Estimation of Hazard Models with Dependence Across Observations: Correlation, Frailty and Contagion.

James Wolter*
Yale University
(Job Market Paper)

November 21, 2011

Abstract

This paper proposes nonparametric and semi-nonparametric estimation of hazard models with various types of dependence between observations. The methods are designed to examine the impact that different macroeconomic and financial conditions have on hazard rates. First, dependence between time-varying covariate processes across observations is examined. Among other possibilities, covariate processes that are common to all observations are permitted. This is motivated by situations where macroeconomic variables such as the interest or unemployment rate affect all hazard rates. Second, I examine a global latent risk factor which increases clustering of defaults. This unobserved risk factor is referred to as frailty. Finally, defaults of certain observations are allowed to directly affect the hazard rate of related observations. This phenomenon is given the general name contagion. The martingale nature of default is preserved in the presence of these types of dependence. Martingale CLT and FCLT results are derived. Two types of estimation are presented, both of which are based on the derived martingale results. First, a kernel approach is taken. Asymptotic results are derived while accounting for dependence between processes using mixing conditions. Second, a point process likelihood approach is taken. Sieve estimation is possible in the presence of dependent process, contagion and frailty. Again, mixing conditions are assumed. The path of the unobserved frailty component and the parametric impact of covariates are consistently estimated. The estimate of the frailty path is then used to estimate the underlying stochastic process the unobserved risk factor follows.

1 Introduction

The hazard rates of random economic events frequently depend on macroeconomic conditions. Just a few examples include mortgage default, corporate default, retirement, investment decisions and labor market decisions. If the macroeconomy impacts hazard rates, then dependence across observations is a fundamental aspect of the analysis. If there were independence, the distribution of observations at risk

*I would like to thank Don Andrews, Phil Haile, Costas Meghir, Ed Vytlacil and participants at the econometric seminars at Yale for helpful comments. I would especially like to thank Xiaohong Chen and Peter Phillips for a heroic amount of advice, guidance, comments and support. Finally, I would like to thank Yale University and the Cowles foundation for financial support. Part of this paper was written while the author was a Carl A. Anderson fellow. All remaining errors are my own.

of default[1] during a recession would be the same as those at risk during an expansionary period. When global economic conditions matter, this cannot be the case.

One approach to modeling the dependence of default on macroeconomic conditions is to include macro variables as covariates in hazard analysis. From an estimation standpoint, these variables cause problems because they are common to all observations observed at a particular calendar time. As a result, observations cannot be *i.i.d.* When estimating the effect of macro variables on the hazard rate, the dependence that results from the use of covariates common to all observations must be incorporated into the statistical analysis.[2] Dependence of defaults may also be driven by correlation between observation-specific covariates. These types of variables often have dependence with macro covariates. Economic hazard models should account for all of these potential correlations.

In unemployment duration analysis, models often assume that global aspects of the economy affect hazard rates. The level of unemployment insurance benefits impacts all observations within a state or country. Government support for the unemployed clearly affects the hazard rate of becoming employed. In the United States, there are interactions between state and federal unemployment benefits. This causes dependence between support levels across states. Statewide unemployment rates are sometimes used as covariates. These variables are identical for all observations within a state at a fixed calendar time. This forces dependence across hazard rates. Other macro variables which impact unemployment durations are the interest rate and GDP growth rate.

Another example with correlation between relevant covariates is mortgage default. Over 75 percent of subprime mortgages originated in the US between 2003 and 2007 have an adjustable interest rate component (see Mayer, Pence and Sherlund (2009)). Interest rates paid on these mortgages are tied to other interest rates or indexes, such as the 3-month US treasury rate or LIBOR. As a result, payment on these mortgages are determined by observable macroeconomic covariates. This causes dependence among mortgage defaults. Other variables which impact mortgage defaults are unemployment rates and housing values. These covariates are likely correlated geographically.

The considerations given above suggest a more general proposition: random economic events are correlated and the econometric analysis of their hazard rates must account for this correlation. There is surprisingly little statistical analysis of hazard rates where dependence between observations is considered. One specification which has received attention is cluster analysis. In these situation, covariates are allowed to be dependent within groups while the groups are independent. See Martinussen and Scheike (2010), Hougaard (2000) or Aalen, Borgan and Gjessing (2010) for an overview of these types of methods. This assumption is not sufficient for our purposes. Hazard models using common macroeconomic variables imply dependence between all observations. In economic situations, there may be no natural way to place observations into groups with statistically independent covariates across groups.

In some situations, the available data will not be enough to explain the character of the observed defaults. For example, in corporate default modeling, defaults may be clustered around recessions or financial crashes. If the clustering is too severe, hazard models may capture the realized data poorly. Some statistical support for the failure of standard hazard models in this context is given in Das, Duffie, Kapadia and Saita (2007). In order to account for this potential time dependent model misspecification,

[1] Throughout the paper, we will refer to random times which have corresponding hazard rates as "default".

[2] Throught the paper, we will refer to any global variables (such as macroeconomic variables) as common variables.

Duffie, Eckner, Horel and Saita (2009) introduce a global latent risk factor into a standard hazard model. This risk factor is similar to an unobserved macroeconomic covariate. As in Duffie et al. (2009), we call this latent variable frailty. At time periods where frailty is high, there is clustering in default beyond what can be explained by the observed covariates. When frailty is low, defaults are suppressed. The additional time varying dependence improves model fit. Frailty has received attention elsewhere (Azizpour, Giesecke and Schwenkler (2011), Koopman, Lucas and Schwaab (2011)). As it is unlikely available data will ever contain all elements that are relevant to hazard rates, this model extension is relevant for many applications.

The path of the frailty variable can be considered a given deterministic function of time. However, without any additional structure on the latent risk factor, forecasting will be ruled out. This is because, with no assumption on how frailty will propagate in the future, we have no probabilistic description of what will happen next. Duffie et al. (2009) proposes assuming that frailty follows a diffusion. If the underlying diffusion can be characterized, frailty can be used to improve default predictions.

While defaults may be correlated as a result of dependence in relevant variables (both observed and unobserved), they may also be correlated as a result of other defaults themselves. The survival of a business will depend on many economic variables, but it will also depend on the existence of rivals. If a competitor fails, the probability of survivors failing will be affected directly by that default. Another prominent example is credit default. Models often assume if one contract defaults, the hazard rate for remaining contracts increases. Other examples of this phenomenon potentially include the adoption of new technology or the decision to retire, among others. I will refer to this phenomenon as contagion. A rigorous definition is given in the sequel. This notion is well developed in the financial default literature. See Giesecke and Kim (2010), Collin-Dufresne, Goldstein and Helwege (2009), Jarrow and Yu (2001) and Davis and Lo (2001).

In this paper I investigate the econometric analysis of continuous-time hazard models whose observations are dependent in the ways outlined above. The single-spell situation is considered throughout. The estimation approaches will generalize to the multiple-spell case. I focus on nonparametric and semi-nonparametric estimation.

Nonparametric kernel estimation of a hazard model with single-spell data and time-varying covariates is achieved in Nielsen and Linton (1995) and Linton, Nielsen and Van de Geer (2003) (hereafter NL (1995) and LNV (2003)). These papers are leading examples in the kernel approach to hazard estimation. However, they assume that covariate stochastic processes across observations are independent (although these processes may be dependent within observations). Therefore, hazard estimation with common covariates is not justified by these papers.

I extend the results of NL (1995) and LNV (2003) to the case where covariate processes may be dependent across observations. In particular, I assume specific α-mixing conditions outlined below. α-mixing is chosen for concreteness and the ideas can be applied to other mixing conditions. This dependence framework allows for global variables such as the S&P 500 or interest rates to affect hazard rates. Variables specific to observations are also possible.

In order to estimate the hazard function in this situation, a cross section of observations is not sufficient. If all observations are observed over the same calendar time interval, then each observation is impacted by the same portion of the common processes. The impact of macro variables can not be consistently estimated in this situation because there is effectively no sampling of these variables. Defaults and the covariates determining their hazard rates must be observed over periods with different realizations of the common time series. When macro variables are of interest, this corresponds to observing defaults over periods with different macroeconomic conditions. The proposed sampling scheme assumes that observations potentially default over a fixed time interval which can be interpreted as the life of a contract. This assumption may also be interpreted as censoring all observations with durations longer than a fixed time length. By sampling from observations at risk of default over different blocks of calendar time, we effectively sample from the common process. From this sampling scheme we can asymptotically recover hazard functions dependent on common variables. The fixed time length assumption may be unreasonable in some situations. This is removed at the expense of stronger assumptions elsewhere.

