $1 \le i \le n$; here $Y_1, \ldots, Y_n$ are the response variables corresponding to the settings $x_1, \ldots, x_n$, respectively, of the input variable. Observations on these response variables can be used for inference concerning $f(\cdot|\cdot)$, $F(\cdot|\cdot)$ and $Q(\cdot|\cdot)$.

The classical approach is to assume a fixed parametric model $f(\cdot|\theta_1, \ldots, \theta_K)$ for the density function of Y and consider fixed parametric forms $\theta_k = h_k(x; \beta_k)$ for the dependence of $\theta_1, \ldots, \theta_K$ on x. Normal linear models and generalized linear models in which Y has a gamma distribution with known shape parameter are of this form [see McCullagh and Nelder (1983)].

A refinement of the classical approach is to assume that $\theta_k = h_k(x)$ for $1 \le k \le K$, where $h_1, \ldots, h_K$ are unknown continuous functions on $\mathscr{X}$, approximate these functions by members of some flexible J-dimensional linear space $\mathscr{H}$ such as a space of polynomial, trigonometric series or polynomial splines, and let $J \to \infty$, as $n \to \infty$.

Received September 1989; revised December 1990.
[1]Research supported in part by NSF Grants DMS-86-00409 and DMS-89-02016.
AMS 1980 *subject classifications*. Primary 62G05; secondary 62F12.
Key words and phrases. Input-response model, exponential families, B-splines, maximum likelihood, rates of convergence.

A further refinement is to choose a flexible K-dimensional linear space $\mathscr{S}$ and a basis $B_1, \ldots, B_K$ of $\mathscr{S}$, approximate $\log f(\cdot|x)$ by $\theta_1 B_1 + \cdots + \theta_K B_K - C(\theta_1, \ldots, \theta_K)$, where $C(\theta_1, \ldots, \theta_K)$ is the normalizing constant, approximate the dependence of $\theta_1, \ldots, \theta_K$ on x by members of some flexible J-dimensional linear space $\mathscr{H}$ and let $J, K \to \infty$ as $n \to \infty$. This *doubly-flexible* approach will be pursued in the present paper, with $\mathscr{S}$ and $\mathscr{H}$ being spaces of polynomial splines.

For theoretical purposes, $\mathscr{X}$ and $\mathscr{Y}$ are required to be compact subintervals of $\mathbb{R}$. Let $\mathscr{S}$ be a standard linear space of spline functions of a given order $q \geq 1$ on $\mathscr{Y}$ having dimension $K \geq 2$. (The functions in $\mathscr{S}$ are piecewise polynomials of degree $q - 1$ or less. If $q = 1$, we choose them to be right-continuous on $\mathscr{Y}$ and continuous at the right endpoint of $\mathscr{Y}$; if $q \geq 2$, they are $(q - 2)$-times continuously differentiable on $\mathscr{Y}$.) Let $B_1, \ldots, B_K$ be a basis of $\mathscr{S}$ consisting of B-splines [see deBoor (1978)]. Then $B_1, \ldots, B_K$ are nonnegative and sum to one on $\mathscr{Y}$.

Given $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_K)^t \in \Theta$, set $s(y; \boldsymbol{\theta}) = \theta_1 B_1(y) + \cdots + \theta_K B_K(y)$ for $y \in \mathscr{Y}$, $C(\boldsymbol{\theta}) = \log \int \exp(s(y; \boldsymbol{\theta})\, dy$ and $f(y; \boldsymbol{\theta}) = \exp(s(y; \boldsymbol{\theta}) - C(\boldsymbol{\theta}))$ for $y \in \mathscr{Y}$. Also, set

$$\Theta = \left\{\boldsymbol{\theta} = (\theta_1, \ldots, \theta_K)^t \in \mathbb{R}^K : \theta_1 + \cdots + \theta_K = 0\right\}.$$

Then $f(\cdot; \boldsymbol{\theta})$, $\boldsymbol{\theta} \in \Theta$, defines an identifiable exponential family; it is referred to as a *logspline model* since $\log f(\cdot; \boldsymbol{\theta}) \in \mathscr{S}$. The theory of such models was developed in Stone (1990), which is a precursor to the present paper. Barron and Sheu (1991) independently obtained results for logspline models as well as for similar models involving polynomials and trigonometric series.

Let $\mathscr{H}$ be a standard linear space of spline functions of a given order (which is not necessarily the same as that of $\mathscr{S}$) having dimension $J \geq 1$ and let $H_1, \ldots, H_J$ be a basis of $\mathscr{H}$ consisting of B-splines. Let $\mathscr{B}$ denote the collection of $J \times K$ matrices $\boldsymbol{\beta} = (\beta_{jk})$ of real numbers β_{jk}, $1 \leq j \leq J$ and $1 \leq k \leq K$, such that $\sum_k \beta_{jk} = 0$ for $1 \leq j \leq J$, which can be regarded as a $[J(K - 1)]$-dimensional subspace of $\mathbb{R}^{JK}$. Let $\boldsymbol{\beta} \in \mathscr{B}$. For $1 \leq k \leq K$, let $h_k(\cdot; \boldsymbol{\beta})$ be the real-valued function on $\mathscr{X}$ defined by $h_k(x; \boldsymbol{\beta}) = \sum_j \beta_{jk} H_j(x)$ for $x \in \mathscr{X}$. Set $\mathbf{h}(x; \boldsymbol{\beta}) = (h_1(x; \boldsymbol{\beta}), \ldots, h_K(x; \boldsymbol{\beta}))^t$ for $x \in \mathscr{X}$ and observe that $\mathbf{h}(\cdot; \boldsymbol{\beta})$ is a Θ-valued function on $\mathscr{X}$. Also, set $f(y|x; \boldsymbol{\beta}) = f(y; \mathbf{h}(x; \boldsymbol{\beta})) = \exp(s(y; \mathbf{h}(x; \boldsymbol{\beta})) - C(\mathbf{h}(x; \boldsymbol{\beta})))$ for $\boldsymbol{\beta} \in \mathscr{B}$, $x \in \mathscr{X}$ and $y \in \mathscr{Y}$. Then $f(\cdot|x; \boldsymbol{\beta})$ is a positive density function on $\mathscr{Y}$ for $\boldsymbol{\beta} \in \mathscr{B}$ and $x \in \mathscr{X}$. We refer to $f(\cdot|x; \boldsymbol{\beta})$, $x \in \mathscr{X}$ and $\boldsymbol{\beta} \in \mathscr{B}$, as a *logspline response model*.

The log-likelihood function $l(\boldsymbol{\beta})$, $\boldsymbol{\beta} \in \mathscr{B}$, is defined by

$$l(\boldsymbol{\beta}) = \sum_i \log f(Y_i|x_i; \boldsymbol{\beta}) = \sum_i \left[s(Y_i; \mathbf{h}(x_i; \boldsymbol{\beta})) - C(\mathbf{h}(x_i; \boldsymbol{\beta}))\right], \qquad \boldsymbol{\beta} \in \mathscr{B}.$$

Set $c(\boldsymbol{\beta}) = \sum_i C(\mathbf{h}(x_i; \boldsymbol{\beta}))$ for $\boldsymbol{\beta} \in \mathscr{B}$. Then $l(\boldsymbol{\beta}) = \sum_i s(Y_i; \mathbf{h}(x_i; \boldsymbol{\beta})) - c(\boldsymbol{\beta})$ for $\boldsymbol{\beta} \in \mathscr{B}$. The expected log-likelihood function $\lambda(\boldsymbol{\beta})$, $\boldsymbol{\beta} \in \mathscr{B}$, is given by

$$\lambda(\boldsymbol{\beta}) = El(\boldsymbol{\beta}) = \sum_i \int s(y; \mathbf{h}(x_i; \boldsymbol{\beta}))\, f(y|x_i)\, dy - c(\boldsymbol{\beta}), \qquad \boldsymbol{\beta} \in \mathscr{B}.$$

The functions $l(\cdot)$, $c(\cdot)$ and $\lambda(\cdot)$ can also be viewed as functions on $\mathbb{R}^{JK}$. For $\boldsymbol{\beta} \in \mathscr{B}$, let $\mathbf{I}(\boldsymbol{\beta})$ denote the corresponding information matrix, which equals the Hessian matrix of $c(\cdot)$ at $\boldsymbol{\beta}$ and is a positive semidefinite symmetric $JK \times JK$ matrix. Thus $c(\cdot)$ is a convex function on $\mathscr{B}$ and $l(\cdot)$ and $\lambda(\cdot)$ are concave functions on $\mathscr{B}$.

It is assumed that $\mathscr{H}$ is nonsingular relative to the design set: If $h \in \mathscr{H}$ and $h(x_1) = \cdots = h(x_n) = 0$, then $h = 0$ on $\mathscr{X}$. Then $\boldsymbol{\tau}^t \mathbf{I}(\boldsymbol{\beta})\boldsymbol{\tau} > 0$ for $\boldsymbol{\beta} \in \mathscr{B}$ and $\boldsymbol{\tau} \in \mathscr{B}$ with $\boldsymbol{\tau} \neq 0$. Thus $c(\cdot)$ is strictly convex and $l(\cdot)$ and $\lambda(\cdot)$ are strictly concave on $\mathscr{B}$.

Let $\hat{\boldsymbol{\beta}}$ denote the maximum likelihood estimate of $\boldsymbol{\beta}$; that is, the value of $\boldsymbol{\beta} \in \mathscr{B}$ that maximizes the log-likelihood function. Then $\hat{\boldsymbol{\beta}}$ may or may not exist. Under the nonsingularity assumption on $\mathscr{H}$, if $\hat{\boldsymbol{\beta}}$ exists, then it is unique. Given $x \in \mathscr{X}$, consider the maximum likelihood estimate $\hat{f}(\cdot|x) = f(\cdot|x; \hat{\boldsymbol{\beta}})$ of $f(\cdot|x)$ and let $\hat{F}(\cdot|x)$ and $\hat{Q}(\cdot|x)$ denote the corresponding maximum likelihood estimates of $F(\cdot|x)$ and $Q(\cdot|x)$.

Similarly, $\lambda(\cdot)$ has at most one maximum on $\mathscr{B}$. It is easily seen that $\lambda(\cdot)$ does have a maximum on $\mathscr{B}$ and hence a unique maximum $\boldsymbol{\beta}^*$ on $\mathscr{B}$. Consider the function $f^*(\cdot|\,\cdot)$ on $\mathscr{X} \times \mathscr{Y}$ defined by $f^*(y|x) = f(y|x; \boldsymbol{\beta}^*)$ for $x \in \mathscr{X}$ and $y \in \mathscr{Y}$. Let $F^*(\cdot|x)$ and $Q^*(\cdot|x)$ denote the distribution function and quantile function, respectively, corresponding to $f^*(\cdot|x)$.

The knot sequences defining $\mathscr{S}$ and $\mathscr{H}$ are allowed depending on n, but it is assumed that they are σ-quasiuniform in the sense of Schumaker [(1981), page 216]: The ratios of the differences between consecutive knots are bounded away from zero and infinity uniformly in n. We make the mild assumption on the design points that there is an $M > 0$ (independent of n) such that, for n sufficiently large ($n \gg 1$),

$$M^{-1}n\int h^2(x)\,dx \leq \sum_i h^2(x_i) \leq Mn\int h^2(x)\,dx, \qquad h \in \mathscr{H}. \tag{1}$$

[The nonsingularity assumption on $\mathscr{H}$ is an immediate consequence of (1).]

