Statistica Sinica **11**(2001), 173-197

# CONCAVE EXTENDED LINEAR MODELING: A THEORETICAL SYNTHESIS

Jianhua Z. Huang

*University of Pennsylvania*

*Abstract:* Extended linear modeling provides a flexible framework for functional estimation problems with multiple covariates. Such problems include ordinary and generalized regression, density and conditional density estimation, hazard regression, spectral density estimation and polychotomous regression. In this paper, we develop a general theory on the rate of convergence of maximum likelihood estimation in extended linear modeling. The role of concavity of the log-likelihood function is highlighted. Both correctly specified and misspecified models are treated in a unified manner. Applications are made to a variety of structural models: saturated models, partly linear models, and functional ANOVA models. Two specific contexts, counting process regression and conditional density estimation, are used to illustrate the general theory.

*Key words and phrases:* Conditional density estimation, counting process, functional ANOVA model, generalized additive model, generalized linear model, hazard regression, method of sieves, nonparametric, partly linear model, rates of convergence, splines, tensor product, time-dependent covariates.

## 1. Introduction

Extended linear modeling (Hansen (1994) and Stone, Hansen, Kooperberg and Truong (1997)) provides a flexible framework for functional estimation problems with multiple covariates. Such problems include ordinary and generalized regression, density and conditional density estimation, hazard regression, spectral density estimation, and polychotomous regression. The purpose of this paper is to develop a general theory on the rate of convergence in estimation. The concavity of the log-likelihood function plays a crucial role.

Let $\eta$ be the function of interest defined on a domain $\mathcal{U}$. In many applications, $\mathcal{U}$ is a finite dimensional Euclidean space. The function $\eta$ is related to the distribution of a (possibly vector-valued) random variable $\mathbf{W}$, taking values in an arbitrary set $\mathcal{W}$. To estimate $\eta$, an i.i.d. sample, $\mathbf{W}_1, \ldots, \mathbf{W}_n$, from the distribution of $\mathbf{W}$ is observed.

For the purpose of imposing specific structures on $\eta$, we introduce a linear space $\mathbb{H}$ of real-valued functions on $\mathcal{U}$ and assume that $\eta \in \mathbb{H}$. We refer to $\mathbb{H}$ as the

model space. By choosing $\mathbb{H}$ appropriately, we get many familiar structural models, such as classical linear models ($\mathbb{H} = \{\eta(\mathbf{u}) = \mathbf{u}^T\beta, \mathbf{u} \in \mathcal{U}\}$), additive models ($\mathbb{H} = \{\eta(\mathbf{u}) = \eta_1(u_1) + \eta_2(u_2) + \cdots + \eta_L(u_L), \mathbf{u} = (u_1, \ldots, u_L) \in \mathcal{U}\}$), partly linear models ($\mathbb{H} = \{\eta(\mathbf{u}) = \mathbf{u}_1^T\beta + \eta_2(\mathbf{u}_2), \mathbf{u} = (\mathbf{u}_1, \mathbf{u}_2) \in \mathcal{U}\}$), partly linear additive models ($\mathbb{H} = \{\eta(\mathbf{u}) = \mathbf{u}_1^T\beta + \eta_2(u_2) + \cdots + \eta_M(u_M), \mathbf{u} = (u_1, \ldots, u_M) \in \mathcal{U}\}$), varying coefficient models ($\mathbb{H} = \{\eta(\mathbf{u}) = \eta_1(u_L)u_1 + \eta_2(u_L)u_2 + \cdots + \eta_{L-1}(u_L)u_{L-1}, \mathbf{u} = (u_1, \ldots, u_L) \in \mathcal{U}\}$), functional ANOVA models ($\mathbb{H} = \{\eta(\mathbf{u}) = \eta_1(u_1) + \eta_2(u_2) + \cdots + \eta_L(u_L) + \text{selected interaction terms}, \mathbf{u} = (u_1, \ldots, u_L) \in \mathcal{U}\}$, here an "interaction term" is a function of two or more variables).

Let $p(\eta, \mathbf{w})$ denote the probability density of $\mathbf{W}$, which depends on the unknown function $\eta$. For a candidate function $h$ of $\eta$, the log-likelihood is given by $l(h, \mathbf{w}) = \log p(h, \mathbf{w})$. Define the expected log-likelihood by $\Lambda(h) = E[l(h, \mathbf{W})]$, where the expectation is taken with respect to the true function $\eta$. We say our model is a *concave extended linear model* if (i) $l(h, \mathbf{w})$ is concave in $h$ for each value of $\mathbf{w} \in \mathcal{W}$, that is, given any two functions $h_1, h_2 \in \mathbb{H}$ whose log-likelihood functions are well-defined, $l(\alpha h_1 + (1-\alpha)h_2, \mathbf{w}) \geq \alpha l(h_1, \mathbf{w}) + (1-\alpha)l(h_2, \mathbf{w})$ for $\alpha \in (0, 1)$ and $\mathbf{w} \in \mathcal{W}$; and (ii) $\Lambda(h)$ is strictly concave in $h$, that is, given any two functions $h_1, h_2 \in \mathbb{H}$ that are not essentially equal and whose expected log-likelihood functions are well-defined, $\Lambda(\alpha h_1 + (1-\alpha)h_2) > \alpha\Lambda(h_1) + (1-\alpha)\Lambda(h_2)$ for $\alpha \in (0, 1)$. (Here two functions on $\mathcal{U}$ are essentially equal if they differ only on a subset of $\mathcal{U}$ having Lebesgue measure zero.) In the above definition, we implicitly assume that the set of functions $h$ such that $l(h, \mathbf{w})$ and $\Lambda(h)$ are well-defined is a convex set.

In many applications, $\eta$ need not totally specify the probability distribution of $\mathbf{W}$. In such applications, we can take $l(h, \mathbf{w})$ to be the logarithm of a conditional likelihood, a pseudo-likelihood, or a partial likelihood, depending on the problem under consideration. From now on, we allow this broad view of $l(h, \mathbf{w})$ in extended linear modeling. For simplicity, we still call $l(h, \mathbf{w})$ 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 $l(h, \mathbf{w})$, the function $\eta$ is the essentially unique function in $\mathbb{H}$ that maximizes the expected log-likelihood. Consider, for example, the estimation of a regression function $\eta(\mathbf{x}) = E(Y|\mathbf{X} = \mathbf{x})$. One can take $l(h, \mathbf{W}) = -[Y - h(\mathbf{X})]^2$ with $\mathbf{W} = (\mathbf{X}, Y)$. If the conditional distribution of $Y$ given $\mathbf{X}$ is normal with constant variance, $l$ is (up to additive and multiplicative constants) the conditional log-likelihood. If this conditional distribution is not normal, we can think of $l$ as the logarithm of a pseudo-likelihood. In either case, the true regression function $\eta$ maximizes the expected log-likelihood.

The class of concave extended linear models is extremely rich, containing many estimation problems as special cases. Here are some examples.

**Hazard regression.** Consider a positive survival time $T$, a positive censoring time $C$, the observed time $T \wedge C = \min(T, C)$, and an $\mathcal{X}$-valued random vector $\mathbf{X}$ of covariates. Let $\delta = \text{ind}(T \leq C)$ be the indicator random variable that equals one or zero according as $T \leq C$ ($T$ is uncensored) or $T > C$ ($T$ is censored). Suppose $T$ and $C$ are conditionally independent given $\mathbf{X}$. Suppose also that $P(C \leq \tau) = 1$ for a known positive constant $\tau$. Let $\eta(\mathbf{x}, t) = \log\{f(t|\mathbf{x})/[1 - F(t|\mathbf{x})]\}$, $t > 0$, denote the logarithm of the conditional hazard function, where $f(t|\mathbf{x})$ and $F(t|\mathbf{x})$ are the conditional density and conditional distribution functions, respectively, of $T$ given that $\mathbf{X} = \mathbf{x}$. The log-likelihood for a candidate $h$ for $\eta$ is given by $l(h, \mathbf{W}) = \delta h(\mathbf{X}, T \wedge C) - \int_0^{T \wedge C} \exp\, h(\mathbf{X}, t)\, dt$. Here, $\mathbf{W} = (\mathbf{X}, T \wedge C, \delta)$ and $\mathcal{U} = \mathcal{X} \times [0, \tau]$.

**Conditional density estimation.** Consider a random pair $(\mathbf{X}, \mathbf{Y})$, where $\mathbf{X}$ is $\mathcal{X}$-valued, $\mathbf{Y}$ is $\mathcal{Y}$-valued, and the conditional distribution of $\mathbf{Y}$ given that $\mathbf{X} = \mathbf{x}$ has a positive density. Since the corresponding log-density $\phi$ satisfies the nonlinear constraint $\int_{\mathcal{Y}} \exp \phi(\mathbf{y}|\mathbf{x})\, d\mathbf{y} = 1$ for $\mathbf{x} \in \mathcal{X}$, it is not natural to model $\phi$ as a member of a linear space. To overcome this difficulty, we write $\phi(\mathbf{y}|\mathbf{x}) = \eta(\mathbf{y}|\mathbf{x}) - c(\mathbf{x}; \eta)$ and model $\eta$ as a member of some linear space; here $c(\mathbf{x}; \eta) = \log \int_{\mathcal{Y}} \exp \eta(\mathbf{y}|\mathbf{x})\, d\mathbf{y}$. By imposing a suitable linear constraint on $\eta$, we can make the map $\sigma : \eta \mapsto \phi$ one-to-one (see Section 4 for the details). Then the problem of estimating $\phi$ is reduced to that of estimating $\eta$ and can thereby be cast into the framework of extended linear modeling. The (conditional) log-likelihood is given by $l(h, \mathbf{X}, \mathbf{Y}) = h(\mathbf{Y}|\mathbf{X}) - c(\mathbf{X}; h)$. Here $\mathbf{W} = (\mathbf{X}, \mathbf{Y})$ and $\mathcal{U} = \mathcal{X} \times \mathcal{Y}$.

Generalized regression, density estimation (Stone et al. (1997)), spectral density estimation (Kooperberg, Stone and Truong (1995b)), polychotomous regression (Kooperberg, Bose and Stone (1997)), event history analysis (Huang and Stone (1998)), proportional hazards regression (Huang, Kooperberg, Stone and Truong (1999)), and extension of hazard regression to counting process regression (Section 3) can also be treated in the framework of concave extended linear models.

Let $\mathbb{G} \subset \mathbb{H}$ be a finite-dimensional space of bounded functions, whose dimension may depend on the sample size. We estimate $\eta$ by using maximum likelihood over $\mathbb{G}$, that is, we take $\hat{\eta} = \text{argmax}_{g \in \mathbb{G}}\, \ell(g)$, where $\ell(g) = (1/n) \sum_{i=1}^n l(g, \mathbf{W}_i)$ is the scaled log-likelihood. Typical choices of $\mathbb{G}$ include spaces of polynomials, trigonometric polynomials, or polynomial splines. When $\mathbb{H}$ has a specific structure, $\mathbb{G}$ is chosen to have the same structure. For example, if $\mathbb{H}$ is a space of additive functions, we can choose $\mathbb{G}$ to be a space of additive splines. See Section 2.3 for more examples of choosing $\mathbb{G}$ for structural models. We refer to $\mathbb{G}$ as the estimation space.

Given the estimate $\hat{\eta}$, the following questions about its asymptotic behavior arise naturally. Is $\hat{\eta}$ consistent? If so, what is the rate of convergence? How does $\hat{\eta}$ behave when the model is misspecified, that is, when $\eta \notin \mathbb{H}$? In the structural models we consider, the target function usually has a decomposition as a sum of certain component functions. For example, in additive models, the target function is a sum of functions of a single variable. Thus any reasonable estimate should have a similar decomposition. Then, under what conditions and at what rate will the components of the estimate converge to the corresponding components of the target function?

