Covariate selection for semiparametric hazard function regression models

Florentina Bunea       Ian W. McKeague

Department of Statistics
Florida State University
Tallahassee, FL 32306-4330, USA
bunea@stat.fsu.edu      mckeague@stat.fsu.edu

Abstract

We study a flexible class of non-proportional hazard function regression models in which the influence of the covariates splits into the sum of a parametric part and a time-dependent nonparametric part. We develop a method of covariate selection for the parametric part by adjusting for the implicit fitting of the nonparametric part. Our approach is based on the general model selection methodology of Barron, Birgé and Massart, suitably adapted to the censored semiparametric regression framework. Asymptotic consistency of the proposed covariate selection method is established, leading to asymptotically normal estimators of both parametric and nonparametric parts of the model in the presence of covariate selection. The approach is applied to a real data set and a simulation study is presented.

1 Introduction

Covariate selection is a form of model selection in which the class of models under consideration is represented by subsets of covariate components to be included in the analysis. Model selection methods are well developed in parametric settings, and in recent years they have been extended to wide classes of nonparametric models (Barron et al. 1999). For applications in survival analysis, however, in which the presence of censoring and the use of

1Key words and phrases. Additive risk model, Cox model, penalized partial likelihood, survival analysis.
complex time-dependent hazard function regression models is becoming increasingly pop-
ular (see, e.g., Andersen et al. 1993), generally applicable and fully validated procedures
have not yet been developed.

In this paper we study covariate selection for conditional hazard function models of the
form

\[
h(t, x, z) = \psi(\beta^T x + f(t)^T z),
\]

where \( \psi \) is a known (non-negative) link function, \((x, z)\) is a partition of the covariates into
a \( q \)-vector \( x \) and a \( p \)-vector \( z \), \( \beta \) is an unknown \( q \)-vector of regression parameters and \( f(t) \)
is an unknown \( p \)-dimensional non-random function of time. We develop a model selection
procedure to find the best subset of \( x \)-covariates and study the asymptotic properties of the
corresponding regression parameter estimates after model selection.

The above model provides a flexible extension of the Cox (1972) proportional hazards
model \( h(t, x) = \exp(\beta^T x + f(t)) \), where \( f(t) \) is the log-baseline hazard function. Our model
is more flexible in the sense that it allows some of the covariates to have a longitudinal (or
time-dependent) influence on survival. For the identity link function, the model reduces
to the partly parametric additive risk model of McKeague and Sasieni (1994). Recently,
Martinussen et al. (2002) studied the model in the case of an exponential link function.

Typically, some covariates are known to have a longitudinal influence on survival, so
those covariates are placed in \( z \). However, only a small (but fixed) number of covariates
can be treated in this way as an additional time-dependent function enters the model for
each component of \( z \). The remaining covariates are placed in \( x \). This creates the need for a
procedure to select a subset of the \( x \)-components that avoids both overfitting and underfitting.
With the non-zero components of \( \beta \) corresponding to an unknown subset \( I = I_0 \) of the \( x \)-
covariates, the statistical problem is to estimate \( I_0 \) and the corresponding components of
\( \beta \).
Numerous covariate selection procedures have been proposed for the Cox model: penalized partial likelihood—henceforth PPL (Senoussi, 1990), a backwards elimination covariate selection method (Fleming and Harrington, 1991), Bayesian model averaging (Raftery et al. 1996, 1997), Bayesian variable selection (Faraggi and Simon, 1998), the lasso method for PPL (Tibshirani, 1997), and non-concave PPL (Fan and Li, 2002). Large sample properties of these procedures are largely unexplored, with the exceptions of Senoussi (1990) and Fan and Li (2002). All these procedures only require parametric model selection techniques because they exploit partial likelihood which does not involve the infinite dimensional part of the semiparametric model (the baseline hazard function). A more sophisticated PPL procedure was developed by Letue (2000) for fitting the general proportional hazards model

\[ h(t, x) = \exp(g(x) + f(t)), \]

where \( g(x) \) is an unknown function of the covariates \( x \) and \( f(t) \) is the log-baseline hazard function. This model may be unsuitable, however, when \( x \) has high dimension because of the curse of dimensionality. None of the above procedures extends beyond the proportional hazards framework.

To study semiparametric models of the form (1.1), in which a partial likelihood for \( \beta \) is not available and \( I_0 \) is also regarded as a parameter, we need a different approach. We consider the following two-stage procedure. The first stage (covariate selection) is to estimate \( I_0 \) by \( \hat{I} \) derived from maximizing a penalized full likelihood. To produce a consistent \( \hat{I} \), the effect of estimating \( f \) nonparametrically needs to be controlled via a penalty that is different from the parametric model selection procedures mentioned above. In Senoussi (1990), for instance, the penalty term has the form \( a_n |I|/n \), with \( a_n \to \infty \) and \( a_n/n \to 0 \), where \(||\) denotes the cardinality of a set; restrictions on \( a_n \) (e.g., \( a_n = \log n \)) then lead to consistent estimators of \( I_0 \). This type of penalty term will not work for full likelihood because the penalty must also balance the bias caused by estimation of the infinite dimensional part of
the model; see Bunea (2002) for a regression example. The second stage of our procedure is to refit the model with the \( x \)-components restricted to those in \( \hat{I} \) using the estimators of \( \beta \) and the cumulative regression function \( \int_0^t f(s) \, ds \) developed by McKeague and Sasieni (1994) and Martinussen et al. (2002). The end result is consistent covariate selection along with asymptotically normal estimators of both parametric and nonparametric parts of the model.

The paper is organized as follows. In Sections 2.1 and 2.2 we introduce the proposed method. Section 2.3 contains the main result giving the consistency of \( \hat{I} \). In Section 3 we present a simulation study comparing the proposed approach with various competitors. An application to real data is discussed in Section 4. The proofs of intermediate results are collected in Section 5.

2 Covariate selection

In this section we present the proposed method of selecting the best subset \( I_0 \) of the \( x \)-covariates based on a penalized full likelihood procedure. The procedure leads to consistent estimates of \( I_0 \). We also establish upper bounds on the convergence rates of the corresponding estimators of \( \beta \) and \( f \).

2.1 Preliminaries

The survival time \( T \) is assumed to be conditionally independent of a censoring time \( C \) given the covariates \((X, Z)\). We observe \( n \) i.i.d. copies of the right censored survival time \( T^{\text{cens}} = \min(T, C) \) and the censoring indicator \( \delta = 1(T \leq C) \), along with \((X, Z)\). The true conditional hazard function \( h(t, X, Z) \) of \( T \) given \((X, Z)\) is specified by (1.1) where \( t \) is restricted to a fixed time interval \([0, \tau]\). The covariates are assumed to be bounded.