Given a subinterval I of $\mathscr{X}$, let $|I|$ denote the length of I and set $N(I) = \#\{i\colon x_i \in I\}$. Under (7) below, in light of the σ-quasiuniformity of the knot sequence defining $\mathscr{H}$, a sufficient condition for (1) is that for every $\delta > 0$, there is an $M > 0$ such that for $n \gg 1$,

$$\begin{aligned} &M^{-1}n|I| \leq N(I) \\ &\qquad \leq Mn|I| \quad \text{for every subinterval } I \text{ of } \mathscr{X} \text{ such that } |I| \geq n^{\delta-1}. \end{aligned} \tag{2}$$

Let $\mathscr{T} = \mathscr{H} \otimes \mathscr{S}$ denote the tensor product of $\mathscr{H}$ and $\mathscr{S}$; that is, the linear space of real-valued functions on $\mathscr{X} \times \mathscr{Y}$ spanned by functions of the form $h \otimes s$ as h and s range over $\mathscr{H}$ and $\mathscr{S}$, respectively; here $(h \otimes s)(x, y) = h(x)s(y)$ for $x \in \mathscr{X}$ and $y \in \mathscr{Y}$. Then $\mathscr{T}$ has dimension JK and the functions $H_j \otimes B_k$, $1 \leq j \leq J$ and $1 \leq k \leq K$, form a basis of $\mathscr{T}$.

Given a real-valued function $g(\cdot|\,\cdot)$ on $\mathscr{X} \times \mathscr{Y}$, set

$$\|g(\cdot|\,\cdot)\|_\infty = \sup_{x \in \mathscr{X}} \sup_{y \in \mathscr{Y}} g(y|x).$$

Also, set $\delta_{\mathscr{T}} = \inf_{t \in \mathscr{T}} \|\log f(\cdot|\cdot) - t\|_\infty$. Under the σ-quasiuniform condition on the knot sequences, $\delta_{\mathscr{T}} \to 0$ as $J, K \to \infty$; see Schumaker [(1981), Theorem 12.8] for this result and for an upper bound to $\delta_{\mathscr{T}}$ in terms of the smoothness of $\log f(\cdot|\cdot)$.

Since $f(\cdot|\cdot)$ is continuous and positive on the compact set $\mathscr{X} \times \mathscr{Y}$, $\log f(\cdot|\cdot)$ is bounded and continuous on this set. Under (1) and the σ-quasiuniform condition on the knot sequences, it was shown in Stone (1989) that

$$\|\log f(\cdot|\cdot) - \log f^*(\cdot|\cdot)\|_\infty = O(\delta_{\mathscr{T}}). \tag{3}$$

It follows from (3) that

$$\|f(\cdot|\cdot) - f^*(\cdot|\cdot)\|_\infty = O(\delta_{\mathscr{T}}), \tag{4}$$

$$\|F(\cdot|\cdot) - F^*(\cdot|\cdot)\|_\infty = O(\delta_{\mathscr{T}}) \tag{5}$$

and

$$\|Q(\cdot|\cdot) - Q^*(\cdot|\cdot)\|_\infty = O(\delta_{\mathscr{T}}). \tag{6}$$

[In (6) the supremum is over p and x with $0 < p < 1$ and $x \in \mathscr{X}$.]

From now on, it is assumed that

$$JK = o(n^{(1/2)-\varepsilon}) \quad \text{for some } \varepsilon \in \left(0, \tfrac{1}{2}\right). \tag{7}$$

This is slightly stronger than the assumption $JK = o(\sqrt{n})$, which arises in Portnoy (1986, 1988).

In Section 2 it will be shown that $\hat{\boldsymbol{\beta}}$ exists except on an event whose probability tends to zero with n. There the asymptotic behavior of $\hat{\boldsymbol{\beta}}$ will also be determined.

In Section 3, it will be shown that

$$\hat{f}(y|x) - f^*(y|x) = O_P\left(\sqrt{JK/n}\right), \tag{8}$$

$$\int \left[\hat{f}(y|x) - f^*(y|x)\right]^2 dy = O_P(JK/n), \tag{9}$$

$$\max_{x,y} \left|\hat{f}(y|x) - f^*(y|x)\right| = O_P\left(\sqrt{JK(\log JK)/n}\right), \tag{10}$$

$$\max_{y} \left|\hat{F}(y|x) - F^*(y|x)\right| = O_P\left(\sqrt{J/n}\right) \tag{11}$$

and

$$\max_{p} \left|\hat{Q}(p|x) - Q^*(p|x)\right| = O_P\left(\sqrt{J/n}\right). \tag{12}$$

In (8), $x \in \mathscr{X}$ and $y \in \mathscr{Y}$ are fixed, while in (9), (11) and (12), $x \in \mathscr{X}$ is fixed. The order of magnitude $\sqrt{JK}$ in (8) is plausible: There are about $n/(JK)$ trials per unknown parameter β_{jk}, so the asymptotic standard deviation of the $\hat{\beta}_{jk}$'s should be proportional to $\sqrt{JK/n}$. In light of the local support of the B-splines, the asymptotic standard deviation of $\hat{f}(y|x)$ should have the same order of magnitude as that of the $\hat{\beta}_{jk}$'s.

Suppose that $\delta_{\mathscr{F}} = O(J^{-p_1} + K^{-p_2})$, where $p_1 > \frac{1}{2}$ and $p_2 > \frac{1}{2}$. Set

$$p = \frac{2}{p_1^{-1} + p_2^{-1}}, \qquad \gamma_1 = \frac{p}{p_1(2p+2)} = \frac{p_2}{p_1 + p_2 + 2p_1p_2}$$

and

$$\gamma_2 = \frac{p}{p_2(2p+2)} = \frac{p_1}{p_1 + p_2 + 2p_1p_2}.$$

Let $a_n \sim b_n$ mean that a_n/b_n is bounded away from zero and infinity as $n \to \infty$.

Suppose that $p > 1$. Choose J and K such that $J \sim n^{\gamma_1}$ and $K \sim n^{\gamma_2}$. Then $\delta_{\mathscr{F}} = O(n^{-p/(2p+2)})$. Also $JK \sim n^{1/(p+1)}$, so (7) holds. By (4) and (8),

$$\hat{f}(y|x) - f(y|x) = O_P(n^{-p/(2p+2)}); \tag{13}$$

by (4) and (9),

$$\int\left[\hat{f}(y|x) - f(y|x)\right]^2 dy = O_P(n^{-2p/(2p+2)}) = O_P(n^{-p/(p+1)}). \tag{14}$$

Choose J and K such that $J \sim (n/\log n)^{\gamma_1}$ and $K \sim (n/\log n)^{\gamma_2}$. Then

$$\delta_{\mathscr{F}} = O\big((n/\log n)^{-p/(2p+2)}\big).$$

Also $JK \sim (n/\log n)^{1/(p+1)}$, so (7) again holds. By (4) and (10),

$$\max_{x,y}\left|\hat{f}(y|x) - f(y|x)\right| = O_P\big((n/\log n)^{-p/(2p+2)}\big). \tag{15}$$

Suppose $p_1 > \frac{1}{2}$ and $p_2 > 2p_1/(2p_1 - 1)$ and set $a = p_1/[p_2(2p_1 + 1)]$. Choose J and K such that $J \sim n^{(1/(2p_1+1)}$, $K^{-1} = O(n^{-a})$ and (7) holds. Then $\delta_{\mathscr{F}} = O(n^{-p_1/(2p_1+1)})$. By (5) and (11),

$$\max_{y}\left|\hat{F}(y|x) - F(y|x)\right| = O_P(n^{-p_1/(2p_1+1)}); \tag{16}$$

by (6) and (12),

$$\max_{p}\left|\hat{Q}(p|x) - Q(p|x)\right| = O_P(n^{-p_1/(2p_1-1)}). \tag{17}$$

Under reasonable further specifications, the rates of convergence in (13)–(17) are optimal [see Stone (1980, 1982) and Hasminskii and Ibragimov (1990)]. For fixed $y \in \mathscr{Y}$, the rate of convergence $\hat{F}(y|x) - F(y|x) = O_P(n^{-p_1/(2p_1+1)})$ can be achieved by using a different estimate under the corresponding smoothness assumption on $F(y|x)$ as a function of x without having to make any smoothness assumption on $F(y|x)$ as a function of y. [Observe that $F(y|x) = E(\text{ind}(Y \le y)|X = x)$ and see Stone (1980).]

Let $J, K \to \infty$ as $n \to \infty$; let τ^* be defined as $f^*(y|x)$, $F^*(y|x)$ with y in the interior of $\mathscr{Y}$, or $Q^*(p|x)$ with $0 < p < 1$; and let $\hat{\tau}$ be defined as the corresponding maximum likelihood estimate $\hat{f}(y|x)$, $\hat{F}(y|x)$ or $\hat{Q}(y|x)$. Let $\text{ASD}(\hat{\tau})$ and $\text{SE}(\hat{\tau})$ denote the asymptotic standard deviation and standard error, respectively, of $\hat{\tau}$, as usually defined in terms of the information matrix

in large-sample parametric inference. Then, uniformly for $x \in \mathscr{X}$, $\mathrm{SE}(\hat{\tau})/\mathrm{ASD}(\hat{\tau}) = 1 + o_p(1)$ and the distributions of $(\hat{\tau} - \tau^*)/\mathrm{ASD}(\hat{\tau})$ and $(\hat{\tau} - \tau^*)/\mathrm{SE}(\hat{\tau})$ converge to the standard normal distribution as $n \to \infty$. These results will be verified in Section 4, where explicit formulas for the various asymptotic standard deviations and standard errors will be given.

According to the last result, for $0 < \alpha < 1$, $\hat{\tau} \pm z_{1-\alpha/2}\mathrm{SE}(\hat{\tau})$ is an asymptotic $100(1 - \alpha)\%$ confidence interval for τ^*; here $z_{1-\alpha/2}$ is the $(1 - \alpha/2)$th quantile of the standard normal distribution. Such confidence intervals are useful in practice, but they must be interpreted with care. Under the additional, but dubious, assumption that $(\tau - \tau^*)/\mathrm{ASD}(\hat{\tau}) = o(1)$, the confidence intervals for τ^* can be interpreted as confidence intervals for τ itself.

The arguments used in Sections 2–4, which are natural outgrowths of those developed in Stone (1985, 1986 and 1990), also apply when the fixed design is replaced by a random sample from the distribution of a random variable X having a density function that is bounded away from zero and infinity on $\mathscr{X}$ [in which case, a suitable probabilistic version of (2) is easily verified]. Alternatively, the joint density function $f(\cdot, \cdot)$ can be estimated by $\hat{f}(x, y) = f(x, y; \hat{\boldsymbol{\beta}}_1) = \exp[\Sigma_j \Sigma_k \hat{\beta}_{1jk} H_j(x) B_k(y) - c(\hat{\boldsymbol{\beta}}_1)]$, where $\hat{\boldsymbol{\beta}}_1 = (\hat{\beta}_{1jk})$ is the maximum likelihood estimate and $c(\hat{\boldsymbol{\beta}}_1)$ is the normalizing constant. The asymptotic behavior of this estimate follows from results in Barron and Sheu (1991) or from the extension of results in Stone (1990) given in Koo (1988). The corresponding estimate of the marginal density function of X is given by $f(x; \hat{\boldsymbol{\beta}}_1) = \int f(x, y; \hat{\boldsymbol{\beta}}_1)\, dy$. This leads to the alternative conditional density function estimate $\hat{f}_1(y|x) = f(x, y; \hat{\boldsymbol{\beta}}_1)/f(x; \hat{\boldsymbol{\beta}}_1)$ which has the same form as the estimate $f(y|x)$ defined above, but with an estimate $\hat{\boldsymbol{\beta}}_1$ that differs somewhat from $\hat{\boldsymbol{\beta}}$. The alternative conditional density estimate inherits the accuracy of the corresponding estimate of the joint density function. [The preceding remarks in this paragraph were suggested by a reviewer. It should be pointed out that the alternative estimate of the conditional density function achieves the rates of convergence obtained in the present paper only under an auxiliary smoothness assumption on the marginal density function of X. In the related context of nonparametric regression, Fan (1990) refers to estimates of the regression function that require such an auxiliarly assumption as being *inadmissible*: They are dominated by estimates that achieve the optimal rate of convergence without requiring the auxiliary smoothness assumption on the marginal density function of X.]