In this paper we provide general answers to these questions in the context of concave extended linear models. The results are applicable to a broad range of estimation problems including those mentioned above. Previous work on asymptotics for extended linear models includes Stone (1985, 1986, 1990, 1991, 1994), Hansen (1994), Kooperberg, Stone and Truong (1995a, b), Huang (1998a, b), Huang and Stone (1998), and Huang et al. (1999). While these works focused on either a specific context (e.g., regression), or a specific model space (e.g., an additive model), or a specific type of estimation space (e.g., an additive spline space), the current paper gives a unified treatment of various contexts and various types of model and estimation spaces. By considering possible combinations of the likelihood (estimation context), model space (structural assumption) and type of estimation space, we can obtain a rich body of results.

The theoretical synthesis in this paper provides insightful understanding of the structure of extended linear modeling. By singling out the "concavity" property of the log-likelihood, we identify the common features of various estimation problems that can be treated effectively within the framework of extended linear modeling. Moreover, by permitting some or all of the components in the ANOVA decomposition of the function of interest to be parametric, the current theory broadens the scope of extended linear modeling as originally considered in Hansen (1994) and Stone et al. (1997). In particular, we obtain a unified treatment of rates of convergence for partly linear additive models and functional ANOVA models in various statistical contexts (Section 2.3). The treatment of model misspecification for partly linear models is also new.

Section 2 presents the general result on rates of convergence (Theorem 2.1), from which consistency is a simple consequence. Both correctly specified and misspecified models are treated in a unified manner. Applications to saturated models, partly linear models, and functional ANOVA models are discussed. In Sections 3 and 4 , two specific contexts are studied in detail to illustrate the power of our general theory. Section 3 deals with counting process regression with possibly internal time-dependent covariates and thus extends previous work of Kooperberg, Stone and Truong (1995a) and Huang and Stone (1998). Section

4 studies the conditional density estimation problem, which provides a novel application of the general theory when the function of interest is subject to nonlinear constraints. All proofs are given in the Appendices.

**Notation.** Let $\#(B)$ denote the cardinality (number of members) of a set $B$. For a function $h$ on $\mathcal{U}$, set $\|h\|_\infty = \sup_{\mathbf{u}\in\mathcal{U}} |h(\mathbf{u})|$. Given positive numbers $a_n$ and $b_n$ for $n \geq 1$, let $a_n \lesssim b_n$ mean that $a_n/b_n$ is bounded and let $a_n \asymp b_n$ mean that $a_n \lesssim b_n$ and $b_n \lesssim a_n$. Given random variables $V_n$ for $n \geq 1$, let $V_n = O_P(b_n)$ mean that $\lim_{c\to\infty} \limsup_n P(|V_n| \geq cb_n) = 0$ and let $V_n = o_P(b_n)$ mean that $\limsup_n P(|V_n| \geq cb_n) = 0$ for all $c > 0$. For a random variable $V$, let $E_n$ denote expectation relative to its empirical distribution; that is, $E_n(V) = n^{-1}\sum_i V_i$, where $V_i$, $1 \leq i \leq n$, is a random sample from the distribution of $V$. We use $M, M_1, M_2, \ldots$ to denote generic constants, which may vary from one context to another.

## 2. Description of Main Results

In this section we present our main results and illustrate their application. The discussion of necessary regularity conditions and technical proofs is postponed to Appendix A.

### 2.1. Rates of convergence

We assume that the log-likelihood $l(h, \mathbf{w})$ 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, $\ell(h, \mathbf{w})$ and $\Lambda(h)$ are well-defined on $\mathbb{G}$.

In typical applications of extended linear models, the model space $\mathbb{H}$ corresponding to a set of structural assumptions is at best an approximation to reality. To define an appropriate target for the estimate when the the model is misspecified (that is, $\eta \notin \mathbb{H}$), let $\eta^* = \arg\max_{h\in\mathbb{H}} \Lambda(h)$ denote the best approximation to $\eta$ in $\mathbb{H}$, where "best" means maximizing the expected log-likelihood. Typically, the best approximation $\eta^*$ exists and is essentially unique when the model is concave. It follows from the information inequality that if $\eta \in \mathbb{H}$, then $\eta^* = \eta$ almost everywhere. Thus we regard $\eta^*$ as our target, which allows us to give a unified treatment of correctly specified and misspecified models.

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

Since $\hat{\eta}$ maximizes the scaled log-likelihood $\ell(g)$, which should be close to the expected log-likelihood $\Lambda(g)$ for $g \in \mathbb{G}$ when sample size is large, it is natural to

think that $\hat{\eta}$ is directly estimating the best approximation $\bar{\eta} = \arg\, max_{g\in\mathbb{G}}\Lambda(g)$ to $\eta$ in $\mathbb{G}$. If $\mathbb{G}$ is chosen such that $\bar{\eta}$ is close to $\eta^*$, then $\hat{\eta}$ should provide a reasonable estimate of $\eta^*$. This motivates the decomposition

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

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

Let $\|\cdot\|$ be a norm on $\mathbb{H}$ such that $\|h\| < \infty$ and $\|h\| \le C_0\|h\|_\infty$ for $h \in \mathbb{H}$ and a positive constant $C_0$. The norm $\|\cdot\|$ is used to measure the distance between two functions in $\mathbb{H}$. Typically, it is chosen to be an $L_2$-norm on $\mathcal{U}$ relative to an appropriate measure that depends on the estimation problem. In the regression context, for example, a natural choice is given by $\|h\|^2 = E[h^2(\mathbf{X})]$. In the following, we assume without loss of generality that $C_0 = 1$ since, otherwise, we can apply the same arguments to the norm $\|\cdot\|/C_0$.

To get mathematically rigorous results, we need some regularity conditions. A set of such conditions (i.e., Conditions A.1, A.2 and A.4) is assumed to hold throughout this section and will be stated explicitly in Appendix A. Roughly speaking, we require that the target function $\eta^*$ exist and that the log-likelihood and the expected log-likelihood be concave. In addition, the norm $\|\cdot\|$ should be connected to the log-likelihood and expected log-likelihood in a suitable manner. These conditions are satisfied in the various estimation contexts discussed in the previous section and will be verified explicitly in Sections 3 and 4 for counting process regression and conditional density estimation.

Set $N_n = \dim(\mathbb{G})$, $A_n = \sup_{g\in\mathbb{G},\|g\|\neq 0}\{\|g\|_\infty/\|g\|\} \ge 1$, and $\rho_n = \inf_{g\in\mathbb{G}}\|g - \eta^*\|_\infty$. The following is the main result of this section. The proof is given in Appendix A.

**Theorem 2.1.** *Suppose* $\lim_n A_n\rho_n = 0$ *and* $\lim_n A_n^2 N_n/n = 0$. *Then* $\bar{\eta}$ *exists uniquely for* $n$ *sufficiently large and* $\|\bar{\eta} - \eta^*\|^2 = O(\rho_n^2)$. *Moreover,* $\hat{\eta}$ *exists uniquely except on an event whose probability tends to zero as* $n \to \infty$, *and* $\|\hat{\eta} - \bar{\eta}\|^2 = O_P(N_n/n)$. *Consequently,* $\|\hat{\eta} - \eta^*\|^2 = O_P(N_n/n + \rho_n^2)$.

Note that consistency of the estimate is an immediate consequence of this theorem. The bounds for the magnitudes of the estimation and approximation errors can be interpreted intuitively as follows: $N_n/n$ is just the inverse of the number of observations per parameter, and $\rho_n$ is the best possible approximation rate to the target function in the estimation space. This result provides considerable insight: while the error bound of the stochastic part can be explained by a heuristic variance calculation, that of the systematic part is reduced to a problem of approximation theory.

The constants $A_n$ and $\rho_n$ above were introduced in Huang (1998a) in developing a general theory on rates of convergence of least squares estimates in the regression context. The magnitude of $A_n$ can be determined by employing results in the approximation theory literature for various commonly used approximating spaces, including polynomials, trigonometric polynomials, splines, wavelets, and finite elements. The constant $\rho_n$ is the minimum $L_\infty$ norm of the error when $\eta^*$ is approximated by a function in $\mathbb{G}$. By using results from approximation theory, the magnitude of $\rho_n$ can be determined for commonly used approximating spaces if a smoothness condition is imposed on $\eta^*$. See the paper just referred to for further discussion of these constants.

In the following subsections, we illustrate how to apply Theorem 2.1. In particular we give the magnitudes of $A_n$ and $\rho_n$ when $\mathbb{G}$ is an appropriately chosen approximating linear space and $\eta$ satisfies some smoothness condition.

### 2.2. Saturated models

Let $\mathcal{U}$ be the Cartesian product of compact intervals $\mathcal{U}_1, \ldots, \mathcal{U}_L$. Consider the saturated model, that is, let $\mathbb{H}$ be the space of all square-integrable functions on $\mathcal{U}$. In this case, there is no structural assumption imposed on $\eta$, so $\eta^* = \eta$.

A commonly used smoothness condition is as follows. Let $0 < \beta \leq 1$. A function $h$ on $\mathcal{U}$ is said to satisfy a Hölder condition with exponent $\beta$ if there is a positive number $\gamma$ such that $|h(\mathbf{u}) - h(\mathbf{u}_0)| \leq \gamma |\mathbf{u} - \mathbf{u}_0|^\beta$ for $\mathbf{u}_0, \mathbf{u} \in \mathcal{U}$; here $|\mathbf{u}| = (\sum_{l=1}^L u_l^2)^{1/2}$ is the Euclidean norm of $\mathbf{u} = (u_1, \ldots, u_L) \in \mathcal{U}$. Given an $L$-tuple $i = (i_1, \ldots, i_L)$ of nonnegative integers, set $[i] = i_1 + \cdots + i_L$ and let $D^i$ denote the differential operator defined by $D^i = \partial^{[i]}/(\partial u_1^{i_1} \cdots \partial u_L^{i_L})$. Let $k$ be a nonnegative integer and set $p = k + \beta$. A function on $\mathcal{U}$ is said to be *p-smooth* if it is $k$ times continuously differentiable on $\mathcal{U}$ and $D^i$ satisfies a Hölder condition with exponent $\beta$ for all $i$ with $[i] = k$.

Let $m$ and $J$ be fixed nonnegative integers. Given an increasing sequence of real numbers $a = t_0 < t_1 < \cdots < t_J < t_{J+1} = b$, a function on $[a, b]$ is a polynomial spline with degree $m$ and $J$ interior knots $\{t_j, 1 \leq j \leq J\}$ if the following holds: (i) it is a polynomial of degree $m$ in the intervals $[t_j, t_{j+1})$, $0 \leq j \leq J - 1$, and $[t_{J-1}, t_J]$; if $m > 0$, then (ii) it has $m - 1$ continuous derivatives on $[a, b]$. The collection of such spline functions constitutes a linear space of dimension $J + m + 1$ and is denoted $\mathrm{Spl}([a, b], m, J)$. In statistical applications, the number of interior knots $J$ may depend on the sample size $n$. In this case, we require that the knots have bounded mesh ratio, that is, that the ratios of the distances between consecutive knots are bounded above by a universal constant.

Suppose $\eta$ is $p$-smooth. For $m \geq p - 1$, set $\mathbb{G}_l = \mathrm{Spl}(\mathcal{U}_l, m, J_n)$ for $1 \leq l \leq L$ and let $\mathbb{G}$ be the tensor product of $\mathbb{G}_1, \ldots, \mathbb{G}_L$, that is, the linear space

spanned by functions $g_1(u_1) \cdot \ldots \cdot g_L(u_L)$ with $g_l \in \mathbb{G}_l$ for $1 \leq l \leq L$. Suppose $p > L/2$, $\lim_n J_n = \infty$, and $\lim_n J_n^{2L}/n = 0$. Then $N_n \asymp J_n^L$, $A_n \asymp J_n^{L/2}$ (see Corollary 2 of Huang (1998a)) and $\rho_n \asymp J_n^{-p}$ (see (13.69) and Theorem 12.8 of Schumaker (1981)). Hence, $\lim_n A_n \rho_n = 0$ and $\lim_n A_n^2 N_n/n = 0$. By Theorem 2.1. $\|\hat{\eta} - \eta\|^2 = O_P(J_n^L/n + J_n^{-2p})$. In particular, for $J_n \asymp n^{1/(2p+L)}$, we have $\|\hat{\eta} - \eta\|^2 = O_P(n^{-2p/(2p+L)})$.

