1

Extended Linear Modeling with Splines

Jianhua Z. Huang and Charles J. Stone [1]

Abstract

Extended linear models form a very general framework for statistical modeling. Many practically important contexts fit into this framework, including regression, logistic and Poisson regression, density estimation, spectral density estimation, and conditional density estimation. Moreover, hazard regression, proportional hazard regression, marked point process regression, and diffusion processes, all perhaps with time-dependent covariates, also fit into this framework. Polynomial splines and their tensor products provide a universal tool for constructing maximum likelihood estimates for extended linear models. The theory of rates of convergence for such estimates as it applies both to fixed knot splines and to free knot splines will be surveyed, and the implications of this theory for the development of corresponding methodology will be discussed briefly.

1.1 Introduction

Polynomail splines are useful in statistical modeling and data analysis. The theoretical framework of extended linear modeling has evolved through a long-term investigation, starting in the mid-eighties, of the properties of spline-based estimates in various contexts. Some early results in this effort are Stone (1985, 1986, 1990, 1991, 1994) and Kooperberg, Stone and Truong (1995b, 1995d). Hansen (1994) expanded and unified the then existing theory. The resulting synthesis played a key role in the Stone, Hansen, Kooperberg and Truong (1997), which reviewed the theory and corresponding methodology. Shortly after that, Huang (1998a, 1998b) substantially simplified and extended the theoretical approach. These improvements helped lead to Huang and Stone (1998) and Huang, Kooperberg, Stone

[1]Jianhua Z. Huang is Assistant Professor, Department of Statistics, The Wharton School, University of Pennsylvania, Philadelphia, PA 19104-6302 (Email: jianhua@wharton.upenn.edu). Charles J. Stone is Professor, Department of Statistics, University of California, Berkeley, CA 94720-3860 (E-mail: stone@stat.berkeley.edu); he was supported in part by National Science Foundation grant DMS-9802071.

and Truong (2000). More recently, Huang (2001) provided a fresh theoretical synthesis of extended linear modeling. To the extent that this work pertained to spline-based methods, it was restricted to fixed knot splines. Still more recently, Stone and Huang (2002a, 2002b) have used the framework of Huang (2001) to investigate the theoretical properties of extended linear modeling with free knot splines.

Section 2 gives a detailed description of the theoretical framework of extended linear models. Some examples of such models are presented in Section 3. Section 4 describes asymptotic results for maximum likelihood estimates, such as consistency and rates of convergence; the focus is on fixed knot spline estimates. Section 5 discusses using functional analysis of variance to construct structural models in order to tame the curse of dimensionality. Free knot splines are studied in Section 6. Some implications of the theory for the development of corresponding methodology will be mentioned in Section 7.

1.2 Extended Linear Models: Theoretical Framework

Consider a $\mathcal{W}$-valued random variable $\boldsymbol{W}$, where $\mathcal{W}$ is an arbitrary set. The probability density $p(\eta, \boldsymbol{w})$ of $\boldsymbol{W}$ depends on an unknown function η. The function η is defined on a domain $\mathcal{U}$, which may or may not be the same as $\mathcal{W}$. We assume that $\mathcal{U}$ is a compact subset of some Euclidean space and that it has positive volume $\text{vol}(\mathcal{U})$. The problem of interest is estimation of η based on a random sample from the distribution of $\boldsymbol{W}$.

Corresponding to a candidate function h for η, the log-likelihood is given by $l(h, \boldsymbol{w}) = \log p(h, \boldsymbol{w})$. The expected log-likelihood is defined by $\Lambda(h) = E[l(h, \boldsymbol{W})]$, where the expectation is taken with respect to the probability measure corresponding to the true function η. There may be some mild restrictions on h for $l(h, \boldsymbol{w})$, $\boldsymbol{w} \in \mathcal{W}$, and $\Lambda(h)$ to be well-defined. It follows from the information inequality that η is the essentially unique function on $\mathcal{U}$ that maximizes the expected log-likelihood. (Here two functions on $\mathcal{U}$ are regarded as essentially equal if their difference equals zero except on a subset of $\mathcal{U}$ having Lebesgue measure zero.)

In many applications, we are interested in a function η that is related to but need not totally specify the probability distribution of $\boldsymbol{W}$. In such applications, we can modify the above setup by taking $l(h, \boldsymbol{w})$ to be the logarithm of a conditional likelihood, a pseudo-likelihood, or a partial likelihood, depending on the problem under consideration.

Consider, for example, the estimation of a regression function $\eta(\boldsymbol{x}) = E(Y|\boldsymbol{X} = \boldsymbol{x})$. In terms of the above notation, $\boldsymbol{W}$ consists of a pair of random variables $\boldsymbol{X}$ and Y, and $\mathcal{U}$ is the range of $\boldsymbol{X}$. We can take $l(h, \boldsymbol{w})$ to be the negative of the residual sum of squares; that is, $l(h, \boldsymbol{W}) = -[Y -$

$h(\boldsymbol{X})]^2$ with $\boldsymbol{W} = (\boldsymbol{X}, Y)$. If the conditional distribution of Y given $\boldsymbol{X}$ is assumed to be normal with constant variance, then l is (up to additive and multiplicative constants) the conditional log-likelihood. Even if this conditional distribution is not assumed to be normal, we can still think of l as the logarithm of a pseudo-likelihood. In either case, the true regression function η maximizes $\Lambda(h) = E[l(h, \boldsymbol{W})] = -E[\eta(\boldsymbol{X}) - h(\boldsymbol{X})]^2$.

From now on, we will adopt this broad view of $l(h, \boldsymbol{w})$. For simplicity, we will still call $l(h, \boldsymbol{w})$ the log-likelihood and $\Lambda(h)$ the expected log-likelihood. To relate the function of interest to the log-likelihood, we assume that, subject to mild conditions on $l(h, \boldsymbol{w})$, the function η is the essentially unique function that maximizes the expected log-likelihood.

Let $\mathbb{H}$ be a linear space of square-integrable functions on $\mathcal{U}$ such that if two functions on $\mathcal{U}$ are essentially equal and one of them is in $\mathbb{H}$, then so is the other one. We refer to $\mathbb{H}$ as the *model space* and to $l(h, \boldsymbol{W})$, $h \in \mathbb{H}$, as forming an *extended linear model*. If $\mathbb{H}$ is the space of all square-integrable functions on $\mathcal{U}$ or differs from this space only by the imposition of some identifiability restrictions as in the context of density estimation (see Section 1.3), we refer to $\mathbb{H}$ as being *saturated*. Otherwise, we refer to this space as being *unsaturated*.

The use of unsaturated spaces allows us to impose structural assumptions on the extended linear model. Suppose $\mathcal{U}$ is the Cartesian product of compact intervals $\mathcal{U}_1, \ldots, \mathcal{U}_L$, each having positive length. We can impose an additive structure by letting $\mathbb{H}$ be the space of functions of the form $h_1(u_1) + \cdots + h_L(u_L)$, where h_l is a square-integrable function on $\mathcal{U}_l$ for $1 \leq l \leq L$. This and more general ANOVA structures will be considered in Section 1.5. Alternatively, we can impose an additive, semilinear structure by letting $\mathbb{H}$ be the space of functions of the form $h_1(u_1) + b_2 u_2 + \cdots + b_L u_L$, where h_1 is a square-integrable function on $\mathcal{U}_1$ and $b_2, \ldots, b_L$ are real numbers.

The extended linear model is said to be *concave* if the following two properties are satisfied: (i) The log-likelihood function is concave; that is, given any two functions $h_1, h_2 \in \mathbb{H}$ whose log-likelihoods are well-defined, $l(\alpha h_1 + (1-\alpha)h_2, \boldsymbol{w}) \geq \alpha l(h_1, \boldsymbol{w}) + (1-\alpha) l(h_2, \boldsymbol{w})$ for $0 < \alpha < 1$ and $\boldsymbol{w} \in \mathcal{W}$. (ii) The expected log-likelihood function is strictly concave; that is, given any two essentially different functions $h_1, h_2 \in \mathbb{H}$ whose expected log-likelihoods are well-defined, $\Lambda(\alpha h_1 + (1-\alpha)h_2) > \alpha\Lambda(h_1) + (1-\alpha)\Lambda(h_2)$ for $0 < \alpha < 1$. Here, we implicitly assume that the set of functions such that $l(h, \boldsymbol{w})$ and $\Lambda(h)$ are well-defined is a convex set.