The numerical and practical aspects of logspline modelling were treated in Stone and Koo (1986) and Kooperberg and Stone (1991). The results to date clearly indicate that logspline modelling is competitive with other approaches such as kernel density estimation. A numerical investigation of logspline response modelling has yet to be carried out. Such a study would undoubtedly go beyond what is mathematically tractable. In particular, $\mathscr{X}$ and $\mathscr{Y}$ could be unbounded if linear restrictions were imposed on the tails of the various splines entering into the model. Also, it would be worthwhile to study stepwise selection of the basis functions of the model, as introduced in Smith (1982) and used successfully by Breiman and Peters (1988), Friedman and Silverman

(1989), Friedman (1991), Breiman (1989, 1991), Kooperberg and Stone (1991) and Jin (1990).

2. Parameter estimation. For $\mathbf{b} = (b_{jk}) \in \mathscr{B}$, let $|\mathbf{b}|$ denote the nonnegative square root of $\Sigma_j \Sigma_k b_{jk}^2$. In the next result, $x \in \mathscr{X}$, $y \in \mathscr{Y}$ and $0 < p < 1$; the quantities j and k in (b) and the quantity j in (c) are allowed to depend on n in an arbitrary deterministic manner.

THEOREM 1. (a) $\hat{\boldsymbol{\beta}}$ *exists except on an event whose probability tends to zero as* $n \to \infty$.

(b) $\hat{\beta}_{jk} - \beta_{jk}^* = O_P(\sqrt{JK/n})$.

(c) $(1/K)\Sigma_k(\hat{\beta}_{jk} - \beta_{jk}^*)^2 = O_P(JK/n)$.

(c) $|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*|^2 = O_P(J^2K^2/n)$.

(e) $\max_{j,k} |\hat{\beta}_{jk} - \beta_{jk}^*| = O_P(\sqrt{JK(\log JK)/n})$.

The proof of Theorem 1 is divided into a number of lemmas. For $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_K)^t \in \Theta$, let $|\boldsymbol{\theta}|$ denote the nonnegative square root of $\Sigma_k \theta_k^2$. Also, let $\|s\|_2$ and $\|s\|_\infty$ be defined in the usual manner for functions s on $\mathscr{Y}$.

LEMMA 1. *Let* $M > 0$. *Then there is an* $M_1 > 0$ *such that if* $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in \Theta$, $\|s(\cdot; \boldsymbol{\theta}_1)\|_\infty \le M$ *and* $\|s(\cdot; \boldsymbol{\theta}_2)\|_\infty \le M$, *then*

$$[C(\boldsymbol{\theta}_2) - C(\boldsymbol{\theta}_1)]^2 \le M_1 \| s(\cdot; \boldsymbol{\theta}_2) - s(\cdot; \boldsymbol{\theta}_1) \|_2^2.$$

PROOF. Since

$$C(\boldsymbol{\theta}_2) - C(\boldsymbol{\theta}_1) = \log \int e^{s(y; \boldsymbol{\theta}_2)}\, dy - \log \int e^{s(y; \boldsymbol{\theta}_1)}\, dy$$

and $0 < \text{length}(\mathscr{Y}) < \infty$, the desired result follows from the Schwarz inequality and elementary properties of the exponential and logarithm functions. □

LEMMA 2. *Let* $M > 0$. *Then there is an* $M_1 > 0$ *such that if* $\boldsymbol{\theta}, \boldsymbol{\theta}^* \in \Theta$, $\|\log f(\cdot; \boldsymbol{\theta}^*)\|_\infty \le M$ *and* $\|s(\cdot; \boldsymbol{\theta}) - s(\cdot; \boldsymbol{\theta}^*)\|_\infty \le M$, *then*

(a) $\|\log f(\cdot; \boldsymbol{\theta}) - \log f(\cdot; \boldsymbol{\theta}^*)\|_\infty \le M_1$

and

(b) $M_1^{-1}K^{-1}|\boldsymbol{\theta} - \boldsymbol{\theta}^*|^2 \le \|\log f(\cdot; \boldsymbol{\theta}) - \log f(\cdot; \boldsymbol{\theta}^*)\|_2^2 \le M_1 K^{-1} |\boldsymbol{\theta} - \boldsymbol{\theta}^*|^2$.

PROOF. By Lemma 3 of Stone (1990), there is an $M_2 > 0$ such that $|C(\boldsymbol{\theta}^*)| \le M_2$. Thus $\|s(\cdot; \boldsymbol{\theta}^*)\|_\infty \le M + M_2$ and hence $\|s(\cdot; \boldsymbol{\theta})\|_\infty \le 2M + M_2$. Consequently, there is an $M_3 > 0$ such that $|C(\boldsymbol{\theta})| \le M_3$ and hence $\|\log f(\cdot; \boldsymbol{\theta}) - \log f(\cdot; \boldsymbol{\theta}^*)\|_\infty \le M + M_2 + M_3$, which yields (a). According to (12) of Stone (1986), there is an $M_4 > 0$ such that $M_4^{-1}K^{-1}|\boldsymbol{\theta} - \boldsymbol{\theta}^*|^2 \le \|s(\cdot; \boldsymbol{\theta}) - s(\cdot; \boldsymbol{\theta}^*)\|_2^2 \le M_4 K^{-1} |\boldsymbol{\theta} - \boldsymbol{\theta}^*|^2$. Hence, by Lemma 1, there is an $M_5 > 0$ such that

$$\| \log f(\cdot; \boldsymbol{\theta}) - \log f(\cdot; \boldsymbol{\theta}^*) \|_2^2 \le M_5 K^{-1} |\boldsymbol{\theta} - \boldsymbol{\theta}^*|^2.$$

Observe that

$$\|\log f(\cdot;\boldsymbol{\theta}) - \log f(\cdot;\boldsymbol{\theta}^2)\|_2^2 = \int\Big(\sum_k (\theta_k - \theta_k^*)B_k(y) - [C(\boldsymbol{\theta}) - C(\boldsymbol{\theta}^*)]\Big)^2 dy$$

$$= \int\Big(\sum_k \{\theta_k - \theta_k^* - [C(\boldsymbol{\theta}) - C(\boldsymbol{\theta}^*)]\}B_k(y)\Big)^2 dy.$$

Thus, by (12) of Stone (1986), there is an $M_6 > 0$ such that

$$\|\log f(\cdot;\boldsymbol{\theta}) - \log f(\cdot;\boldsymbol{\theta}^*)\|_2^2 \geq M_6^{-1}K^{-1}\sum_k \{\theta_k - \theta_k^* - [C(\boldsymbol{\theta}) - C(\boldsymbol{\theta}^*)]\}^2.$$

Now $\boldsymbol{\theta}, \boldsymbol{\theta}^* \in \Theta$, so $\sum_k(\theta_k - \theta_k^*) = 0$ and hence

$$\sum_k \{\theta_k - \theta_k^* - [C(\boldsymbol{\theta}) - C(\boldsymbol{\theta}^*)]\}^2 \geq \sum_k (\theta_k - \theta_k^*)^2 = |\boldsymbol{\theta} - \boldsymbol{\theta}^*|^2.$$

Consequently, $\|\log f(\cdot;\boldsymbol{\theta}) - \log f(\cdot;\boldsymbol{\theta})\|_2^2 \geq M_6^{-1}K^{-1}|\boldsymbol{\theta} - \boldsymbol{\theta}^*|^2$. Therefore, (b) is valid. □

LEMMA 3. $\|s(\cdot;\mathbf{h}(\cdot;\boldsymbol{\beta}_2)) - s(\cdot;\mathbf{h}(\cdot;\boldsymbol{\beta}_1))\|_\infty \leq |\boldsymbol{\beta}_2 - \boldsymbol{\beta}_1|$ *for* $\boldsymbol{\beta}_1, \boldsymbol{\beta}_2 \in \mathscr{B}$.

PROOF. By the properties of B-splines, $\sum_j\sum_k H_j^2(x)B_k^2(y) \leq \sum_j\sum_k H_j(x)B_k(y) = 1$. Thus, by the Schwarz inequality,

$$[s(y;\mathbf{h}(x;\boldsymbol{\beta}_2)) - s(y;\mathbf{h}(x;\boldsymbol{\beta}_1))]^2 = \Big(\sum_j\sum_k (\beta_{2jk} - \beta_{1jk})H_j(x)B_k(y)\Big)^2$$

$$\leq |\boldsymbol{\beta}_2 - \boldsymbol{\beta}_1|^2,$$

where $\boldsymbol{\beta}_1 = (\beta_{1jk})$ and $\boldsymbol{\beta}_2 = (\beta_{2jk})$. □

The next result follows from (1), the σ-quasiuniformity of the knot sequence defining $\mathscr{H}$ and (viii) of de Boor [(1978), page 155] [see the proof of (12) of Stone (1986)].

LEMMA 4. *There is an* $M > 0$ *such that, for* $n \gg 1$,

$$M^{-1}nJ^{-1}|\boldsymbol{\beta}_2 - \boldsymbol{\beta}_1|^2 \leq \sum_i |\mathbf{h}(x_i;\boldsymbol{\beta}_2) - \mathbf{h}(x_i;\boldsymbol{\beta}_1)|^2$$

$$\leq MnJ^{-1}|\boldsymbol{\beta}_2 - \boldsymbol{\beta}_1|^2, \qquad \boldsymbol{\beta}_1, \boldsymbol{\beta}_2 \in \mathscr{B}.$$

Let $\boldsymbol{\beta} \in \mathscr{B}$. Then, by a direct computation,

$$\boldsymbol{\tau}^t\mathbf{I}(\boldsymbol{\beta})\boldsymbol{\tau} = \sum_i \int [s(y;\mathbf{h}(x_i;\boldsymbol{\tau})) - a_i]^2 f(y|x_i;\boldsymbol{\beta})\,dy, \qquad \boldsymbol{\tau} \in \mathscr{B}, \tag{18}$$

where $a_i = \int s(y; \mathbf{h}(x_i; \tau)) f(y|x_i; \beta)\, dy$. Let $\beta \in \mathscr{B}$. Then $(d/dt)\lambda(\beta^* + t(\beta - \beta^*))|_{t=0} = 0$ and

$$\frac{d^2}{dt^2}\lambda(\beta^* + t(\beta - \beta^*)) = -(\beta - \beta^*)^t \mathbf{I}(\beta^* + t(\beta - \beta^*))(\beta - \beta^*).$$

Thus, by (18),

$$\lambda(\beta) - \lambda(\beta^*) = -\int_0^1 (1 - t)\Bigg(\sum_i \int \big[s(y; \mathbf{h}(x_i; \beta - \beta^*)) - a_i(t) \big]^2 \times f(y|x_i; \beta + t(\beta - \beta^*))\, dy \Bigg) dt, \tag{19}$$

where $a_i(t) = \int s(y; \mathbf{h}(x_i; \beta - \beta^*)) f(y|x_i; \beta + t(\beta - \beta^*))\, dy$ for $1 \le i \le n$.