We suppose that the link function \( \psi \) is positive, continuous and strictly increasing on some (sufficiently large) known bounded interval \([a, b]\) for which \( \psi(a) \leq h(t, x, z) \leq \psi(b) \) for all \( t, x, z \). This means that \( h(t, x, z) \) has known uniform bounds in terms of values of the
given link function. For the identity link function, \(a > 0\) and \(b\) represent prespecified bounds on the hazard function; in practice, \(a\) can be chosen arbitrarily small and \(b\) arbitrarily large, so they have no effect on the estimation procedure. For the exponential link function, \(a\) and \(b\) are bounds on the log-hazard function. The inverse of \(\psi\) is denoted \(\psi^{-1} : [\psi(a), \psi(b)] \to [a, b]\).

The function
\[
    r(t, x, z) = \psi^{-1}(h(t, x, z)) = \beta^T x + f(t)^T z
\]
plays a central role in our approach; in the case of the exponential link, \(r\) is simply the log-hazard function.

The following set of conditions is assumed throughout.

**CONDITIONS**

(A1) \(a \leq r(t, X, Z) \leq b\).

(A2) \(\psi\) and \(\psi^{-1}\) are Lipschitz on \([a, b]\) and \([\psi(a), \psi(b)]\), respectively.

(A3) \(P(C \geq \tau | X, Z)\) is bounded away from zero.

(A4) \(\text{Var}(l^T X | Z = z) > 0\) for any nonzero \(l \in \mathbb{R}^q\) and any \(z\).

(A5) \(\text{Var}(d^T Z | X = x) > 0\) for any nonzero \(d \in \mathbb{R}^p\) and any \(x\).

(A6) The components of \(f\) belong to \(B_\infty^\alpha(L^2)\), for some \(1/2 < \alpha \leq 1\).

(A7) \(X\) and \(Z\) are uniformly bounded.

Here \(B_\infty^\alpha(L^2)\) is the Besov space of order \(\alpha\) corresponding to the \(L^2\)-space of square-integrable functions on \([0, \tau]\); see, e.g., DeVore and Lorentz (1993) for the precise definition and properties.

Conditions A4 and A5 are identifiability assumptions that allow us to make separate inferences on the parametric and nonparametric parts of the model, and can be checked in
practice by inspecting scatterplots of the components of $X$ with respect to the components of $Z$.

### 2.2 Sieves and selection criterion

We now introduce suitable parametric submodels (sieves) consisting of the functions $u(t, x, z)$ that will be used to approximate the true $r(t, x, z)$. The sieves also naturally provide approximations to the true conditional hazard function.

Define the sieve

$$S_I = \bar{S}_I \cap \{u(\cdot): a \leq u(\cdot) \leq b\}$$

indexed by a given subset $I = \{i_1, \ldots, i_l\}$ of the $x$-covariate indices, where $\bar{S}_I$ is the finite-dimensional linear approximating space

$$\bar{S}_I = \langle x_{i_1}, \ldots, x_{i_l}, \phi_{n,1}(t)z_1, \ldots, \phi_{n, N_n}(t)z_1, \ldots, \phi_{n,1}(t)z_p, \ldots, \phi_{n, N_n}(t)z_p \rangle,$$  \hspace{1cm} (2.3)

with $\phi_{n,i}(t) \equiv 1_{[i-1)N_n,i/N_n]}(t/\tau)$, for $i = 1, \ldots, N_n$, and $N_n = \lceil n^{1/(2\alpha+1)} \rceil$, where $\lceil \cdot \rceil$ denotes the integer part. Thus, within each $S_I$, the components of $f$ are approximated by step functions based on a regular partition of $[0, \tau]$. The mesh of the partition depends on $\alpha$ and the sample size $n$. We note that although minimax adaptive estimation of these functions is possible, see, e.g, Barron et al. (1999) for a very general approach, this would require the construction of a much richer set of approximating spaces which would result in a very involved algorithm that may become computationally intractable. Since at this stage of our estimation procedure we only need a good initial estimate of the nonparametric part of our model, we content ourselves with the simple construction above.

To select amongst the subsets $I \subseteq \{1, \ldots, q\}$ we use the re-normalized log-likelihood as a contrast function:

$$\gamma_n(u) = -n^{-1} \sum_{i=1}^{n} \left\{ \int_0^\tau \log \psi(u(t, X_i, Z_i)) \, dN_i(t) - \int_0^\tau Y_i(t) \psi(u(t, X_i, Z_i)) \, dt \right\},$$ \hspace{1cm} (2.4)
see (5.17). The subscript $i$ above refers to the $i$th individual in the sample, $N_i(t) = I(T_i^{\text{cens}} \leq t, \delta_i = 1)$, and $Y_i(t) = I(T_i^{\text{cens}} \geq t)$. We declare $\hat{r} \in S_I$ a penalized maximum likelihood sieve estimator if
\[
\gamma_n(\hat{r}) + \text{pen}(\hat{r}) = \inf_I \left[ \inf_{u \in S_I} (\gamma_n(u) + \text{pen}(I)) \right],
\]
where $\text{pen}(I)$ is a penalty term which will be defined in the next subsection.

### 2.3 Consistency of the selection

In this subsection we show that the method proposed above consistently estimates $I_0$. The choice of the penalty term is crucial for this result. We begin by giving the motivation behind our choice and we defer the full proof to Theorem 2. Note that
\[
P(\hat{I} \neq I_0) = P(I_0 \subseteq \hat{I}) + P(I_0 \nsubseteq \hat{I}).
\]
We show in Theorem 1 that our procedure leads to consistent estimators of $\beta$. Then, it is easy to show that $P(I_0 \nsubseteq \hat{I}) \to 0$. To see this assume that, for example, $\beta = (1, 2, 0, 3)$. Then we cannot consistently estimate it by, say, $\hat{\beta} = (\hat{\beta}_1, \hat{\beta}_2, 0, 0)$. The study of the second inclusion is more delicate. One cannot rely on the consistency of $\hat{\beta}$ alone to show that $P(I_0 \nsubseteq \hat{I}) \to 0$, as we can always consistently estimate zeros. Thus, one needs a different argument, which will, in turn, lead to restrictions on the penalty term. Note that
\[
P(I_0 \subseteq \hat{I}) = \sum_{I \supseteq I_0} P(\hat{I} = I).
\]
Let $\mathbb{D}_n \equiv \mathbb{P}_n - P$, where the measure $P$ corresponds to the hazard function $h = \psi \circ r$ and $\mathbb{P}_n$ is the empirical measure that puts mass $1/n$ at each observation. Then, as in the course of the proof of Theorem 2, by the definition of our estimator, for any $I \supseteq I_0$ and an appropriately defined constant $B$ and function $r_I$, an upper bound for $P(\hat{I} = I)$ is given by
\[
P \left( \sup_{v \in \hat{S}_I} [\gamma(r_I) - \gamma(v)] - \|v - r_I\|^2_v > \text{pen}(I) - \text{pen}(I_0) - BpN_n^{-2\alpha} \right).
\]
Thus, a first restriction on the penalty term is