As mentioned above, the model space $\mathbb{H}$ incorporates structural assumptions (e.g., additivity) on the true function of interest. Such structural assumptions are not necessarily true and are considered rather as approximations. Thus, it is natural to think that any estimation procedure will estimate the best approximation to the true function with the imposed structure. This "best approximation" can be defined formally using the expected log-likelihood. Observe that if $\eta \in \mathbb{H}$, then $\eta = \operatorname{argmax}_{h \in \mathbb{H}} \Lambda(h)$

since η maximizes the expected log-likelihood by assumption. More generally, we think of $\eta^* = \text{argmax}_{h\in\mathbb{H}}\,\Lambda(h)$ as the "best approximation" in $\mathbb{H}$ to η. Typically, when the expected log-likelihood function is strictly concave, such a best approximation exists and is essentially unique. If $\eta \in \mathbb{H}$, then η^* is essentially equal to η.

In the regression context η^* is the orthogonal projection of η onto $\mathbb{H}$ with respect to the L_2 norm on $\mathbb{H}$ given by $\|h\|^2 = E[h^2(\boldsymbol{X})]$; that is, $\eta^* = \text{argmin}_{h\in\mathbb{H}}\,\|h-\eta\|^2$. Here, to guarantee the existence of η^*, we need to assume that $\mathbb{H}$ is a Hilbert space; that is, it is closed in the metric corresponding to the indicated norm.

We now turn to estimation. Let $\boldsymbol{W}_1, \ldots, \boldsymbol{W}_n$ be a random sample of size n from the distribution of $\boldsymbol{W}$. Let $\mathbb{G} \subset \mathbb{H}$ be a finite-dimensional linear space of bounded functions, whose dimension may depend on the sample size. We estimate η by using maximum likelihood over $\mathbb{G}$, that is, we take $\widehat{\eta} = \text{argmax}_{g\in\mathbb{G}}\,\ell(g)$, where $\ell(g) = (1/n)\sum_{i=1}^{n} l(g, \boldsymbol{W}_i)$ is the normalized log-likelihood. Here the space $\mathbb{G}$ should be chosen such that the function of interest η can be approximated well by some function in $\mathbb{G}$. Thus $\mathbb{G}$ will be called the approximation space. Since $\mathbb{G}$ is where the maximum likelihood estimation is carried out, it will also be called the estimation space. In this setup we do not specify the form of $\mathbb{G}$; any linear function space with good approximation properties can be used. When $\mathbb{H}$ has a specific structure, $\mathbb{G}$ should be chosen to have the same structure. For example, if $\mathbb{H}$ consists of all square-integrable additive functions, then $\mathbb{G}$ should not contain any non-additive functions. A detailed discussion of constructing the model and estimation spaces using functional ANOVA decompositions to incorporate structural assumptions will be given in Section 1.5. In our application, $\mathbb{G}$ will be chosen as a space built by polynomial splines and their tensor products. That polynomial splines and their tensor products enjoy good approximation power has been extensively studied and documented; see de Boor (1978), Schumaker (1980), and DeVore and Lorentz (1993).

1.3 Examples of Concave Extended Linear Models

The log-likelihood function in an extended linear model takes into account the probability structure of the estimation problem. It is advantagous that the log-likelihood of an extended linear model be concave (and, in fact, suitably strictly concave with probability close to one). The maximum likelihood estimate in a finite-dimensional estimation space is then unique if it exists. Moreover, the Newton–Raphson algorithm when suitably adjusted (for example, by step-halving) is guaranteed to converge to the global maximum. Although the concavity restriction on the log-likehood may look restrictive, it turns out that the collection of intrinsically suitably concave extended linear models is very rich, including of course ordinary regres-

sion. We present in this section a number of other contexts in which the concavity assumption on the log-likelihood is automatically satisfied.

1.3.1 Generalized Regression.

The ordinary (least squares) regression model, which was introduced in the last section, does not impose any structure on the conditional variance of the response given the covariates. On the other hand, in generalized regression the conditional variance depends on the conditional mean in some specified way as in an exponential family.

Consider a random pair $\boldsymbol{W} = (\boldsymbol{X}, Y)$, where the random vector $\boldsymbol{X}$ of covariates is $\mathcal{X}$-valued with $\mathcal{X} = \mathcal{U}$ and the response Y is real-valued. Suppose the conditional distribution of Y given that $\boldsymbol{X} = \boldsymbol{x} \in \mathcal{X}$ has the form of an exponential family

$$P(Y \in dy|\boldsymbol{X} = \boldsymbol{x}) = \exp[B(\eta(\boldsymbol{x}))y - C(\eta(\boldsymbol{x}))]\Psi(dy), \tag{1.1}$$

where $B(\cdot)$ is a known, twice continuously differentiable function on $\mathbb{R}$ whose first derivative is strictly positive on $\mathbb{R}$, Ψ is a nonzero measure on $\mathbb{R}$ that is not concentrated at a single point, and $C(\eta) = \log \int_{\mathbb{R}} \exp[(B(\eta)y]\Psi(dy) < \infty$ for $\eta \in \mathbb{R}$. Observe that $B(\cdot)$ is strictly increasing and $C(\cdot)$ is twice continuously differentiable on $\mathbb{R}$.

Here the function of interest is the response function $\eta(\cdot)$, which specifies the dependence on $\boldsymbol{x}$ of the conditional distribution of the response Y given that the value of the vector $\boldsymbol{X}$ of covariates equals $\boldsymbol{x}$. The mean of this conditional distribution is given by

$$\mu(\boldsymbol{x}) = E(Y|\boldsymbol{X} = \boldsymbol{x}) = A(\eta(\boldsymbol{x})) = \frac{C'(\eta(\boldsymbol{x}))}{B'(\eta(\boldsymbol{x}))} \qquad \boldsymbol{x} \in \mathcal{X}. \tag{1.2}$$

The (conditional) log-likelihood is given by

$$l(h, \boldsymbol{X}, Y) = B(h(\boldsymbol{X}))Y - C(h(\boldsymbol{X})),$$

and its expected value is given by

$$\Lambda(h) = E[B(h(\boldsymbol{X}))\mu(\boldsymbol{X}) - C(h(\boldsymbol{X}))],$$

which is essentially uniquely maximized at $h = \eta$. The log-likelihood is automatically concave if $B(\eta)$ is a linear function of η and in certain other specific cases as well (e.g., in probit models).

The family (1.1) includes as special cases many useful distributions such as Bernoulli, Possion, Gaussian, gamma, and inverse-Gaussian.

When the underlying exponential family is the Bernoulli distribution with parameter $\pi(\boldsymbol{x})$ and the function of interest is $\eta(\boldsymbol{x}) = \text{logit}(\pi(\boldsymbol{x})) = \log(\pi(\boldsymbol{x})/(1-\pi(\boldsymbol{x})))$, we get logistic regression, which is closely connected to classification. Here $P(Y = 1|\boldsymbol{X} = \boldsymbol{x}) = \pi(\boldsymbol{x})$, $P(Y = 0|\boldsymbol{X} = \boldsymbol{x}) = 1-\pi(\boldsymbol{x})$, Ψ is concentrated on $\{0, 1\}$ with $\Psi(\{0\}) = \Psi(\{1\}) = 1$, $B(\eta) = \eta$, $C(\eta) = \log(1 + \exp(\eta))$, and $\mu(\boldsymbol{x}) = \pi(\boldsymbol{x}) = \exp \eta(\boldsymbol{x})/(1 + \exp \eta(\boldsymbol{x}))$.

When the underlying exponential family is the Poisson distribution with parameter $\lambda(\boldsymbol{x})$ and the function of interest is $\eta(\boldsymbol{x}) = \log \lambda(\boldsymbol{x})$, we get Poisson regression. Here $P(Y = y|\boldsymbol{X} = \boldsymbol{x}) = \lambda^y \exp(-\lambda(\boldsymbol{x})/y!$ for $y \in \mathcal{Y} = \{0, 1, 2, \ldots\}$, Ψ is concentrated on $\mathcal{Y}$ with $\Psi(\{y\}) = 1/y!$ for $y \in \mathcal{Y}$, $B(\eta) = \eta$, $C(\eta) = \exp \eta$, and $\mu(\boldsymbol{x}) = \lambda(\boldsymbol{x}) = \exp \eta(\boldsymbol{x})$.