The results for nonparametric kernel estimation include a derivation of asymptotic bias and confidence bands for an estimate of the hazard function. A uniform rate of convergence is presented. Finally, by assuming a multiplicative or additive structure on the hazard function, the curse of dimensionality is circumvented. In these results, we do not incorporate contagion or frailty. These dependence structures are explored with a second estimator.

A second set of results in this paper extend the point process likelihood estimation approach presented in Karr (1987). In this portion of the paper, we assume a Cox proportional hazard specification. Karr (1987) examines sieve estimation of the baseline hazard in a proportional hazard model. In that work, the observations are assumed *i.i.d.* and the covariates' coefficients are assumed known.

Frailty is a global unobserved risk factor which can be thought of as an unobserved macro covariate. Frailty is defined as an unobserved, strictly positive function of time. In a proportional hazard specification, frailty takes the place of the baseline hazard in the model. When observations are at risk over different calendar time intervals their corresponding baseline hazards will be different. Their baseline hazard will correspond to the portion of the frailty function coincident with the calendar time. When the frailty path is above one, there is additional default above that explained by the covariates. When the frailty path is below one, the opposite is true. Duffie et al. (2009) adds additional structure to the situation by assuming the frailty path is determined by the realization of a mean-reverting diffusion process. The idea is that unobserved or difficult to quantify common risk factors are driving default clustering. These risk factors are mean reverting to a baseline level of risk.

In Duffie et al. (2009), a Bayesian approach is taken in estimating the diffusion. The consistency result in that paper relies on an *i.i.d.* assumption which is violated in situations that are the focus of the current paper. Here, I derive consistent estimates using a point process likelihood approach. In a first step, the realized path of the frailty diffusion is estimated using sieves. Once the diffusion path is estimated, statistical methods which incorporate continuous observations of a diffusion can be used to estimate the underlying frailty dynamics. This estimation of the frailty process is done in the presence of dependent covariates.

When common macro variables are present, sampling a single cross section over a block of calendar time will not lead to consistent estimation. This is for reasons discussed above - there is effectively no sampling of the common processes in this case. All that is observed are the common processes over a single block of time. Consistent estimation of the coefficients for the common processes is a prerequisite for recovering the frailty path. Therefore, cross sectional sampling does not recover the corresponding portion of the frailty path. We must observe common processes over an increasingly large number of calendar time intervals in order to estimate the coefficients of these covariates consistently. In order to estimate the frailty function, the number of observations over each calendar time interval must also approach infinity. An appropriate sampling scheme for this situation incorporates both these requirements.

Assuming such a sampling scheme, I derive conditions under which the covariate coefficients and the portion of the frailty path coincident with each block of calendar time can be estimated consistently. This produces an estimate of the frailty path. In addition, the current value of frailty is a by-product of estimation. This is important for forecasting applications.

In a final model, a notion of contagion is allowed for in a likelihood estimation approach. Contagion is present when observations have covariate stochastic processes which depend on the default of other observations. Contagion is assumed to be confined within groups. Defaults of group members can only affect the covariates of other group members. The specification also allows for covariates that do not depend on other observations defaulting. These non-contagious covariates are not restricted to be dependent only within groups. As a result, there is still dependence in defaults across groups. This setup allows for a simple statement of the needed mixing conditions. When contagious effects are not confined within groups the needed mixing conditions become more complicated and difficult to verify.

When contagion is present, I assume cross sectional sampling of observations that are at risk of default over a fixed time interval $[0, T]$. This sampling scheme is chosen for simplicity and rules out common macroeconomic processes as covariates. The estimation in Karr (1987) is extended so the covariate coefficients are estimated instead of assumed known. This is done in the presence of dependence across covariates and contagion. In this result, consistent estimates of the baseline hazard and covariate coefficients are obtained. Widely used results from Anderson and Gill (1982) are extended to this dependence situation as well. From these results, we are able to derive confidence bands for the covariate coefficients. Similar estimation can be conducted with common processes if frailty is not present. The introduction of frailty rules out standard estimation approaches like Anderson and Gill (1982).

The remainder of the paper is organized as follows. Section 2 presents the model and derives martingale results required for estimation. Section 3 presents extensions of the kernel methods of NL (1995) and LNV (2003) to the dependent case. Section 4 presents sieve estimation using a point process likelihood approach. Contagion and frailty are defined and incorporated into the analysis. Several other related results are presented in this section. Section 5 concludes. Some results and proofs are presented in the appendices.

2 Models and Martingale Preliminaries

This section describes how the random times used in this paper are constructed. Special attention is given to incorporating common covariates across observations. In addition, dependence between observation-specific variables is incorporated. Our focus will be on deriving the martingale nature of random times with the appropriate dependence. A martingale CLT and FCLT are derived. These are preliminary results needed to facilitate estimation in later sections. Throughout this section, the example of credit default will be used to motivate the specification. The results are general, but keeping an example in mind makes the presentation easier to follow.

I do not allow for frailty or contagion here. The same type of martingale structure holds in a similar specification with these forms of dependence. These notions of dependence are rigorously defined in Section 4. Estimation with these model elements is postponed until that section.

First, I present the covariate processes which are assigned to each observation. Then, I show how those processes are used to construct the random times at which observations default. The hazard function of these random times will arise naturally in the construction. The hazard model given here has the same distributional properties as almost all hazard models in the literature. However, the exact construction is invaluable when dealing with mixing conditions, as we do throughout the paper.

The main sampling scheme considered will be referred to as block/step sampling. It has three important elements for each observation: a set of covariate stochastic processes $X^i(t)$ which are specific to the observation i; a set of covariate stochastic processes which are common $Y(t)$; and a calendar time G^i at which observation i is "born" or becomes at risk of default. More specifics are given below. Related sampling schemes are also discussed.

2.1 Block/Step Sampling

We index each observation by $i \in \mathbb{N}_0$. In addition, each observation has a deterministic constant G^i which corresponds to the calendar time at which it begins to be at risk of default. Each observation is at risk of default over a fixed time interval of length T. However, the period over which they are at risk corresponds to the calendar time interval $\left[G^i, G^i + T\right]$. These calendar time intervals are allowed to overlap. Throughout, G^i are assumed to be deterministic.

The time interval $\left[G^i, G^i + T\right]$ may be contractually specified, such as the duration of a loan. The default situation may also have a natural time interval. For example, a model for school dropout. Another possibility is that observations are censored after a specified time interval. In Section 3.2 below we relax this assumption at the expense of stronger assumptions elsewhere.

Let $\left\{ X^i(t) \middle| t \in [0,T] \right\}$ be d covariate stochastic processes specific to each observation. $X^i(t)$ for $t \in [0,T]$ corresponds to the value of these covariates over the calendar time interval $\left[G^i, G^i + T\right]$. Defining $X^i(t)$ on $[0,T]$ instead of $\left[G^i, G^i + T\right]$ is done for notational simplicity. We make the important additional assumption on $X^i(t)$ that its paths are left-continuous with right-hand-limits (càglàd for short)[3]. Assume the distribution of the variables $X^i(t)$ has support equal to the compact set $\mathcal{X} = \mathcal{X}_1 \times \cdots \times \mathcal{X}_d$ for each $t \in [0,T]$. I emphasize that we are not making any stationarity assumption on $X^i(s)$ within observations. We only assume the support of the covariates at each time is the same and rectangular. The length of time the observation has been at risk of default t will also be a variable in the sequel.

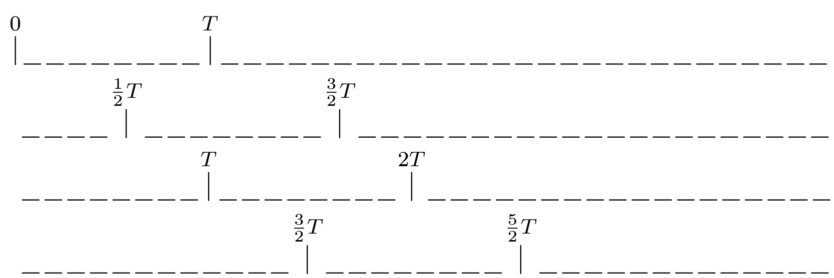
In addition, there is another set of j covariate processes $Y(t)$. $Y(t)$ is assumed to be stationary with càglàd paths and to have compact support $\mathcal{Y} = \mathcal{Y}_1 \times \cdots \times \mathcal{Y}_j$ for each $t \in [0,\infty)$. For simplicity, the support of all covariate processes can be thought of as $[0,T] \times [0,1]^{d+j}$. The covariate processes $Y(t)$ are common to all observations in that, for each observation i, the portion of $Y(t)$ corresponding to the calendar time the observation is at risk $\left[G^i, G^i + T\right]$ affects the hazard rate for that observation. The process $Y(t)$ correspond to macro or other global variables such as the S&P 500 or unemployment rate.