Choose $\varepsilon \in (0, \frac{1}{2})$ satisfying (7): $JK = o(n^{(1/2)-\varepsilon})$.

LEMMA 5. *There is a $\delta > 0$ such that if $n \gg 1$, $\beta \in \mathscr{B}$ and $|\beta - \beta^*| = n^{\varepsilon} JK/\sqrt{n}$, then $\lambda(\beta) - \lambda(\beta^*) \le -\delta n^{2\varepsilon} JK$.*

PROOF. It follows from (3), (19), Lemma 2(a) and Lemma 3 [see the proof of Lemma 4 in Stone (1990)] that there is a $\delta_1 > 0$ such that

$$\lambda(\beta) - \lambda(\beta^*) \le -\delta_1 \sum_i \int \big[\log f(y|x_i; \beta) - \log f(y|x_i; \beta^*) \big]^2 dy.$$

By lemmas 2(b), 3 and 4, there is a $\delta_2 > 0$ such that

$$\sum_i \int \big[\log f(y|x_i; \beta) - \log f(y|x_i; \beta^*) \big]^2 dy \ge \delta_2 n J^{-1} K^{-1} |\beta - \beta^*|^2.$$

Consequently, the desired result holds. □

LEMMA 6. *Let $\delta > 0$. Then there is a $\delta_1 > 0$ such that if $n \gg 1$, $\beta \in \mathscr{B}$ and $|\beta - \beta^*| \le n^{\varepsilon} JK/\sqrt{n}$, then*

$$P\Big(l(\beta) - l(\beta^*) - [\lambda(\beta) - \lambda(\beta^*)] \ge \frac{\delta}{2} n^{2\varepsilon} JK \Big) \le \exp\big(-\delta_1 n^{2\varepsilon} JK \big).$$

PROOF. Write $l(\beta) - l(\beta^*) - [\lambda(\beta) - \lambda(\beta^*)] = \sum_i Z_i$, where

$$Z_i = \log f(Y_i|x_i; \beta) - \log f(Y_i|x_i; \beta^*) - E\big[\log f(Y_i|x_i; \beta) - \log f(Y_i|x_i; \beta^*) \big].$$

It follows from (3), Lemma 2(a) and Lemma 3 that there is an $M_1 > 0$ such

that $P(|Z_i| \le M_1) = 1$ for $1 \le i \le n$. Observe that, for $1 \le i \le n$, $EZ_i = 0$ and

$$\operatorname{var}(Z_i) \le E\left\{\left[\log f(Y_i|x_i;\beta) - \log f(Y_i|x_i;\beta^*)\right]^2\right\}$$

$$= \int \left[\log f(y|x_i;\beta) - \log f(y|x_i;\beta^*)\right]^2 f(y|x_i)\,dy.$$

We now conclude from Lemmas 2(b), 3 and 4 and the boundedness of $f(\cdot|\cdot)$ that there is an $M_2 > 0$ such that $\sum_i \operatorname{var}(Z_i) \le M_2 n^{2\varepsilon} JK$. The desired result now follows from Bernstein's inequality [see (2.13) in Hoeffding, (1963)]. □

LEMMA 7. *Let $\delta > 0$. Then there is a $\delta_1 > 0$ such that*

$$|l(\beta_2) - l(\beta_1) - [\lambda(\beta_2) - \lambda(\beta_1)]| \le \frac{\delta}{2} n^{2\varepsilon} JK$$

for $n \gg 1$, $\beta_1, \beta_2 \in \mathscr{B}$, $|\beta_1 - \beta^*| = n^{\varepsilon} JK/\sqrt{n}$, $|\beta_2 - \beta^*| = n^{\varepsilon} JK/\sqrt{n}$ and $|\beta_2 - \beta_1| \le \delta_1 n^{2\varepsilon - 1} JK$.

PROOF. Observe that

$$|l(\beta_2) - l(\beta_1) - [\lambda(\beta_2) - \lambda(\beta_1)]| \le 2n \|\log f(\cdot|\cdot;\beta_2) - \log f(\cdot|\cdot;\beta_1)\|_\infty.$$

By Lemmas 1 and 3, there is an $M_1 > 0$ such that $\|\log f(\cdot|\cdot;\beta_2) - \log f(\cdot|\cdot;\beta_1)\|_\infty \le M_1|\beta_2 - \beta_1|$. Thus the desired result is valid. □

The diameter of a subset B of $\mathscr{B}$ is defined as $\sup\{|\beta_2 - \beta_1|: \beta_1, \beta_2 \in B\}$. The next result is easily established by considering suitable inscribed and circumscribed JK-dimensional cubes.

LEMMA 8. *Let $\delta_1 > 0$. Then there is an $M > 0$ such that, for $n \gg 1$,*

$$\{\beta \in \mathscr{B}: |\beta - \beta^*| = n^{\varepsilon} JK/\sqrt{n}\}$$

can be covered by $\exp(MJK(\log n))$ subsets of $\mathscr{B}$ each having diameter at most $\delta_1 n^{2\varepsilon - 1} JK$.

LEMMA 9. (a) *$\hat{\beta}$ exists except on an event whose probability tends to zero as $n \to \infty$.*

(b) $|\hat{\beta} - \beta^*| = O_P(n^{\varepsilon} JK/\sqrt{n})$.

PROOF. Set $\mathscr{B}_1 = \{\beta \in \mathscr{B}: |\beta - \beta^*| \le n^{\varepsilon} JK/\sqrt{n}\}$. Then $\mathscr{B}_1$ is a compact set whose boundary relative to $\mathscr{B}$ is $\mathscr{B}_2 = \{\beta \in \mathscr{B}: |\beta - \beta^*| = n^{\varepsilon} JK\sqrt{n}\}$. By Lemma 5 there is a $\delta > 0$ such that $\lambda(\beta) - \lambda(\beta^*) \le -\delta n^{2\varepsilon} JK$ for $\beta \in \mathscr{B}_2$. Thus it follows from Lemmas 6–8 that, except on an event whose probability tends to zero as $n \to \infty$, $l(\beta) < l(\beta^*)$ for $\beta \in \mathscr{B}_2$, so $l(\cdot)$ has a local maximum in the interior of $\mathscr{B}_1$ relative to $\mathscr{B}$. The desired results now follows from the strict concavity of $l(\cdot)$ on $\mathscr{B}$. □

The next result follows from (4), (7) and Lemmas 2a, 3 and 9.

LEMMA 10. *There is a positive constant M such that, except on an event whose probability tends to zero as $n \to \infty$, $M^{-1} \le f(\cdot \mid \cdot\,; \boldsymbol{\beta}^* + t(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)) \le M$ for $0 \le t \le 1$.*

LEMMA 11. *There is an $M > 0$ such that, for $n \gg 1$, $\boldsymbol{\beta} \in \mathscr{B}$ and $\boldsymbol{\tau} \in \mathscr{B}$,*

$$M^{-1} n J^{-1} K^{-1} |\boldsymbol{\tau}|^2 \min f(\cdot \mid \cdot\,; \boldsymbol{\beta}) \le \boldsymbol{\tau}^t \mathbf{I}(\boldsymbol{\beta}) \boldsymbol{\tau} \le M n J^{-1} K^{-1} |\boldsymbol{\tau}|^2 \max f(\cdot \mid \cdot\,; \boldsymbol{\beta}).$$

PROOF. Set min = min $f(\cdot \mid \cdot\,; \boldsymbol{\beta})$ and max = max $f(\cdot \mid \cdot\,; \boldsymbol{\beta})$. Using (1), (18) and (12) of Stone (1986), we see that there are positive numbers δ_1, δ_2 and δ such that, for $n \gg 1$, $\boldsymbol{\beta} \in \mathscr{B}$ and $\boldsymbol{\tau} \in \mathscr{B}$,

$$\begin{aligned}
\boldsymbol{\tau}^t \mathbf{I}(\boldsymbol{\beta}) \boldsymbol{\tau} &\ge \delta_1(\min) \sum_i \int \left[s(y; \mathbf{h}(x_i; \boldsymbol{\tau})) - a_i \right]^2 dy \\
&= \delta_1(\min) \sum_i \int \left(\sum_k \left[h_k(x_i; \boldsymbol{\tau}) - a_i \right] B_k(y) \right)^2 dy \\
&\ge \delta_2(\min) K^{-1} \sum_i \sum_k \left[h_k(x_i; \boldsymbol{\tau}) - a_i \right]^2 \\
&= \delta_2(\min) K^{-1} \sum_i \sum_k \left(\sum_j (\tau_{jk} - a_i) H_j(x_i) \right)^2 \\
&\ge \delta_2(\min) K^{-1} \sum_i \sum_k \left(\sum_j \tau_{jk} H_j(x_i) \right)^2 \\
&\ge \delta(\min) n J^{-1} K^{-1} |\boldsymbol{\tau}|^2.
\end{aligned}$$

Similarly, we conclude from (18) that $\boldsymbol{\tau}^t \mathbf{I}(\boldsymbol{\beta}) \boldsymbol{\tau} \le (\max) \sum_i \int [s(y\,; \mathbf{h}(x_i; \boldsymbol{\tau}))]^2 dy$ and hence that there is an $M > 0$ such that, for $n \gg 1$, $\boldsymbol{\beta} \in \mathscr{B}$ and $\boldsymbol{\tau} \in \mathscr{B}$, $\boldsymbol{\tau}^t \mathbf{I}(\boldsymbol{\beta}) \boldsymbol{\tau} \le M(\max) n J^{-1} K^{-1} |\boldsymbol{\tau}|^2$. Thus the desired result is valid. □

Let $\mathbf{S}(\boldsymbol{\beta}) \in \mathscr{B}$ denote the score at $\boldsymbol{\beta}$; that is, the JK-dimensional matrix the entry in row j and column k of which is

$$\frac{\partial l(\boldsymbol{\beta})}{\partial \beta_{ij}} = \sum_i H_j(x_j) \left(B_k(Y_i) - \frac{\partial C}{\partial \theta_k}(\mathbf{h}(x_i; \boldsymbol{\beta})) \right).$$

[In computing $\partial C(\boldsymbol{\theta}) / \partial \theta_k$, we let $\boldsymbol{\theta}$ range over $\mathbb{R}^K$.] Then $E\mathbf{S}(\boldsymbol{\beta}^*) = \mathbf{0}$ and

$$\begin{aligned}
E|\mathbf{S}(\boldsymbol{\beta}^*)|^2 &= \sum_i \sum_j \sum_k H_j^2(x_i) \operatorname{var}(B_k(Y_i)) \\
&\le \sum_i \left(\sum_j H_j^2(x_i) \right) E\left(\sum_k B_k^2(Y_i) \right) \le n.
\end{aligned}$$

Consequently, the following result is valid.

LEMMA 12. $|\mathbf{S}(\boldsymbol{\beta}^*)|^2 = O_P(n)$.