When the underlying exponential family is the normal distribution with mean $\mu(\boldsymbol{x})$ and known variance σ^2 and the function of interest is the regression function $\mu(\boldsymbol{x})$, we get ordinary regression as discussed in Section 1.2. Here $P(Y \in (y, y + dy)|\boldsymbol{X} = \boldsymbol{x}) = (1/\sqrt{2\pi\sigma^2}) \exp\{-(y - \mu(\boldsymbol{x}))^2/\sigma^2\}\, dy$ for $y \in \mathbb{R}$, $B(\eta) = \eta/\sigma^2$, $C(\eta) = -\eta^2/\sigma^2$, and $\eta(\boldsymbol{x}) = \mu(\boldsymbol{x})$.

If the conditional distribution of Y is not fully specified as in (1.1), $l(h, \boldsymbol{X}, Y)$ can be thought of as a quasi log-likelihood. In connecting the unknown function to the log-likelihood, we assume that (1.2) holds. This assumption guarantees that the expected log-likelihood is essentially maximized at the true function η of interest and thereby validates maximizing the sample log-likelihood.

1.3.2 Polychotomous Regression

Polychotomous regression, which is closely connected to multiple classification, is an extension of logistic regression. Let Y be a qualitative random variable having $K + 1$ possible values. Without loss of generality, we can think of this random variable as ranging over $\mathcal{Y} = \{1, \ldots, K + 1\}$. Suppose that $P(Y = k|\boldsymbol{X} = \boldsymbol{x}) > 0$ for $\boldsymbol{x} \in \mathcal{X}$ and $k \in \mathcal{Y}$. Set

$$\eta_k(\boldsymbol{x}) = \log \frac{P(Y = k|\boldsymbol{X} = \boldsymbol{x})}{P(Y = K + 1|\boldsymbol{X} = \boldsymbol{x})}, \qquad 1 \le k \le K.$$

The log-likelihood is given by

$$\begin{aligned} l(h, \boldsymbol{X}, Y) \;=\; & h_1(\boldsymbol{X})I_1(Y) + \cdots + h_K(\boldsymbol{x})I_K(Y) \\ & - \log(1 + \exp h_1(\boldsymbol{X}) + \cdots + \exp h_K(\boldsymbol{X})), \end{aligned}$$

where $I_k(Y)$ equals one or zero according as $Y = k$ or $Y \neq k$ and $h = (h_1, \ldots, h_K)$ is a candidate for $\eta = (\eta_1, \ldots, \eta_K)$. Here $\boldsymbol{W} = (\boldsymbol{X}, Y)$ and $\mathcal{U} = \mathcal{X}$.

1.3.3 Hazard Estimation and Regression

Estimation of a hazard function and its dependence on covariates is important in survival analysis. Consider a positive survival time T, a positive censoring time C, the observed time $\min(T, C)$, and an $\mathcal{X}$-valued random vector $\boldsymbol{X}$ of covariates. Let $\delta = \text{ind}(T \le C)$ be the indicator random variable that equals one or zero according as $T \le C$ (T is uncensored) or $T > C$ (T is censored), and set $Y = \min(T, C)$ and $\boldsymbol{W} = (\boldsymbol{X}, Y, \delta)$. Suppose T and C are conditionally independent given $\boldsymbol{X}$. Suppose also that $P(C \le \tau) = 1$

for a known positive constant τ. Let

$$\eta(\boldsymbol{x}, t) = \log \frac{f(t|\boldsymbol{x})}{1 - F(t|\boldsymbol{x})}, \qquad t > 0,$$

denote the logarithm of the conditional hazard function, where $f(t|\boldsymbol{x})$ and $F(t|\boldsymbol{x})$ are the conditional density function and conditional distribution function, respectively, of T given that $\boldsymbol{X} = \boldsymbol{x}$. Then

$$1 - F(t|\boldsymbol{x}) = \exp\Big(- \int_0^t \exp \eta(\boldsymbol{x}, u)\, du\Big), \qquad t > 0,$$

and hence

$$f(t|\boldsymbol{x}) = \exp\Big(\eta(\boldsymbol{x}, t) - \int_0^t \exp \eta(\boldsymbol{x}, u)\, du\Big), \qquad t > 0.$$

Hazard regression concerns estimation of the conditional hazard function $\eta(\boldsymbol{x}, t)$. Since the likelihood equals $f(T \wedge C|\boldsymbol{X})$ for an uncensored case and $1 - F(T \wedge C|\boldsymbol{X})$ for a censored case, it can be written as

$$\begin{aligned} & [f(T \wedge C|\boldsymbol{X})]^{\delta} [1 - F(T \wedge C|\boldsymbol{X})]^{1-\delta} \\ = \; & \left(\frac{f(T \wedge C|\boldsymbol{X})}{1 - F(T \wedge C|\boldsymbol{X})} \right)^{\delta} [1 - F(T \wedge C|\boldsymbol{X})] \\ = \; & [\exp\, \eta(\boldsymbol{X}, T \wedge C)]^{\delta} \exp\left(- \int_0^{T \wedge C} \exp\, \eta(\boldsymbol{X}, t)\, dt \right). \end{aligned}$$

Thus the log-likelihood for a candidate h for η is given by

$$l(h, \boldsymbol{W}) = \delta h(\boldsymbol{X}, Y) - \int_0^Y \exp\, h(\boldsymbol{X}, t)\, dt.$$

Here, $\mathcal{U} = \mathcal{X} \times [0, \tau]$. Hazard estimation corresponds to the above setup with the random vector $\boldsymbol{X}$ of covariates ignored.

1.3.4 Density Estimation

Let $\boldsymbol{Y}$ have an unknown positive density function on $\mathcal{Y}$. Suppose we want to estimate the log-density ϕ. Since ϕ is subject to the intrinsic non-linear constraint $\int_{\mathcal{Y}} \exp \phi(\boldsymbol{y})\, d\boldsymbol{y} = 1$, it is convenient to write $\phi = \eta - C(\eta)$ and model η as a member of some linear space; here $C(h) = \log \int_{\mathcal{Y}} \exp h(\boldsymbol{y})\, d\boldsymbol{y}$. Note that η is determined up to an arbitrary constant as the log-density function ϕ. By imposing a linear constraint such as $\int_{\mathcal{Y}} \eta(\boldsymbol{y})\, d\boldsymbol{y} = 0$, we can determine η uniquely and thus make the map $\sigma : \eta \mapsto \phi$ one-to-one. The log-likelihood is given by $l(h, \boldsymbol{Y}) = h(\boldsymbol{Y}) - C(h)$. Here $\boldsymbol{W} = \boldsymbol{Y}$ and $\mathcal{U} = \mathcal{Y}$.

1.3.5 Conditional Density Estimation

Consider a random pair $(\boldsymbol{X}, \boldsymbol{Y})$, where $\boldsymbol{X}$ is $\mathcal{X}$-valued, $\boldsymbol{Y}$ is $\mathcal{Y}$-valued, and the conditional distribution of $\boldsymbol{Y}$ given that $\boldsymbol{X} = \boldsymbol{x}$ has a positive density. Since the corresponding log-density ϕ satisfies the nonlinear constraint $\int_{\mathcal{Y}} \exp \phi(\boldsymbol{y}|\boldsymbol{x})\, d\boldsymbol{y} = 1$ for $\boldsymbol{x} \in \mathcal{X}$, it is not natural to model ϕ as a member of a linear space. To overcome this difficulty, we write $\phi(\boldsymbol{y}|\boldsymbol{x}) = \eta(\boldsymbol{y}|\boldsymbol{x}) - C(\boldsymbol{x};\eta)$ and model η as a member of some linear space; here $C(\boldsymbol{x};\eta) = \log \int_{\mathcal{Y}} \exp \eta(\boldsymbol{y}|\boldsymbol{x})\, d\boldsymbol{y}$. By imposing a suitable linear constraint on η (such as $\int_{\mathcal{Y}} \eta(\boldsymbol{y}|\boldsymbol{x})h(\boldsymbol{x})\, d\boldsymbol{x} = 0$ for all square-integrable functions h of $\boldsymbol{x}$) we can make the map $\sigma : \eta \mapsto \phi$ one-to-one. Then the problem of estimating ϕ is reduced to that of estimating η and can thereby be cast into the framework of extended linear modeling. The (conditional) log-likelihood is given by $l(h, \boldsymbol{X}, \boldsymbol{Y}) = h(\boldsymbol{Y}|\boldsymbol{X}) - C(\boldsymbol{X}; h)$. Here $\boldsymbol{W} = (\boldsymbol{X}, \boldsymbol{Y})$ and $\mathcal{U} = \mathcal{X} \times \mathcal{Y}$.