[3]Left continuity implies that the processes $X^i(t)$ and processes based on $X^i(t)$ used below are predictable, an important technical property for our results. See Jacod and Shiryaev (2003) for a definition of predictability and discussion of its importance.

The assumption that all processes have the same support through time could be relaxed. This would be necessary for including variables such as total work experience in an employment hazard model. However, the support of the hazard function would have to become irregular and this would complicate notation. We focus on the rectangular case in the sequel. In section 4, we allow our covariate processes to have discrete supports or supports that change through time. The required assumption is that the covariates have support *contained* in a compact rectangle for all $t \in [0, T]$.

To sum up, the relevant covariates for observation i are $\left\{ Z^i(t) = \left(t, X^i(t), Y\left(G^i + t\right)\right) \middle| t \in [0, T] \right\}$. Note that the entire analysis below can be conducted under the assumption that there are no common $Y(t)$ processes. Dependence between the covariate processes $X^i(t)$ will be assumed. Therefore, even in this simpler situation, the results that follow are an extension of the existing literature.

Although the martingale limit theorem we present below holds more generally, we focus on a particular sampling structure. I will refer to this sampling structure as "block/step sampling". Assume that $G^i = c_i \delta$ where $c_i \in \mathbb{N}$, $c_i \leq c_{i+1}$ and $k\delta = T$ for some $k \in \mathbb{N}$. The sampling corresponds to the following diagram.



Here, $\delta = \frac{T}{2}$. The calendar times where observations start are multiples of δ. δ is the "step". All observations starting at the same time form a "block". There can be any finite number of observations in each block, including zero. The blocks are of equal length $[0, T]$ and are a distance $\delta > 0$ apart, where $\delta < T$ is possible. For asymptotic results to hold, we must assume $c_i \to \infty$ as $i \to \infty$.

Block/step sampling corresponds to a number of relevant economic situations. The previously given examples of unemployment duration, mortgage default and corporate default all fit into this sampling scheme. This is because, in these examples, different observations begin to be a risk of default starting at different calendar times. People become unemployed or sign a mortgage at different calendar times and corporations take on debt at different calendar times. We gather more data as more relevant economic relationships are started. As we get more observations, this happens at increasingly larger calendar times. A large number of other economic situations fall into this setup. Because $c_{i+1} - c_i$ can take on any positive integer, the sampling allows for flexibility in the spacing of observations. Observations can start at any integer multiple of δ. If δ is taken to be small, most irregularly spaced sampling situations can be captured by block/step sampling.

Another situation constitutes a second standard form of sampling that is examined below. This is a standard cross section with no block/step structure. In this sampling, we consider a cross section observed over a time interval $[0, T]$. This corresponds to the situation where $G^i = 0$ for all i. Here, there are no common processes $Y(t)$. We assume observations are separated into groups of finite size. Groups are comprised of successive members of the sample ordering i. The size of the groups need not be equal. Each group has its own set of m covariate processes $R^l(t)$. Here, l indexes the group number. Assume the distribution of the variables $R^l(t)$ has compact support $\mathcal{R} = \mathcal{R}_1 \times \cdots \times \mathcal{R}_m$ for each $t \in [0, T]$. The covariates $R^l(t)$ are allowed to be dependent across groups. In the sequel, we will abuse notation and write $R^i(t)$ for the group-specific covariates corresponding to observation i's group.

This set up could be used, for example, in a spacial model where geographically specific covariates are observed. Asymptotics would be interpreted as holding when the number of specific geographical locations goes to infinity. It is simple to extend block/step sampling to incorporate block specific covariate processes $R^i(t)$. $R^i(t)$ could be used, for example, to account for covariates specific to cohorts. In this extension, $R^i(0)$ corresponds to the value of these covariates at calendar time G^i. The variables $R^i(t)$ are defined on $[0, T]$ instead of $\left[G^i, G^i + T\right]$ for notational simplicity, similarly to $X^i(t)$ described above. For the most part, the variables $R^i(t)$ are left out of the specification to keep notation manageable.

2.2 Construction and Martingale Structure

Our construction of the random times follows Bielecki and Rutkowski (2004) Example 9.1.5. We present results for the general case of block/step sampling with block specific covariates. In this situation, the relevant covariates for observation i are $\left(t, X_t^i, R_t^i, Y_{t+G^i}\right)$. Recall that we have defined X_t^i and R_t^i on $[0, T]$ instead of $\left[G^i, G^i + T\right]$ for notational simplicity. It is trivial to remove any of these groups of processes and deal with a simpler model. In what follows, $\alpha(\cdot)$ will be the hazard function in the construction, with covariate stochastic processes taken as arguments. We make the following assumption

(A1): $\alpha : [0, T] \times \mathcal{X} \times \mathcal{R} \times \mathcal{Y} \to \mathbb{R}$ is a continuous function such that

$$\begin{aligned} \inf_{(t,x,r,y) \in [0,T] \times \mathcal{X} \times \mathcal{R} \times \mathcal{Y}} \alpha(t, x, r, y) &= \underline{C} > 0, \\ \sup_{(t,x,r,y) \in [0,T] \times \mathcal{X} \times \mathcal{R} \times \mathcal{Y}} \alpha(t, x, r, y) &= \overline{C} < \infty. \end{aligned}$$

for all $i \in \mathbb{N}_0$.

We assume A1 throughout the paper. Random times τ_i are defined as

$$\begin{aligned} \Gamma_t^i &\equiv \int_0^t \alpha\left(s, X_s^i, R_s^i, Y_{s+G^i}\right) ds \\ \tau_i &\equiv \inf\left\{ t \in \mathbb{R}_+ \middle| \Gamma_t^i \geq \eta_i \right\} \end{aligned}$$

where η_i is an independent standard exponentially distributed random variable. The η_i variables are independent of all covariates and each other. Notice that the portion of the process $Y(t)$ corresponding to the calendar time the observation is at risk is used in the definition of Γ_t^i. This is also true for $\left(X_t^i, R_t^i\right)$ because their time intervals are adjusted from $\left[G^i, G^i + T\right]$ to $[0, T]$. In this model, the function $\alpha(\cdot)$ is the hazard function. Specifically:

$$\alpha\left(t, X_t^i, R_t^i, Y_{t+G^i}\right) = \lim_{\Delta \to 0} \frac{1}{\Delta} P\left\{ t \leq \tau_i < t + \Delta \middle| \tau_i \geq t \right\}.$$

The distributions of defaults in this model are equivalent to those in most hazard models encountered in the literature, including all *i.i.d.* cases. A notable exception is models where dependence between observations is derived from copulas. One approach is to put a copula on the η_i (see Cherubini et al. (2004)). Here, this is ruled out by assuming η_i are independent.

We observe the covariates $\left(t, X_t^i, R_t^i, Y_{t+G^i}\right)$ and the random times τ_i. Conditionally independent censoring can be added to the model. This type of censoring is independent of the random default variable, conditional on the covariates. This is assumed in many papers which exploit the martingale structure of default in estimation. Most stochastic processes shown to be martingales in the sequel retain this martingale structure in the presence of conditionally independent censoring. See Martinussen and Scheike (2010) or Anderson et al. (1994) for an overview of this type of censoring. We do not include censoring in what follows for notational simplicity. All results in the sequel follow with conditionally independent censoring unless specifically noted otherwise.

We need a little more notation:

$$\Lambda_t^i = \int_0^t \alpha\left(s, X_s^i, R_s^i, Y_{s+G^i}\right) \mathbf{1}_{\{\tau_i \geq s\}} ds,$$

$$M_t^i = \mathbf{1}_{\{\tau_i \leq t\}} - \Lambda_t^i.$$

Many of the following asymptotic results depend on M_t^i being continuous-time martingales. The needed martingale structure is verified below. Λ_t^i is a "compensator". It is strictly increasing and $-\Lambda_t^i$ subtracts off just enough to make M_t^i a mean-zero martingale under certain technical conditions. The relationship between M_t^i, $\mathbf{1}_{\{\tau_i \leq t\}}$ and Λ_t^i is known as the Doob-Meyer decomposition. A brief account of the martingale results needed for this paper are presented in Appendix D. In the notation from Appendix D, $\Lambda_t^i = \left\langle M^i, M^i \right\rangle_t$. See Fleming and Harrington (1991) for an excellent and very thorough account of the martingale theory of counting processes used in this paper.