$$\text{pen}(I) - \text{pen}(I_0) > B p N_n^{-2\gamma}. \quad (2.7)$$

Note that $B p N_n^{-2\gamma}$ is a bias term introduced through the approximation of the infinite-dimensional part of the model within a space $S_I$ of finite dimension $p N_n$. We emphasize that a similar derivation for a fully parametric model would not contain the bias term $B p N_n^{-2\gamma}$ and that a requirement of the type (2.7) is not a byproduct of our method of proof, but it is intrinsic to the nature of a semiparametric model; see also Bunea (2002) for a regression example. More generally, a penalty term satisfying (2.7) can be used in any semiparametric model selection procedure in which the criterion $\gamma$ satisfies the same bound as in (2.6).

For our choice of $N_n$ and with $|I|$ denoting the cardinality of $I$, the following penalty term

$$\text{pen}(I) = C \left( \frac{|I| + 1}{n} \right)^p N_n \log n, \quad (2.8)$$

satisfies (2.7), for $n$ large enough. We discuss the choice of $C$ for large and small samples in Section 5.1. Notice that a penalty term in which the dimensions of the two parts of the approximating space are added rather than multiplied does not satisfy (2.8); thus, a direct extension of the penalty terms developed for selection methods in which the parameter of interest belongs to one of the candidate spaces fails in this context. For this choice of penalty, we prove in Theorem 2 that, asymptotically, we cannot underestimate or overestimate $I_0$. As we discussed above, to show that we cannot underestimate $I_0$, we shall first show that the selected $\hat{\beta}$ is consistent. We prove this in Corollary 1, which is an immediate consequence of the much stronger result of Theorem 1, in which we give finite sample upper bounds on the risk of our estimators.

Let $r_I$ denote the orthogonal projection of $r$ onto $\tilde{S}_I$ in the $L^2$ space corresponding to the measure $\nu = \text{Leb} \times \mu_{X,Z}$ on $[0, \tau] \times \mathbb{R}^q \times \mathbb{R}^p$ and $\mu_{X,Z}$ is the distribution of $(X, Z)$. Notice
that, by Lemma 3, \( r_I \in S_I \). The \( L^2_{\nu} \) and \( L^2_{\nu,1} \)-norms, respectively, are denoted \( \| \cdot \|_{\nu} \) and \( \| \cdot \|_2 \). The Euclidean norm is denoted \( | \cdot |_2 \).

**Theorem 1** Under Conditions A1–A7, for the estimators relative to the collection of approximating spaces (2.3) and for the penalty term (2.8) there exist positive constants \( C_1 \) and \( C_2 \) such that

\[
E_P \| r - \hat{r} \|^2_{\nu} \leq C_1 \inf_I \left[ \| r - r_I \|^2_{\nu} + \text{pen}(I) + C_2/n \right].
\]  

(2.9)

The following corollary provides rates of convergence of the estimators \( \hat{\beta} \) and \( \hat{f} \) corresponding to \( \hat{r} \). Let \( \mathcal{D}_\alpha \) denote the family of functions \( f = (f_1, \ldots, f_p)^T \) with each component \( f_j \) belonging to a fixed bounded subset of the Besov space in Condition A6.

**Corollary 1** Under the conditions of Theorem 1, we obtain:

1. \( \| \hat{f}_j - f_j \|_2^2 = O_P \left( \frac{\log n}{n^{2/(2+1)}} \right) \), uniformly over \( f \in \mathcal{D}_\alpha \).
2. \( \| \hat{\beta} - \beta \|_2^2 = O_P \left( \frac{\log n}{n^{2/(2+1)}} \right) \), uniformly over \( \beta \) in any compact \( \mathcal{K} \subset \mathbb{R}^q \).

In the above result, the rate for \( \hat{f} \) is the minimax optimal nonparametric rate, up to a \( \log n \) factor, but the rate for \( \hat{\beta} \) is not the optimal \( \sqrt{n} \)-rate. The optimal rate for estimating \( \beta \) can be retrieved by using an existing \( \sqrt{n} \)-consistent estimator in the model with \( \hat{I} \) as the \( x \)-covariate components, as in Theorem 5.2 of Bunea (2002). The following consistency result for \( \hat{I} \) validates this procedure.

**Theorem 2** Under the conditions of Theorem 1, we have \( P(\hat{I} \neq I_0) \to 0 \) as \( n \to \infty \).

**Proof** Note that

\[
P(\hat{I} \neq I_0) = P(I_0 \subsetneq \hat{I}) + P(I_0 \supsetneq \hat{I})
\]

(2.10)

We show that each term in the right hand side of (2.10) converges to zero.

1. \( P(I_0 \subsetneq \hat{I}) \to 0 \) as \( n \to \infty \).
Notice that if \( I_0 = \{1, \ldots, q\} \), \( P(I_0 \subseteq \hat{I}) = 0 \), so it is enough to consider \( I_0 \subsetneq \{1, \ldots, q\} \).