The maximum likelihood equation $\mathbf{S}(\hat{\boldsymbol{\beta}}) = \mathbf{0}$ for $\hat{\boldsymbol{\beta}}$ can be written as

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

Thus it can be rewritten as $\mathbf{D}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*) = \mathbf{S}(\boldsymbol{\beta}^*)$, where $\mathbf{D}$ is the $JK \times JK$ matrix given by $\mathbf{D} = \int_0^1 \mathbf{I}(\boldsymbol{\beta}^* + t(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*))\, dt$.

LEMMA 13. $|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*| = O_P(JK/\sqrt{n})$.

PROOF. According to the maximum likelihood equation for $\hat{\boldsymbol{\beta}}$,

$$(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)^t \mathbf{D}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*) = (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)^t \mathbf{S}(\boldsymbol{\beta}^*). \tag{20}$$

It follows from Lemma 12 that

$$(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)^t \mathbf{S}(\boldsymbol{\beta}^*) = O_P(|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*|\sqrt{n}). \tag{21}$$

According to Lemmas 10 and 11, there is an $M > 0$ such that, except on an event whose probability tends to zero as $n \to \infty$,

$$(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^t \mathbf{D}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \geq M^{-1} n J^{-1} K^{-1} |\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*|^2. \tag{22}$$

We conclude from (20)–(22) that $nJ^{-1}K^{-1}|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*|^2 = O_P(|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*|\sqrt{n})$, which yields the desired result. □

Let $\mathrm{VC}(\mathbf{S}(\boldsymbol{\beta}^*))$ denote the variance–covariance matrix of $\mathbf{S}(\boldsymbol{\beta}^*)$.

LEMMA 14. *There is an $M > 0$ such that, for $n \gg 1$ and $\boldsymbol{\tau} \in \mathscr{B}$,*

$$M^{-1} n J^{-1} K^{-1} |\boldsymbol{\tau}|^2 \leq \boldsymbol{\tau}^t \mathrm{VC}(\mathbf{S}(\boldsymbol{\beta}^*))\boldsymbol{\tau} \leq M n J^{-1} K^{-1} |\boldsymbol{\tau}|^2.$$

PROOF. Since

$$\boldsymbol{\tau}^t\, \mathrm{VC}(\mathbf{S}(\boldsymbol{\beta}^*))\boldsymbol{\tau} = \sum_i \int \big[s(y; \mathbf{h}(x_i; \boldsymbol{\tau})) - a_i\big]^2 f(y|x_i)\, dy, \qquad \boldsymbol{\tau} \in \mathscr{B}, \tag{23}$$

where $a_i = \int s(y; \mathbf{h}(x_i; \boldsymbol{\tau})) f(y|x_i)\, dy$ for $1 \leq i \leq n$, the desired result follows from the argument used to prove Lemma 11. □

Let $\boldsymbol{\beta} \in \mathscr{B}$. Then there is a positive semidefinite symmetric $JK \times JK$ matrix $[\mathbf{I}(\boldsymbol{\beta})]^-$ having range $\mathscr{B}$ such that $\mathbf{I}(\boldsymbol{\beta})[\mathbf{I}(\boldsymbol{\beta})]^- \boldsymbol{\tau} = [\mathbf{I}(\boldsymbol{\beta})]^- \mathbf{I}(\boldsymbol{\beta})\boldsymbol{\tau} = \boldsymbol{\tau}$ for $\boldsymbol{\tau} \in \mathscr{B}$. [Consider the orthogonal diagonalization of $\mathbf{I}(\boldsymbol{\beta})$.] The matrix $[\mathbf{I}(\boldsymbol{\beta})]^-$ is referred to as the generalized inverse of $\mathbf{I}(\boldsymbol{\beta})$.

Let $\hat{\boldsymbol{\varphi}} \in \mathscr{B}$ be the approximation to $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*$ defined by $\mathbf{I}(\boldsymbol{\beta}^*)\hat{\boldsymbol{\varphi}} = \mathbf{S}(\boldsymbol{\beta}^*)$. Then $\hat{\boldsymbol{\varphi}} = [\mathbf{I}(\boldsymbol{\beta}^*)]^- \mathbf{S}(\boldsymbol{\beta}^*)$, $E\hat{\boldsymbol{\varphi}} = \mathbf{0}$ and

$$\mathrm{VC}(\hat{\boldsymbol{\varphi}}) = [\mathbf{I}(\boldsymbol{\beta}^*)]^- \mathrm{VC}(\mathbf{S}(\boldsymbol{\beta}^*))[\mathbf{I}(\boldsymbol{\beta}^*)]^-. \tag{24}$$

The next result follows from (24) and Lemmas 10, 11 and 14. (Consider a symmetric square root of $[\mathbf{I}(\boldsymbol{\beta}^*)]^-$.)

LEMMA 15. *There is an $M > 0$ such that if $n \gg 1$, then*

$$M^{-1}n^{-1}JK|\boldsymbol{\tau}|^2 \le \boldsymbol{\tau}^t[\mathbf{I}(\boldsymbol{\beta}^*)]^-\boldsymbol{\tau} \le Mn^{-1}JK|\boldsymbol{\tau}|^2, \qquad \boldsymbol{\tau} \in \mathscr{B}, \tag{25}$$

and

$$M^{-1}n^{-1}JK|\boldsymbol{\tau}|^2 \le \mathrm{var}(\boldsymbol{\tau}^t\hat{\boldsymbol{\varphi}}) \le Mn^{-1}JK|\boldsymbol{\tau}|^2, \qquad \boldsymbol{\tau} \in \mathscr{B}. \tag{26}$$

Given $y \in \mathscr{Y}$ and $\boldsymbol{\beta} \in \mathscr{B}$, let $\mathbf{G}(y|x;\boldsymbol{\beta}) \in \mathscr{B}$ denote the gradient of $\log f(y|x;\cdot)$ at $\boldsymbol{\beta}$: the $J \times K$ matrix the entry in row j and column k of which is

$$H_j(x)\left(B_k(y) - \frac{\partial C}{\partial \theta_k}(\mathbf{h}(x;\boldsymbol{\beta}))\right).$$

It follows from Lemma 3 of Stone (1990) and Lemma 10 that

$$\max_k \left|\frac{\partial C}{\partial \theta_k}(\mathbf{h}(x;\boldsymbol{\beta}^*))\right| = O(K^{-1}). \tag{27}$$

Thus there is an $M > 0$ such that

$$|\mathbf{G}(y|x;\boldsymbol{\beta}^*)| \le M, \qquad x \in \mathscr{X} \text{ and } y \in \mathscr{Y}. \tag{28}$$

Observe that $\mathbf{S}(\boldsymbol{\beta}) = \sum_i \mathbf{G}(Y_i|x_i;\boldsymbol{\beta})$ for $\boldsymbol{\beta} \in \mathscr{B}$ and hence that

$$\hat{\varphi}_{jk} = \sum_i \left([\mathbf{I}(\boldsymbol{\beta}^*)]^- \mathbf{G}(Y_i|x_i;\boldsymbol{\beta}^*)\right)_{jk}. \tag{29}$$

The quantities j and k in (a) of the next result and the quantity j in (b) and (d) are allowed to depend on n in an arbitrary deterministic manner.

LEMMA 16. (a) $\hat{\varphi}_{jk} = O_P(\sqrt{JK/n})$.
(b) $(1/K)\sum_k \hat{\varphi}_{jk}^2 = O_P(JK/n)$.
(c) $|\hat{\boldsymbol{\varphi}}|^2 = O_P(J^2K^2/n)$.
(d) $\max_k |\hat{\varphi}_{jk}| = O_P(\sqrt{JK(\log K)/n})$.
(e) $\max_{j,k} |\hat{\varphi}_{jk}| = O_P(\sqrt{JK(\log JK)/n})$.

PROOF. Now $\max_{j,k} E\hat{\varphi}_{jk}^2 = \max_{j,k} \mathrm{var}(\hat{\varphi}_{jk}) = O(JK/n)$ by (26), so (a)–(c) hold. By (25) and (28), there is an $M > 0$ such that, for $n \gg 1$,

$$\max_{x,y} \left|\boldsymbol{\tau}^t[\mathbf{I}(\boldsymbol{\beta}^*)]^- \mathbf{G}(y|x;\boldsymbol{\beta}^*)\right| \le Mn^{-1}JK|\boldsymbol{\tau}|, \qquad \boldsymbol{\tau} \in \mathscr{B},$$

and hence

$$\max_{x,y}\max_{j,k}\left|\left([\mathbf{I}(\boldsymbol{\beta}^*)]^{-}\mathbf{G}(y|x;\boldsymbol{\beta}^*)\right)_{jk}\right| \le MJK/n.$$

Parts (d) and (e) now follow from (7), (29) and Bernstein's inequality. □

LEMMA 17. $\left|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^* - \hat{\boldsymbol{\varphi}}\right|^2 = O_P(n^{-2}J^3K^2(\log JK))$.

PROOF. It follows from the maximum likelihood equation that

$$\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^* = \hat{\boldsymbol{\varphi}} - [\mathbf{I}(\boldsymbol{\beta}^*)]^{-}[\mathbf{D} - \mathbf{I}(\boldsymbol{\beta}^*)](\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*).$$

By (25),

$$\left|[\mathbf{I}(\boldsymbol{\beta}^*)]^{-}[\mathbf{D} - \mathbf{I}(\boldsymbol{\beta}^*)](\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)\right|^2 = O_P\left(n^{-2}(JK)^2\left|[\mathbf{D} - \mathbf{I}(\boldsymbol{\beta}^*)](\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)\right|^2\right).$$

The entry in row (j, k) and column (j', k') of $\mathbf{D} - \mathbf{I}(\boldsymbol{\beta}^*)$ can be written as

$$\sum_{j''}\sum_{k''} A_{jkj'j'j''k''}\left(\hat{\beta}_{j''k''} - \beta^*_{j''k''}\right),$$

where

$$A_{jkj'k'j''k''} = \sum_i \int_0^1 (1-t)H_j(x_i)H_{j'}(x_i)H_{j''}(x_i) \times \frac{\partial^3 C}{\partial\theta_k\,\partial\theta_{k'}\,\partial\theta_{k''}}\left(\mathbf{h}\left(x_i;\boldsymbol{\beta}^* + t(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)\right)\right).$$

Thus the entry in row (j, k) of $[\mathbf{D} - \mathbf{I}(\boldsymbol{\beta}^*)](\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)$ is

$$\sum_{j'}\sum_{k'}\sum_{j''}\sum_{k''} A_{jkj'k'j''k''}\left(\hat{\beta}_{j'k'} - \beta^*_{j'k'}\right)\left(\hat{\beta}_{j''k''} - \beta^*_{j''k''}\right),$$

which is dominated in absolute value by

$$\max_{j,k}\left(\hat{\beta}_{jk} - \beta^*_{jk}\right)^2 \sum_{j'}\sum_{k'}\sum_{j''}\sum_{k''}\left|A_{jkj'k'j''k''}\right|.$$

There is a positive integer J_0 such that $A_{jkj'k'j''k''} = 0$ unless $|j' - j| \le J_0$ and $|j'' - j| \le J_0$. Thus, by (8) of Stone (1989), there is an $M_1 > 0$ such that

$$\sum_{j'}\sum_{j''}\left|A_{jkj'k'j''k''}\right| \le M_1 J^{-1} n \sup_{x\in\mathscr{X}}\max_{0\le t\le 1}\left|\frac{\partial^3 C}{\partial\theta_k\,\partial\theta_{k'}\,\partial\theta_{k''}}\left(\mathbf{h}\left(x;\boldsymbol{\beta}^* + t(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)\right)\right)\right|.$$

Consequently [see the proof of Lemma 15 of Stone (1990)],

$$\sum_j\sum_k\left(\sum_{j'}\sum_{k'}\sum_{j''}\sum_{k''}\left|A_{jkj'k'j''k''}\right|\right)^2 = O_P(nK^{-1})$$

and hence $|[\mathbf{D} - I(\boldsymbol{\beta}^*)](\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)|^2 = O_P(nK^{-1} \max_{j,k}(\hat{\beta}_{jk} - \beta_{jk}^*)^4)$. Therefore,

$$\left|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^* - \boldsymbol{\varphi}\right|^2 = O_P\left(n^{-1}J^2K \max_{j,k}\left(\hat{\beta}_{jk} - \beta_{jk}^*\right)^4\right).$$

We now conclude from Lemma 16(e) that

$$\max_{j,k}\left(\hat{\beta}_{jk} - \beta_{jk}^*\right)^2 = O_P\left(n^{-1}JK(\log JK) + n^{-1}J^2K \max_{j,k}\left(\hat{\beta}_{jk} - \beta_{jk}^*\right)^4\right).$$

Thus $\max_{j,k}(\hat{\beta}_{jk} - \beta_{jk}^*)^2 = O_P(n^{-1}JK(\log JK))$ by (7) and Lemma 9(b), so

$$\left|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^* - \hat{\boldsymbol{\varphi}}\right|^2 = O_P(n^{-2}J^3K^2(\log JK)). \qquad \square$$

Theorem 1(a) coincides with Lemma 9(a) and the remaining parts of Theorem 1 follow from (7) and Lemmas 16 and 17.