1.3.6 Diffusion Process Regression

Diffusion type processes form a large class of continuous time processes that are widely used for stochastic modeling with application to physical, biological, medical, economic, and social sciences; see Prakasa Rao (1999a, 1999b). Consider a one-dimensional diffusion type process $Y(t)$ that satisfies the stochastic differential equation

$$dY(t) = \eta(t, \boldsymbol{X}(t)) + \sigma(t)\, d\boldsymbol{W}(t), \qquad 0 \le t \le \tau,$$

where $0 < \tau < \infty$ and $\boldsymbol{W}(t)$ is a Wiener process. It is assumed that the diffusion coefficient $\sigma^2(t)$ at time t is a known, predictable, random function of time. It is also assumed that the value at time t of the drift coefficient is an unknown function $\eta(t, \boldsymbol{X}(t))$ of t and the value at time t of a predictable covariate process $\boldsymbol{X}(t) = (X_1(t), \ldots, X_L(t))$, $0 \le t \le \tau$. We refer to η as the *regression function*. Let $Z(t)$, $0 \le t \le \tau$, be a predictable $\{0, 1\}$-valued process. The process $Z(t)$ can be thought of as a censoring indicator: the processes $\boldsymbol{X}(t)$ and $Y(t)$ are only observed when $Z(t) = 1$. The problem of interest is to estimate the function η based on a random sample of n realizations of $\boldsymbol{W} = \{(\boldsymbol{X}(t), Y(t)) : 0 \le t \le \tau \text{ and } Z(t) = 1\}$. The (partial) log-likelihood corresponding to a candidate h for η based on a single observation is given by

$$l(h) = \int Z(t) \frac{h(t, \boldsymbol{X}(t))}{\sigma^2(t)}\, dY(t) - \frac{1}{2} \int Z(t) \frac{h^2(t, \boldsymbol{X}(t))}{\sigma^2(t)}\, dt.$$

This can be seen either by passing to the limit from a discrete-time approximation or by modeling $(\boldsymbol{X}(t), Y(t))$, $0 \le t \le \tau$, as a multidimensional diffusion process and determining the appropriate partial log-likelihood.

1.3.7 Other Contexts

Counting process regression (Huang 2001), event history analysis (Huang and Stone 1998), marked point process regression (Li 2001), proportional hazards regression (Huang, Kooperberg, Stone and Truong 2001), robust regression (Stone 2001), and spectral density estimation (Kooperberg, Stone and Truong 1995d), can also be cast into the framework of concave extended linear models.

1.4 Consistency and Rate of Convergence

In this section we present results on the asymptotic properties of the maximum likelihood estimate $\widehat{\eta}$ in concave extended linear models. As discussed in Section 1.2, the best approximation η^* in $\mathbb{H}$ to the function η of interest can be thought as a general target of estimation whether or not $\eta \in \mathbb{H}$. The existence of η^* has been established in various contexts in papers cited in Section 1.1. We say that $\widehat{\eta}$ is consistent in estimating η^* if $\|\widehat{\eta} - \eta^*\| \to 0$ in probability for some norm $\|\cdot\|$. We will state conditions that ensure consistency and also determine the rates of convergence of $\widehat{\eta}$ to η^*. In the asymptotic analysis, it is natural to let the dimension N_n of the estimation space $\mathbb{G}$ grow with the sample size. The growing dimensionality of the estimation space introduces improved approximation power of this space for increasing sample size.

We assume that the log-likelihood $l(h, \boldsymbol{w})$ and expected log-likelihood $\Lambda(h)$ are well-defined and finite for every bounded function h on $\mathcal{U}$. Since the estimation space $\mathbb{G} \subset \mathbb{H}$ is a finite-dimensional linear space of bounded functions, $\ell(h, \boldsymbol{w})$ and $\Lambda(h)$ are well-defined on $\mathbb{G}$.

Since $\widehat{\eta}$ maximizes the normalized log-likelihood $\ell(g)$, which should be close to the expected log-likelihood $\Lambda(g)$ for $g \in \mathbb{G}$ when the sample size is large, it is natural to think that $\widehat{\eta}$ is directly estimating the best approximation $\bar{\eta} = \text{argmax}_{g \in \mathbb{G}} \Lambda(g)$ in $\mathbb{G}$ to η. If $\mathbb{G}$ is chosen such that $\bar{\eta}$ is close to η^*, then $\widehat{\eta}$ should provide a reasonable estimate of η^*. This motivates the decomposition

$$\widehat{\eta} - \eta^* = (\bar{\eta} - \eta^*) + (\widehat{\eta} - \bar{\eta}),$$

where $\bar{\eta} - \eta^*$ and $\widehat{\eta} - \bar{\eta}$ are referred to, respectively, as the *approximation error* and the *estimation error*.

Given a function h on $\mathcal{U}$, let $\|h\|_\infty = \sup_{\boldsymbol{u} \in \mathcal{U}} |h(\boldsymbol{u})|$ denote its L_∞ norm. Let $\|\cdot\|$ be the normalized L_2 norm relative to Lebesgue measure on $\mathcal{U}$; that is, $\|h\| = \{\int_{\mathcal{U}} h^2(\boldsymbol{u})\, d\boldsymbol{u}/\text{vol}(\mathcal{U})\}^{1/2}$. Note that $\|h\| \leq \|h\|_\infty$. In our asymptotic theory we will use $\|\widehat{\eta} - \eta^*\|$ to measure the discrepancy between $\widehat{\eta}$ and η^*.

Sometimes it is more natural to use other norms to measure the discrepancy. In the regression context, for example, one would use $\|h\|_0^2 =$

$E[h^2(\boldsymbol{X})]$ where the expectation is with respect to the distribution of the covariates $\boldsymbol{X}$. Such a norm is closely related to the mean prediction error. Precisely, the mean prediction error of a candidate h for the regression function η is defined by $\mathrm{PE}(h) = E\{[Y^* - h(\boldsymbol{X}^*)]^2\}$, where $(\boldsymbol{X}^*, Y^*)$ is a pair of observations independent of the observed data and having the same distribution as $(\boldsymbol{X}, Y)$. It is easily seen that

$$\mathrm{PE}(h) = E[\mathrm{var}(Y|\boldsymbol{X})] + \|h - \eta\|_0^2.$$

Under mild conditions (for example, if the density of $\boldsymbol{X}$ is bounded away from zero and infinity), the norm $\|\cdot\|_0$ is equivalent to the normalized L_2-norm $\|\cdot\|$, that is, there are positive constants c_1 and c_2 such that $c_1\|h\| \le \|h\|_0 \le c_2\|h\|$ for any square-integrable function h on $\mathcal{U}$. Thus our asymptotic results, presented for the norm $\|\cdot\|$, can also be stated in terms of the more natural norm $\|\cdot\|_0$. (This is generally true for other concave extended linear models, though we illustrated this point only in the regression case.)

Given random variables V_n for $n \ge 1$, let $V_n = O_P(b_n)$ mean that $\lim_{c\to\infty} \limsup_n P(|V_n| \ge cb_n) = 0$ and let $V_n = o_P(b_n)$ mean that $\lim_n P(|V_n| \ge cb_n) = 0$ for $c > 0$. Set

$$\rho_n = \inf_{g\in\mathbb{G}} \|g - \eta^*\|_\infty.$$

Under various mild assumptions on η^* and $\mathbb{G}$, $\rho_n \to 0$ as $n \to \infty$.

Proposition 1.1. [Huang 2001] *Under appropriate conditions, $\bar{\eta}$ exists uniquely for n sufficiently large and*

$$\|\bar{\eta} - \eta^*\|^2 = O(\rho_n^2).$$

Moreover, $\widehat{\eta}$ exists uniquely except on an event whose probability tends to zero as $n \to \infty$ and

$$\|\widehat{\eta} - \bar{\eta}\|^2 = O_P\Big(\frac{N_n}{n}\Big).$$

Consequently,

$$\|\widehat{\eta} - \eta^*\|^2 = O_P\Big(\frac{N_n}{n} + \rho_n^2\Big).$$

In particular, $\widehat{\eta}$ is consistent in estimating η^*; that is, $\|\widehat{\eta} - \eta^*\| = o_P(1)$.

According to this result, the squared norm of the estimation error is bounded in probability by the inverse of the number of observations per parameter, while the squared norm of the approximation error is bounded above by a multiple of the best obtainable approximation rate in the estimation space to the target function.