The rate of convergence $n^{-2p/(2p+L)}$ is actually optimal; there is no estimate that can have a faster rate of convergence uniformly over the class of $p$-smooth functions (see Stone (1982)). This illustrates the curse of dimensionality: given the same smoothness condition, the larger the dimension $L$, the slower the rate.

### 2.3. Structural models

Many structural models can be specified by constructing an appropriate model space using linear function spaces and their tensor products. In such a case, it is natural to require that the estimation spaces have the same structure as the model space. We show in this section how Theorem 2.1 can be applied in this situation to obtain the rate of convergence of our estimate. We also show under what conditions and at what rate the components of the estimate converge to those of the target function. Suppose $\mathcal{U}$ is the Cartesian product of $\mathcal{U}_1, \ldots, \mathcal{U}_L$, where each $\mathcal{U}_l$ is a compact subset of some Euclidean space.

**Model space.** Let $\mathbb{H}_\emptyset$ denote the space of constant functions on $\mathcal{U}$. Given $1 \leq l \leq L$, let $\mathbb{H}_l \supset \mathbb{H}_\emptyset$ denote a closed subspace of the space of all square-integrable functions on $\mathcal{U}_l$, which can be finite- or infinite-dimensional. Given a nonempty subset $s = \{s_1, \ldots, s_k\}$ of $\{1, \ldots, L\}$, let $\mathbb{H}_s$ denote the tensor product of $\mathbb{H}_{s_1}, \ldots, \mathbb{H}_{s_k}$, which is the closure of the space of functions on $\mathcal{U}$ spanned by the functions $f$ of the form $f(\mathbf{u}) = \prod_{i=1}^k f_{s_i}(u_{s_i})$, where $f_{s_i} \in \mathbb{H}_{s_i}$ for $1 \leq i \leq k$. Let $\mathcal{S}$ denote a hierarchical collection of subsets of $\{1, \ldots, L\}$, where hierarchical means that if $s$ is a member of $\mathcal{S}$ and $r$ is a subset of $s$, then $r$ is a member of $\mathcal{S}$. The corresponding model space $\mathbb{H}$ is defined by $\mathbb{H} = \{\sum_{s \in \mathcal{S}} h_s : h_s \in \mathbb{H}_s \text{ for } s \in \mathcal{S}\}$. Note that the various structural models discussed in Section 1 can all be put into this framework by making suitable choices of $\mathbb{H}_l$ and $\mathcal{S}$. For example, taking $\mathbb{H}_l = L^2(\mathcal{U}_l)$ for $1 \leq l \leq L$ and $\mathcal{S} = \{\emptyset, \{1\}, \ldots, \{L\}\}$, we obtain an additive model.

**Estimation space.** Let $\mathbb{G}_\emptyset$ denote the space of constant functions on $\mathcal{U}$, dimension $N_\emptyset = 1$. Given $1 \leq l \leq L$, let $\mathbb{G}_l$ ($\mathbb{G}_\emptyset \subset \mathbb{G}_l \subset \mathbb{H}_l$) denote a linear space of bounded, real-valued functions on $\mathcal{U}_l$, which may vary with the sample size $n$, with finite, positive dimension $N_l = N_{ln}$. Given a nonempty subset $s = \{l_1, \ldots, l_k\}$ of $\{1, \ldots, L\}$, let $\mathbb{G}_s$ be the tensor product of $\mathbb{G}_{l_1}, \ldots, \mathbb{G}_{l_k}$. The estimation space $\mathbb{G}$ is defined by $\mathbb{G} = \{\sum_{s \in \mathcal{S}} g_s : g_s \in \mathbb{G}_s \text{ for } s \in \mathcal{S}\}$. Note that

the dimension of $\mathbb{G}_s$ is $N_s = \prod_{i=1}^k N_{l_i}$ and that the dimension $N_n$ of $\mathbb{G}$ satisfies $N_n \asymp \sum_{s\in\mathcal{S}} N_s$. If $\mathbb{H}_l$ is finite-dimensional, we can choose $\mathbb{G}_l = \mathbb{H}_l$.

The functions in $\mathbb{H}$ and $\mathbb{G}$ can have a number of representations as a sum of components. To obtain a unique such representation, we require that each non-constant component be orthogonal to all possible values of the proper lower-order components relative to an appropriate inner product, which leads to the notion of a functional ANOVA decomposition. Usually, one uses a theoretical inner product on the model space and an empirical inner product on the estimation space. 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 needs to be totally determined by the data since it will be used to decompose the estimate. A systematic account of functional ANOVA decompositions is provided in Huang (2000). Some of the results of that paper are summarized in Appendix D.

Theorem 2.1 can be used to study the convergence properties of our estimates when the model and estimation spaces are constructed as above. Set $A_s = A_{sn}(\mathbb{G}_s) = \sup_{g\in\mathbb{G}_s, \|g\|\neq 0}(\|g\|_\infty/\|g\|)$ and $\rho_s = \rho_{sn}(\eta_s^*, \mathbb{G}_s) = \inf_{g\in\mathbb{G}_s} \|g - \eta_s^*\|_\infty$ for $s \in \mathcal{S}$. The constants $A_s$ and $\rho_s$, which are analogs of the constants $A_n$ and $\rho_n$, are defined on the tensor product spaces that constitute the estimation space $\mathbb{G}$ and thus are more straightforward to determine. Suppose Condition D.1 in Appendix D holds. The following result is a consequence of Theorem 2.1 and Lemma D.1.

**Corollary 2.1.** *Suppose* $\lim_n A_s\rho_{s'} = 0$ *and* $\lim_n A_s^2 N_{s'}/n = 0$ *for each pair* $s, s' \in \mathcal{S}$. *Then* $\|\hat\eta - \eta^*\|^2 = O_P(\sum_{s\in\mathcal{S}}(N_s/n + \rho_s^2))$.

Suppose now that $\eta^*$ and $\hat\eta$, as members of $\mathbb{H}$ and $\mathbb{G}$ respectively, have the ANOVA decompositions $\eta^* = \sum_{s\in\mathcal{S}} \eta_s^*$ and $\hat\eta = \sum_{s\in\mathcal{S}} \hat\eta_s$ (See Appendix D for formal definitions.). Suppose Conditions D.1–D.3 in Appendix D hold. Corollary 2.1 and Theorem D.1 together yield the following result.

**Corollary 2.2.** *Suppose* $\lim_n A_s\rho_{s'} = 0$ *and* $\lim_n A_s^2 N_{s'}/n = 0$ *for each pair* $s, s' \in \mathcal{S}$. *Then* $\|\hat\eta_s - \eta_s^*\|^2 = O_P(\sum_{s\in\mathcal{S}}(N_s/n + \rho_s^2))$ *for* $s \in \mathcal{S}$.

We now give some examples of structural models. For simplicity, suppose that each $\mathcal{U}_l$, $1 \le l \le L$, is a compact interval. Set $\mathbb{H}_l^0 = \{\alpha + \beta u_l, u_l \in \mathcal{U}_l : \alpha, \beta \in \mathbb{R}\}$ and $\mathbb{H}_l^2 = L^2(\mathcal{U}_l)$ for $l = 1, \ldots, L$. In the following examples, when $\mathbb{H}_l^0$ is used as a building block for the model space, we choose $\mathbb{G}_l = \mathbb{H}_l^0$ as the corresponding building block for the estimation space; when $\mathbb{H}_l^2$ is used as a building block for the model space, we choose $\mathbb{G}_l = \mathrm{Spl}(\mathcal{U}_l, m, J_n)$ as the corresponding building block for the estimation space.

**Linear models.** $\mathbb{G} = \mathbb{H} = \mathbb{H}_1^0 + \cdots + \mathbb{H}_L^0$. In this case, $A_n$ is independent of $n$ and $\rho_n = 0$. It follows from Theorem 2.1 that $\|\hat\eta - \eta^*\|^2 = O_P(1/n)$. Write

$\hat{\eta} = \hat{\alpha}+\hat{\beta}_1 u_1+\cdots \hat{\beta}_L u_L$ and $\eta^* = \alpha^*+\beta_1^* u_1+\cdots+\beta_L^* u_L$. Then $|\hat{\alpha}-\alpha^*|^2 = O_P(1/n)$ and $|\hat{\beta}_l - \beta_l^*|^2 = O_P(1/n)$ for $1 \le l \le L$.

**Partly linear additive models.** $\mathbb{H} = \mathbb{H}_1^0 + \cdots + \mathbb{H}_K^0 + \mathbb{H}_{K+1}^2 + \cdots + \mathbb{H}_L^2$, where $1 \le K < L$. Set $\mathbb{G} = \mathbb{G}_1 + \cdots + \mathbb{G}_L$, where $\mathbb{G}_l = \mathbb{H}_l^0$ for $l = 1, \ldots, K$ and $\mathbb{G}_l = \mathrm{Spl}(\mathcal{U}_l, m, J_n)$ for $l = K+1, \ldots, L$. Then $A_{\{l\}}$ is bounded and independent of $n$ and $N_{\{l\}} = 2$ for $1 \le l \le K$; $A_{\{l\}} \asymp J_n^{1/2}$ and $N_{\{l\}} \asymp J_n$ for $K < l \le L$. Let $\eta^* = \eta_\emptyset + \eta_1^*(u_1) + \cdots + \eta_L^*(u_L)$ be the ANOVA decomposition of $\eta^*$. Note that $\eta_l^* = \beta_l^*(u_l - \langle u_l, 1\rangle)$ for $1 \le l \le K$. Suppose the functions $\eta_l^*$, $K < l \le L$, are $p$-smooth. Let $m \ge p-1$. Suppose further that $p > 1/2$, $J_n \to \infty$ as $n \to \infty$, and $J_n^2 = o(n)$. Then $\rho_{\{l\}} = 0$ for $1 \le l \le K$ and $\rho_{\{l\}} \asymp J_n^{-p}$ for $K < l \le L$. Hence, $\lim_n A_{\{l\}}^2 N_{\{l'\}}/n = 0$ and $\lim_n A_{\{l\}}\rho_{\{l'\}} = 0$ for $1 \le l, l' \le L$. It follows from Corollary 2.1 that $\|\hat{\eta} - \eta^*\|^2 = O_P(J_n/n + J_n^{-2p})$. Furthermore, let $\hat{\eta} = \hat{\eta}_\emptyset + \hat{\eta}_1(u_1) + \cdots + \hat{\eta}_L(u_L)$ be the ANOVA decomposition of $\hat{\eta}$. Then, by Corollary 2.2, $\|\hat{\eta}_l - \eta_l^*\|^2 = O_P(J_n/n + J_n^{-2p})$ for $1 \le l \le L$. Taking $J_n \asymp n^{1/(2p+1)}$, we find $\|\hat{\eta} - \eta^*\|^2 = O_P(n^{-2p/(2p+1)})$ and $\|\hat{\eta}_l - \eta_l^*\|^2 = O_P(n^{-2p/(2p+1)})$ for $1 \le l \le L$. In particular, for $1 \le l \le K$, write $\hat{\eta}_l(u_l) = \hat{\beta}_l(u_l - \langle u_l, 1\rangle_n)$. Then $|\hat{\beta}_l - \beta_l^*| = O_P(n^{-p/(2p+1)})$.

The rates of convergence obtained here are in parallel to Theorems 1 and 2 in Chen (1995). While Chen's results were obtained in the context of generalized regression, our results apply to a broad range of estimation problems. Moreover, our results apply to misspecified models, which were not considered in his paper.