3. Functional estimation. In this section, (8)–(12) will be verified. To help the reader follow the details, we first indicate the proof of (10). It follows from (3) that $\|\log f^*(\cdot|\cdot)\|_\infty = O(1)$. Thus, to prove (10), it suffices to show that

$$\max_{x,y}\left|\log f\left(y|x;\hat{\boldsymbol{\beta}}\right) - \log f(y|x;\boldsymbol{\beta}^*)\right| = O_P\left(\sqrt{JK(\log JK)/n}\right). \tag{30}$$

Let $\nabla C(\boldsymbol{\theta})$ denote the gradient vector of $C(\cdot)$ at $\boldsymbol{\theta}$, whose kth entry $\partial C(\boldsymbol{\theta})/\partial\theta_k$ is computed as $\boldsymbol{\theta}$ ranges over $\mathbb{R}^K$. Observe that

$$\begin{aligned}&\log f\left(y|x;\hat{\boldsymbol{\beta}}\right) - \log f(y|x;\boldsymbol{\beta}^*)\\ &\quad= [\mathbf{G}(y|x;\boldsymbol{\beta}^*)]^t\hat{\boldsymbol{\varphi}} + [\mathbf{G}(y|x;\boldsymbol{\beta}^*)]^t\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^* - \hat{\boldsymbol{\varphi}}\right) - R,\end{aligned} \tag{31}$$

where

$$R = C\left(\mathbf{h}\left(x;\hat{\boldsymbol{\beta}}\right)\right) - C(\mathbf{h}(x;\boldsymbol{\beta}^*)) - [\nabla C(\mathbf{h}(x;\boldsymbol{\beta}^*))]^t\mathbf{h}\left(x;\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\right). \tag{32}$$

According to Lemma 18(c) below,

$$\max_{x,y}\left|[\mathbf{G}(y|x;\boldsymbol{\beta}^*)]^t\hat{\boldsymbol{\varphi}}\right| = O_P\left(\sqrt{JK(\log JK)/n}\right); \tag{33}$$

according to (7), (28) and Lemma 17,

$$\max_{x,y}\left|[\mathbf{G}(y|x;\boldsymbol{\beta}^*)]^t\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^* - \hat{\boldsymbol{\varphi}}\right)\right| = o_P\left(\sqrt{JK/n}\right); \tag{34}$$

and, according to (7) and Lemma 19 below,

$$\max_x|R| = o_P\left(\sqrt{JK/n}\right). \tag{35}$$

The desired result (30) follows from (31)–(35).

LEMMA 18. (a) $[\mathbf{G}(y|x;\boldsymbol{\beta}^*)]^t\hat{\boldsymbol{\varphi}} = O_P(\sqrt{JK/n})$.
(b) $\int\{[\mathbf{G}(y|x;\boldsymbol{\beta}^*)]^t\hat{\boldsymbol{\varphi}}\}^2\,dy = O_P(JK/n)$.
(c) $\max_{x,y}|[\mathbf{G}(y|x;\boldsymbol{\beta}^*)]^t\hat{\boldsymbol{\varphi}}| = O_P(\sqrt{JK(\log JK)/n})$.

PROOF. Part (a) follows from (7) and Lemma 17. In order to verify (b), choose $x \in \mathscr{X}$. There are at most J_0 values of j such that $H_j(x) > 0$. For any such j,

$$\left(\frac{1}{K}\sum_k \left|\hat{\beta}_{jk} - \beta^*_{jk} - \hat{\varphi}_{jk}\right|\right)^2 \le \frac{1}{K}\sum_k \left(\hat{\beta}_{jk} - \beta^*_{jk} - \hat{\varphi}_{jk}\right)^2 \le \frac{1}{K}\sum_j \sum_k \left(\hat{\beta}_{jk} - \beta^*_{jk} - \hat{\varphi}_{jk}\right)^2.$$

Thus, by (7) and Lemma 17,

$$\max_j \frac{1}{K}\sum_k \left|\hat{\beta}_{jk} - \beta^*_{jk} - \hat{\varphi}_{jk}\right| = O_P\left(n^{-1}J^{3/2}K^{1/2}\sqrt{\log JK}\right) = o_P\left(\sqrt{J/n}\right),$$

so (b) holds. Part (c) follows from (27) and Lemma 16(a, b); (d) follows from (27), Lemma 16(b) and (12) of Stone (1986); (e) follows from (27) and Lemma 16(e); and (f) follows from (7) and (e). □

LEMMA 19. *The following is valid*:

$$\max_x \left|C\left(\mathbf{h}\left(x;\hat{\boldsymbol{\beta}}\right)\right) - C\left(\mathbf{h}\left(x;\boldsymbol{\beta}^*\right)\right) - \left[\nabla C\left(\mathbf{h}\left(x;\boldsymbol{\beta}^*\right)\right)\right]^t \mathbf{h}\left(x;\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\right)\right| = O_P\left(JK(\log JK)/n\right).$$

PROOF. Observe that

$$C\left(\mathbf{h}\left(x;\hat{\boldsymbol{\beta}}\right)\right) - C\left(\mathbf{h}\left(x;\boldsymbol{\beta}^*\right)\right) = \left[\nabla C\left(\mathbf{h}\left(x;\boldsymbol{\beta}^*\right)\right)\right]^t \mathbf{h}\left(x;\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\right) + \int_0^1 (1-t)\left(\int \left[s\left(y;\mathbf{h}\left(x;\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\right)\right) - a(t)\right]^2 \times f\left(y|x;\boldsymbol{\beta}^* + t\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\right)\right) dy\right) dt,$$

where $a(t) = \int s(y;\mathbf{h}(x;\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)) f(y|x;\boldsymbol{\beta}^* + t(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*))\, dy$ [see (18)]. The desired result now follows from Lemma 10, Theorem 1(e), and (12) of Stone (1986). □

LEMMA 20. (a) $\max_{x,y} |\log f(y|x;\hat{\boldsymbol{\beta}}) - \log f(y|x;\boldsymbol{\beta}^*) - [\mathbf{G}(y|x;\boldsymbol{\beta}^*)]^t\hat{\boldsymbol{\varphi}}| = o_P(\sqrt{JK/n})$.

(b)
$$\max_{x,y}\left|F\left(y|x;\hat{\boldsymbol{\beta}}\right) - F\left(y|x;\boldsymbol{\beta}^*\right) - \int_{y'\le y} f\left(y'|x;\boldsymbol{\beta}^*\right)\left[\mathbf{G}\left(y'|x;\boldsymbol{\beta}^*\right)\right]^t\hat{\boldsymbol{\varphi}}\, dy'\right| = o_P\left(\sqrt{J/n}\right).$$

PROOF. Part (a) follows from (31), (32), (34) and (35). Since

$$F(y|x;\hat{\boldsymbol{\beta}}) - F(y|x;\boldsymbol{\beta}^*) = \int_0^y (e^{\log f(y'|x;\hat{\boldsymbol{\beta}}) - \log f(y'|x;\boldsymbol{\beta}^*)} - 1) f(y'|x;\boldsymbol{\beta}^*)\, dy',$$

(b) follows from (a) and Lemma 18(c). □

LEMMA 21.

$$\max_y \left| \int_{y' \le y} f(y'|x;\boldsymbol{\beta}^*)[\mathbf{G}(y'|x;\boldsymbol{\beta}^*)]^t \hat{\boldsymbol{\varphi}}\, dy' \right| = O_P(\sqrt{J/n}).$$

PROOF. Let $\boldsymbol{\tau}$ be the member of $\mathscr{B}$, the entry in row j and column k of which is

$$H_j(x) \frac{\partial C}{\partial \theta_k}(\mathbf{h}(x;\boldsymbol{\beta}^*)).$$

Then $E(\boldsymbol{\tau}^t\hat{\boldsymbol{\varphi}}) = 0$ since $E\hat{\boldsymbol{\varphi}} = \mathbf{0}$ and $\operatorname{var}(\boldsymbol{\tau}^t\hat{\boldsymbol{\varphi}}) = O(J/n)$ by (26) since $|\boldsymbol{\tau}|^2 = O(K^{-1})$ by (27), so $\boldsymbol{\tau}^t\hat{\boldsymbol{\varphi}} = O_P(\sqrt{J/n})$. Thus to prove the desired result it suffices to verify that

$$\max_y \left| \sum_j H_j(x) \sum_k \hat{\varphi}_{jk} \int_{y' \le y} f(y'|x;\boldsymbol{\beta}^*) B_k(y')\, dy' \right| = O_P(\sqrt{J/n}). \tag{36}$$

For any given value of y, all but a bounded number of terms $\int_{y' \le y} f(y'|x;\boldsymbol{\beta}^*)B_k(y')\,dy'$ are equal to $\int f(y'|x;\boldsymbol{\beta}^*)B_k(y')\,dy'$ or to zero. By (4), (7) and Lemma 16(d), the total contribution of the bounded number of exceptional terms is $O_P(K^{-1}\sqrt{JK(\log K)/n}) = O_P(\sqrt{J/n})$. Let the B-splines $B_1, \ldots, B_k$ be ordered according to the right endpoints of their supports. In order to verify (36), it suffices to show that

$$\max_{k'} \left| \sum_{k \le k'} \hat{\varphi}_{jk} \int f(y|x;\boldsymbol{\beta}^*) B_k(y)\, dy \right| = O_P(\sqrt{J/n}). \tag{37}$$