Those readers not interested in specifics about the required martingale structure of M_t^i and related asymptotic results should skip the rest of this section, perhaps taking a quick look at Lemma 1 and Proposition 3. We define the following sequence of filtrations, which correspond to observations $i = 1, \ldots, n$.

$$\mathcal{F}^n = \sigma\left\{ X^i(u), R^i(v), Y(s) \middle| 0 \leq u, v \leq T; s \in [0, \infty), i = 1, \ldots, n \right\},$$

$$\mathcal{H}_t^i = \sigma\left\{ \mathbf{1}_{\{\tau_i \leq u\}} \middle| 0 \leq u \leq t \right\},$$

$$\mathcal{G}_t^n = \mathcal{F}^n \vee \mathcal{H}_t^1 \vee \cdots \vee \mathcal{H}_t^n.$$

Notice also that the processes Γ_t^i are adapted to the filtration $\mathcal{G}_t^n$.[4]

These constructs are well known in the literature on point processes. However, what is different about this set up is the temporal adjustment. Although the covariates relevant for observation i come from its corresponding calendar time interval $\left[G^i, G^i + T\right]$, we shift these variables through time and define M_t^i on the interval $[0, T]$. This is done for all observations. Because of the sampling structure, different M_t^i are derived from different portions of the common processes $Y(t)$. The result is that portions of the $Y(t)$ processes corresponding to different calendar time intervals are put "on top of each other" in the interval $[0, T]$. Depending on the filtration chosen, this may cause problems. It is not obvious that the needed

[4]For all $t > T$, we define $\Gamma_t^i \equiv \Gamma_T^i$ and $\mathcal{G}_t^n \equiv \mathcal{G}_T^n$. All other processes will be defined for $t > T$ with the same convention unless otherwise stated.

martingale structure will hold. Our choice of filtration will avoid these problems and a certain technical issue discussed below.

A martingale structure for M_t^i is a required preliminary result for many statistics regarding point processes. We establish this property below. In addition, the "usual conditions" on the underlying filtration are necessary for many results related to martingales.[5]. We establish the usual conditions for the completion of the filtration $\mathcal{G}_t^n$ for any n. This is a nontrivial property, particularly in the block/step sampling situation where temporal adjustment is required.[6] The usual conditions are a requirement for the FCLT result we derive below. This FCLT is used to discuss specification testing in Section 4.

What is critical about the following lemma is that, for each n, all $M_t^i = \mathbf{1}_{\{\tau_i \leq t\}} - \Lambda_t^i$, $i = 1, \ldots, n$ are martingales with respect to the same filtration. This happens despite the fact that different blocks of $Y(t)$ are "on top of each other" in the block/step case.

Lemma 1 (1) *The completion of the filtration $\mathcal{G}_t^n$ with respect to the null sets given by $\mathcal{G}_\infty^n = \cup_{t\in[0,\infty)}\mathcal{G}_t^n$ and the underlying probability measure (we will write this as $\overline{\mathcal{G}}_t^n$) is a right continuous filtration for all $n \in \mathbb{N}$.* (2) *$\mathbf{1}_{\{\tau_i \leq t\}} - \Lambda_t^i$ are $\overline{\mathcal{G}}_t^n$-martingales for all $i = 1, \ldots, n$.*

Proof. See Appendix B. ■

Remark 2 *Lemma 1 is based on results from Brémaud (1981), which verify that filtrations derived from right continuous piecewise constant processes are right continuous. The right continuous and piecewise constant processes used in the proof are $\mathbf{1}_{\{\tau_i \leq t\}}$. However, by including the entire paths of the covariates $X^i(s)$ and the common covariates $R^i(s), Y(s)$ in $\mathcal{F}_t^n = \mathcal{F}^n$, we are able to avoid the assumption that these processes are right continuous and piecewise constant. As a result, we can consider examples where the covariate processes follow diffusions or other continuously changing processes. This is particularly important given that we need to assume covariates have left-continuous paths for many of our arguments. The Brémaud (1981) results do not follow from left-continuous piecewise constant processes. Frequently, covariate processes will be left-continuous and piecewise constant in applications.*

I modify the martingale CLT used in Nielsen and Linton (1995) in order to obtain results relevant to the current setting. The martingale CLT used in NL (1995) is only justified in the case where the covariates are *i.i.d.*[7] However, similar results follow from a Rebolledo-type martingale CLT which holds for general martingales. Our construction of defaults will be important in verifying that these asymptotic results hold in specific situations. This is the main reason for insisting on this construction.

Some of these results require the underlying filtration to satisfy the usual hypothesis, which we established in Lemma 1 for the derived filtration. We follow the results of Liptser and Shiryayev (1980), (1989) in deriving the martingale CLT presented below in Proposition 3. These results are related to those in Hall and Heyde (1980). However, our approach allows for simple conditions under which a FCLT holds.

The next result looks superficially the same as the martingale CLT given in NL (1995). However, our results allows for the type of dependence between covariates outlined above.

[5] A filtration satisfies the usual conditions if it is right continuous and complete. Right continuoity of a filtration $\mathcal{F}_t$ means $\mathcal{F}_t = \cap_{h>0}\mathcal{F}_{t+h}$ for any $t \in [0,\infty)$. Completion means $\mathcal{F}_0$ contains all the P null sets of $\mathcal{F}$. See any standard text such as Protter (2005).

[6] Any filtration can be made right continuous, but this may destroy the martingale nature of M_t^i.

[7] This also rules out contagion as defined in Section 4.

Proposition 3 *Let* $\left\{g_1^{(n)},\ldots,g_n^{(n)}\right\}$ *be an array of predictable processes w.r.t.* $\overline{\mathcal{G}}_t^n$ *with a uniform bound across rows over the interval* $[0,T]$. *Suppose* σ *is a constant. If, as* $n\to\infty$,

$$\sum_{i=1}^{n}\int_0^T\left[g_i^{(n)}(s)\right]^2 d\Lambda_t^i(s)\to^p\sigma^2, \tag{1}$$

$$\sum_{i=1}^{n}\int_0^T\left[g_i^{(n)}(s)\right]^2\mathbf{1}_{\left\{\left|g_i^{(n)}(s)\right|>\epsilon\right\}}d\Lambda_t^i(s)\to^p 0,\qquad\forall\epsilon>0. \tag{2}$$

Then

$$\sum_{i=1}^{n}\int_0^T g_i^{(n)}(s)\,dM_i(s)\Rightarrow N\left(0,\sigma^2\right).$$

If (1) and (2) hold for all $t\in[0,T]$, *with* $\sigma^2(t)$ *a deterministic function, then*

$$\sum_{i=1}^{n}\int_0^T g_i^{(n)}(s)\,dM_i(s)\Rightarrow^{D[0,T]}M(t)$$

where $M(t)$ *is a continuous Gaussian martingale with variance structure* $\sigma^2(t)$. *If (1) converges to a random variable* σ^2 *instead of a constant, weak convergence will be to a mixed normal with mean zero and variance* σ^2.

Proof. See Appendix B. ■

3 Kernel Estimation

In this section, I derive kernel estimates which extend the results of Nielsen and Linton (1995) and Linton, Nielsen and Van de Geer (2001) to the dependent case. I focus on the block/step sampling case with no block specific covariate processes $R^i(t)$. The model can easily be extended to account for the processes $R^i(t)$ and all results below hold when they are present. However, the notation is already complicated and we remove $R^i(t)$ to focus on the ideas.

Let k be a continuous one-dimensional probability density and $k_b(\cdot)=\frac{1}{b}k(\cdot/b)$ for some bandwidth $b>0$. $K(u)=\Pi_{j=1}^d k(u_j)$ where $u=(u_1,\ldots,u_d)$ and $K_b(u)=\Pi_{j=1}^d k_b(u_j)$. We use product kernels and a single bandwidth throughout. This can be modified in practice. Assume we are in the block/step sampling case. Recall that we write $Z^i(t)=\left(t,X^i(t),Y\left(G^i+t\right)\right)$. The following results will be pointwise and we assume that $z=(t,x,y)$ is an interior point in the support $[0,T]\times\mathcal{X}\times\mathcal{Y}$. The following notation is used

$$\begin{aligned}K_b\left(z-Z^i(s)\right) &= \frac{1}{b}k\left(\frac{t-s}{b}\right)\\ &\quad\times\frac{1}{b^d}k\left(\frac{x_1-X^{i1}(s)}{b}\right)\cdots k\left(\frac{x_d-X^{id}(s)}{b}\right)\\ &\quad\times\frac{1}{b^j}k\left(\frac{y_1-Y^1\left(s+G^i\right)}{b}\right)\cdots k\left(\frac{y_j-Y^j\left(s+G^i\right)}{b}\right).\end{aligned}$$

In conducting estimation, NL (1995) and LNV (2003) are followed. The focus is on the differences between results due to the assumed dependence between observations. As much as possible, the notation of these previous papers is used.