**Functional ANOVA models.** Given a hierarchical collection $\mathcal{S}$, a functional ANOVA model can be specified by defining the model space $\mathbb{H}$ as at the beginning of this subsection with $\mathbb{H}_l = \mathbb{H}_l^2$ for $l = 1, \ldots, L$. For instance, if $\mathcal{S} = \{\emptyset, \{1\}, \{2\}, \{3\}, \{1,2\}\}$, one has a functional ANOVA model with one two-factor interaction component, where the function $\eta$ of interest is modeled to have the form

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

To get an estimate of $\eta$, we construct the estimation space $\mathbb{G}$ corresponding to $\mathcal{S}$ with $\mathbb{G}_l = \mathrm{Spl}(\mathcal{U}_l, m, J_n)$ for $l = 1, \ldots, L$. Suppose the functions $\eta_s^*$, $s \in \mathcal{S}$, are $p$-smooth. Let $m \ge p-1$. Set $d = \max_{s\in\mathcal{S}} \#(s)$ and suppose that $p > d/2$, $J_n \to \infty$ as $n \to \infty$, and $J_n^{2d} = o(n)$. Observe that $A_s \asymp J_n^{\#(s)/2}$, $N_s \asymp J_n^{\#(s)}$, and $\rho_s \asymp J_n^{-p}$ for $s \in \mathcal{S}$. (See the example following Theorem 2.1.) Hence, $\lim_n A_s^2 N_{s'}/n = 0$ and $\lim_n A_s \rho_{s'} = 0$ for $s, s' \in \mathcal{S}$. It follows from Corollary 2.1 that $\|\hat{\eta} - \eta^*\|^2 = O_P(J_n^d/n + J_n^{-2p})$. Furthermore, let $\hat{\eta} = \sum_{s\in\mathcal{S}} \hat{\eta}_s$ and $\eta^* = \sum_{s\in\mathcal{S}} \eta_s^*$ be the ANOVA decompositions of $\hat{\eta}$ and $\eta^*$. Then, by Corollary 2.2, $\|\hat{\eta}_s - \eta_s^*\|^2 = O_P(J_n^d/n + J_n^{-2p})$ for $s \in \mathcal{S}$. Taking $J_n \asymp n^{1/(2p+d)}$, we get that $\|\hat{\eta} - \eta^*\|^2 = O_P(n^{-2p/(2p+d)})$ and $\|\hat{\eta}_s - \eta_s^*\|^2 = O_P(n^{-2p/(2p+d)})$ for $s \in \mathcal{S}$.

The rate of convergence $n^{-2p/(2p+d)}$ here should be compared with the rate $n^{-2p/(2p+L)}$ for the saturated model. Note that $d$ is the maximum order of interaction between the components of the argument variable $\mathbf{u}$. Thus, by using models with only main effects and low-order interactions ($d < L$), we can obtain faster rates of convergence than by using the saturated model and thereby ameliorate the curse of dimensionality. In particular, the rate $n^{-2p/(2p+1)}$ for an additive model is the same as that for estimating a one-dimensional function. Similarly, the rate $n^{-2p/(2p+2)}$ for a model involving two-factor interactions is the same as that for estimating a two-dimensional function.

**Remark 2.1.** In the examples above, we use univariate splines to build the estimation spaces. Theorem 2.1 and Corollaries 2.1 and 2.2 also apply when the estimation spaces are built from polynomials, trigonometric polynomials, or bivariate or multivariate splines. One can proceed as in Huang (1998) to determine the constants $A_n$, $\rho_n$, $A_s$, and $\rho_s$.

**Remark 2.2.** In some situations, it is not natural to model the function of interest as a member of a linear space due to inherent restrictions on this function. Our result can still be applied after a slight modification. Let $\phi$ denote a function on $\mathcal{U}$ of interest that is subject to some nonlinear constraints. (Linear constraints can be easily incorporated into the model space $\mathbb{H}$.) Suppose we can relate $\phi$ to a function $\eta$ through a one-to-one map $\sigma : \phi = \sigma(\eta)$, where $\eta$ is naturally modeled as a member of the model space $\mathbb{H}$. Then $\hat{\phi} = \sigma(\hat{\eta})$, the maximum likelihood estimate in $\sigma(\mathbb{G})$, can be used to estimate $\phi = \sigma(\eta)$. When the model is misspecified, we think of $\hat{\phi}$ as estimating $\phi^* = \sigma(\eta^*)$, the best approximation to $\phi$ in $\sigma(\mathbb{H})$, where "best" means maximizing the expected log-likelihood. Usually, $\sigma(\cdot)$ satisfies a Lipschitz condition in some neighborhood of zero: for every positive constant $M_1$, there is a positive constant $M_2$ such that

$$\|\sigma(h_1) - \sigma(h_2)\| \le M_2\|h_1 - h_2\| \qquad \text{for } h_1, h_2 \in \mathbb{H} \text{ with } \|h_1\|_\infty, \|h_2\|_\infty \le M_1. \tag{2.1}$$

In this case, the results in Theorem 2.1 can be easily translated to results about $\hat{\phi}$. See Section 4 for an illustration in the context of conditional density estimation.

## 3. Counting Process Regression

In this section, we apply the general theory for extended linear modeling to the context of counting process regression, which includes hazard regression for right-censored survival data as an important special case. In this context, the intensity of a counting process is related to a vector of time-dependent covariates. We will show that the conditions on the log-likelihood in the general results in Section 2 are implied by more primitive and statistically more natural conditions.

It is well known that the counting process formulation provides a general framework for survival analysis, and more generally for event history analysis; see Andersen, Borgan, Gill and Keiding (1993). Under this formulation, the nonparametric kernel method was studied by McKeague and Utikal (1990), Nielsen and Linton (1995), and Dabrowska (1997). However, the ease of incorporating various structural assumptions to achieve dimensionality reduction makes the extended linear modeling approach attractive in this context. Kooperberg, Stone and Truong (1995a) and Huang and Stone (1998) have applied extended linear modeling to hazard regression for right-censored survival data, a special case of counting process regression. Their treatment requires either that the covariates be time-independent or that the time-dependent covariates be external. The counting process formulation employed in this paper allows us to treat time-dependent covariates that need not be external (commonly referred to as internal).

Let $\mathcal{T} = [0, \tau]$ for some positive constant $\tau$. Suppose $(\Omega, \mathcal{F}, P)$ is a complete probability space and that $\{\mathcal{F}_t : t \in \mathcal{T}\}$, $\mathcal{F}_t \subset \mathcal{F}$, is a filtration satisfying the "usual conditions", that is, $\mathcal{F}_t$ is a family of right-continuous, increasing $\sigma$-algebras and $\mathcal{F}_0$ contains the $P$-null sets of $\mathcal{F}$. We assume $\{N(t) : t \in \mathcal{T}\}$ is an adapted (see Andersen et al. (1993)) counting process with intensity

$$E[N(dt)|\mathcal{F}_{t-}] = Y(t) \exp \eta(t, \mathbf{X}(t))\, dt, \tag{3.1}$$

where $Y(t)$ is a $\{0,1\}$-valued, predictable process, indicating the times at which the process $N$ is under observation, and $\mathbf{X}(t)$ is an $\mathcal{X}$-valued, predictable covariate process. The interest lies in estimating the log-hazard function $\eta$ based on a random sample.

For the special case of hazard regression with right-censored survival data, one observes $(T \wedge C, \mathrm{ind}(T \le C))$, where $T$ is the survival time of an individual and $C$ is the censoring time. Suppose $T$ and $C$ are conditionally independent given the process $\mathbf{X} = (\mathbf{X}(t))$, and that the conditional hazard of $T$ given $(\mathbf{X}(s) : s \le t)$ is $\exp \eta(t, \mathbf{X}(t))$. Let $N(t) = \mathrm{ind}(T \le C \wedge t)$ be the counting process with a single jump at an uncensored survival time. Then $N(\cdot)$ has intensity given by (3.1), where $Y(t) = \mathrm{ind}(T \wedge C \ge t)$ is the indicator that the individual is observed to be at risk at time $t$. (See Example 1 of McKeague and Utikal (1990).)

Returning to the counting process framework, we write the scaled log-likelihood for a candidate function $h$ for $\eta$ based on the random sample, $(N_i, Y_i, \mathbf{X}_i)$, $1 \le i \le n$, as

$$\ell(h) = \frac{1}{n} \sum_i \Big( \int_{\mathcal{T}} h(t, \mathbf{X}_i(t)) N_i(dt) - \int_{\mathcal{T}} Y_i(t) \exp h(t, \mathbf{X}_i(t))\, dt \Big)$$

(see Jacod (1975) and page 1512 of Dabrowska (1997)). Usually, the covariates $\mathbf{X}_i(t)$ are observed only for the times $t$ such that $Y_i(t) = 1$. The current formulation still works in this case, however, since the log-likelihood does not involve the unobserved values of $\mathbf{X}_i(t)$. The expected log-likelihood is given by

$$\Lambda(h) = E\Big( \int_{\mathcal{T}} h(t, \mathbf{X}(t)) N(dt) - \int_{\mathcal{T}} Y(t) \exp h(t, \mathbf{X}(t))\, dt \Big).$$

Note that $\Lambda(h)$ is well-defined for any integrable function $h$ when the log-hazard $\eta(t, \mathbf{x})$ is bounded. The present setup falls into the framework of extended linear modeling described in Section 1 with $\mathbf{W} = (N, Y, \mathbf{X})$ and $\mathcal{U} = \mathcal{T} \times \mathcal{X}$.

For square-integrable functions $h_1$ and $h_2$ on $\mathcal{T} \times \mathcal{X}$, define the empirical inner product and norm by $\langle h_1, h_2 \rangle_n = E_n \int_{\mathcal{T}} Y(t) h_1(t, \mathbf{X}(t)) h_2(t, \mathbf{X}(t))\, dt$ and $\|h_1\|_n^2 = \langle h_1, h_1 \rangle_n$. The theoretical versions of these quantities are given by $\langle h_1, h_2 \rangle = E \int_{\mathcal{T}} Y(t) h_1(t, \mathbf{X}(t)) h_2(t, \mathbf{X}(t))\, dt$ and $\|h_1\|^2 = \langle h_1, h_1 \rangle$.

**Condition 3.1.** There is a bounded function $\eta^* \in \mathbb{H}$ that maximizes $\Lambda(\cdot)$ over $\mathbb{H}$.

**Condition 3.2.** The function $\eta(t, \mathbf{x})$ is bounded on $\mathcal{T} \times \mathcal{X}$.

**Condition 3.3.** For each $t \in \mathcal{T}$, the measure $P(Y(t) = 1, \mathbf{X}(t) \in \cdot)$ has a density $f_{Y(t)=1, \mathbf{X}(t)}(t, \mathbf{x})$ with respect to the Lebesgue measure on $\mathcal{X}$. Moreover, $f_{Y(t)=1, \mathbf{X}(t)}(t, \mathbf{x})$ is bounded away from zero and infinity uniformly in $t \in \mathcal{T}$ and $\mathbf{x} \in \mathcal{X}$.

The above are mild regularity conditions. Condition 3.1 is necessary to define the target when the model is misspecified. The same argument as in Theorem 1 of Kooperberg, Stone and Truong (1995) can be used to verify the existence of $\eta^*$. Conditions similar to Condition 3.3 have been used in McKeague and Utikal (1990), Nielson and Linton (1995), and Dabrowska (1997).

Let the model and estimation spaces be defined as in Theorem 2.1. The theoretical norm defined in this section is used as the distance measure. The following is the main result of this section. The proof is given in Appendix B.

**Theorem 3.1.** *Suppose Conditions* 3.1-3.3 *hold and that* $\lim_n A_n N_n^2 / n = 0$ *and* $\lim_n A_n \rho_n = 0$. *Then the conclusions of Theorem* 2.1 *hold.*

**Remark 3.1.** In the marker dependent hazard model (Nielsen and Linton (1995)), the log-hazard $\eta(t, \mathbf{X}(t))$ depends only on the marker process $(\mathbf{X}(t))$, that is, $\eta(t, \mathbf{X}(t)) = \eta(\mathbf{X}(t))$. Theorem 3.1 still holds in such a situation with the spaces $\mathbb{H}$ and $\mathbb{G}$ being appropriate spaces of functions on $\mathcal{X}$. In addition, the result does not rely on the assumption that the log-hazard depends only on the marker. When such an assumption is invalid, $\hat{\eta}$ can be viewed as an estimate of the function $\eta^*$ depending only on $\mathbf{X}(t)$ that maximizes the expected log-likelihood among functions on $\mathcal{X}$.