The main technical requirement for Proposition 1.1 is that the log-likelihood be suitably concave. (See the cited paper for details.) Specifically, it is required that the following two conditions hold:

C1. For any positive constant K, there are positive numbers M_1 and M_2 such that

$$-M_1\|h_2-h_1\|^2 \le \frac{d^2}{d\alpha^2}\Lambda(h_1+\alpha(h_2-h_1)) \le -M_2\|h_2-h_1\|^2$$

for $h_1, h_2 \in \mathbb{H}$ with $\|h_1\|_\infty \le K$ and $\|h_2\|_\infty \le K$ and $0 \le \alpha \le 1$.

C2. (i)

$$\sup_{g\in\mathbb{G}} \frac{\left|\frac{d}{d\alpha}\ell(\bar\eta+\alpha g)\Big|_{\alpha=0}\right|}{\|g\|} = O_P\left(\left(\frac{N_n}{n}\right)^{1/2}\right);$$

(ii) for any positive constant K, there is a positive number M such that

$$\frac{d^2}{d\alpha^2}\ell(g_1+\alpha(g_2-g_1)) \le -M\|g_2-g_1\|^2, \qquad 0\le\alpha\le 1,$$

for $g_1, g_2 \in \mathbb{G}$ with $\|g_1\|_\infty \le K$ and $\|g_2\|_\infty \le K$, except on an event whose probability tends to zero as $n \to \infty$.

Proposition 1.1 treats a general estimation space $\mathbb{G}$. It is readily applicable to fixed knot spline estimates when the knot positions are prespecified but the number of knots is allowed to increase with the sample size. Suppose $\mathcal{U}$ is the Cartesian product of compact intervals $\mathcal{U}_1, \ldots, \mathcal{U}_L$. Consider the saturated model, in which η is a bounded function and no structural assumptions are imposed on η (that is, $\mathbb{H}$ is the space of all square-integrable functions on $\mathcal{U}$). Correspondingly, $\eta^* = \eta$. Suppose η has bounded p-th derivative (for an integer p), or more generally, suppose that η is p-smooth for a specified positive number p; that is, η is k times continuously differentiable on $\mathcal{U}$, where k is the greatest integer less than p, and all the kth-order mixed partial derivatives of η satisfy a Hölder condition with exponent $p-k$.

Let $\mathbb{G}_l$ be a linear space of splines having degree $q \ge p-1$ for $1 \le l \le L$ and let $\mathbb{G}$ be the tensor product of $\mathbb{G}_1, \ldots, \mathbb{G}_L$, which is the space spanned by functions of the form $g_1(u_1)\cdots g_L(u_L)$, where g_l runs over $\mathbb{G}_l$ for $1 \le l \le L$. Suppose the knots have bounded mesh ratio (that is, the ratios of the differences between consecutive knots are bounded away from zero and infinity uniformly in n). Let a_n denote the smallest distance between two consecutive knots. For two sequences of positive numbers b_{1n} and b_{2n}, let $b_{1n} \asymp b_{2n}$ mean that the ratio b_{1n}/b_{2n} is bounded away from 0 and infinity. Then $N_n \asymp a_n^{-L}$ and $\rho_n \asymp a_n^p \asymp N_n^{-p/L}$. According to Proposition 1.1,

$$\|\widehat\eta-\eta\|^2 = O_P\Big(\frac{1}{na_n^L} + a_n^{2p}\Big).$$

In particular, for $a_n \asymp n^{-1/(2p+L)}$, we have that

$$\|\widehat\eta-\eta\|^2 = O_P(n^{-2p/(2p+L)}).$$

The choice of $a_n \asymp n^{-1/(2p+L)}$ balances the contributions to the error bound from the estimation error and the approximation error, that is, $1/(na_n^L) \asymp a_n^{2p}$. The resulting rate of convergence $n^{-2p/(2p+L)}$ actually is optimal: no estimate has a faster rate of convergence uniformly over the class of p-smooth functions [Stone (1982)]. The rate of convergence depends on two quantities: the specified smoothness p of the target function and the dimension L of the domain on which the target function is defined. Note the dependence of the rate of convergence on the dimension L: given the smoothness p, the larger the dimension, the slower the rate of convergence; moreover, the rate of convergence tends to zero as the dimension tends to infinity. This provides a mathematical description of a phenomenon commonly known as the "curse of dimensionality."

The following question arises naturally. What if we restrict attention to additive estimates

$$\widehat{\eta}(\boldsymbol{x}) = \widehat{\eta}_1(x_1) + \cdots + \widehat{\eta}_L(x_L)$$

of the best additive approximation

$$\eta^*(\boldsymbol{x}) = \eta_1^*(x_1) + \cdots + \eta_L^*(x_L)$$

to η? Can we now achieve the rate of convergence $n^{-p/(2p+1)}$?

1.5 Functional ANOVA

Imposing structures such as additivity on an unknown multivariate function indeed can imply faster rates of convergence of the corresponding estimate and thus tame the curse of dimensionality. In this section we will discuss how to impose general structural assumptions using functional ANOVA decompositions. We will also study rates of convergence of the maximum likelihood estimates when the model and estimation spaces are constructed in some structured way.

To introduce the notion of functional ANOVA, it is helpful to look at a simple example. Suppose that $\mathcal{U} = \mathcal{U}_1 \times \mathcal{U}_2 \times \mathcal{U}_3$, where $\mathcal{U}_1$, $\mathcal{U}_2$ and $\mathcal{U}_3$ are compact intervals, each having positive length. Any square-integrable function on $\mathcal{U}$ can be decomposed as

$$\begin{aligned}\eta(\boldsymbol{u}) \quad = \quad & \eta_\emptyset + \eta_{\{1\}}(u_1) + \eta_{\{2\}}(u_2) + \eta_{\{3\}}(u_3) + \eta_{\{1,2\}}(u_1,u_2) \\ & + \eta_{\{1,3\}}(u_1,u_3) + \eta_{\{2,3\}}(u_2,u_3) + \eta_{\{1,2,3\}}(u_1,u_2,u_3).\end{aligned} \tag{1.3}$$

For identifiability, we require that each nonconstant component be orthogonal to all possible values of the corresponding lower-order components relative to an appropriate inner product. The expression (1.3) can then be viewed as a functional analysis of variance (ANOVA) decomposition. Correspondingly, we call $\eta_\emptyset$ the constant component; $\eta_{\{1\}}(u_1), \eta_{\{2\}}(u_2)$ and $\eta_{\{3\}}(u_3)$ the main effect components; $\eta_{\{1,2\}}(u_1,u_2)$, $\eta_{\{1,3\}}(u_1,u_3)$ and

$\eta_{\{2,3\}}(u_2, u_3)$ the two-factor interaction components; and $\eta_{\{1,2,3\}}(u_1, u_2, u_3)$ the three-factor interaction component. The right side of (1.3) is referred to as the ANOVA decomposition of η.

If no structural assumption is imposed on η, we need to consider all the components in the above ANOVA decomposition. The resulting model is saturated. However, the desire to tame the curse of dimensionality leads us to employ unsaturated models, which discard some terms in the ANOVA decomposition. For example, removing all the interaction components in the above ANOVA decomposition of η, we get the additive model

$$\eta(\boldsymbol{u}) = \eta_\emptyset + \eta_{\{1\}}(u_1) + \eta_{\{2\}}(u_2) + \eta_{\{3\}}(u_3). \tag{1.4}$$

We can also include some selected interactions in the model and still keep the model manageable. For example, the following model includes just the interaction between u_1 and u_2:

$$\eta(\boldsymbol{u}) = \eta_\emptyset + \eta_{\{1\}}(u_1) + \eta_{\{2\}}(u_2) + \eta_{\{3\}}(u_3) + \eta_{\{1,2\}}(u_1, u_2). \tag{1.5}$$

To fit these models using maximum likelihood, it is necessary to choose the estimation space $\mathbb{G}$ to respect the imposed structure on η. As a result the estimate will have the same structure. For example, by choosing $\mathbb{G}$ appropriately, the maximum likelihood estimate will have the forms

$$\widehat{\eta}(\boldsymbol{u}) = \widehat{\eta}_\emptyset + \widehat{\eta}_{\{1\}}(u_1) + \widehat{\eta}_{\{2\}}(u_2) + \widehat{\eta}_{\{3\}}(u_3)$$

and

$$\widehat{\eta}(\boldsymbol{u}) = \widehat{\eta}_\emptyset + \widehat{\eta}_{\{1\}}(u_1) + \widehat{\eta}_{\{2\}}(u_2) + \widehat{\eta}_{\{3\}}(u_3) + \widehat{\eta}_{\{1,2\}}(u_1, u_2).$$

for models (1.4) and (1.5), respectively.