We can write

\[
P(I_0 \subset \hat{I}) = \sum_{I \supset I_0} P(\hat{I} = I)
\]

where \( I \subset \{1, \ldots, q\} \). Define

\[
f_n(I) \equiv \inf_{u \in S_I} (\gamma_n(u) + \text{pen}(I)).
\]

By the definition of the estimator we have

\[
P(\hat{I} = I) = P(f_n(I) - f_n(I') < 0, \text{ for all } I' \neq I) \leq P(f_n(I) - f_n(I_0) < 0).
\]

With notation (2.12), by adding and subtracting \( \gamma_n(r_I) + \text{pen}(I_0) \) , we have

\[
f_n(I) - f_n(I_0) = -\sup_{v \in S_{n(I)}} [\gamma_n(r_I) - \gamma_n(v) - \text{pen}(I) + \text{pen}(I_0)]
\]

\[-\inf_{u \in S_{I_0}} [\gamma_n(u) - \gamma_n(r_I) + \text{pen}(I_0) - \text{pen}(I_0)].
\]

By (2.11), we restrict attention only to \( I \supset I_0 \). By Lemma 3, \( r_I \in S_I \) for any \( I \supset I_0 \). Then

\[
P(f_n(I) - f_n(I_0) < 0)
\]

\[
< P(\sup_{v \in S_I} [\gamma_n(r_I) - \gamma_n(v) - \text{pen}(I) + \text{pen}(I_0)] > 0)
\]

\[+ P(\inf_{u \in S_{I_0}} [\gamma_n(u) - \gamma_n(r_I)] > 0)
\]

\[
< P(\sup_{v \in S_I} [\gamma_n(r_I) - \gamma_n(v) - \text{pen}(I) + \text{pen}(I_0)] > 0)
\]

\[+ P(\gamma_n(r_I) - \gamma_n(r_I) > 0)
\]

\[= P(\sup_{v \in S_I} [\gamma_n(r_I) - \gamma_n(v) - \text{pen}(I) + \text{pen}(I_0)] > 0).
\]

Notice now that, for \( c_1 \) and \( c_2 \) given by Lemma 1 in Section 5, we obtain

\[
\gamma_n(r_I) - \gamma_n(v) = \mathbb{P}_n[\gamma(r_I) - \gamma(v)] + E_P(\gamma(r_I) - \gamma(r)) - E_P(\gamma(v) - \gamma(r))
\]
\[
\begin{align*}
\leq & \; \mathbb{D}_n[\gamma(r_I) - \gamma(v)] + c_2\|r_I - r\|_v^2 - c_1\|v - r\|_v^2 \\
\leq & \; \mathbb{D}_n[\gamma(r_I) - \gamma(v)] - c_1\|v - r_I\|_v^2/2 + (c_1 + c_2)\|r_I - r\|_v^2 \\
\leq & \; \mathbb{D}_n[\gamma(r_I) - \gamma(v)] - c_1\|v - r_I\|_v^2/2 + B_1(c_1 + c_2)n^{-2\alpha/(2\alpha + 1)},
\end{align*}
\]

since \(\|v - r_I\|_v^2 \leq 2(\|v - r\|_v^2 + \|r_I - r\|_v^2)\) and for \(B_1\) given in the proof of Corollary 1. Then, with \(\text{pen}(I) = C(|I| + 1)n^{-2\alpha/(2\alpha + 1)}\log n\), for \(n\) large enough and a dominating constant \(L^*\), we obtain

\[
\begin{align*}
\gamma_n(r_I) - \gamma_n(v) &= \text{pen}(I) + \text{pen}(I_0) \\
\leq & \; \mathbb{D}_n[\gamma(r_I) - \gamma(v)] - c_1\|v - r_I\|_v^2/2 + B_1(c_1 + c_2)n^{-2\alpha/(2\alpha + 1)} - Cn^{-2\alpha/(2\alpha + 1)} \log n \\
\leq & \; \mathbb{D}_n[\gamma(r_I) - \gamma(v)] - c_1\|v - r_I\|_v^2/2 - L^*n^{-2\alpha/(2\alpha + 1)} \log n.
\end{align*}
\]

Notice now that \(2n^{1/(2\alpha + 1)} \geq (|I| + pn^{1/(2\alpha + 1)})/p(q + 1)\). Let

\[
\sigma_n \equiv (|I| + pn^{1/(2\alpha + 1)}) \log n / 2np(q + 1).
\]

Let \(A \equiv L^*/2p(q + 1)\) and \(A^* \equiv \min(c_1/2, A)\). Then, noting that \(|I| + p[n^{1/(2\alpha + 1)}] \) is the dimension of \(S_I\), we apply Theorem 5 of Birgé and Massart (1998). For positive constants \(C_3\) and \(C_4\) given by this theorem, we obtain, for any \(I \supseteq I_0\), that

\[
\begin{align*}
P \left( \sup_{v \in S_I} [\gamma_n(r_I) - \gamma_n(v) - \text{pen}(I) + \text{pen}(I_0)] > 0 \right) \\
\leq & \; P \left( \sup_{v \in S_I} \mathbb{D}_n[\gamma(r_I) - \gamma(v)] > A^*\|r_n - v\|_v^2 + \sigma_n \right) \\
\leq & \; P \left( \sup_{v \in S_I} \mathbb{D}_n[\gamma(r_I) - \gamma(v)] > 1/A^* \right) \leq C_3 \exp(-C_4n\sigma_n) \to 0.
\end{align*}
\]

Note that the hypotheses of Theorem 5 of Birgé and Massart (1998) are verified in the course of the proof of our Theorem 1. Then, from (2.11), (2.13), (2.14) and (2.15), we have \(P(I_0 \subset \hat{I}) \to 0\), which completes the proof of this step.

2. \(P(I_0 \not\subset \hat{I}) \to 0\).

\[
P(I_0 \not\subset \hat{I}) = P(j \notin \hat{I} \text{ for all } j \in I_0) \leq P(j_0 \notin \hat{I} \text{ for some } j_0 \in I_0)
\]
\[ P(j_0 \in I_0 - \hat{I}) \leq P(\beta_{j_0} \neq 0, \hat{\beta}_{j_0} = 0) \]
\[ \leq P(|\hat{\beta}_{j_0} - \beta_{j_0}| = |\beta_{j_0}| > 0) \rightarrow 0, \]

by the component-wise consistency of \( \hat{\beta} \). This completes the proof of this theorem.

3 Simulation study

This section reports some simulation results designed to compare the proposed approach with various competitors.

We compare our penalized full likelihood (proposed-PFL) procedure with the penalized partial likelihood (PPL) procedure having penalty term \( \text{pen}(I) = C \log n |I|/n \), as used by Senoussi (1990). Note that, in the special case of the Cox model, this is a comparison between two asymptotically consistent methods. We also compare with the performance of an alternative PFL procedure (naive-PFL) having penalty term \( \text{pen}(I) = C \log n (|I| + pN_n)/n \), which may be regarded as a naive adjustment for the bias caused by estimating the nonparametric part of the model. To give a fair comparison between the three procedures we restrict our simulation study to the case of a Cox model (exponential link function and \( p = 1 \)).