Define $N^i(t)$ as

$$N^i(t) = \mathbf{1}_{\{\tau_i \leq t\}}.$$

The hazard rate $\alpha(t,x,y)$ is estimated by

$$\widehat{\alpha}(t,x,y) = \frac{\frac{1}{n}\sum_{i=1}^n \int_0^T K_b\left(z - Z^i(s)\right) dN_i(s)}{\frac{1}{n}\sum_{i=1}^n \int_0^T K_b\left(z - Z^i(s)\right) \mathbf{1}_{\{\tau_i \geq s\}} ds} \equiv \frac{\widehat{o}(z)}{\widehat{e}(z)}.$$

Define

$$\alpha^*(t,x,y) = \frac{\frac{1}{n}\sum_{i=1}^n \int_0^T K_b\left(z - Z^i(s)\right) \alpha\left(Z^i(s)\right) \mathbf{1}_{\{\tau_i \geq s\}} ds}{\frac{1}{n}\sum_{i=1}^n \int_0^T K_b\left(z - Z^i(s)\right) \mathbf{1}_{\{\tau_i \geq s\}} ds},$$

and decompose $\widehat{\alpha} - \alpha_0$ as

$$\left(\widehat{\alpha} - \alpha_0\right)(z) = \left(\widehat{\alpha} - \alpha^*\right)(z) + \left(\alpha^* - \alpha_0\right)(z) = \frac{\mathcal{V}_z}{\mathcal{E}_z} + \frac{\mathcal{B}_z}{\mathcal{E}_z}.$$

Here,

$$\begin{aligned} \mathcal{E}_z &\equiv \frac{1}{n}\sum_{i=1}^n \int_0^T K_b\left(z - Z^i(s)\right) \mathbf{1}_{\{\tau_i \geq s\}} ds, \qquad (3) \\ \mathcal{V}_z &\equiv \frac{1}{n}\sum_{i=1}^n \int_0^T K_b\left(z - Z^i(s)\right) dM_i(s), \\ \mathcal{B}_z &\equiv \frac{1}{n}\sum_{i=1}^n \int_0^T K_b\left(z - Z^i(s)\right) \left[\alpha\left(Z^i(s)\right) - \alpha_0(z)\right] \mathbf{1}_{\{\tau_i \geq s\}} ds. \end{aligned}$$

We will also need the following

$$\begin{aligned} \widehat{\sigma}_z^2 &\equiv \frac{1}{\mathcal{E}_z^2} \frac{b^{d+j+1}}{n} \sum_{i=1}^n \int_0^T K_b^2\left(z - Z^i(s)\right) dN_i(s), \qquad (4) \\ \mathcal{K}_z &\equiv \frac{b^{d+j+1}}{n} \sum_{i=1}^n \int_0^T K_b^2\left(z - Z^i(s)\right) \alpha\left(Z^i(s)\right) \mathbf{1}_{\{\tau_i \geq s\}} ds, \\ \mathcal{H}_z &\equiv \frac{b^{2(d+j+1)}}{n^2} \sum_{i=1}^n \int_0^T K_b^4\left(z - Z^i(s)\right) \alpha\left(Z^i(s)\right) \mathbf{1}_{\{\tau_i \geq s\}} ds. \end{aligned}$$

Assume that each observation

$$\left\{ Z^i(t) = \left(t, X^i(t), Y\left(G^i + t\right)\right) \middle| \, t \in [0,T] \right\}$$

has the same functional distribution. Specifically, let $X^i(t+)$ and $Y\left(G^i + t+\right)$ be the right continuous versions of these processes[8].

[8] The right continuous version of a process is defined as $X(t+) = \lim_{s \downarrow t} X(s)$ for each $t \in \mathbb{R}$.

(**A2**): Assume, for each i

$$\left\{ Z^i(t+) = \left(t, X^i(t+), Y\left(G^i + t+\right)\right) \middle| t \in [0,T] \right\}$$

has the same functional distribution.[9,10]

Additionally, we assume the following distributions and densities are the same for each observation.[11]

(**A3**): Each observation has the same mean functional

$$w(t) = \mathbb{E}\left[\mathbf{1}_{\{\tau_i \geq t\}}\right],$$

and conditional distribution function

$$F_t(x,y) = \mathbb{P}\left[X^i(t) \leq x, Y\left(G^i + t\right) \leq y \middle| \mathbf{1}_{\{\tau_i \geq t\}} = 1\right].$$

This distribution is assumed to have a corresponding density $f_t(x,y)$ with support equal to $\mathcal{X}_1 \times \cdots \times \mathcal{X}_d \times \mathcal{Y}_1 \times \cdots \times \mathcal{Y}_j$, the compact support assumed above.

Define

$$e(z) = f_t(x,y)\, w(t).$$

Assumption (A3) is the main stumbling block to applying kernel estimation to hazard models with contagion. I define rigorously what is meant by contagion in Section 4. The assumption implies that the covariates are continuously distributed for all $t \in [0,T]$. Many specifications of contagion will not satisfy this. However, it may be possible to satisfy these conditions in some contagion situations.

We use assumption (A4) from LNV (2003) in the sequel. It is reproduced here for convenience.

(**K2**): The kernel k has support $[-1,1]$, is symmetric about 0 and is of order r, that is, $\int_{-1}^{1} k(u)\, u^r du \in (0,\infty)$, where $r \geq 2$ is an even integer. The kernel is also $r-1$ times continuously differentiable on $[-1,1]$ with lipschitz remainder; that is, there exists a finite constant k_{lip} such that $\left|k^{(r-1)}(u) - k^{(r-1)}(u')\right| \leq k_{lip}\left|u-u'\right|$ for all u, u'. Finally, $k^{(j)}(\pm 1) = 0$ for $j = 0, \ldots, r-1$.

In the following result, we follow NL (1995) Theorem 1. However, we add additional high level assumptions necessary because of the dependence in our model. Later in this section, I discuss some assumptions which imply those given in Theorem 4 below.

[9] Specifically, assume the observations i have the same distribution in the Skorokhod space. See Billingsley (1999) or Jacod and Shiryaev (2003) for more on the Skorokhod space. We need to consider right continuous versions of the above processes because the Skorokhod space is defined on $D[0,T]$, the space of all right continuous paths with left hand limits on $[0,T]$. We had previously assumed $Z^i(s)$ has càglàd paths. The purpose of A2 is to preserve the values of certain expectations of integrals that arise in the *i.i.d.* case. Because the relevant processes are used in integrals, making them right continuous does not affect the expectation of those integrals. Consideration of the right continuous versions of the processes is purely for the convenience of using the Skorohod space for functional distributions.

[10] Càglàd processes can only have a countable number of discontinuities. These are the only values that change when we make the process right continuous. See Ethier and Kurtz (1986) pg.116 Lemma 5.1.

[11] This would follow directly from an assumption on the functional distribution of Z^i on a space that contains càglàd paths. Such a space is not standard, so we make the additional assumption.

Theorem 4 *Assume (A1)-(A3) and (K2). Assume (S): $e(z) > 0$ on a neighborhood of z; $\alpha, e \in C^r$ in a neighborhood of z where $r \geq 2$. Define the constants $\kappa_1 = \int_{-1}^{1} v^2 k(v)\, dv$ and $\kappa_2 = \int_{-1}^{1} k^2(v)\, dv$. (B): $nb^{d+j+1} \to \infty$ and $b \to 0$. Assume*

$$\mathcal{E}_z - \mathbb{E}[\mathcal{E}_z] \to^p 0. \tag{5}$$

$$\mathcal{K}_z - \mathbb{E}[\mathcal{K}_z] \to^p 0. \tag{6}$$

$$\mathcal{B}_z - \mathbb{E}[\mathcal{B}_z] \to^p 0. \tag{7}$$

$$\mathcal{H}_z - \mathbb{E}[\mathcal{H}_z] \to^p 0. \tag{8}$$

Then the following holds. $C(z)$ is a constant that depends on z.

$$n^{1/2} b^{(d+j+1)/2} \left[\widehat{\alpha}(z) - \alpha^*(z)\right] \Rightarrow N\left[0, \kappa_2^{d+j+1} \frac{\alpha(z)}{e(z)}\right], \tag{9}$$

$$b^{-r}\left[\alpha^*(z) - \alpha_0(z)\right] \to^p C(z) \tag{10}$$

$$\widehat{\sigma}_z^2 \to^p \sigma_z^2 \equiv \kappa_2^{d+j+1} \frac{\alpha(z)}{e(z)} \tag{11}$$

In particular, if we choose the bandwidth such that $b \sim n^{-1/(d+j+1+2r)}$, then the asymptotic bias is given by $C(z)$.