## 4. Conditional Density Estimation

In this section, we apply the general theory for extended linear modeling to the context of conditional density estimation. See Hansen (1994) for a previous (more complicated) treatment of this problem, where the estimation spaces are required to be built from polynomial splines.

Let $\mathcal{X}$ and $\mathcal{Y}$ be compact sets in possibly different Euclidean spaces. Consider a random pair $(\mathbf{X}, \mathbf{Y})$, where $\mathbf{X}$ is $\mathcal{X}$-valued and $\mathbf{Y}$ is $\mathcal{Y}$-valued. Suppose $\mathbf{X}$ and $\mathbf{Y}$ have a positive joint density $f_{\mathbf{X},\mathbf{Y}}$. Let $f_{\mathbf{X}}$ denote the density of $\mathbf{X}$, and let $f_{\mathbf{Y}|\mathbf{X}}$ denote the conditional density of $\mathbf{Y}$ given $\mathbf{X}$. Then $f_{\mathbf{X},\mathbf{Y}}(\mathbf{x}, \mathbf{y}) = f_{\mathbf{X}}(\mathbf{x}) f_{\mathbf{Y}|\mathbf{X}}(\mathbf{y}|\mathbf{x})$ for $\mathbf{x} \in \mathcal{X}$ and $\mathbf{y} \in \mathcal{Y}$. Our interest lies in estimating $f_{\mathbf{Y}|\mathbf{X}}(\mathbf{y}|\mathbf{x})$ based on a random sample of size $n$ from the joint distribution of $(\mathbf{X}, \mathbf{Y})$.

Our approach is to model the logarithm $\phi = \log f_{\mathbf{Y}|\mathbf{X}}$ of the conditional density. This parameterization has the advantage that $f_{\mathbf{Y}|\mathbf{X}}$ is guaranteed to be positive while $\phi$ ranges freely over $\mathbb{R}$. Given a function $h$ on $\mathcal{X} \times \mathcal{Y}$ and given $\mathbf{x} \in \mathcal{X}$, set $c(\mathbf{x}; h) = \log \int_{\mathcal{Y}} \exp h(\mathbf{y}|\mathbf{x})\, d\mathbf{y}$; if $c(\mathbf{x}; h) < \infty$, then $\exp(h(\mathbf{y}|\mathbf{x}) - c(\mathbf{x}; h))$ is a density on $\mathcal{Y}$. For any function $h$ on $\mathcal{X} \times \mathcal{Y}$, the (conditional) log-likelihood is given by $l(h, \mathbf{x}, \mathbf{y}) = h(\mathbf{y}|\mathbf{x}) - c(\mathbf{x}; h)$; in particular, if $c(\mathbf{x}; h) = \infty$, then $l(h, \mathbf{x}, \mathbf{y}) = -\infty$. The expected log-likelihood is given by $\Lambda(h) = E[h(\mathbf{Y}|\mathbf{X}) - c(\mathbf{X}; h)]$ when the relevant expectation exists.

Note that, for any two functions $h$ and $h_0$ defined respectively on $\mathcal{X} \times \mathcal{Y}$ and $\mathcal{X}$, we have that $l(h + h_0) = l(h)$. Hence, any function of the form $\phi(\mathbf{y}|\mathbf{x}) + h_0(\mathbf{x})$, where $h_0$ is a function depending only on the variable $\mathbf{x}$ such that $c(\mathbf{x}; \phi + h_0) < \infty$, maximizes $\Lambda(\cdot)$. However, if we restrict our search to valid conditional densities, then $\phi(\mathbf{y}|\mathbf{x})$ is the essentially unique function maximizing $\Lambda(\cdot)$.

**Lemma 4.1.** *If both $\phi_1(\mathbf{y}|\mathbf{x})$ and $\phi_2(\mathbf{y}|\mathbf{x})$ maximize $\Lambda(\cdot)$ over some convex set of functions on $\mathcal{X} \times \mathcal{Y}$, then $\phi_2(\mathbf{y}|\mathbf{x}) - \phi_1(\mathbf{y}|\mathbf{x}) = \phi_0(\mathbf{x})$ almost everywhere for some function $\phi_0$ depending only on $\mathbf{x}$. If we further require that $\phi_1$ and $\phi_2$ be conditional densities, then it is necessary that $\phi_1 = \phi_2$ almost everywhere.*

**Proof.** Let $h_1$ and $h_2$ denote any two functions such that $c(\mathbf{x}; h_1) < \infty$ and $c(\mathbf{x}; h_2) < \infty$ for almost all $\mathbf{x} \in \mathcal{X}$. For $\alpha \in (0, 1)$, set $h_\alpha = h_1 + \alpha(h_2 - h_1)$. Then

$$\frac{d}{d\alpha} l(h_\alpha; \mathbf{x}, \mathbf{y}) = h_2(\mathbf{y}|\mathbf{x}) - h_1(\mathbf{y}|\mathbf{x}) - E[h_2(\mathbf{Y}_\alpha|\mathbf{X}) - h_1(\mathbf{Y}_\alpha|\mathbf{X})|\mathbf{X} = \mathbf{x}],$$

$$\frac{d^2}{d\alpha^2} l(h_\alpha; \mathbf{x}, \mathbf{y}) = -\mathrm{Var}\,[h_2(\mathbf{Y}_\alpha|\mathbf{X}) - h_1(\mathbf{Y}_\alpha|\mathbf{X})|\mathbf{X} = \mathbf{x}],$$

where, given $\mathbf{X} = \mathbf{x}$, $\mathbf{Y}_\alpha$ has the density $f_{\mathbf{Y}_\alpha|\mathbf{X}}(\mathbf{y}|\mathbf{x}) = \exp(h_\alpha(\mathbf{y}|\mathbf{x}) - c(\mathbf{x}; h_\alpha))$. (It follows by a standard argument in the context of one-parameter exponential

families that $c(\mathbf{x}; h_\alpha) < \infty$ and the conditional expectations and conditional variances appearing above are finite.) Moreover,

$$\frac{d^2}{d\alpha^2}\Lambda(h_1 + \alpha(h_2 - h_1)) = -E\{\mathrm{Var}\,[h_2(\mathbf{Y}_\alpha|\mathbf{X}) - h_1(\mathbf{Y}_\alpha|\mathbf{X})|\mathbf{X}]\}. \tag{4.1}$$

The lemma is a simple consequence of (4.1).

Observe that $c(\mathbf{x}; \phi) = 0$ for any conditional density $\phi$. If we model $\phi$ as lying in some linear space of functions, we need to consider the nonlinear constraint $c(\mathbf{x}; \phi) = 0$. Alternatively, write $\phi = \sigma(\eta) := \eta - c(\mathbf{x}; \eta)$ and model $\eta$ as a member of some linear space $\mathbb{H}$. Then the constraint $c(\mathbf{x}; \phi) = 0$ is automatically satisfied. The identifiability of $\eta$ is ensured under a suitable linear constraint, which can be incorporated into the structure of the space $\mathbb{H}$.

For ease in expressing the identifiability constraint on $\eta$, we employ orthogonality relative to an appropriate inner product. In addition, two ancillary function spaces are introduced to describe the structure of $\mathbb{H}$. We will see later on that such a formulation is very convenient in dealing with the ANOVA structure of a function.

Specifically, let $\psi$ denote the uniform distribution on $\mathcal{X} \times \mathcal{Y}$. Denote the corresponding inner product and norm by $\langle \cdot, \cdot \rangle_\psi$ and $\| \cdot \|_\psi$. Let $\perp_\psi$ denote orthogonality relative to $\langle \cdot, \cdot \rangle_\psi$. Let $\mathbb{H}_1$ be a Hilbert space of functions on $\mathcal{X} \times \mathcal{Y}$ equipped with the inner product $\langle \cdot, \cdot \rangle_\psi$, and let $\mathbb{H}_0$ be the space of functions in $\mathbb{H}_1$ that depend only on variable $\mathbf{x}$. We require that, given any $h \in \mathbb{H}_1$, $h(\mathbf{y}_0|\mathbf{x}) \in \mathbb{H}_0$ for any fixed $\mathbf{y}_0 \in \mathcal{Y}$. Set $\mathbb{H} = \{h \in \mathbb{H}_1 : h \perp_\psi \mathbb{H}_0\}$. Then $\eta$ is identifiable when modeled as a member of $\mathbb{H}$. Indeed, suppose that $\phi = \eta_1 - c(\mathbf{x}; \eta_1) = \eta_2 - c(\mathbf{x}; \eta_2)$ for $\eta_1, \eta_2 \in \mathbb{H}$. Then $\eta_1 - \eta_2 = c(\mathbf{x}; \eta_2) - c(\mathbf{x}; \eta_1)$ depends only on $\mathbf{x}$ and thus it belongs to $\mathbb{H}_0$. On the other hand, $\eta_1 - \eta_2 \perp_\psi \mathbb{H}_0$. Hence $\eta_1 = \eta_2$ almost everywhere.

**Condition 4.1.** There is a bounded conditional density function $\phi^* \in \mathbb{H}_1$ that maximizes $\Lambda(\cdot)$ over $\mathbb{H}_1$.

According to Lemma 4.1, the conditional density $\phi^*$ in this condition is uniquely defined. If the true conditional density $\phi$ is a member of $\mathbb{H}_1$, then $\phi^* = \phi$ and Condition 4.1 is just the requirement that $\phi$ be bounded. On the other hand, Condition 4.1 guarantees that the maximizer of $\Lambda(\cdot)$ over $\mathbb{H}$ exists. To verify this claim, for $h \in \mathbb{H}_1$, set

$$(P_\psi h)(\mathbf{y}|\mathbf{x}) = h(\mathbf{y}|\mathbf{x}) - \frac{1}{|\mathcal{Y}|}\int_{\mathcal{Y}} h(\mathbf{y}|\mathbf{x})\, d\mathbf{y}, \qquad \mathbf{x} \in \mathcal{X},\, \mathbf{y} \in \mathcal{Y}.$$

Then $P_\psi$ is the orthogonal projection onto $\mathbb{H}$ relative to $\langle \cdot, \cdot \rangle_\psi$. Set $\eta^* = P_\psi \phi^*$. We have the unique decomposition $\phi^* = \eta^* + (\phi^* - \eta^*)$, where $\eta^* \in \mathbb{H}$ and

$\phi^* - \eta^* \in \mathbb{H}_0$. Since $\phi^*$ is a conditional density, $c(\mathbf{x}; \phi^*) = 0$. It is easily seen that $\eta^* - c(\mathbf{x}; \eta^*) = \phi^*$ and $\eta^*$ is the unique function in $\mathbb{H}$ that maximizes the expected log-likelihood. We also have $\|\eta^*\|_\infty \le 2\|\phi^*\|_\infty < \infty$.

We construct the estimation spaces to have the same structure as the model space. Let $\mathbb{G}_1$ be a finite-dimensional subspace of $\mathbb{H}_1$ and let $\mathbb{G}_0$ be the space of functions in $\mathbb{G}_1$ that depend only on variable $\mathbf{x}$. We require that, for each $g \in \mathbb{G}_1$, $g(\mathbf{y}_0|\mathbf{x}) \in \mathbb{G}_0$ for any fixed $\mathbf{y}_0 \in \mathcal{Y}$. Set $\mathbb{G} = \{g \in \mathbb{G}_1 : g \perp_\psi \mathbb{G}_0\}$. It is easily shown that $g \in \mathbb{G}$ if and only if $g \in \mathbb{G}_1$ and $\int_{\mathcal{Y}} g(\mathbf{y}|\mathbf{x})\, d\mathbf{y} = 0$ for almost all $\mathbf{x} \in \mathcal{X}$. Hence, $\mathbb{G} \subset \mathbb{H}$.