Let $\mathscr{K}$ be a subset of consecutive integers in $\{1, \ldots, K\}$ and let K' denote the cardinality of $\mathscr{K}$. Let $\boldsymbol{\tau}$ denote the $J \times K$ matrix having entry $\int f(y|x;\boldsymbol{\beta}^*)B_k(y)\,dy$ in row j and column k for $k \in \mathscr{K}$ and all other entries equal to zero. Then

$$|\boldsymbol{\tau}|^2 = O_P(K'/K^2). \tag{38}$$

Since

$$\operatorname{var}\left(\sum_{k \in \mathscr{K}} \hat{\varphi}_{jk} \int f(y|x;\boldsymbol{\beta}^*) B_k(y)\, dy \right) = \boldsymbol{\tau}^t \operatorname{VC}(\hat{\boldsymbol{\varphi}})\boldsymbol{\tau},$$

it follows from (26) and (38) that

$$\operatorname{var}\left(\sum_{k \in \mathscr{K}} \hat{\varphi}_{jk} \int f(y|x;\boldsymbol{\beta}^*) B_k(y)\, dy \right) = O\left(\frac{JK'}{nK}\right). \tag{39}$$

Observe next that

$$\sum_{k\in\mathscr{K}} \hat{\varphi}_{jk}\int f(y|x;\boldsymbol{\beta}^*)\,B_k(y)\,dy = \sum_i Z_{\mathscr{K}i}, \tag{40}$$

where

$$Z_{\mathscr{K}i} = \Big(\sum_{k\in\mathscr{K}} \big([\mathbf{I}(\boldsymbol{\beta}^*)]^{-}\,\mathbf{G}(Y_i|x_i;\boldsymbol{\beta}^*)\big)_{jk}\Big)\int f(y|x;\boldsymbol{\beta}^*)\,B_k(y)\,dy, \tag{41}$$

$$1 \le i \le n,$$

are independent random variables whose sum has mean zero. By (4), (25) and (28),

$$|Z_{\mathscr{K}i}| \le b = O\big(n^{-1}J\sqrt{K}\big). \tag{42}$$

It follows from (39)–(42) and Bernstein's inequality that there is a $\delta > 0$ such that

$$\begin{aligned} &P\left(\left|\sum_{k\in\mathscr{K}} \hat{\varphi}_{jk}\int f(y|x;\boldsymbol{\beta}^*)\,B_k(y)\,dy\right| \ge A\sqrt{J/n}\,(K'/K)^{\alpha}\right) \\ &\quad\le 2\Big\{\exp\big[-\delta A\sqrt{n/(JK)}\,(K'/K)^{\alpha}\big] + \exp\big[-\delta A^2(K/K')^{1-2\alpha}\big]\Big\} \end{aligned} \tag{43}$$

for $A > 0$ and $0 < \alpha < \frac{1}{2}$.

Set $R = \min[r\colon 2^r \ge K]$. For $0 \le r \le R$, let $\mathscr{M}_r$ denote the collection of all sets of integers of the form $\{(m-1)2^r + 1, \ldots, m2^r\}$, where $1 \le m \le K/2^r$ and note that the cardinality of $\mathscr{M}_r$ is at most $K/2^r$ and that each set in $\mathscr{M}_r$ has cardinality 2^r. It follows from (7) and (43) that, for any $\alpha > 0$, A can be chosen sufficiently large so that, for $0 < \alpha < \frac{1}{4}$ and $n \gg 1$,

$$\begin{aligned} &P\left(\left|\sum_{k\in\mathscr{K}} \hat{\varphi}_{jk}\int f(y|x;\boldsymbol{\beta}^*)\,B_k(y)\,dy\right| \ge A\sqrt{J/n}\,(K'/K)^{\alpha}\right. \\ &\qquad\left.\text{for some } \mathscr{K}\in\mathscr{M}_0\cup\cdots\cup\mathscr{M}_R\right) \le \alpha. \end{aligned} \tag{44}$$

For $1 \le k' \le K$, $\{1, \ldots, k'\}$ can be written as a disjoint union of sets $\mathscr{K}\in \mathscr{M}_0 \cup \cdots \cup \mathscr{M}_R$ such that for $0 \le r \le R$, there is at most one such $\mathscr{K}\in\mathscr{M}_r$. Thus it follows from (44) that (37) holds. □

Equations (8), (9), (11) and (12) follow from (4), (7) and Lemmas 18, 20 and 21.

4. Asymptotic normality. In this section the asymptotic normality of $(\hat{\tau} - \tau^*)/\mathrm{ASD}(\hat{\tau})$ and $(\hat{\tau} - \tau^*)/\mathrm{SE}(\hat{\tau})$ will be established, where $\hat{\tau}$ is $\hat{f}(y|x)$, $\hat{F}(y|x)$ or $\hat{Q}(p|x)$.

The next result follows from (1), (4), (18), (23) and (12) of Stone (1986).

LEMMA 22. *There is an $M > 0$ such that if $n \gg 1$, then*

$$\big|\boldsymbol{\tau}^t\mathrm{VC}(\mathbf{S}(\boldsymbol{\beta}^*))\boldsymbol{\tau} - \boldsymbol{\tau}^t\mathbf{I}(\boldsymbol{\beta}^*)\boldsymbol{\tau}\big| \le MnJ^{-1}K^{-1}\delta_{\mathscr{T}}|\boldsymbol{\tau}|^2, \qquad \boldsymbol{\tau}\in\mathscr{B}.$$

According to (24), (25), (28) and Lemma 22, there is an $M > 0$ such that if $n \gg 1$, then

$$\begin{aligned}(45)\qquad &\left|\operatorname{var}\left([\mathbf{G}(y|x;\boldsymbol{\beta}^*)]^t\hat{\boldsymbol{\varphi}}\right) - [\mathbf{G}(y|x;\boldsymbol{\beta}^*)]^t[\mathbf{I}(\boldsymbol{\beta}^*)]^- \mathbf{G}(y|x;\boldsymbol{\beta}^*)\right| \\ &\quad \le Mn^{-1}JK\delta_{\mathscr{F}}, \qquad x \in \mathscr{X} \text{ and } y \in \mathscr{Y}.\end{aligned}$$

Throughout the remainder of the section, it is assumed that $J, K \to \infty$ as $n \to \infty$. Under this assumption, $\delta_{\mathscr{F}} \to 0$ as $n \to \infty$. Also, it follows from (27) that there is an $M > 0$ such that if $n \gg 1$, then

$$(46)\qquad M^{-1} \le \left|\mathbf{G}(y|x;\boldsymbol{\beta}^*)\right| \le M, \qquad x \in \mathscr{X} \text{ and } y \in \mathscr{Y}.$$

If follows from (26) and (46) that if $n \gg 1$, then

$$(47)\qquad M^{-1}n^{-1}JK \le \operatorname{var}\left([\mathbf{G}(y|x;\boldsymbol{\beta}^*)]^t\hat{\boldsymbol{\varphi}}\right) \le Mn^{-1}JK, \qquad x \in \mathscr{X} \text{ and } y \in \mathscr{Y}.$$

LEMMA 23. *Uniformly for $x \in \mathscr{X}$ and $y \in \mathscr{Y}$, the distribution of*

$$\frac{[\mathbf{G}(y|x;\boldsymbol{\beta}^*)]^t\hat{\boldsymbol{\varphi}}}{\operatorname{SD}\left([\mathbf{G}(y|x;\boldsymbol{\beta}^*)]^t\hat{\boldsymbol{\varphi}}\right)}$$

converges to the standard normal distribution as $n \to \infty$.

PROOF. Observe that $[\mathbf{G}(y|x;\boldsymbol{\beta}^*)]^t\hat{\boldsymbol{\varphi}} = \sum_i Z_i$, where

$$Z_i = [\mathbf{G}(y|x;\boldsymbol{\beta}^*)]^t[\mathbf{I}(\boldsymbol{\beta}^*)]^- \mathbf{G}(Y_i|x_i;\boldsymbol{\beta}^*), \qquad 1 \le i \le n.$$

The random variables $Z_1, \ldots, Z_n$ are independent and their sum has mean zero. Moreover, by (25) and (28), there is an $M > 0$ such that $|Z_i| \le Mn^{-1}JK$ for $1 \le i \le n$. The desired result now follows from (7), (47) and the central limit theorem [see the corollary on page 201 of Chung (1974)]. □

Sets $\mathbf{G}^*(y|x) = \mathbf{G}(y|x;\boldsymbol{\beta}^*)$ and $\hat{\mathbf{G}}(y|x) = \mathbf{G}(y|x;\hat{\boldsymbol{\beta}})$. Then

$$\operatorname{ASD}\left(\hat{f}(y|x)\right) = f^*(y|x)\left\{[\mathbf{G}^*(y|x)]^t[\mathbf{I}(\boldsymbol{\beta}^*)]^- \mathbf{G}^*(y|x)\right\}^{1/2},$$

$$\operatorname{SE}\left(\hat{f}(y|x)\right) = \hat{f}(y|x)\left\{\left[\hat{\mathbf{G}}(y|x)\right]^t\left[\mathbf{I}(\hat{\boldsymbol{\beta}})\right]^- \hat{\mathbf{G}}(y|x)\right\}^{1/2},$$

$$\operatorname{ASD}\left(\hat{F}(y|x)\right) = \left[\left(\int_{y' \le y} f^*(y'|x)\mathbf{G}^*(y'|x)\,dy'\right)^t [\mathbf{I}(\boldsymbol{\beta}^*)]^- \times \int_{y' \le y} f^*(y'|x)\mathbf{G}^*(y'|x)\,dy'\right]^{1/2},$$

$$\operatorname{SE}\left(\hat{F}(y|x)\right) = \left[\left(\int_{y' \le y} \hat{f}(y'|x)\hat{\mathbf{G}}(y'|x)\,dy'\right)^t \left[\mathbf{I}(\hat{\boldsymbol{\beta}})\right]^- \int_{y' \le y} \hat{f}(y'|x)\hat{\mathbf{G}}(y'|x)\,dy'\right]^{1/2},$$

$$\operatorname{ASD}\left(\hat{Q}(p|x)\right) = \left.\frac{\operatorname{ASD}\left(\hat{F}(y|x)\right)}{f^*(y|x)}\right|_{y=Q^*(p|x)}$$

and

$$\mathrm{SE}\big(\hat{Q}(p|x)\big) = \frac{\mathrm{SE}\big(\hat{F}(y|x)\big)}{\hat{f}(y|x)}\bigg|_{y=\hat{Q}(p|x)}.$$

It follows from (47) and Lemmas 20(a) and 23 that, uniformly for $x \in \mathscr{X}$ and $y \in \mathscr{Y}$, the distribution of

$$\frac{\log f(y|x;\hat{\boldsymbol{\beta}}) - \log f(y|x;\boldsymbol{\beta}^*)}{\mathrm{SD}\big(\mathbf{G}(y|x;\hat{\boldsymbol{\beta}})\big)}$$

converges to the standard normal distribution as $n \to \infty$. It now follows easily from (45) and (47) that the distribution of

$$\frac{f(y|x;\hat{\boldsymbol{\beta}}) - f(y|\mathbf{x};\boldsymbol{\beta}^*)}{\mathrm{ASD}\big(f(y|x;\hat{\boldsymbol{\beta}})\big)}$$

converges to the standard normal distribution as $n \to \infty$ uniformly for $x \in \mathscr{X}$ and $y \in \mathscr{Y}$.

The next result follows from (1), (4), (10), (18) and (12) of Stone (1986).