Proof. See Appendix B. ■

The assumptions (5)-(8) each state that a particular array of mean zero random variables converges to zero in probability. All of these arrays are derived from the underlying covariates and random variables η_i. As a result, an assumption on the dependence of the processes $Z^i(t)$ determines the dependence properties for these arrays. The next lemma gives an important preliminary result to establishing mixing conditions on the summands in (5)-(8).

Lemma 5 *Let $f(x, y, t)$ be a bounded continuous function on $\mathbb{R}^{d+j+1}$. Define $W^1 = \int_0^T f(X_s, Y_{G^i+s}, s)\, ds$ and $W^2 = \int_0^T f(X_s, Y_{G^i+s}, s)\, \mathbf{1}_{\{\tau_i \geq s\}} ds$. Then*

$$\sigma\left(W^1\right), \sigma\left(W^2\right) \subset \sigma\left\{\eta_i, X^i(s), Y\left(G^i + s\right) \middle| 0 \leq s \leq T\right\}. \tag{12}$$

Proof. See Appendix B. ■

Lemma 5 allows us to convert mixing conditions on the underlying covariate processes into mixing conditions on the integrals we use for estimation. This will facilitate verification of the conditions in Theorem 4. Each of the arrays used in Theorem 4 will inherit the mixing conditions of the underlying processes $Z^i(t)$, where mixing is in the dimension $i = 1, \ldots, n$. Therefore, (5)-(8) can all be established with a single mixing condition on the covariates. This is far more natural an economic assumption than directly making dependence or convergence assumptions on the rather complicated sums (5)-(8). Directly assuming mixing conditions for the appropriate arrays will also verify (5)-(8). Below, we will assume α-mixing for concreteness. However, Lemma 5 can accommodate a variety of different mixing conditions.

The interpretation of Lemma 5 is simple. If we know the paths of the processes $X^i(s)$, $Y\left(G^i + s\right)$

and we know η_i, then we can derive the values of W^1 and W^2. If we only know the values of W^1 and W^2, we can not necessarily recover the paths of $X^i(s)$, $Y\left(G^i+s\right)$ or η_i. Information is potentially lost in the integration. Putting mixing conditions on the right hand side of (12) is stronger than needed for the results of Theorem 4 to hold. Only the arrays in (5)-(8) need to satisfy the required mixing conditions.

Our approach to verifying the conditions in Theorem 4 is to make mixing assumption on the underlying variables $Z^i(t)$ and η_i. Then, by Lemma 5, these mixing assumptions are transferred onto the summands of integrals in $\mathcal{E}_z$, $\mathcal{K}_z$, $\mathcal{B}_z$ and $\mathcal{H}_z$. Once their expectations are subtracted off, the rows of these arrays are sums of mean zero random variables which satisfy mixing conditions. Finally, we show (5)-(8) with a Bernstein inequality based on the mixing properties.

The Bernstein inequality from Bosq (1999) pg. 27 will be used. This requires α-mixing. Different Bernstein inequalities require other conditions. In particular, several other Bernstein inequalities which can facilitate the proof require stationarity assumptions. In these situations, we need the rows of the arrays $\mathcal{E}_z$, $\mathcal{K}_z$, $\mathcal{B}_z$ and $\mathcal{H}_z$ to be stationarity. This requires a stationary functional distribution assumption on the underlying covariates. See Modha and Masry (1996) or Chen and Shen (1998) for alternate Bernstein inequalities requiring stationarity. Other choices have various drawbacks and advantages.

Recall, for a sequence of random variables W_t which generate σ-fields

$$\mathcal{F}_i^j = \sigma\left\{W_t \middle| t=i,\ldots,j\right\},$$

the α-mixing coefficients are

$$\alpha(n) = \sup_{k\in\mathbb{N}} \sup\left\{\left|P(A\cap B) - P(A)P(B)\right|; A\in\mathcal{F}_1^k, B\in\mathcal{F}_{k+n}^\infty\right\}.$$

Proposition 6 *Assume (A1)-(A3). Define the sequence of σ-fields*

$$\mathcal{H}_l^m = \vee_{i=l}^m \sigma\left\{\eta_i, X^i(s), Y\left(G^i+u\right)\middle| 0\le s,u\le T\right\}.$$

Assume the system of σ-fields $\mathcal{H}_l^m$ has an α-mixing rate $\alpha(n)$ such that for some $\overline{\alpha}>0$ and $\xi>0$,

$$\alpha(n) \le \overline{\alpha} n^{-\xi}$$

If, for some $0<\gamma<1$,

$$\begin{aligned} b^{2(d+j+1)} n^{\gamma} &\to \infty, \\ n^{-(1-\gamma)\xi-\gamma} b^{-(d+j+1)/2} &\to 0, \end{aligned}$$

then the conditions (5)-(8) of Theorem 4 hold. Assume $\mathcal{H}_l^m$ has an α-mixing rate $\alpha(n)$ such that for some $\overline{\alpha}>0$ and $c>0$,

$$\alpha(n) \le \overline{\alpha}\exp(-cn).$$

If, for some $0 < \gamma < 1$*,*

$$\begin{aligned} b^{2(d+j+1)}n^{\gamma} &\rightarrow \infty, \\ b^{-(d+j+1)/2}n^{\gamma}\exp\left(-\frac{c}{2}n^{(1-\gamma)}\right) &\rightarrow 0, \end{aligned}$$

then the conditions (5)-(8) of Theorem 4 hold.

The variables η_i do not affect the mixing rate because they are *i.i.d.* and independent of the covariates. If β-mixing holds with the same inequalities given in Proposition 6, the results also holds because of the well known inequality $2\alpha(n) \leq \beta(n)$. The mixing rate needed for Proposition 6 reduces to controlling the mixing rate for $Y(t)$ if the processes $X^i(t)$ are *i.i.d.* and independent of $Y(t)$. Dependence between the sets of variables $X^i(t)$ and between $X^i(t)$ and $Y(t)$ is also possible. This allows for the most relevant economic situation where all variables are correlated with each other. Continuous time mixing conditions on $X^i(t)$ and $Y(t)$ easily transfer onto the relevant system of σ-fields $\mathcal{H}_l^m$.

In order to have an asymptotically normal estimator with a finite asymptotic bias, the requirements for the bandwidth are $n^{1/2}b^{(d+j+1)/2} \leq b^{-r}$ and those conditions given in Proposition 6. If the inequality $n^{1/2}b^{(d+j+1)/2} \leq b^{-r}$ is weak, there will be no asymptotic bias. Comparing $b^{2(d+j+1)}n^{\gamma} \rightarrow \infty$ to the analogous condition in the *i.i.d.* case $b^{(d+j+1)}n \rightarrow \infty$, we see that the rate at which $b \rightarrow 0$ must be slowed down in the dependence case. A similar reduced rate for the bandwidth is required if other Bernstein inequalities are used. Successive observations in the ordering provide less additional information than in the *i.i.d.* case because they are correlated with previous observations. Therefore, given a fixed data size, observations further from the desired estimation point in pointwise estimation must be more intensively incorporated into estimators compared with the *i.i.d.* case. This requires the bandwidth to converge more slowly to zero.

Theorem 7 *Let* I *be any compact set such that* $I \subset [0,T] \times \mathcal{X} \times \mathcal{Y}$*. Make the assumptions (A1)-(A3) and (K2). Assume* $\mathcal{H}_l^m$ *from Proposition 6 has an* α*-mixing rate* $\alpha(n)$ *such that for some* $\overline{\alpha} > 0$ *and* $c > 0$*,*

$$\alpha(n) \leq \overline{\alpha}\exp(-cn).$$

In addition, we assume (B) from Theorem 4. Assume $w(t)$ *and* $f_t(x,y)$ *are continuous on* $[0,T]$ *for all* (x,y)*. Finally, assume* $\alpha(\cdot)$ *and* $e(\cdot)$ *are* $r \geq 2$ *time continuously differentiable and* $\inf_{(t,x,y)\in[0,T]\times\mathcal{X}\times\mathcal{Y}}[e(z)] > 0$*. Then, for any* $0 < \gamma < 1$ *which satisfies*

$$\frac{n^{(9/4)\gamma}\exp\left(-\frac{c}{2}n^{(1-\gamma)}\right)}{\log n} \rightarrow 0,$$

the following holds:

$$\sup_{x\in I}|\widehat{\alpha}(x) - \alpha(x)| = O(b^r) + O_p\left\{\sqrt{\frac{\log n}{n^{\gamma}b^{2(d+j+1)}}}\right\} \tag{13}$$

Proof. See appendix B. ■

A similar theorem holds assuming polynomial alpha mixing decay as in Proposition 6. Comparing the above theorem with Lemma 3 from LNV (2003), we see that the dependence slows down the uniform convergence rate by replacing $nb^{(d+j+1)}$ with $n^{\gamma}b^{2(d+j+1)}$ in (13). This is a similar adjustment as required for Proposition 6. The rate of uniform convergence (13) is the result of the particular Bernstein inequality used in the proofs. Other Bernstein inequalities will facilitate the same proof with different assumptions and resulting convergence rates.