For square-integrable functions $h_1$ and $h_2$ on $\mathcal{X} \times \mathcal{Y}$, define the empirical inner product and norm by $\langle h_1, h_2 \rangle_n = E_n[h_1(\mathbf{Y}|\mathbf{X}) h_2(\mathbf{Y}|\mathbf{X})]$ and $\|h_1\|_n^2 = \langle h_1, h_1 \rangle_n$. The theoretical versions of these quantities are given by $\langle h_1, h_2 \rangle = E[h_1(\mathbf{Y}|\mathbf{X}) h_2(\mathbf{Y}|\mathbf{X})]$ and $\|h_1\|^2 = \langle h_1, h_1 \rangle$.

**Condition 4.2.** The joint density $f_{\mathbf{X},\mathbf{Y}}$ is bounded away from zero and infinity on $\mathcal{X} \times \mathcal{Y}$.

Let $\hat{\eta} = \operatorname{argmax}_{g \in \mathbb{G}} \ell(g)$ denote the maximum likelihood estimate of $\eta^*$ in $\mathbb{G}$. Set $\hat{\phi} = \sigma(\hat{\eta})$. Then $\hat{\phi}$ is the maximum likelihood estimate of $\phi^* = \sigma(\eta^*)$. Note that $\sigma(\cdot)$ satisfies the Lipschitz condition (2.1). We apply Theorem 2.1 to obtain the rate of convergence of $\hat{\eta}$ to $\eta^*$ and subsequently the rate of convergence of $\hat{\phi}$ to $\phi^*$.

Set $A_n = \sup_{g \in \mathbb{G}} \{\|g\|_\infty / \|g\|\}$ and $N_n = \dim(\mathbb{G})$. Define $A_{0n}$ and $N_{0n}$ similarly by replacing $\mathbb{G}$ with $\mathbb{G}_0$ in the definition of $A_n$ and $N_n$. Set $\rho_{1n} = \inf_{g \in \mathbb{G}_1} \|g - \phi^*\|_\infty$. The following is the main result of this section. The proof is given in Appendix C.

**Theorem 4.1.** *Suppose Conditions* 4.1 *and* 4.2 *hold. In addition, suppose that* $\lim_n A_n^2 N_n / n = 0$, $\lim_n A_{0n}^2 N_{0n} / n = 0$, *and* $\lim_n A_n \rho_{1n} = 0$. *Then* $\hat{\phi}$ *exists except on an event whose probability tends to zero as* $n \to \infty$, *and* $\|\hat{\phi} - \phi^*\|^2 = O_P(N_n/n + \rho_{1n}^2)$.

**Remark 4.1.** In Theorem 4.1, we can replace $\rho_{1n}$ by $\rho_n = \inf_{g \in \mathbb{G}} \|g - \eta^*\|_\infty$. However, we prefer to use $\rho_{1n}$ in the statement of this theorem since $\rho_{1n}$ is more straightforward to determine than $\rho_n$. In comparison with Theorem 2.1, the additional requirement $\lim_n A_{0n}^2 N_{0n}/n = 0$ is used to ensure that the theoretical and empirical inner products are close on $\mathbb{G}_0$, that is, $\sup_{g \in \mathbb{G}_0} |\|g\|_n / \|g\| - 1| = o_P(1)$.

**Remark 4.2.** Under the conditions of Theorem 4.1, $\|\hat{\eta} - \eta^*\| = O_P(N_n/n + \rho_n^2)$, where $\rho_n = \inf_{g \in \mathbb{G}} \|g - \eta^*\|_\infty$. See the proof of Theorem 4.1.

## 5. Conclusion

This work is inspired by the pioneering work of Stone (1985, 1986, 1994). In his rejoinder, Stone (1994) states that

> Thus, distinct but closely related theories have been or are being developed for regression, logistic and Poisson regression, polychotomous regression, hazard regression and the estimation of hazard, density, conditional density and spectral density estimation. It would be worthwhile to synthesis this theoretical work.

In this paper we achieve this synthesis by identifying an important common feature of various extended linear models — concavity of the log-likelihood. We obtain general asymptotic results that can treat simultaneously a broad range of estimation problems, a variety of structural models, and various types of estimation spaces. In our general framework, the structural assumption on the unknown function is specified by choosing an appropriate model space, and the estimation space is built accordingly. Our results are applicable when the structural assumption is either correctly specified or misspecified. These results are given under very broad conditions which need to be verified in each specific context. We illustrate, in the contexts of counting process regression and conditional density estimation, how to verify these broad conditions by using more primitive and statistically more natural conditions. As an important byproduct of the theoretical insight, we often obtain stronger results with simpler proofs. For example, our result on counting process regression extends in many ways Kooperberg et al. (1995a), but with a substantially simpler proof. Recently, the theoretical framework in this paper has been used as the foundation to study free knot splines in various extended linear models (see Stone and Huang (2000)).

## Acknowledgement

The author is grateful to Chuck Stone for continuous encouragement and many valuable suggestions during this work. The author also wishes to thank the editor, Ker-Chau Li, and an associate editor for constructive suggestions that led to significant improvement of the presentation and made the paper more focused.

## Appendix A. General results on rates of convergence

In this appendix, we provide two general results on convergence rates in extended linear modeling, one for handling the approximation error and one for handling the estimation error. These two results together yield Theorem 2.1.

### Approximation Error

**Condition A.1.** The best approximation $\eta^*$ in $\mathbb{H}$ to $\eta$ exists and there is

a positive constant $K_0$ such that $\|\eta^*\|_\infty \le K_0$.

In the regression context, $\eta^*$ is just the orthogonal projection of $\eta$ onto $\mathbb{H}$ relative to a certain inner product, which obviously exists; see Huang (1998a). In general, the existence of $\eta^*$ can be established by taking into account the specific properties of the log-likelihood; see, for example, Theorems 4.1 and 5.1 of Stone (1994), Theorem 1 of Kooperberg, Stone and Truong (1995a), and Theorem 2.1 of Huang and Stone (1998). When the model is concave, $\eta^*$ is essentially uniquely defined when it exists.

**Condition A.2.** For each pair of bounded functions $h_1, h_2 \in \mathbb{H}$, $\Lambda(h_1 + \alpha(h_2 - h_1))$ is twice continuously differentiable with respect to $\alpha$. For any positive constant $K$, there are positive numbers $M_1$ and $M_2$ such that

$$-M_1\|h_2 - h_1\|^2 \le \frac{d^2}{d\alpha^2}\Lambda(h_1 + \alpha(h_2 - h_1)) \le -M_2\|h_2 - h_1\|^2, \qquad 0 \le \alpha \le 1, \tag{A.1}$$

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

**Lemma A.1.** *Suppose Conditions* A.1 *and* A.2 *hold. Let $K_1$ be a positive constant such that $K_1 > K_0$ with $K_0$ as in Condition* A.1. *Then there are positive numbers $M_3$ and $M_4$ such that $-M_3\|h - \eta^*\|^2 \le \Lambda(h) - \Lambda(\eta^*) \le -M_4\|h - \eta^*\|^2$ for all $h \in \mathbb{H}$ with $\|h\|_\infty \le K_1$.*

**Proof.** Let $h \in \mathbb{H}$ with $\|h\|_\infty \le K_1$. Since $\eta^*$ maximizes $\Lambda(\cdot)$, $(d/d\alpha)\Lambda((1 - \alpha)\eta^* + \alpha h)|_{\alpha=0} = 0$. Integrating by parts, we get that

$$\Lambda(h) - \Lambda(\eta^*) = \int_0^1 (1 - \alpha)\frac{d^2}{d\alpha^2}\Lambda((1 - \alpha)\eta^* + \alpha h)\, d\alpha.$$

The desired result now follows from Condition A.2.

**Theorem A.1.** (Approximation Error) *Suppose Conditions* A.1 *and* A.2 *hold and* $\lim_n A_n\rho_n = 0$. *Let $K_1$ be a positive constant such that $K_1 > K_0$ with $K_0$ as in Condition* A.1. *Then $\bar\eta$ exists uniquely and $\|\bar\eta\|_\infty \le K_1$ for $n$ sufficiently large. Moreover, $\|\bar\eta - \eta^*\|^2 = O(\rho_n^2)$.*

**Proof.** Since $\mathbb{G}$ is finite-dimensional, it follows by a compactness argument that there is a function $g^* \in \mathbb{G}$ such that $\|g^* - \eta^*\|_\infty = \rho_n$. Let $a > 1$ denote a positive constant (to be determined later). Choose $g \in \mathbb{G}$ with $\|g - \eta^*\| \le a\rho_n$. Then by the definition of $A_n$, $\|g - g^*\|_\infty \le A_n\|g - g^*\| \le A_n(\|g - \eta^*\| + \|\eta^* - g^*\|) \le A_n\rho_n(a + 1)$ and $\|g\|_\infty \le \|g - g^*\|_\infty + \|g^* - \eta^*\|_\infty + \|\eta^*\|_\infty \le A_n\rho_n(a + 1) + \rho_n + \|\eta^*\|_\infty$. Since $\lim_n A_n\rho_n = 0$, we obtain that, for $n$ sufficiently large, $\|g\|_\infty \le K_1$

for all $g \in \mathbb{G}$ with $\|g - \eta^*\| \le a\rho_n$. It now follows from Lemma A.1 that, for $n$ sufficiently large,

$$\Lambda(g) - \Lambda(\eta^*) \le -M_4 a^2 \rho_n^2 \quad \text{for all } g \in \mathbb{G} \text{ with } \|g - \eta^*\| = a\rho_n, \tag{A.2}$$

$$\Lambda(g^*) - \Lambda(\eta^*) \ge -M_3 \rho_n^2. \tag{A.3}$$

Let $a$ be chosen such that $a > \sqrt{M_3/M_4}$. Then it follows from (A.2) and (A.3) that, for $n$ sufficiently large, $\Lambda(g) < \Lambda(g^*)$ for all $g \in \mathbb{G}$ with $\|g-\eta^*\| = a\rho_n$. Since $\|g^* - \eta^*\| < a\rho_n$, we conclude from the definition of $\bar{\eta}$ and the strict concavity of $\Lambda(\cdot)$ that $\bar{\eta}$ exists uniquely and $\|\bar{\eta} - \eta^*\| < a\rho_n$ for $n$ sufficiently large. Hence $\|\bar{\eta}\|_\infty \le K_1$ and $\|\bar{\eta} - \eta^*\|^2 = O(\rho_n^2)$.

### Estimation Error

**Condition A.3.** There is a positive constant $K_0$ such that, for $n$ sufficiently large, the best approximation $\bar{\eta}$ in $\mathbb{G}$ to $\eta$ exists uniquely and $\|\bar{\eta}\|_\infty \le K_0$.

This condition is, in fact, a consequence of Theorem A.1. It is convenient to state it as a condition so that, in the theorem below, conditions on the expected log-likelihood need not be specified.

**Condition A.4.** For any pair $g_1, g_2 \in \mathbb{G}$, $\ell(g_1 + \alpha(g_2 - g_1))$ is twice continuously differentiable with respect to $\alpha$.

(i)

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

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

**Theorem A.2.** (Extimation Error) *Suppose Conditions* A.3 *and* A.4 *hold and* $\lim_n A_n^2 N_n/n = 0$. *Let* $K_1$ *be a positive constant such that* $K_1 > K_0$ *with* $K_0$ *as in Condition* A.3. *Then* $\hat{\eta}$ *exists uniquely and* $\|\hat{\eta}\|_\infty \le K_1$, *except on an event whose probability tends to zero as* $n \to \infty$. *Moreover,* $\|\hat{\eta} - \bar{\eta}\|^2 = O_P(N_n/n)$.