The data were simulated using the conditional hazard function (1.1) with exponential link, \( q = 7, \beta = (1, 1, 0, 0, 0, 1, 0)^T, p = 1 \) and \( f(t) = 0 \), which is a Cox model with constant baseline hazard function, so both PPL and proposed-PFL are consistent. The covariates \( X \) were i.i.d. uniform \((0, 1)\), and \( Z \equiv 1 \). The censoring time was taken as exponential with rate .5, and the end of follow-up \( \tau = .3 \). Note that the true \( x \)-covariate indices are \( I_0 = \{1, 2, 6\} \).

We began by simulating 50 datasets each with 50 observations and estimated \( \beta \) via each procedure. We chose \( \alpha = 1 \) in the case of proposed-PFL. To calibrate the tuning constant \( C \) in each case, we varied \( C \) over a fine grid and examined boxplots of the estimates of \( \beta \); the best results were obtained with \( C = 0.7 \) for proposed-PFL and \( C = 0.3 \) for the other two
Figure 1: Componentwise boxplots of estimates of $\beta$. Top row: proposed-PFL ($n = 50, 200$). Middle row: PPL ($n = 50, 200$). Bottom row: naive-PFL ($n = 50, 200$).

procedures, see the left panels of Figure 1. All three procedures perform well and there is no clear winner at this sample size.

Next we simulated 50 datasets each with 200 observations and applied the procedures using the tuning constant $C$ calibrated for $n = 50$; see the right panels of Figure 1. It is clear that the proposed procedure outperforms both PPL and naive-PFL, and correctly identifies
the zero coefficients of $\beta$ in almost every case. The results strongly suggest that our approach achieves consistency of $\hat{I}$ at a faster rate than both the PPL and naive-PFL methods.

4 Example

The data come from a Mayo Clinic trial in primary biliary cirrhosis of the liver, see Fleming and Harrington (1991). Times between registration and death (possibly right censored) are available for 312 patients; we only consider the 276 patients for whom complete covariate information is available at registration. Nine of the 17 covariates clearly have no effect and are excluded. We restrict attention to the following eight:

- age: Age in years
- edema: Presence of edema (0 = no, .5 resolved, 1 = unresolved with therapy)
- bili: Serum bilirubin, in mg/dl.
- albu: Albumin, in gm/dl.
- copp: Urine copper, in $\mu$g/day.
- SGOT: SGOT, in U/ml.
- thromb: Prothrombine time, in seconds.
- hist: Histologic stage of disease, graded 1, 2, 3, or 4.

Of these covariates, bili, albu and thromb were log-transformed. We used an exponential link, $\tau = 3500$ days, $\alpha = 1$, $p = 1$, $Z \equiv 1$, and the same penalty constants $C$ as in the simulation study above. The results are displayed in Table 1.

We find the best subset of covariates to be $\hat{I} = \{\text{bili, albu, copp, thromb, hist}\}$ using the proposed-PFL method. In contrast, the PPL method gives $\{\text{age,bili,albu,copp,hist}\}$, and this is in essential agreement with the lasso solution of Tibshirani (1997). On the other hand, Fleming and Harrington (1991), using a backwards elimination method, concluded that the
best selection was \{age, edema, bili, albu, thromb\}, as did Raftery et al. (1996) using Bayesian model averaging.

Our approach yields a different result to these previous analyses. In particular, the variable thromb has a high \(Z\)-score under our procedure, in marked disagreement with PPL which does not include that covariate in the selected model. The naive-PFL method only excludes one of the covariates (SGOT), and on the basis of the simulation study we may explain this as overfitting due to the incorrect penalty term.

**Table 1.** Comparison of estimates of \(\beta\) for the Mayo Clinic data.

<table>
<thead>
<tr>
<th>Covariate</th>
<th>PPL</th>
<th></th>
<th></th>
<th>Naive-PFL</th>
<th></th>
<th></th>
<th>Proposed-PFL</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Coeff</td>
<td>SE</td>
<td>(Z)-score</td>
<td>Coeff</td>
<td>SE</td>
<td>(Z)-score</td>
<td>Coeff</td>
<td>SE</td>
<td>(Z)-score</td>
</tr>
<tr>
<td>age</td>
<td>0.03</td>
<td>0.009</td>
<td>2.92</td>
<td>0.02</td>
<td>0.009</td>
<td>2.16</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>edema</td>
<td>0.98</td>
<td>0.39</td>
<td>3.08</td>
<td>1.12</td>
<td>0.33</td>
<td>3.39</td>
<td>1.05</td>
<td>0.32</td>
<td>3.21</td>
</tr>
<tr>
<td>bili</td>
<td>0.78</td>
<td>0.11</td>
<td>6.67</td>
<td>0.76</td>
<td>0.12</td>
<td>6.32</td>
<td>0.75</td>
<td>0.12</td>
<td>6.23</td>
</tr>
<tr>
<td>albu</td>
<td>–2.25</td>
<td>0.83</td>
<td>–2.71</td>
<td>–3.58</td>
<td>0.74</td>
<td>–4.81</td>
<td>–3.77</td>
<td>0.73</td>
<td>–5.12</td>
</tr>
<tr>
<td>copp</td>
<td>0.002</td>
<td>0.001</td>
<td>2.36</td>
<td>0.002</td>
<td>0.001</td>
<td>2.46</td>
<td>0.003</td>
<td>0.001</td>
<td>2.65</td>
</tr>
<tr>
<td>SGOT</td>
<td>0.</td>
<td>–</td>
<td>–</td>
<td>0.</td>
<td>–</td>
<td>–</td>
<td>0.</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>thromb</td>
<td>0.</td>
<td>–</td>
<td>–</td>
<td>–3.08</td>
<td>0.62</td>
<td>–4.90</td>
<td>–2.63</td>
<td>0.59</td>
<td>–4.42</td>
</tr>
<tr>
<td>hist</td>
<td>0.33</td>
<td>0.14</td>
<td>2.28</td>
<td>0.35</td>
<td>0.15</td>
<td>2.28</td>
<td>0.41</td>
<td>0.15</td>
<td>2.67</td>
</tr>
</tbody>
</table>

5 Proofs

We first give some counting process notation used in the proofs. Let \(N(t) = I(T^{\text{cens}} \leq t, \delta = 1)\) be the single-jump counting process that registers whether an uncensored failure has occurred by time \(t\), and \(Y(t) = I(T^{\text{cens}} \geq t)\) the corresponding “at risk” indicator. Define the filtration \(\mathcal{F}_t = \mathcal{F}_0 \vee \sigma\{N(s): s \leq t\}\), where \(\mathcal{F}_0 = \sigma(X, Z)\). Under the true probability measure \(P\) on \(\mathcal{F} \equiv \mathcal{F}_\tau\), the counting process \(N(t)\) has intensity process \(\lambda(t) = Y(t)h(t, X, Z)\), which means that

\[
M(t) = N(t) - \int_0^t \lambda(s) \, ds, \quad t \in [0, \tau]
\]  

(5.16)
is an $\mathcal{F}_t$-martingale under $P$.