The result (9) from Theorem 4 states that, provided all the regulatory conditions are satisfied, the asymptotic variance of the estimators is the same as in the *i.i.d.* case. The consistent estimator of the asymptotic variance (11) is the same one used in the *i.i.d.* case. However, several simulation studies in the literature show that when the same estimator of the asymptotic variance for the *i.i.d.* case is used in the dependent case, the estimator preforms poorly in small samples. See Chen, Liao and Sun (2011) for an overview of this problem and more citations. Other approaches to estimating the asymptotic variance in finite samples have been shown to preform better. Possible approaches include bootstrapping and long run variance. Again, see Chen, Liao and Sun (2011). Incorporating these extensions into our estimation methods is a topic of future research.

3.1 The Curse of Dimensionality

Throughout this section, we face a standard curse of dimensionality problem as the number of covariates increases. As in LNV (2003), if we further restrict the form of the hazard function we are able to greatly improve the rate of convergence in our dependent case. Specifically, we may assume the hazard is either additively or multiplicatively separable:

$$\alpha(z) = c_A + \sum_{l=1}^{d+j+1} g_j(z_j) \tag{14}$$

$$\alpha(z) = c_M \prod_{l=1}^{d+j+1} h_j(z_j) \tag{15}$$

where c_A and c_M are constants. We still assume (A1). In particular, $0 < \underline{C} \leq \alpha(z)$. The individual functions in (14)-(15) are not separately identified. We need to define a probability measure over the compact rectangle I in order to identify them. Let Q be an arbitrary cdf with probability only on I and with marginal cdfs $Q_l(z_l) = (\infty, \ldots, \infty, z_l, \infty, \ldots, \infty)$ and $Q_{-l}(z_{-l}) = (z_1, \ldots, z_{l-1}, \infty, z_{l+1}, \ldots, z_{d+j+1})$. The functions in (14)-(15) are identified by assuming, for all $l = 1, \ldots, d+j+1$

$$\int g_l(z_l)\, dQ_l(z_l) = 0, \tag{16}$$

or

$$\int h_l(z_l)\, dQ_l(z_l) = 1. \tag{17}$$

LNV (2003) allow for very general Q and for Q to be estimated from data. For simplicity, we assume Q is known with the following form

(**A4**): Q is continuous with respect to Lebesgue measure. Its has density equal to $\frac{1}{q_1} \cdots \frac{1}{q_{d+j+1}}$ where q_l is the length of the compact interval corresponding to variable l in the compact rectangle I. It has marginal densities $\frac{1}{q_l}$ for Q_l and $\frac{1}{q_{-l}} = \frac{1}{q_1} \cdots \frac{1}{q_{l-1}} \frac{1}{q_{l+1}} \cdots \frac{1}{q_{d+j+1}}$ for Q_{-l}.

Under this assumption Q satisfies assumption (A2) from LNV (2003). The following results could be extended to more complicated probability measures. Note also that $\int \alpha(z)\, dQ(z) = c_A$ and $\int \alpha(z)\, dQ(z) = c_M$ in each model respectively. I write c generically when the specific model is unimportant.

Make the following definitions:

$$\begin{aligned}
\alpha_{Q_{-j}}(z_j) &= \int \alpha(z)\, dQ_{-j}(z_{-j}), \\
\alpha^A_{Q_{-j}}(z_j) &= \alpha_{Q_{-j}}(z_j) - c = g_j(z_j), \\
\alpha^M_{Q_{-j}}(z_j) &= \frac{\alpha_{Q_{-j}}(z_j)}{c} = h_j(z_j).
\end{aligned}$$

and define the corresponding estimators

$$\begin{aligned}
\widehat{c} &= \int \widehat{\alpha}(z)\, dQ(z), \\
\widehat{\alpha}_{Q_{-j}}(z_j) &= \int \widehat{\alpha}(z)\, dQ_{-j}(z_{-j}), \\
\widehat{\alpha}^A_{Q_{-j}}(z_j) &= \widehat{\alpha}_{Q_{-j}}(z_j) - \widehat{c}, \\
\widehat{\alpha}^M_{Q_{-j}}(z_j) &= \frac{\widehat{\alpha}_{Q_{-j}}(z_j)}{\widehat{c}}.
\end{aligned}$$

Theorem 8 *Assume all the conditions of Theorem 7, (A4) and the identification assumption (16) or (17). Let $C > 0$ and $0 < \theta \leq r/(2r+1)$. Assume*

$$b = \left[\frac{n^{\theta-1/2}}{C}\right]^2. \tag{18}$$

With $\theta = r/(2r+1)$ and large enough γ and r, all of the following conditions will hold assuming the form of b in (18). Other bandwidth choices are possible. Note that it must be the case that $\gamma > 1/2$.

$$n^{1/2} b^{2r-(d+j+1)/2} \to 0, \tag{19}$$

$$\begin{gathered}
n^{1/2-\gamma/2} \log^{1/2}(n)\, b^{r-3(d+j+1)/2} \to 0, \\
n^{1/2-\gamma} \log(n)\, b^{-5(d+j+1)/2} \to 0, \\
\frac{\log n}{n^{\gamma} b^{2(d+j+1)+1}} \to 0, \\
n b^{2(d+j+1)+1} \to \infty.
\end{gathered}$$

If these conditions hold, there exists functions $m_j(\cdot)$, $v_j(\cdot)$ which are bounded and continuous on I_j such that for all $z_j \in I_j$

$$n^{\theta}\left(\widehat{\alpha}_{Q_{-j}}(z_j) - \alpha_{Q_{-j}}(z_j)\right) \Rightarrow N\left[m_j(z_j), v_j(z_j)\right].$$

$v_j(z_j)$ is equal to

$$v_j(z_j) = C^2 \kappa_2 \int_{I_{-j}} \frac{\alpha(z)\, \frac{1}{q_{-j}}}{e(z)} dz_{-j}.$$

If $\theta < r/(2r+1)$*, then the estimator is asymptotically unbiased (i.e.* $m_j(z_j) = 0$*). Assuming we compute* $\widehat{c}$ *using bandwidths such that* $\widehat{c} - c = O_p\left(n^{-1/2}\right)$*, then*

$$n^{\theta}\left(\widehat{\alpha}^{A}_{Q_{-j}}(z_j) - g_j(z_j)\right) \Rightarrow N\left[m_j(z_j), v_j(z_j)\right], \quad (20)$$

$$n^{\theta}\left(\widehat{\alpha}^{M}_{Q_{-j}}(z_j) - h_j(z_j)\right) \Rightarrow N\left[m_j(z_j)/c, v_j(z_j)/c^2\right]. \quad (21)$$

All that is needed for (20)-(21) is that $\widehat{c} - c$ converges to zero faster than n^{θ}. 1/2 is an upper bound for $r/(2r+1)$, which corresponds to infinite differentiability. The result always holds if $\widehat{c} - c = O_p\left(n^{-1/2}\right)$, which is what is presented in the theorem. The fact that we can always choose γ and r such that the conditions (19) hold shows how smoothness of the hazard function and weakening of dependence between covariates facilitates the results. A similar theorem can be derived assuming polynomial mixing decay. In this case, the trade off is more important because the dependence is stronger.

Extending this result for different Bernstein inequalities is possible. Care needs to be taken in modifying the proof as specifics of the inequality become entangled with the needed rates of convergence for the bandwidth.

A problem in any implementation of these results is how to choose the bandwidth. A practical approach is to use a version of cross validation. This has been considered in the hazard case in Ramlau-Hansen (1981), Nielsen (1990), Anderson et al. (1994) and NL (1995). b is chosen to minimize the criterion function

$$M(b) = \sum_{i=1}^{n}\int_0^1 \widehat{\alpha}^2_{-i}\left(Z^i(s)\right)\mathbf{1}_{\{\tau_i \geq s\}}ds - 2\sum_{i=1}^{n}\int_0^1 \widehat{\alpha}_{-i}\left(Z^i(s)\right)dN^i_s$$

where $\widehat{\alpha}_{-i}$ is the leave-one-out version of the estimator. Asymptotically, this is equivalent to minimizing

$$\sum_{i=1}^{n}\int\left[\widehat{\alpha}_{-i}\left(Z^i(s)\right) - \alpha\left(Z^i(s)\right)\right]^2\mathbf{1}_{\{\tau_i \geq s\}}ds.$$

3.2 Unbounded Estimation

The results above restrict the time interval for which observations are at risk to be bounded with length T. In many situations, there is no natural upper bound on the time interval observations are at risk. In this subsection, I discuss how to remove this restriction at the price of stronger assumptions.