In general, suppose that $\mathcal{U} = \mathcal{U}_1 \times \cdots \times \mathcal{U}_L$ for some positive integer L, where each $\mathcal{U}_l$ is a compact subset of some Euclidean space and it has positive volume in that space. If η is square-integrable, we can define its ANOVA decomposition in a similar manner as above. Selecting certain terms in its ANOVA decomposition in the modeling process corresponds to imposing a particular structural assumption on η. Specifically, let $\mathcal{S}$ be a hierarchical collection of subsets of $\{1, \ldots, L\}$; by hierarchical we mean that if $s \in \mathcal{S}$, then $r \in \mathcal{S}$ for every subset r of s. Consider the model space

$$\mathbb{H} = \left\{ \sum_{s \in \mathcal{S}} h_s : h_s \in \mathbb{H}_s \text{ for } s \in \mathcal{S} \right\}, \tag{1.6}$$

where $\mathbb{H}_s$ is the space of square-integrable functions that depends only on u_l, $l \in s$. Note that the set $\mathcal{S}$ describes precisely which interaction terms are included in the model. For example, the additive model (1.4) and the model (1.5) with a single interaction component correspond to $\mathcal{S} = \{\emptyset, \{1\}, \{2\}, \{3\}\}$ and $\mathcal{S} = \{\emptyset, \{1\}, \{2\}, \{3\}, \{1, 2\}\}$, respectively. Note that the best approximation η^* in $\mathbb{H}$ to η, which is the general target for

estimation in an extended linear model, has the form

$$\eta^* = \sum_{s \in \mathcal{S}} \eta_s^*, \tag{1.7}$$

where η_s^* is a member of $\mathbb{H}_s$ for $s \in \mathcal{S}$ and $\eta_\emptyset^*$ is a constant.

Again, the maximum likelihood method can be used to do the estimation and the estimation space $\mathbb{G}$ is chosen to take the form

$$\mathbb{G} = \Big\{ \sum_{s \in \mathcal{S}} g_s : g_s \in \mathbb{G}_s \text{ for } s \in \mathcal{S} \Big\}, \tag{1.8}$$

where $\mathbb{G}_s$ is the tensor product space of $\mathbb{G}_l, l \in s$, and for each $1 \le l \le L$, $\mathbb{G}_l$ is an appropriate finite-dimensional space of functions of u_l that contains all constant functions. The resulting estimate should have the form

$$\widehat{\eta} = \sum_{s \in \mathcal{S}} \widehat{\eta}_s, \tag{1.9}$$

where $\widehat{\eta}_s$ is a member of $\mathbb{G}_s$ for $s \in \mathcal{S}$ and $\widehat{\eta}_\emptyset$ is a constant.

The asymptotic theory concerns the consistency of $\widehat{\eta}$ in estimating η^* and the consistency of the components $\widehat{\eta}_s$ of $\widehat{\eta}$ in estimating the corresponding components of η_s^* of η^*. (Suitable identifiability constraints must be imposed on the components of $\widehat{\eta}$ and η^* as indicated in the discussion following (1.3) and in the discussion at the end of this section.) The consistency of the estimated components is desirable since one would like examination of the components of $\widehat{\eta}$ to shed light on the shape of η^* and its components.

We should allow the dimensions of $\mathbb{G}_l$ and thus those of $\mathbb{G}_s$ to depend on the sample size. Set $N_s = \dim(\mathbb{G}_s)$ and $\rho_s = \inf_{g \in \mathbb{G}_s} \|g - \eta_s^*\|_\infty$ for $s \in \mathcal{S}$.

Proposition 1.2. [Huang 2001] *Under appropriate conditions,*

$$\|\widehat{\eta} - \eta^*\|^2 = O_P\Big(\sum_{s \in \mathcal{S}} \Big(\frac{N_s}{n} + \rho_s^2 \Big) \Big)$$

and

$$\|\widehat{\eta}_s - \eta_s^*\|^2 = O_P\Big(\sum_{s \in \mathcal{S}} \Big(\frac{N_s}{n} + \rho_s^2 \Big) \Big), \qquad s \in \mathcal{S}.$$

Suppose each η_s^*, $s \in \mathcal{S}$, in (1.7) is p-smooth. Let the estimation space be given by (1.8) with each $\mathbb{G}_l$ being a linear space of degree q splines on $\mathcal{U}_l$ as in the previous section with $q \ge p - 1$. Let $\#(B)$ denote the cardinality (number of members) of a set B. Then $N_s \asymp a_n^{-\#(s)}$ and $\rho_s \asymp a_n^p$, $s \in \mathcal{S}$. Set $d = \max_{s \in \mathcal{S}} \#(s)$. According to Proposition 1.2,

$$\|\widehat{\eta} - \eta^*\|^2 = O_P\Big(\frac{1}{n a_n^d} + a_n^{2p} \Big)$$

and

$$\|\widehat{\eta}_s - \eta_s^*\|^2 = O_P\Big(\frac{1}{na_n^d} + a_n^{2p}\Big), \qquad s \in \mathcal{S}.$$

In particular, for $a_n \asymp n^{-1/(2p+d)}$, we have that

$$\|\widehat{\eta} - \eta^*\|^2 = O_P(n^{-2p/(2p+d)})$$

and

$$\|\widehat{\eta}_s - \eta_s^*\|^2 = O_P(n^{-2p/(2p+d)}), \qquad s \in \mathcal{S}.$$

The rate of convergence $n^{-2p/(2p+d)}$ for an unsaturated model should be compared with the rate $n^{-2p/(2p+L)}$ for the saturated model. Note here that d is the maximum order of interaction among the components of η^* and $\widehat{\eta}$. For the additive model, we have $d = 1$, so the rate is $n^{-2p/(2p+1)}$, which is the same as that for estimating a one-dimensional target function. For models with interaction components of order two and no higher-order interaction components, we have $d = 2$ and corresponding rate of convergence $n^{-2p/(2p+2)}$, the same rate for estimating a two-dimensional target function. Hence, for large L, we can achieve much faster rates of convergence by considering structural models involving only low-order interactions in the ANOVA decomposition of the target function and thereby tame the curse of dimensionality. (Of course, if η does not itself possess the imposed structure, then the faster rates of convergence are obtained at the expense of estimating η^* rather than η.)

As discussed at the beginning of this section, we define the ANOVA decomposition of a function by forcing each nonconstant component to be orthogonal to all possible values of the corresponding lower-order components relative to an appropriate inner product. Usually, one uses a theoretical inner product to decompose η^* and an empirical inner product to decompose $\widehat{\eta}$. For example, in the regression case, it is natural to define the theoretical and empirical inner products by $\langle h_1, h_2\rangle = E[h_1(\boldsymbol{X})h_2(\boldsymbol{X})]$ and $\langle h_1, h_2\rangle_n = (1/n)\sum_i [h_1(\boldsymbol{X}_i)h_2(\boldsymbol{X}_i)]$. The reason for using different inner products is that the theoretical inner product is often defined in terms of the data-generating distribution and hence depends on unknown quantities, while the empirical inner product must be totally determined by the data since it will be used to decompose the estimate. An important necessary condition for Proposition 1.2 is that the two inner products or the corresponding norms are close; precisely, $\sup_{g\in\mathbb{G}} |\|g\|_n/\|g\| - 1| = o_P(1)$.

1.6 Free Knot Splines in Extended Linear Models

In this section we extend the results in Section 1.4 to handle free knot splines; that is, the knot positions are treated as free parameters to be determined from the data.

Consider a concave extended linear model specified by the log-likelihood $l(h, \boldsymbol{W})$ and model space $\mathbb{H}$. Let $\mathbb{G}_{\boldsymbol{\gamma}}$, $\boldsymbol{\gamma} \in \Gamma$, be a collection of finite-dimensional linear subspaces of $\mathbb{H}$. We assume that the functions in each such space $\mathbb{G}_{\boldsymbol{\gamma}}$ are bounded and call $\mathbb{G}_{\boldsymbol{\gamma}}$ an estimation space. Here $\boldsymbol{\gamma}$ can be thought of as the knot positions when the estimation space consists of spline functions, and our interest lies in choosing the knot positions using the data. It is assumed that the spaces $\mathbb{G}_{\boldsymbol{\gamma}}$, $\boldsymbol{\gamma} \in \Gamma$, have a common dimension and that the index set Γ is a compact subset of $\mathbb{R}^J$ for some positive integer J. We allow $\dim(\mathbb{G}_{\boldsymbol{\gamma}})$, Γ, and J to vary with the sample size n.