If the conditional hazard function changes from $h$ to $h'$ and the distribution of the covariates is unchanged, then we write the new probability measure on $\mathcal{F}$ as $P'$. The intensity of $N$ under $P'$ is then $\lambda'(t) = Y(t)h'(t, X, Z)$. If $h(t, X, Z)$ and $h'(t, X, Z)$ are bounded and bounded away from zero over $t \in [0, \tau]$ a.s., then the restrictions $P'_t$ and $P_t$ of $P'$ and $P$ to $\mathcal{F}_t$ are mutually absolutely continuous and the log-likelihood ratio is

$$
\log \frac{dP'_t}{dP_t} = \int_0^t \log \left( \frac{\lambda'(s)}{\lambda(s)} \right) \, dN(s) - \int_0^t (\lambda'(s) - \lambda(s)) \, ds,
$$

where $\log 0/0 = 0$, see Andersen et al. (1993, p. 98).

The Hellinger distance between two probability measures $P$ and $P'$ is defined by

$$
\rho^2(P, P') = \frac{1}{2} E_Q(\sqrt{V} - \sqrt{V'})^2
$$

where $Q = (P + P')/2$, $V = dP/dQ$, and $V' = dP'/dQ$. Note that $\rho^2(P, P')$ does not depend on the choice of the dominating measure $Q$. The Kullback–Liebler information number between $P$ and $P'$ is $K(P, P') = \int \log(dP/dP') \, dP$ when $P$ is absolutely continuous with respect to $P'$, otherwise $K(P, P') = \infty$.

### 5.1 Proof of Theorem 1


Lemma 1 provides the “closing argument” of Barron et al. (condition C, page 377), and gives an equivalence (up to constants) between the Kullback–Leibler information number

$$
K(P, P') = E(\gamma_n(u) - \gamma_n(r))
$$

and the $L^2$-norm between $u$ and $r$ (cf., Birgé and Massart, 1998, Section 4.2). This equivalence is established using a result of Jacod and Shiryaev (1987) which allows us to express the Hellinger distance between two counting processes in terms of their intensities.
In Lemma 2 we check condition $M$ (6.4). This amounts to checking a Lipschitz type condition on the process $u \mapsto \gamma(u)$:

$$
\gamma(u) = \gamma(u, T^{\text{obs}}, \delta, X, Z) = \int_0^\tau Y(t) \psi(u(t, X, Z)) \, dt - \int_0^\tau \log \psi(u(t, X, Z)) \, dN(t).
$$

(5.18)

Condition $M$ (6.5) holds by Lemma 9, page 400 of Barron et al. (1999, page 372). Assumption $M$ (6.6) follows by the definition of $\gamma$ in (5.18), our Conditions A and the Lipschitz property of log on $[\psi(a), \psi(b)]$. Then, for some constant $a_1 > 0$, we have

$$
\| \Delta(\cdot, u, v) \|_\infty \leq a_1 \| v - u \|_\infty.
$$

Lastly, note that Theorem 8 of Barron et al. (1999) applies for any penalty term greater than $\tilde{C}(|I| + pN_n)/n$, where $|I| + pN_n$ is the dimension of the approximating space and $\tilde{C}$ is a positive constant given by their Theorem. Notice that, since $|I| + 1, pN_n \geq 1$, then $2(|I| + 1)pN_n \geq (|I| + pN_n)$ and so $2(|I| + 1)pN_n \log n/n > \tilde{C}(|I| + pN_n)/n$, for $n$ large enough. Thus, in the definition of our penalty term, one can take $C = 2$. However, other choices are possible, and for small sample sizes it is easy to calibrate the value of $C$ via simulation, as we did in Section 3.

**Lemma 1** Suppose $u(t, x, z)$ satisfies Condition A1 in place of $r$. Let $P$ and $P'$ correspond to the conditional hazard functions $h = \psi \circ r$ and $h' = \psi \circ u$, respectively. Then there exist constants $0 < c_1 < c_2$ such that

$$
c_1 ||r - u||_2^2 \leq K(P, P') \leq c_2 ||r - u||_2^2.
$$

**Proof** First note that $P_0 = P'_0$, so $\rho^2(P_0, P'_0) = 0$. Proposition 1.27 and Theorem 4.2 of Jacod and Shiryaev (1987, p.197 and p.237), applied with $Q = P$, then give

$$
\rho^2(P, P') = \frac{1}{2} EP \int_0^\tau \sqrt{V'_s} \left( \sqrt{\lambda(s)} - \sqrt{\lambda'(s)} \right)^2 \, ds
$$

17
where \( V'_t = dP'_t/dP_t \) is given by (5.17). Using (5.17), the bounds on \( r \) and \( u \) and the fact that \( N \) has at most a single jump, it can be easily seen that \( V'_t \) is bounded and bounded away from zero by constants that only depend on \( a, b, \epsilon \) and \( \tau \). Thus, in the sense of the conclusion of the lemma,

\[
\rho^2(P, P') \simeq E_P \int_0^\tau \left( \sqrt{\lambda(s)} - \sqrt{\lambda'(s)} \right)^2 ds. \tag{5.19}
\]

By Birgé and Massart (1998, (7.5) and (7.6)) we have

\[
2\rho^2(P, P') \leq K(P, P') \leq \rho^2(P, P')(4 + 2 \log ||V||_{\infty}) \tag{5.20}
\]

where \( ||V||_{\infty} \) is the supremum norm of \( V = dP/dP' \). Combining (5.19) and (5.20) we find that

\[
K(P, P') \simeq E_P \int_0^\tau \left( \sqrt{\lambda(s)} - \sqrt{\lambda'(s)} \right)^2 ds. \tag{5.21}
\]