**Proof.** By Taylor's expansion,

$$\ell(g) = \ell(\bar{\eta}) + \frac{d}{d\alpha} \ell(\bar{\eta} + \alpha(g - \bar{\eta}))\Big|_{\alpha=0} + \int_0^1 (1-\alpha) \frac{d^2}{d\alpha^2} \ell(\bar{\eta} + \alpha(g - \bar{\eta}))\, d\alpha, \qquad g \in \mathbb{G}. \tag{A.4}$$

Let $a$ be a positive number (to be determined later). Choose $g \in \mathbb{G}$ such that $\|g - \bar{\eta}\| \le a(N_n/n)^{1/2}$. Then by the definition of $A_n$, $\|g - \bar{\eta}\|_\infty \le A_n\|g - \bar{\eta}\| \le a(A_n^2 N_n/n)^{1/2} = o(1)$. Thus, for $n$ sufficiently large, $\|g\|_\infty \le K_1$ for all $g \in \mathbb{G}$ with $\|g - \bar{\eta}\| \le a(N_n/n)^{1/2}$. Consequently it follows from Condition A.4(ii) that, except on an event whose probability tends to zero as $n \to \infty$,

$$\int_0^1 (1-\alpha)\frac{d^2}{d\alpha^2}\ell(\bar{\eta} + \alpha(g - \bar{\eta}))\, d\alpha \le -\frac{M}{2}a^2\Big(\frac{N_n}{n}\Big) \tag{A.5}$$

for all $g \in \mathbb{G}$ with $\|g - \bar{\eta}\| = a(N_n/n)^{1/2}$. Fix an arbitrary positive constant $\epsilon$. By Condition A.4(i), we can choose $a$ sufficiently large such that, except on an event whose probability is less than $\epsilon$,

$$\left|\frac{d}{d\alpha}\ell(\bar{\eta} + \alpha(g - \bar{\eta}))\Big|_{\alpha=0}\right| < \frac{M}{2}a^2\Big(\frac{N_n}{n}\Big) \tag{A.6}$$

for all $g \in \mathbb{G}$ with $\|g - \bar{\eta}\| = a(N_n/n)^{1/2}$. Suppose (A.5) and (A.6) hold. Then, by (A.4), $\ell(g) < \ell(\bar{\eta})$ for all $g \in \mathbb{G}$ with $\|g - \bar{\eta}\| = a(N_n/n)^{1/2}$. Hence by the strict concavity of $\ell(\cdot)$ (which follows from Condition A.4(ii)), $\hat{\eta}$ exists uniquely and $\|\hat{\eta} - \bar{\eta}\| \le a(N_n/n)^{1/2}$. Since $\epsilon$ is arbitrary, the theorem follows.

**Remark A.1.** We give a sufficient condition for Condition A.4(i) when the norm $\|\cdot\|$ is associated with an inner product $\langle\cdot,\cdot\rangle$ defined on $\mathbb{G}$. Let $\{\phi_j : 1 \le j \le N_n\}$ be an orthonormal basis for $\mathbb{G}$ with respect to $\langle\cdot,\cdot\rangle$. Then each function $g \in \mathbb{G}$ can be represented uniquely as $g = \sum_j \beta_j\phi_j$, where $\beta_j = \langle g, \phi_j\rangle$ for $j = 1, \ldots, N_n$. Let $\beta$ denote the $N_n$-dimensional column vector with entries $\beta_j$. To indicate the dependence of $g$ on $\beta$, write $g(\cdot) = g(\cdot;\beta)$ and $\ell(g(\cdot;\beta)) = \ell(\beta)$. Let $\mathbf{S}(\beta) = \partial\ell(\beta)/\partial\beta$ denote the score at $\beta$, that is, the $N_n$-dimensional column vector having entries $\partial\ell(\beta)/\partial\beta_j$. Let $\bar{\beta}$ be the column vector with entries $\bar{\beta}_j = \langle\bar{\eta}, \phi_j\rangle$. Then $(d/d\alpha)\ell(\bar{\eta} + \alpha g)|_{\alpha=0} = [\mathbf{S}(\bar{\beta})]^T(\beta - \bar{\beta})$ and hence

$$\sup_{g\in\mathbb{G}} \frac{\left|\frac{d}{d\alpha}\ell(\bar{\eta} + \alpha g)\Big|_{\alpha=0}\right|}{\|g\|} \le |\mathbf{S}(\bar{\beta})|.$$

Consequently, a sufficient condition for Condition A.4(i) is that $|\mathbf{S}(\bar{\beta})| = O_P(N_n/n)^{1/2}$.

**Remark A.2.** The results in this section can be easily extended to the case where $l(h, \mathbf{w})$ and $\Lambda(h)$ are only defined when $h$ takes values on a proper open subinterval $\mathcal{I}$ of $\mathbb{R}$. This extension is useful in handling the estimation of functions taking values only in a restricted subdomain of $\mathbb{R}$, in particular, the conditional mean of the Poisson distribution in the generalized regression context (see Huang (1998b)).

## Appendix B. Proof of Theorem 3.1: counting process regression

To apply Theorems 2.1, we need only check Conditions A.1, A.2 and A.4. Condition A.1 is a consequence of Condition 3.1. Condition A.2 follows immediately from the definitions of $\Lambda$ and the theoretical norm. It remains to check Condition A.4, assuming that $\bar{\eta}$ exists and is bounded uniformly in $n$ (see the discussion below Condition A.3). Using the fact that $\bar{\eta} \in \mathbb{G}$ maximizes $\Lambda(g)$ in $g \in \mathbb{G}$, we obtain that

$$\begin{aligned}\frac{d}{d\alpha}\ell(\bar{\eta}+\alpha(g-\bar{\eta}))\Big|_{\alpha=0} &= (E_n - E)\Big(\textstyle\int_{\mathcal{T}}[g(t,\mathbf{X}(t)) - \bar{\eta}(t,\mathbf{X}(t))]\,dN(t)\Big) \\ &\quad -(\langle g-\bar{\eta}, \exp\bar{\eta}\rangle_n - \langle g-\bar{\eta}, \exp\bar{\eta}\rangle), \qquad g \in \mathbb{G}.\end{aligned}$$

Since $\bar{\eta} \in \mathbb{G}$ and $\|\bar{\eta}\|_\infty \le M < \infty$, we conclude from Lemma 11 of Huang (1998a) that

$$\sup_{g\in\mathbb{G}} \frac{|\langle g-\bar{\eta}, \exp\bar{\eta}\rangle_n - \langle g-\bar{\eta}, \exp\bar{\eta}\rangle|}{\|g-\bar{\eta}\|} = O_P\Big(\Big(\frac{N_n}{n}\Big)^{1/2}\Big). \tag{B.1}$$

We claim next that

$$\sup_{g\in\mathbb{G}} \frac{|(E_n - E)[\int_{\mathcal{T}} g(t,\mathbf{X}(t))\,dN(t)]|}{\{E\int_{\mathcal{T}} g^2(t,\mathbf{X}(t))\,dN(t)\}^{1/2}} = O_P\Big(\Big(\frac{N_n}{n}\Big)^{1/2}\Big). \tag{B.2}$$

Since $\bar{\eta} \in \mathbb{G}$, it follows from (B.2) that

$$\sup_{g\in\mathbb{G}} \frac{|(E_n - E)(\int_{\mathcal{T}}[g(t,\mathbf{X}(t)) - \bar{\eta}(t,\mathbf{X}(t))]\,dN(t))|}{\{E\int_{\mathcal{T}}[g(t,\mathbf{X}(t)) - \bar{\eta}(t,\mathbf{X}(t))]^2\,dN(t)\}^{1/2}} = O_P\Big(\Big(\frac{N_n}{n}\Big)^{1/2}\Big). \tag{B.3}$$

By conditioning and using Condition 3.2, we obtain that

$$E\int_{\mathcal{T}}[g(t,\mathbf{X}(t)) - \bar{\eta}(t,\mathbf{X}(t))]^2\,N(dt) \asymp \|g-\bar{\eta}\|^2, \qquad \text{uniformly in } g \in \mathbb{G}. \tag{B.4}$$

Hence Condition A.4(i) follows from (B.1) and (B.3). Condition A.4(ii) follows from the definition of the empirical norm and Lemma 10 of Huang (1998a).

We now prove (B.2). It follows from Condition 3.2 and the Cauchy–Schwartz inequality that

$$E\Big[\Big(\int_{\mathcal{T}} g(t,\mathbf{X}(t))\,E(N(dt)|\mathcal{F}_{t-})\Big)^2\Big] \lesssim \|g\|^2. \tag{B.5}$$

Now $M(\cdot) = N(\cdot) - \int_0^{\cdot} E(N(dt)|\mathcal{F}_{t-})$ is a square integrable martingale with predictable variation process $\langle M\rangle = \int_0^{\cdot} E(N(dt)|\mathcal{F}_{t-})$ (the square integrability of $M(\cdot)$ follows from Condition 3.2). Thus, the process $(\int_0^{\cdot} g(t,\mathbf{X}(t))\,dM(t))^2 - \int_0^{\cdot} g^2(t,\mathbf{X}(t))\,E(N(dt)|\mathcal{F}_{t-})$ is a martingale (see Theorem II.3.1 of Andersen, Borgan, Gill and Keidin (1993)). Consequently,

$$E\Big[\Big(\int_{\mathcal{T}} g(t,\mathbf{X}(t))\,dM(t)\Big)^2\Big] = E\int_{\mathcal{T}} g^2(t,\mathbf{X}(t))\,E(N(dt)|\mathcal{F}_{t-}) \asymp \|g\|^2. \tag{B.6}$$

Since $dN(t) = dM(t)+E(N(dt)|\mathcal{F}_{t-})$, we conclude from (B.5), (B.6), and the triangle inequality that $\mathrm{Var}\,(\int_{\mathcal{T}} g(t,\mathbf{X}(t))\,dN(t)) \le E[(\int_{\mathcal{T}} g(t,\mathbf{X}(t))\,dN(t))^2] \lesssim \|g\|^2$. On the other hand, the same argument as in (B.4) yields that $E\int_{\mathcal{T}} g^2(t,\mathbf{X}(t))\,dN(t) \asymp \|g\|^2$. Consequently, $\mathrm{Var}\,(\int_{\mathcal{T}} g(t,\mathbf{X}(t))\,dN(t)) \lesssim E\int_{\mathcal{T}} g^2(t,\mathbf{X}(t))\,dN(t)$, uniformly in $g \in \mathbb{G}$. Thus the same argument as in the proof of Lemma 11 of Huang (1998a) can be used to show that (B.2) holds. The proof of Theorem 3.1 is complete.

## Appendix C. Proof of Theorem 4.1: conditional density estimation

Recall that $\mathbb{G} \subset \mathbb{H}$, $\eta^*$ maximizes the expected log-likelihood over $\mathbb{H}$, and that $\hat{\eta}$ is the maximum likelihood estimate in $\mathbb{G}$. Suppose the conditions in Theorem 2.1 hold. It then follows that $\hat{\eta}$ exists except on an event whose probability tends to zero as $n \to \infty$, and $\|\hat{\eta}-\eta^*\|^2 = O_P(N_n/n+\rho_n^2)$ where $\rho_n = \inf_{g\in\mathbb{G}} \|g-\eta^*\|_\infty$. On the other hand, $\eta^*$ is bounded and $\hat{\eta}$ is bounded except on an event whose probability tends to zero as $n \to \infty$ (see Theorem A.2). Therefore, since $\sigma(\cdot)$ satisfies the Lipschitz condition (2.1), we find $\|\hat{\phi} - \phi^*\|^2 = \|\sigma(\hat{\eta}) - \sigma(\eta^*)\|^2 = O_P(N_n/n + \rho_n^2)$. Now, since $\rho_{1n} \le \|\phi^*\|_\infty < \infty$, there is a function $g_1^* \in \mathbb{G}_1$ such that $\|g_1^* - \phi^*\|_\infty = \rho_{1n}$. Set $\tilde{g} = P_\psi g_1^*$. Then $\tilde{g} \in \mathbb{G}$ and $\|\tilde{g} - \eta^*\|_\infty = \|P_\psi(g_1^*-\phi^*)\|_\infty \le 2\|g_1^*-\phi^*\|_\infty$, so $\rho_n \lesssim \rho_{1n}$ and hence $\|\hat{\phi}-\phi^*\|^2 = O_P(N_n/n+\rho_{1n}^2)$.