For each fixed $\boldsymbol{\gamma} \in \Gamma$, the maximum likelihood estimate is given by $\widehat{\eta}_{\boldsymbol{\gamma}} = \max_{g \in \mathbb{G}_{\boldsymbol{\gamma}}} \ell(g)$. In order to let the data select which estimation space to use, we choose $\widehat{\boldsymbol{\gamma}} \in \Gamma$ such that $\ell(\widehat{\eta}_{\widehat{\boldsymbol{\gamma}}}) = \max_{\boldsymbol{\gamma} \in \Gamma} \ell(\widehat{\eta}_{\boldsymbol{\gamma}})$. (Such a $\widehat{\boldsymbol{\gamma}}$ exists under mild conditions.) We will study the benefit of allowing the flexibility to select the estimation space from among a big collection of such spaces. Specifically we will study the rate of convergence to zero of $\widehat{\eta}_{\widehat{\boldsymbol{\gamma}}} - \eta^*$, where η^* is the best approximation to η in $\mathbb{H}$. For $\boldsymbol{\gamma} \in \Gamma$, set $N_n = \dim(\mathbb{G}_{\boldsymbol{\gamma}})$ and $\rho_{n\boldsymbol{\gamma}} = \inf_{g \in \mathbb{G}_{\boldsymbol{\gamma}}} \|g - \eta^*\|_\infty$.

It follows from Proposition 1.1 that, $\|\widehat{\eta}_{\boldsymbol{\gamma}} - \eta^*\|^2 = O_P(\rho_{n\boldsymbol{\gamma}}^2 + N_n/n)$ for each fixed $\boldsymbol{\gamma} \in \Gamma$. Let $\boldsymbol{\gamma}^*$ be such that $\rho_{n\boldsymbol{\gamma}^*} = \inf_{\boldsymbol{\gamma} \in \Gamma} \rho_{n\boldsymbol{\gamma}}$. (Such a $\boldsymbol{\gamma}^*$ exists under mild conditions.) Then

$$\|\widehat{\eta}_{\boldsymbol{\gamma}^*} - \eta^*\|^2 = O_P\Big(\rho_{n\boldsymbol{\gamma}^*}^2 + \frac{N_n}{n}\Big) = O_P\Big(\inf_{\boldsymbol{\gamma} \in \Gamma} \rho_{n\boldsymbol{\gamma}}^2 + \frac{N_n}{n}\Big).$$

Thus,

$$\inf_{\boldsymbol{\gamma} \in \Gamma} \|\widehat{\eta}_{\boldsymbol{\gamma}} - \eta^*\|^2 \le \|\widehat{\eta}_{\boldsymbol{\gamma}^*} - \eta^*\|^2 = O_P\Big(\inf_{\boldsymbol{\gamma} \in \Gamma} \rho_{n\boldsymbol{\gamma}}^2 + \frac{N_n}{n}\Big).$$

It is natural to expect that, with $\boldsymbol{\gamma}$ estimated by $\widehat{\boldsymbol{\gamma}}$, the squared L_2 norm $\|\widehat{\eta}_{\widehat{\boldsymbol{\gamma}}} - \eta^*\|^2$ of the difference between the estimate and the target will be not much larger than the ideal quantity $\inf_{\boldsymbol{\gamma} \in \Gamma} \|\widehat{\eta}_{\boldsymbol{\gamma}} - \eta^*\|^2$. Hence we hope that $\|\widehat{\eta}_{\widehat{\boldsymbol{\gamma}}} - \eta^*\|^2$ will be not much larger than $\inf_{\boldsymbol{\gamma} \in \Gamma} \rho_{n\boldsymbol{\gamma}}^2 + N_n/n$ in probability. This is confirmed by the next result.

Let $V_{n\boldsymbol{\gamma}} = O_P(b_{n\boldsymbol{\gamma}})$ uniformly over $\boldsymbol{\gamma} \in \Gamma$ mean that

$$\lim_{c \to \infty} \limsup_n P(|V_{n\boldsymbol{\gamma}}| \ge c b_{n\boldsymbol{\gamma}} \text{ for some } \boldsymbol{\gamma} \in \Gamma) = 0,$$

where $b_{n\boldsymbol{\gamma}} > 0$ for $n \ge 1$ and $\boldsymbol{\gamma} \in \Gamma$. As in Section 1.4, it is enlightening to decompose the error into a stochastic part and a systematic part for each fixed $\boldsymbol{\gamma} \in \Gamma$:

$$\widehat{\eta}_{\boldsymbol{\gamma}} - \eta^* = (\widehat{\eta}_{\boldsymbol{\gamma}} - \bar{\eta}_{\boldsymbol{\gamma}}) + (\bar{\eta}_{\boldsymbol{\gamma}} - \eta^*),$$

where $\widehat{\eta}_{\boldsymbol{\gamma}} - \bar{\eta}_{\boldsymbol{\gamma}}$ is referred to as the *estimation error* and $\bar{\eta}_{\boldsymbol{\gamma}} - \eta$ as the approximation error.

Proposition 1.3. [Stone and Huang 2002a] *Under appropriate conditions, for n sufficiently large, $\bar{\eta}_{\boldsymbol{\gamma}}$ exists uniquely for $\boldsymbol{\gamma} \in \Gamma$ and*

$$\|\bar{\eta}_{\boldsymbol{\gamma}} - \eta^*\|^2 = O(\rho^2_{n\boldsymbol{\gamma}})$$

uniformly over $\boldsymbol{\gamma} \in \Gamma$. Moreover, except on an event whose probability tends to zero as $n \to \infty$, $\widehat{\eta}_{\boldsymbol{\gamma}}$ exists uniquely for $\boldsymbol{\gamma} \in \Gamma$ and

$$\sup_{\boldsymbol{\gamma}\in\Gamma} \|\widehat{\eta}_{\boldsymbol{\gamma}} - \bar{\eta}_{\boldsymbol{\gamma}}\|^2 = O_P\Big(\frac{N_n}{n}\Big).$$

Consequently,

$$\|\widehat{\eta}_{\boldsymbol{\gamma}} - \eta^*\|^2 = O_P\Big(\rho^2_{n\boldsymbol{\gamma}} + \frac{N_n}{n}\Big)$$

uniformly over $\boldsymbol{\gamma} \in \Gamma$. In addition,

$$\|\widehat{\eta}_{\widehat{\boldsymbol{\gamma}}} - \eta^*\|^2 = O_P\Big(\inf_{\boldsymbol{\gamma}\in\Gamma} \rho^2_{n\boldsymbol{\gamma}} + (\log n)\frac{N_n}{n}\Big).$$

In the previous theoretical results for fixed knot splines, the squared norms of the approximation error and the estimation error were shown to be bounded above by multiples of $\rho^2_{n\boldsymbol{\gamma}}$ and N_n/n, respectively. Here these results are shown to hold uniformly over the free knot sequences $\boldsymbol{\gamma} \in \Gamma$. Finally, combining the results for the approximation error and the estimation error and incorporating a corresponding result for the maximum likelihood estimation of the knot positions, we get the rate of convergence for the free-knot spline estimate..

The benefit of using free-knot splines is that a smaller approximation error can be achieved; presumably, for certain functions the best approximation rate obtainable for free-knot splines for a collection of knot positions (i.e., $\inf_{\boldsymbol{\gamma}\in\Gamma} \rho_{n\boldsymbol{\gamma}}$) would be much smaller than the best approximation rate obtainable for fixed-knot splines (i.e., $\rho_{n\boldsymbol{\gamma}}$ for a fixed $\boldsymbol{\gamma}$). The cost is a small inflation of the variance; there is an extra $\log n$ term in the variance bound for the free-knot spline estimate. We are currently unable, in any context, to verify either that this extra $\log n$ factor is necessary or that it is unnecessary.