Using the fact that \( C \) is conditionally independent of \( T \) given the covariates \((X, Z)\), as well as the upper bound \( \psi(b) \) on \( h \) and the lower bound on \( P(C \geq t|X, Z) \) in Condition A3, we have

\[
E_P(Y(t)|X, Z) = P(C \geq t|X, Z)E_P(T \geq t|X, Z) \geq \epsilon \exp(-\tau\psi(b))
\]

almost surely, for some \( \epsilon > 0 \), so conditioning on \((X, Z)\) we find that

\[
E_P \int_0^\tau \left( \sqrt{\lambda(s)} - \sqrt{\lambda'(s)} \right)^2 ds = E_P \int_0^\tau \left( \sqrt{h(s, X, Z)} - \sqrt{\psi(u(s, X, Z))} \right)^2 Y(s) ds \\
\simeq E_P \int_0^\tau \left( \sqrt{h(t, X, Z)} - \sqrt{\psi(u(t, X, Z))} \right)^2 dt \\
\simeq E_P \int_0^\tau (\psi(r(t, X, Z)) - \psi(u(t, X, Z)))^2 dt \\
\simeq ||r - u||_p^2.
\]

The penultimate line above follows from the bounds on \( r \) and \( u \), and the Lipschitz property of the square-root function on \([\psi(a), \psi(b)]\), where here \( \psi(a) > 0 \). The last line above follows from the Lipschitz assumptions on \( \psi \) and \( \psi^{-1} \). This combined with (5.21) completes the proof.
Lemma 2  Suppose $u(t, x, z)$ and $v(t, x, z)$ satisfy Condition A1 in place of $r$. Then there exists a constant $c_3 > 0$ only depending on $a$, $b$ and $\tau$ such that

$$E_P(\gamma(u) - \gamma(v))^2 \leq c_3||u - v||^2_v.$$ 

Proof  First note that, in terms of the martingale (5.16), we have

$$\gamma(u) - \gamma(v) = \int_0^\tau Y(t) [\psi(u(t, X, Z)) - \psi(v(t, X, Z))] \ dt - \int_0^\tau \log \frac{\psi(u(t, X, Z))}{\psi(v(t, X, Z))} \ dM(t)$$

$$- \int_0^\tau \log \frac{\psi(u(t, X, Z))}{\psi(v(t, X, Z))} Y(t) h(t, X, Z) \ dt.$$ (5.22)

Consider the first term on the r.h.s. above. By the Cauchy–Schwarz inequality applied to the integral over $t$, we have

$$E_P \left( \int_0^\tau Y(t) [\psi(u(t, X, Z)) - \psi(v(t, X, Z))] \ dt \right)^2 \leq \tau||\psi \circ u - \psi \circ v||^2_v.$$ 

Using standard results on martingale integrals (see, e.g., Andersen et al. 1993, p.78), the second term on the rhs in (5.22) has second moment

$$E_P \left( \int_0^\tau \log \frac{\psi(u(t, X, Z))}{\psi(v(t, X, Z))} \ dM(t) \right)^2 = E_P \int_0^\tau \left( \log \frac{\psi(u(t, X, Z))}{\psi(v(t, X, Z))} \right)^2 Y(t) h(t, X, Z) \ dt$$

$$\leq c_4||\psi \circ u - \psi \circ v||^2_v$$

for some constant $c_4 > 0$, where the last inequality uses the bounds on $u, v, h$, and the Lipschitz property of $\log$ on $[\psi(a), \psi(b)]$. The second moment of the last term in (5.22) can be handled in a similar way to the first. The result now follows by applying the inequality

$$(x + y + z)^2 \leq 3(x^2 + y^2 + z^2)$$

to the square of the rhs of (5.22), using the second moment bounds already established and the Lipschitz assumption on $\psi$. ■

5.2  Proof of Corollary 1

We begin by giving some properties of the orthogonal projection operators that we consider.
Lemma 3 Let \( r_I \equiv \pi(r) \) denote the orthogonal \( L^2 \)-projection of \( r \) onto \( S_I \). Let \( f_{j,n} = \pi_n(f_j) \), where \( \pi_n \) denotes the \( L^2 \) projection onto \( \langle \phi_{n,i} : i = 1, \ldots, N_n \rangle, \ j = 1, \ldots, p \).

(i) \( r_I(t, x, z) = \beta^T x + \sum_{j=1}^p z_j f_{j,n}(t) \), for any \( I \supseteq I_0 \).

(ii) Under A1, \( r_I \in S_I \), for any \( I \).

Proof (i) By the linearity of \( \pi \) and since \( \beta^T x \in S_I \), for any \( I \supseteq I_0 \), \( r_I(t, x, z) = \beta^T x + \sum_{j=1}^p \pi(z_j f_j(t)) \). Hence, we need to show that \( \pi(z_j f_j(t)) = z_j(\pi_n f_j)(t) \).

Recall from the uniqueness of the (Riesz) orthogonal decomposition that for any \( g \in L^2 \), \( \pi(g) \) is the unique element in \( S_I \) such that \( g = \pi(g) \) is orthogonal to \( S_I \) (note that \( S_I \) is finite dimensional and thus closed).

By applying the above property to \( g(t, x, z) = z_j f_j(t) \) we see that it suffices to show that the \( L^2 \)-inner product between \( z_j f_j(t) - z_j(\pi_n f_j)(t) \) and each generating function in \( S_I \) is zero. But the generator of the form \( z_k \phi_{n,j}(t) \) has \( L^2 \)-inner product with \( z_j f_j(t) - z_j(\pi_n f_j)(t) \) given by (where we separate the variables using Fubini’s theorem)

\[
E_p(Z_j Z_k) \times \int_0^\tau (f_j(t) - (\pi_n f_j)(t)) \phi_{n,j}(t) \, dt.
\]

Notice that the second factor above is the \( L^2 \) inner product, so it vanishes by the orthogonal decomposition for the projection \( \pi_n \). The same argument works for the generators of the form \( x_k \).