LEMMA 24. *Uniformly for* $\boldsymbol{\tau} \in \mathscr{B}$,

$$\big|\big[\mathbf{I}(\hat{\boldsymbol{\beta}}) - \mathbf{I}(\boldsymbol{\beta}^*)\big]\boldsymbol{\tau}\big|^2 = O_p(nJ^{-1}K^{-1}(\log JK))|\boldsymbol{\tau}|^2.$$

Since

$$\big\{\big[\mathbf{I}(\hat{\boldsymbol{\beta}})\big]^{-}[\mathbf{I}(\boldsymbol{\beta}^*)]^{-}\big\}\boldsymbol{\tau} = [\mathbf{I}(\boldsymbol{\beta}^*)]^{-}\big[\mathbf{I}(\boldsymbol{\beta}^*) - \mathbf{I}(\hat{\boldsymbol{\beta}})\big]\big[\mathbf{I}(\hat{\boldsymbol{\beta}})\big]^{-}\boldsymbol{\tau}, \qquad \boldsymbol{\tau} \in \mathscr{B},$$

the next result follows from Lemmas 10, 11 and 24.

LEMMA 25. *Uniformly for* $\boldsymbol{\tau} \in \mathscr{B}$,

$$\big|\big\{\big[\mathbf{I}(\hat{\boldsymbol{\beta}})\big]^{-} - [\mathbf{I}(\boldsymbol{\beta}^*)]^{-}\big\}\boldsymbol{\tau}\big|^2 = O_p\big(n^{-3}(JK)^3(\log JK)\big)|\boldsymbol{\tau}|^2.$$

LEMMA 26.

$$\max_{x,y}\big|\mathbf{G}(y|x;\hat{\boldsymbol{\beta}}) - \mathbf{G}(y|x;\boldsymbol{\beta}^*)\big|^2 = O_p\big(n^{-1}JK^{-1}(\log JK)\big).$$

PROOF. Observe that $\mathbf{G}(y|x;\hat{\boldsymbol{\beta}}) - \mathbf{G}(y|x;\boldsymbol{\beta}^*)$ is the $J \times K$ matrix the entry in row j and column k of which is $H_j(x)\int B_k(y)[f(y|x;\boldsymbol{\beta}^*) - f(y|x;\hat{\boldsymbol{\beta}})]\,dy$. The desired result now follows easily from (10). □

It follows easily from (7), (10), (45)–(47) and Lemmas 25 and 26 that, uniformly for $x \in \mathscr{X}$ and $y \in \mathscr{Y}$,

$$\frac{\mathrm{SE}\big(f(y|x;\hat{\boldsymbol{\beta}})\big)}{\mathrm{ASD}\big(f(y|x;\hat{\boldsymbol{\beta}})\big)} = 1 + o_p(1) \quad \text{as } n \to \infty.$$

This completes the proof of asymptotic normality for $\hat{\tau} = f(y|x;\hat{\boldsymbol{\beta}})$.

Let y be in the interior of $\mathscr{Y}$.

LEMMA 27. *There is an $M > 0$ such that if $n \gg 1$, then*

$$M^{-1}K^{-1} \le \left| \int_{y' \le y} f(y'|x;\boldsymbol{\beta}^*)\mathbf{G}(y'|x;\boldsymbol{\beta}^*)\, dy' \right|^2 \le MK^{-1}, \qquad x \in \mathscr{X}.$$

PROOF. The entry in row j and column k of $\int_{y' \le y} f(y'|x;\boldsymbol{\beta}^*)\mathbf{G}(y'|x;\boldsymbol{\beta}^*)\, dy'$ is

$$\int_{y' \le y} \left(H_j(x) B_k(y') - H_j(x) \frac{\partial C}{\partial \theta_k}(\mathbf{h}(x;\boldsymbol{\beta}^*)) \right) f(y'|x;\boldsymbol{\beta}^*)\, dy'$$

$$= H_j(x) \left(\int_{y' \le y} B_k(y') f(y'|x;\boldsymbol{\beta}^*)\, dy' - \int_{y' \le y} f(y'|x;\boldsymbol{\beta}^*)\, dy' \int B_k(y') f(y'|x;\boldsymbol{\beta}^*)\, dy' \right).$$

The upper bound in the conclusion of the lemma now obviously holds, so to complete the proof it suffices to show that there is an $M > 0$ such that if $n \gg 1$, then

$$M^{-1}K^{-1} \le \sum_k \left(\int_{y' \le y} B_k(y') f(y'|x;\boldsymbol{\beta}^*)\, dy' - \int_{y' \le y} f(y'|x;\boldsymbol{\beta}^*)\, dy' \int B_k(y') f(y'|x;\boldsymbol{\beta}^*)\, dy' \right)^2$$

for $x \in \mathscr{X}$. This lower bound is easily established by noting that if the support of $B_k(\cdot)$ is to the right of y, then $\int_{y' \le y} B_k(y') f(y'|x;\boldsymbol{\beta}^*)\, dy' = 0$ for $x \in \mathscr{X}$. □

By (26) and Lemma 27, there is an $M > 0$ such that if $n \gg 1$, then

$$M^{-1}n^{-1}J \le \mathrm{var}\left(\int_{y' \le y} f(y'|x;\boldsymbol{\beta}^*) [\mathbf{G}(y'|x;\boldsymbol{\beta}^*)]^t \hat{\boldsymbol{\varphi}}\, dy' \right) \le Mn^{-1}J, \qquad x \in \mathscr{X}. \tag{48}$$

LEMMA 28. *Uniformly for $x \in \mathscr{X}$, the distribution of*

$$\frac{\int_{y' \le y} f(y'|x;\boldsymbol{\beta}^*) [\mathbf{G}(y'|x;\boldsymbol{\beta}^*)]^t \hat{\boldsymbol{\varphi}}\, dy'}{\mathrm{SD}\left(\int_{y' \le y} f(y'|x;\boldsymbol{\beta}^*) [\mathbf{G}(y'|x;\boldsymbol{\beta}^*)]^t \hat{\boldsymbol{\varphi}}\, dy' \right)}$$

converges to the standard normal distribution as $n \to \infty$.

PROOF. Observe that $\int_{y' \le y} f(y'|x;\beta^*)[\mathbf{G}(y'|x;\beta^*)]^t \hat{\varphi}\, dy' = \sum_i Z_i$, where

$$Z_i = \int_{y' \le y} f(y'|x;\beta^*)\,[\mathbf{G}(y'|x;\beta^*)]^t [\mathbf{I}(\beta^*)]^{-} \mathbf{G}(Y_i|x_i;\beta^*)\, dy', \qquad 1 \le i \le n.$$

By (25), (28) and Lemma 27, there is an $M > 0$ such that $|Z_i| \le Mn^{-1}J\sqrt{K}$ for $1 \le i \le n$. The remainder of the proof is as in that of Lemma 23, with (48) used instead of (47). □

The next result follows from (24), (25), (48) and Lemmas 22 and 27.

LEMMA 29. *Uniformly for* $x \in \mathscr{X}$,

$$\frac{\mathrm{ASD}\big(F(y|x;\hat{\beta})\big)}{\mathrm{SD}\Big(\int_{y' \le y} f(y'|x;\beta^*)\,[\mathbf{G}(y'|x;\beta^*)]^t \hat{\varphi}\, dy\Big)} \to 1 \quad \textit{as } n \to \infty.$$

It follows from (48) and Lemmas 20(b), 28 and 29 that, uniformly for $x \in \mathscr{X}$, the distribution of

$$\frac{F(y|x;\hat{\beta}) - F(y|x;\beta^*)}{\mathrm{ASD}\big(F(y|x;\hat{\beta})\big)}$$

converge to the standard normal distribution as $n \to \infty$. It follows from (4), (7), (10), (25), and (48) and Lemmas 25–27 and 29 that, uniformly for $x \in \mathscr{X}$,

$$\frac{\mathrm{SE}\big(F(y|x;\hat{\beta})\big)}{\mathrm{ASD}\big(F(y|x;\hat{\beta})\big)} = 1 + o_p(1) \quad \text{as } n \to \infty.$$

This completes the proof of asymptotic normality for $\hat{\tau} = F(y|x;\hat{\beta})$.

Observe that $F^*(Q^*(p|x)|x) = p$ and $\hat{F}(\hat{Q}(p|x)|x) = p$. Thus

$$\begin{aligned}\big[\hat{F}(y|x) - F^*(y|x)\big]_{y=Q^*(p|x)} &= -\big[\hat{F}\big(\hat{Q}(p|x)|x\big) - \hat{F}(Q^*(p|x)|x)\big] \\ &= \int_{Q^*(p|x)}^{\hat{Q}(p|x)} \hat{f}(y|x)\, dy.\end{aligned}$$

We now conclude from (4) and (10) that, uniformly for $x \in \mathscr{X}$,

$$\hat{Q}(p|x) - Q^*(p|x) = \big[1 + o_p(1)\big] \frac{\big[\hat{F}(y|x) - F^*(y|x)\big]}{f^*(y|x)}\bigg|_{y=Q^*(p|x)}.$$

The argument used to establish asymptotic normality for $\hat{\tau} = \hat{F}(y|x)$ applies with $y = Q^*(p|x)$ (even though this value of y depends on n). Thus asymptotic normality for $\hat{\tau} = \hat{Q}(p|x)$ is valid.

REFERENCES

BARRON, A. R. and SHEU, C.-H. (1991). Approximation of density functions by sequences of exponential families. *Ann. Statist.* **19** 1347–1369.

BREIMAN, L. (1989). Fitting additive models to regression data. Technical Report 209, Dept. Statistics, Univ. California, Berkeley.

BREIMAN, L. (1991). The Π-method for estimating multivariate functions (with discussion). *Technometrics* **33** 125–162.

BREIMAN, L. and PETERS, S. (1988). Comparing automatic smoothers. Technical Report 161, Dept. Statistics, Univ. California, Berkeley.

CHUNG, K. L. (1974). *A Course in Probability Theory*, 2nd ed. Academic, New York.

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

FAN, J. (1990). A remedy to regression estimators and nonparametric minimax efficiency. Technical Report, Dept. Statistics, Univ. North Carolina, Chapel Hill.

FRIEDMAN, J. H. (1991). Multivariate adaptive regression splines (with discussion). *Ann. Statist.* **19** 1–141.

FRIEDMAN, J. H. and SILVERMAN, B. W. (1989). Flexible parsimonious smoothing and additive modeling (with discussion). *Technometrics* **31** 3–39.

HASMINSKII, R. and IBRAGIMOV, I. A. (1990). On density estimation in the view of Kolmogorov's ideas in approximation theory. *Ann. Statist.* **18** 999–1010.

HOEFFDING, W. (1963). Probability inequalities for sums of bounded random variables. *J. Amer. Statist. Assoc.* **58** 13–30.

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

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

KOOPERBERG, C. and STONE, C. J. (1991). A study of logspline density estimation. *Comput. Statist. Data Anal.* To appear.

MCCULLAGH, P. and NELDER, J. A. (1983). *Generalized Linear Models*. Chapman and Hall, London.

PORTNOY, S. (1986). On the central limit in $\mathbb{R}^p$ when $p \to \infty$. *Probab. Theory Related Fields* **73** 571–583.

PORTNOY, S. (1988). Asymptotic behavior of likelihood methods for exponential families when the number of parameters tends to infinity. *Ann. Statist.* **16** 356–366.

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

SMITH, P. L. (1982). Curve fitting and modeling with splines using statistical variable selection techniques. NASA Report CR-166034. NASA, Langley Research Center, Hampton, Va.

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. (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, Boston.

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

STONE, C. J. and KOO, C.-Y. (1986). Logspline density estimation. *Contemp. Math.* **59** 1–15.

DEPARTMENT OF STATISTICS
UNIVERSITY OF CALIFORNIA
BERKELEY, CALIFORNIA 94720