As a simple illustration of the improved rate of convergence of free-knot spline estimate over fixed-knot spline estimate, consider estimating the regression function $\eta(x)$ in the regression model $Y = \eta(X) + \epsilon$, where X has a uniform distribution on $[-1, 1]$ and ϵ has a normal distribution with mean 0 and variance σ^2. Suppose $\eta(x) = x^2 + |x|$. Clearly, the first derivative of η at 0 does not exist, which implies a slow fixed-knot spline approximation rate if there is no knot very close to 0. Specifically, let the estimation space $\mathbb{G}$ be the space of linear splines on $[-1, 1]$ with $2J_n$ equally spaced knots located at $\pm(2k-1)/(2J_n-1)$, $k = 1, \ldots, J_n$, and let $\widehat{\eta}$ be the least squares estimate on $\mathbb{G}$. It can be shown that for some positive constant c and large n, $P(\|\widehat{\eta} - \eta\|^2 > cJ_n/n + cJ_n^{-3}) > 1/3$ and thus for any

choice of J_n, $P(\|\widehat{\eta} - \eta\|^2 > cn^{-3/4}) > 1/3$. On the other hand, let $\mathbb{G}_\gamma$ be the space of linear splines on $[-1, 1]$ with $2J_n - 1$ knots located at γ and $\pm(2k-1)/(2J_n-1)$, $k = 2, \ldots, J_n$, where $-1/(2J_n-1) \le \gamma \le 1/(2J_n-1)$. Here, we simply replace two fixed knots $\pm 1/(2J_n-1)$ in the previous setup by one free-knot at γ. Using Proposition 1.3, we can show that $\|\widehat{\eta}_{\widehat{\gamma}} - \eta\|^2 = O_P(J_n^{-4} + J_n \log n/n)$. Hence, for $J_n \asymp (n/\log n)^{1/5}$, the convergence rate of the free-knot spline estimate satisfies $\|\widehat{\eta}_{\widehat{\gamma}} - \eta\|^2 = O_P((n^{-1}\log n)^{4/5})$, which is faster than that of the fixed-knot spline estimate. Technical details of this example can be found in Stone and Huang (2002b).

The main technical requirement for Proposition 1.3 is that Conditions C1 and C2 in Section 1.4 be strenthened to hold uniformly over $\boldsymbol{\gamma} \in \Gamma$. For data driven choices of $\boldsymbol{\gamma}$, it is also required that

$$|\ell(\bar{\eta}_{\boldsymbol{\gamma}}) - \ell(\eta^*) - [\Lambda(\bar{\eta}_{\boldsymbol{\gamma}}) - \Lambda(\eta^*)]| = O_P\left(\log^{1/2} n \Big[\|\bar{\eta}_{\boldsymbol{\gamma}} - \eta^*\| \Big(\frac{N_n}{n} \Big)^{1/2} + \frac{N_n}{n} \Big] \right)$$

uniformly over $\boldsymbol{\gamma} \in \Gamma$.

These conditions have been verified separately in each of various contexts—regression, logistic regression, density estimation in Stone and Huang (2002a), diffusion process regression in Stone and Huang (2002b), and marked point process regression in Li (2001).

1.7 Methodological Implications

The successful development of theory, such as that surveyed in this paper, suggests that closely related adaptive methodology based on stepwise selection of basis functions should be worth pursuing in practice. Similarly, the successful development of a theoretical synthesis that applies to a number of seemingly different contexts suggests that the corresponding adaptive methodologies for these contexts should have similar performance.

Since concavity is crucial in the theory that has been presented here, it is presumably also helpful in developing corresponding methodologies, for example, to avoid getting stuck in local optima that are far from being globally optimal. Since additive and more general models containing only low-order interactions have faster rates of convergence than models containing higher-order interactions, the corresponding methodologies are also presumably worthy of pursuit.

These implications of theory for methodology have largely been borne out in practice. For various methodological developments that have been influenced, to one extent or another, by the theory surveyed in this paper, see Friedman and Silverman (1989); Friedman (1991); Kooperberg and Stone (1992); Kooperberg, Stone and Truong (1995a, 1995c); Kooperberg, Bose and Stone (1997); and Kooperberg and Stone (1999).

So far, there has been no statistical theory that applies rigorously to stepwise selection of basis functions. Now that there is a statistical theory for modeling with free knot splines, it is reasonable to view stepwise knot selection, which can be thought of as a special case of the stepwise selection of basis functions, as a computationally efficient shortcut to modeling with free knot splines as in Lindstrom (1999).

References

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

DeVore, R. A. and Lorentz, G. G. (1993). *Constructive Approximation*. Springer-Verlag, Berlin.

Friedman, J. H. (1991). Multivariate adaptive regression splines (with discussion). *The Annals of Statistics* **19**, 1–141.

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

Hansen, M. H. (1994). *Extended Linear Models, Multivariate Splines, and ANOVA*. Ph. D. Dissertation, Department of Statistics, University of California at Berkeley.

Huang, J. Z. (1998a). Projection estimation in multiple regression with application to functional ANOVA models. *The Annals of Statistics* **26**, 242–272.

Huang, J. Z. (1998b). Functional ANOVA models for generalized regression. *Journal of Multivariate Analysis* **67**, 49–71.

Huang, J. Z. (2001). Concave extended linear modeling: a theoretical synthesis. *Statistica Sinica* **11**, 173–197.

Huang, J. Z. and Stone, C. J. (1998). The L_2 rate of convergence for event history regression with time-dependent covariates. *Scandinavian Journal of Statistics* **25**, 603–620.

Huang, J. Z., Kooperberg, C., Stone, C. J. and Truong, Y. K. (2000). Functional ANOVA modeling for proportional hazards regression. *The Annals of Statistics* **28**, 960–999.

Kooperberg, C. and Stone, C. J. (1992). Logspline density estimation for censored data. *Journal of Computational and Graphical Statistics* **1**, 301–328.

Kooperberg, C., Bose, S. and Stone, C. J. (1997). Polychotomous regression. *Journal of the American Statistical Association* **92**, 117–127.

Kooperberg, C. and Stone, C. J. (1999). Stochastic optimization methods for fitting polyclass and feed-forward neural network models. *Journal of Computational and Graphical Statistics* **8**, 169–189.

Kooperberg, C., Stone, C. J. and Truong, Y. K. (1995a). Hazard regression. *Journal of the American Statistical Association* **90**, 78–94.

Kooperberg, C., Stone, C. J. and Truong, Y. K. (1995b). The L_2 rate of convergence for hazard regression. *Scandinavian Journal of Statistics* **22**, 143–157.

Kooperberg, C., Stone, C. J. and Truong, Y. K. (1995c). Logspline estimation of a possibly mixed spectral distribution. *Journal of Time Series Analysis* **16**, 359–388.

Kooperberg, C., Stone, C. J. and Truong, Y. K. (1995d). Rate of convergence for logspline spectral density estimation. *Journal of Time Series Analysis* **16**, 389–401.

Li, W. (2000). *Modeling Marked Point Processes with an Application to Currency Exchange Rates*. Ph. D. Dissertation, Department of Statistics, University of California at Berkeley.

Lindstrom, M. J. (1999). Penalized estimation of free-knot splines. *Journal of Computational and Graphical Statistics* **8**, 333–352.

Prakasa Rao, B. L. S. (1999a). *Semimartingales and their Statistical Inference*. Chapman & Hall/CRC, Boca Raton, Florida.

Prakasa Rao, B. L. S. (1999b). *Statistical Inference for Diffusion Type Processes*. Oxford University Press, New York.

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

Stone, C. J. (1982). Optimal global rates of convergence for nonparametric regression. *The Annals of Statistics* **10**, 1348–1360.

Stone, C. J. (1985). Additive regression and other nonparametric models. *The Annals of Statistics* **13**, 689–705.

Stone, C. J. (1986). The dimensionality reduction principle for generalized additive models. *The Annals of Statistics* **14**, 590–606.

Stone, C. J. (1990). Large-sample inference for log-spline models. *The Annals of Statistics* **18**, 717–741.

Stone, C. J. (1991). Asymptotics for doubly flexible logspline response models. *The Annals of Statistics* **19**, 1832–1854.

Stone, C. J. (1994). The use of polynomial splines and their tensor products in multivariate function estimation (with discussion). *The Annals of Statistics* **22**, 118–184.

Stone, C. J. (2001). Rate of convergence for robust nonparametric regression. Manuscript.

Stone, C. J., Hansen, M., Kooperberg, C. and Truong, Y. K. (1997). Polynomial splines and their tensor products in extended linear modeling (with discussion). *The Annals of Statistics* **25**, 1371–1470.

Stone, C. J. and Huang, J. Z. (2002a). Free knot splines in concave extended linear modeling. *Journal of Statistical Planning and Inference C. R. Rao Volume, Part II*. To appear.

Stone, C. J. and Huang, J. Z. (2002b). Statistical modeling of diffusion processes with free knot splines. Manuscript.