(ii) Notice that by the argument used in (i) and by regarding \( r = \beta^T x + f^T z \) as a function of \( t \) with \( x \) and \( z \) fixed, we have

\[
\pi(r) = \beta^T x + \pi_n(f)^T z = \pi_n(\beta^T x + f^T z) = \pi_n(r).
\]

Then, since the projection operator \( \pi_n \) acts onto histogram sieves and, by A1, \( a \leq r \leq b \), we have that \( a \leq \pi_n(r) \leq b \), which completes the proof of this Lemma. \( \blacksquare \)
Proof of Corollary 1. First note that for $f \in D_\alpha$, the bound on the approximation error given by Theorem 2.4, page 358 in DeVore and Lorentz (1993) is $\|f_f - f_{j,n}\|_2^2 \leq B(\alpha)n^{-2\alpha/(2\alpha+1)}$, for a constant $B(\alpha) > 0$ given by their theorem. By Lemma 3, for any $I \supseteq I_0$, the $L^2_\nu$-projection of $r$ onto $S_I$ is $r_I(t, x, z) = \beta^T x + \sum_{j=1}^{p} z_j f_{j,n}(t)$ and $r_I \in S_I$. Then

$$\|r - r_I\|_\nu^2 = \|\sum_{j=1}^{p} z_j (f_j - f_{j,n})\|_\nu^2 \leq p M^2 \sum_{j=1}^{p} \|f_j - f_{j,n}\|_2^2 \leq p^2 M^2 B(\alpha)n^{-2\alpha/(2\alpha+1)} \equiv B_1 n^{-2\alpha/(2\alpha+1)},$$

where $M$ is a uniform bound on the absolute value of the components of $Z$.

Hence, by Theorem 1, for a dominating constant $C^* > 0$ (depending on $\alpha, B, a, b, \epsilon, \tau, M, p$), we have

$$E_P \|r - \hat{r}\|_\nu^2 \leq C_1 \inf_{I \supseteq I_0} (\|r - r_I\|_\nu^2 + \text{pen}(I) + C_2/n) \leq C_1 \left( \frac{B_1}{n^{2\alpha/(2\alpha+1)}} + \frac{C(|I_0|+1)p\log n}{n^{2\alpha/(2\alpha+1)}} \right) \leq C^* \frac{\log n}{n^{2\alpha/(2\alpha+1)}}. \quad (5.23)$$

Let now $(\tilde{X}, \tilde{Z}) \sim \mu_{X,Z}$, with $(\tilde{X}, \tilde{Z})$ independent of $(X_1, Z_1), \ldots, (X_n, Z_n)$ and notice that $\tilde{X} - E(\tilde{X}|\tilde{Z}) \perp_{\mu_{X,Z}} \tilde{Z}$. Then, writing $E_\nu$ for integration w.r.t. $\nu$, by Pythagoras

$$\|r - \hat{r}\|_\nu^2 = E_\nu \|r - \hat{r}\|^2 = E_\nu \|f - \hat{f}\|^2 + (\hat{\beta} - \beta)^T E_\nu (\tilde{X}|\tilde{Z}) + (\hat{\beta} - \beta)^T (\tilde{X} - E_\nu (\tilde{X}|\tilde{Z}))^2 \leq E_\nu \|f - \hat{f}\|^2 + (\hat{\beta} - \beta)^T E_\nu (\tilde{X}|\tilde{Z})^2 + (\hat{\beta} - \beta)^T (\tilde{X} - E_\nu (\tilde{X}|\tilde{Z}))^2.$$

Since $\|r - \hat{r}\|_\nu^2 = O_P \left( \frac{\log n}{n^{2\alpha/(2\alpha+1)}} \right)$ by (5.23), we have

$$E_\nu [(\hat{\beta} - \beta)^T (\tilde{X} - E_\nu (\tilde{X}|\tilde{Z}))^2] = O_P \left( \frac{\log n}{n^{2\alpha/(2\alpha+1)}} \right). \quad (5.24)$$

Define $\Sigma = (\sigma_{ij})_{q \times q}$, with

$$\sigma_{ij} = \text{Cov}(X_i - \theta_i(Z), X_j - \theta_j(Z)) = \text{Cov}(X_i, X_j) - \text{Cov}(\theta_i(Z), \theta_j(Z)), \quad (5.25)$$

21
for $\theta_i(z) \equiv E_{\nu}(X_i|Z = z)$. Since $(\tilde{X}, \tilde{Z})$ is independent of $\hat{\beta}$ by construction, we also have

$$E_{\nu}[(\hat{\beta} - \beta)^T (\tilde{X} - E(\tilde{X}|\tilde{Z}))] = (\hat{\beta} - \beta)^T \Sigma(\hat{\beta} - \beta) \geq \lambda_{\min}|\hat{\beta} - \beta|^2, \quad (5.26)$$

where we denoted by $\lambda_{\min}$ the smallest eigenvalue of $\Sigma$. Since, by Condition A4, for any nonzero $l \in \mathbb{R}^q$

$$l'\Sigma l = \text{Var}(l'(X - E(X|Z))) = \int \text{Var}(l'X|z) d\mu_Z(z) > 0, \quad (5.27)$$

$\Sigma$ is positive definite and so $\lambda_{\min} > 0$. Thus, from (5.24) and (5.26), we have that for any $\beta \in \mathcal{K}$

$$|\hat{\beta} - \beta|^2 = O_p \left( \frac{\log n}{n^{2\alpha/(2\alpha + 1)}} \right). \quad (5.28)$$

In a similar fashion, observing that $\tilde{Z} - E(\tilde{Z}|\tilde{X}) \perp_{X,z} \tilde{X}$, we obtain now

$$\|r - \hat{r}\|^2_{\nu} = E_{\nu}[(\beta - \hat{\beta})^T \tilde{X} + (\hat{f} - f)^T E(\tilde{Z}|\tilde{X})]^2$$

$$+ E_{\nu}[(\hat{f} - f)^T (\tilde{Z} - E(\tilde{Z}|\tilde{X}))]^2,$$

hence

$$E_{\nu}[(\hat{f} - f)^T (\tilde{Z} - E(\tilde{Z}|\tilde{X}))]^2 = O_p \left( \frac{\log n}{n^{2\alpha/(2\alpha + 1)}} \right).$$

Define now $V = (v_{ij})_{p \times p}$, with

$$v_{ij} = \text{Cov}(Z_i - \eta_i(X), Z_j - \eta_j(X)) = \text{Cov}(Z_i, Z_j) - \text{Cov}(\eta_i(X), \eta_j(X)), \quad (5.29)$$

for $\eta_i(x) \equiv E(Z_i|X = x)$. As before, under Condition A5 this time, $V$ is positive definite, so

$$E_{\nu}\|\hat{f} - f\|^2_2 = O_p \left( \frac{\log n}{n^{2\alpha/(2\alpha + 1)}} \right),$$

which implies that

$$\|\hat{f}_j - f_j\|^2_2 = O_p \left( \frac{\log n}{n^{2\alpha/(2\alpha + 1)}} \right),$$

for $j = 1, \ldots, p$. ■
References