We need only check the conditions in Theorem 2.1. Condition A.1 with $\mathcal{I} = \mathbb{R}$ follows from the discussion after Condition 4.1. To check Condition A.2, suppose that $h_1$ and $h_2$ are in $\mathbb{H}$ and bounded. Using the fact that $f_{\mathbf{Y}_\alpha|\mathbf{X}}(\mathbf{y}|\mathbf{x})$ is bounded away from zero and infinity, we find

$$\mathrm{Var}\,\{[(h_2(\mathbf{Y}_\alpha|\mathbf{X}) - h_1(\mathbf{Y}_\alpha|\mathbf{X})]|\mathbf{X}=\mathbf{x}\} \asymp \int_{\mathcal{Y}} [h_2(\mathbf{y}|\mathbf{x}) - h_1(\mathbf{y}|\mathbf{x})]^2\,d\mathbf{y} \qquad \text{(C.1)}$$

uniformly in $\alpha \in (0,1)$ and $\mathbf{x} \in \mathcal{X}$. On the other hand, Condition 4.2 implies that the density of $\mathbf{X}$ is bounded away from zero and infinity. Condition A.2 thus follows from (4.1), (C.1), and Condition 4.2.

We now check Condition A.4. We can assume that $\bar{\eta}$ exists and is bounded uniformly in $n$ (see the discussion below Condition A.3). We have

$$\frac{\frac{d}{d\alpha}\ell(\bar{\eta}+\alpha(g-\bar{\eta}))\Big|_{\alpha=0}}{\|g-\bar{\eta}\|} = \frac{\langle\tilde{g},1\rangle_n - \langle\tilde{g},1\rangle}{\|g-\bar{\eta}\|},$$

where $\tilde{g}(\mathbf{x},\mathbf{y}) = g(\mathbf{y}|\mathbf{x}) - \bar{\eta}(\mathbf{y}|\mathbf{x}) - E[g(\bar{\mathbf{Y}}|\mathbf{X}) - \bar{\eta}(\bar{\mathbf{Y}}|\mathbf{X})|\mathbf{X}=\mathbf{x}]$. Since $\bar{\eta} \in \mathbb{G}$, $\{\tilde{g} : g \in \mathbb{G}\}$ is a linear space with dimension at most $\dim(\mathbb{G}) = N_n$. The same argument as in Lemma 11 of Huang (1998a) yields $\sup_{\tilde{g}:g\in\mathbb{G}}\{|\langle\tilde{g},1\rangle_n - \langle\tilde{g},1\rangle|/\|\tilde{g}\|\} = O_P((N_n/n)^{1/2})$. Since the joint density of $\mathbf{X}$ and $\mathbf{Y}$ is bounded away from zero

and infinity (Condition 4.2), $\|\tilde{g}\| \lesssim \|\tilde{g}\|_\psi$ and $\|g - \bar{\eta}\|_\psi \lesssim \|g - \bar{\eta}\|$. On the other hand, since the conditional density of $\bar{\mathbf{Y}}$ given $\mathbf{X} = \mathbf{x}$ is bounded away from zero and infinity, $\|\tilde{g}\|_\psi \lesssim \|g - \bar{\eta}\|_\psi$ uniformly in $g \in \mathbb{G}$. Condition A.4(i) then follows.

Observe that, for $g_1, g_2 \in \mathbb{G}$, $(d^2/d\alpha^2)\ell(g_1 + \alpha(g_2 - g_1)) = -E_n\{\mathrm{Var}\,[g_2(\mathbf{Y}_\alpha|\mathbf{X}) - g_1(\mathbf{Y}_\alpha|\mathbf{X})|\mathbf{X}]\}$. Thus, by (C.1), $(d^2/d\alpha^2)\ell(g_1 + \alpha(g_2 - g_1)) \asymp -\int_{\mathcal{Y}} E_n\{[g_2(\mathbf{y}|\mathbf{X}) - g_1(\mathbf{y}|\mathbf{X})]^2\}\, d\mathbf{y}$. Since $\lim_n A_{0n}^2 N_{0n}/n = 0$ by Lemma 10 of Huang (1998a), $\sup_{g \in \mathbb{G}_0} |\|g\|_n/\|g\| - 1| = o_P(1)$. Note that $g_2(\mathbf{y}|\mathbf{x}) - g_1(\mathbf{y}|\mathbf{x}) \in \mathbb{G}_0$ for fixed $\mathbf{y} \in \mathcal{Y}$. Hence, the right-hand side of the above display is bounded above and below by positive multiples of $-\int_{\mathcal{Y}} E\{[g_2(\mathbf{y}|\mathbf{X}) - g_1(\mathbf{y}|\mathbf{X})]^2\}\, d\mathbf{y}$. Condition A.4(ii) then follows from Condition 4.2. This completes the proof.

## Appendix D. Summary of some results on functional ANOVA

In this appendix, we summarize some results on functional ANOVA decompositions that are needed in Section 2.3 in studying structural models. See Huang (2000) for the details.

Given a hierarchical collection $\mathcal{S}$ of indices, define the model space $\mathbb{H}$ and the estimation space $\mathbb{G}$ as at the beginning of Section 2.3. We first give a formal definition of functional ANOVA decomposition.

Let $\langle \cdot, \cdot \rangle$ be a theoretical inner product defined on the space of Lebesgue square-integrable functions on $\mathcal{U}$, and let $\|\cdot\|$ denote the associated norm. Set $\mathbb{H}_\emptyset^0 = \mathbb{H}_\emptyset$ and, for each nonempty set $s \in \mathcal{S}$, let $\mathbb{H}_s^0$ denote the space of functions in $\mathbb{H}_s$ that are orthogonal (relative to the theoretical inner product) to each function in $\mathbb{H}_r$ for every proper subset $r$ of $s$. Under suitable conditions, each function $h \in \mathbb{H}$ can be written uniquely in the form $\sum_{s \in \mathcal{S}} h_s$, where $h_s \in \mathbb{H}_s^0$ for $s \in \mathcal{S}$. We refer to $\sum_{s \in \mathcal{S}} h_s$ as the *theoretical ANOVA decomposition* of $h$, and to $h_s$, $s \in \mathcal{S}$, as the components of $h$ in this decomposition.

Let $\langle \cdot, \cdot \rangle_n$ denote an empirical inner product that is determined by the data and let $\|\cdot\|_n$ denote the associated norm. Set $\mathbb{G}_\emptyset^0 = \mathbb{G}_\emptyset$ and, for each nonempty set $s \in \mathcal{S}$, let $\mathbb{G}_s^0$ denote the space of functions in $\mathbb{G}_s$ that are orthogonal (relative to the empirical inner product) to each function in $\mathbb{G}_r$ for every proper subset $r$ of $s$. Under suitable conditions, each function $g \in \mathbb{G}$ can be written uniquely in the form $\sum_{s \in \mathcal{S}} g_s$, where $g_s \in \mathbb{G}_s^0$ for $s \in \mathcal{S}$. We refer to $\sum_{s \in \mathcal{S}} g_s$ as the *empirical ANOVA decomposition* of $g$, and we refer to $g_s$, $s \in \mathcal{S}$, as the components of $g$.

**Condition D.1.** Let $\|\!|h|\!\|$ denote the $L_2$-norm on $\mathcal{U}$ relative to the Lebesgue measure. There are positive numbers $M_1$ and $M_2 \geq M_1$ such that $M_1 \|\!|h|\!\| \leq \|h\| \leq M_2 \|\!|h|\!\|$ for any Lebesgue square-integrable function $h$.

**Condition D.2.** $\sup_{g \in \mathbb{G}} |\|g\|_n/\|g\| - 1| = o_P(1)$.

**Condition D.3.** Fix any subspace $\widetilde{\mathbb{G}}$ of $\mathbb{G}$ with dimension $\widetilde{N_n}$. Then for any fixed sequence $h_n$, $n \geq 1$, of uniformly bounded functions on $\mathcal{U}$,

$$\sup_{g \in \widetilde{\mathbb{G}}} \frac{|\langle h_n, g\rangle_n - \langle h_n, g\rangle|}{\|g\|} = O_P\Big(\Big(\frac{\widetilde{N_n}}{n}\Big)^{1/2}\Big).$$

**Lemma D.1.** *Under Condition* D.1, $A_n \lesssim (\sum_{s\in\mathcal{S}} A_s^2)^{1/2}$.

Suppose now that $\eta^*$ and $\hat{\eta}$, as members of $\mathbb{H}$ and $\mathbb{G}$ respectively, have the ANOVA decompositions $\eta^* = \sum_{s\in\mathcal{S}} \eta_s^*$ and $\hat{\eta} = \sum_{s\in\mathcal{S}} \hat{\eta}_s$, where $\eta_s^* \in \mathbb{H}_s^0$ and $\hat{\eta}_s \in \mathbb{G}_s^0$ for $s \in \mathcal{S}$.

**Theorem D.1.** *Suppose Conditions* D.1–D.3 *hold. Then* $\|\hat{\eta}_s - \eta_s^*\|^2 = O_P(\|\hat{\eta} - \eta^*\|^2 + \sum_{s\in\mathcal{S}}(N_s/n + \rho_s^2))$.

## References


Andersen, P. K., Borgan, Ø., Gill, R. D. and Keiding, N. (1993). *Statistical Models Based on Counting Processes*. Springer-Verlag, New York.

Chen, H. (1995). Asymptotically efficient estimation in semiparametric generalized linear models. *Ann. Statist.* **23**, 1102-1129.

Dabrowska, D. M. (1997). Smoothed Cox regression. *Ann. Statist.* **23**, 1510-1540.

Hansen, M. (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 for multiple regression with application to functional ANOVA models. *Ann. Statist.* **26**, 242-272.

Huang, J. Z. (1998b). Functional ANOVA models for generalized regression. *J. Multivariate Anal.* **67**, 49-71.

Huang, J. Z. (2000). Convergence of components of an estimate in functional ANOVA models. Manuscript.

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

Huang, J. Z., Kooperberg, C., Stone, C. J. and Truong, Y. K. (2000). Functional ANOVA modeling for proportional hazards regression. *Ann. Statist.* To appear.

Jacod, J. (1975). Multivariate point processes: predictable projection, Radon–Nikodym derivatives, representation of martingales. *Z. Wahrsch. Verw. Gebiete.* **31**, 235-254.

Kooperberg, C., Stone, C. J. and Truong, Y. K. (1995a). The $L_2$ rate of convergence for hazard regression. *Scand. J. Statist.* **22**, 143-157.

Kooperberg, C., Stone, C. J. and Truong, Y. K. (1995b). Rate of convergence for logspline spectral density estimation. *J. Time Ser. Anal.* **16**, 389-401.

McKeague, I. W. and Utikal, K. J. (1990). Inference for a nonlinear counting process regression model. *Ann. Statist.* **18**, 1172-1187.

Nielsen, J. P. and Linton, O. B. (1995). Kernel estimation in a nonparametric marker dependent hazard model. *Ann. Statist.* **23**, 1735-1749.

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

Stone, C. J. (1982). Optimal global rates of convergence for nonparametric regression. *Ann. Statist.* **10**, 1348-1360.

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

Stone, C. J. (1991). Asymptotics for doubly flexible logspline response models. *Ann. Statist.* **19**, 1832-1854.

Stone, C. J. (1994). The use of polynomial splines and their tensor products in multivariate function estimation (with discussion). *Ann. Statist.* **22**, 118-171.

Stone, C. J., Hansen, M., Kooperberg, C. and Truong, Y. (1997). Polynomial splines and their tensor products in extended linear modeling (with discussion). *Ann. Statist.* **25**, 1371-1470.

Stone, C. J. and Huang, J. Z. (2000). Free knot splines in concave extended linear modeling. Submitted.

Department of Statistics, the Wharton School, University of Pennsylvania, Philadelphia, PA 19104-6302, U.S.A.

E-mail: jianhua@wharton.upenn.edu

(Received August 1998; accepted June 2000)