Appendix A shows how to extend the construction given in Section 2 to allow for martingales defined on $[0, \infty]$. Throughout this section, we assume the conditions needed for this extension are satisfied. In addition, we assume that time is not a covariate for simplicity.

All of the arrays in (3) and (4) can be defined at infinity by taking pointwise limits as $T \to \infty$. The values $\widehat{\alpha}(z)$ and $\alpha^*(z)$ can similarly be defined by taking pointwise limits as $T \to \infty$. For the remainder of this section, we assume all relevant definitions have been extended to infinity by taking pointwise limits. We are now in a position to extend Theorem 4. Only a weak additional condition is required.

Proposition 9 *Assume all of the conditions of Theorem 4. In particular, conditions (5)-(8) are assumed directly. The corresponding arrays have been extended by taking pointwise limits as* $T \to \infty$*. Assume,* *for all $\epsilon > 0$*

$$X_{ni} = \int_0^\infty \frac{1}{n^{1/2} b^{(d+j+1)/2}} K\left(z - Z^i(s)\right) dM_i(s),$$

$$\sum_{i=1}^n X_{ni}^2 \mathbf{1}\left\{|X_{ni}| > \epsilon\right\} \to^p 0. \tag{22}$$

Then the results (9)-(11) from Theorem 4 hold.

Proof. Simple extension of the proof of Theorem 4. See Appendix B. ■

Unbounded versions of Theorems 7 and 8 are possible with appropriate additional assumptions. These results are omitted for brevity. The martingale central limit theorem used here is Hall and Heyde (1980) Corollary 3.1. This is basically equivalent to the one used in Theorem 4. The additional Lindberg condition (22) is needed because, when there is no upper bound T, (22) is not trivially satisfied.

We directly assume conditions (5)-(8) instead of deriving them from mixing conditions on the underlying processes. Another approach is to assume the rows of the relevant arrays satisfy the required mixing conditions. For example, the mixing conditions from Proposition 6. This is weaker than making assumptions on the underlying covariate processes.

More specific conditions on covariate processes that imply the required mixing conditions in the unbounded case would be useful. One potential approach is to examine the information structure

$$\mathcal{H}_l^m = \vee_{i=l}^m \sigma\left\{\eta_i, \left(\widetilde{X}^i(s), Y(s)\right) \mathbf{1}_{\{\tau_i \geq s\}} \mathbf{1}_{\{G^i \leq s\}} \middle| 0 \leq s \leq \infty\right\}, \tag{23}$$

where the notation is specified in Appendix A. Here, $\widetilde{X}^i(s)$ is a version of $X^i(s)$ which is defined on $[0, \infty)$ instead of just $\left[G^i, G^i + T\right]$.[12] In (23), the indicators truncate the information before the observation is at risk and after it has defaulted. This truncation is what will facilitate the convergence of the extended versions of (5)-(8). Although it is possible that all observation are impacted by the entire right tail of $Y(t)$, this is not what happens. The information relevant for each observation is truncated by the random default times. The values of $Y(t)$ after the calendar time of default are not relevant for the observation. Further examination of this information structure is beyond the scope of this paper.

4 Point Process Likelihood Estimation: Contagion and Frailty

In this section, contagion and frailty are rigorously defined and incorporated into a hazard model. I give conditions under which a hazard model can be estimated in the presence of these additional sources of dependence. As in previous sections, dependence between the covariate processes across observations is also permitted. A point process likelihood approach is used to semi-nonparametrically estimate the hazard model using sieves. I also discuss specification testing.

A different approach to estimation than the kernel methods presented in Section 3 is considered because those methods encounter difficulty when contagion and frailty are present. First, a standard estimation approach based on Anderson and Gill (1982) is shown to hold when contagion is present.

[12] In Appendix A, H^i is used instead of G^i in order to distinguish the case where $X^i(t)$ is defined on $[0, \infty)$ with the case where it is defined on $\left[G^i, G^i + T\right]$.

Second, I propose a different type of likelihood estimation which uses sieves. This method is similar to Karr (1987). The second estimator can handle frailty.

There is already a large literature on likelihood estimation of point processes. See Martinussen and Scheike (2010) or Anderson et al. (1994) for a review. Our model is also related to methods designed for clustered failure time data. See Martinussen and Scheike (2010) or Hougaard (2000) for an outline of these methods. The nature of dependence across observations in our model is different than in these cases. Our estimation approach also differs, as we consider sieves.

4.1 Contagion

In most of what follows, observations $i \in \mathbb{N}_0$ are put into groups of equal size w. The sampling is sampling of groups. The equal size assumption is not necessary for the results to hold. However, various issues arise when groups are of different sizes. For simplicity of exposition and notation, focus is on groups of the same size.

I now construct random times with contagion. Again, as in Section 3, a good example to keep in mind is credit default. First I describe the covariate processes which are used to define default. Each observation $i \in \mathbb{N}_0$ has a set of d covariate stochastic processes $\left\{ X^i(t) \middle| t \in [0,T] \right\}$. Unlike Section 2, we do not include common processes $Y(t)$ or the length of time an observations has been at risk of default t in the covariates. The main situation we consider is a cross section over the time interval $[0,T]$, corresponding to all observations having $G^i = 0$. Assume the distribution of the variables $X^i(t)$ has support *contained in* the compact set $\mathcal{X} = \mathcal{X}_1 \times \cdots \times \mathcal{X}_d$ for each $t \in [0,T]$. This allows for covariate stochastic processes with discrete support. Each observation in a group also corresponds to a set of j group specific covariate processes $\left\{ R^l(t) \middle| t \in [0,T] \right\}$ corresponding to group l. We continue to abuse notation and write $R^i(t)$ for $R^l(t)$. Assume the distribution of the variables $R^i(t)$ has support *contained in* the compact set $\mathcal{R} = \mathcal{R}_1 \times \cdots \times \mathcal{R}_m$ for each $t \in [0,T]$.

For each observation $i \in \mathbb{N}_0$, we also define a set of covariates which are derived from other defaults in the group. This is where contagion originates. For each observation, define the set of processes

$$V^i(t) = \left\{ \mathbf{1}_{\{\tau_j < s\}} \middle| j \neq i,\ j \text{ in the same group as } i \right\}.^{13}$$

These are processes that indicate other members in the group have defaulted. Defaults have not been defined yet, but will be shortly. The specification is well defined.

Let $g(\cdot)$ be a known bounded measurable function used to define another set of k covariates for observation i: $g\left(t, X_t^i, X_t^{-i}, R_t^i, V_t^i\right)$. Here, X_t^{-i} are the X_t^j covariate processes for the other members of the group: $j \neq i$. The covariate processes defined by $g(\cdot)$ are functions of the processes corresponding to the group ; $\left(X_t^i, X_t^{-i}, R_t^i\right)$, and the other defaults in the group; V_t^i. Note that $g\left(t, X_t^i, X_t^{-i}, R_t^i, V_t^i\right)$ is defined when $V_t^i = 0$. Finally, assume that the paths of all covariates are left-continuous with right-hand-limits (càglàd).

[13] We must use the processes $\mathbf{1}_{\{\tau_j < s\}}$ instead of $\mathbf{1}_{\{\tau_j \leq s\}}$ for technical reasons.

Now I rigorously define default. Random times τ_i are defined as in Section 2 by

$$\begin{aligned}\Gamma_t^i &\equiv \int_0^t h_0(t)\exp\left\{\beta_{01}'\left(X_s^i, R_s^i\right) + \beta_{02}' g\left(t, X_t^i, X_t^{-i}, R_t^i, V_t^i\right)\right\} ds \\ \tau_i &\equiv \inf\left\{t \in \mathbb{R}_+ \mid \Gamma_t^i \geq \eta_i\right\}\end{aligned}$$

where η_i is an independent standard exponentially distributed random variable. $h_0(t)$ is a continuous strictly positive function. Because the hazard for each observation only depends on a finite number of other defaults, the specification is well defined. At time zero, where there are no defaults, $g(\cdot)$ takes on its value with $V_t^i = 0$ until the first default arrives. Then, the remaining observations have their process Γ_t^i (and therefore their hazard) updated to reflect the default. This recursive definition is not circular. It is possible to make similar definitions without a group structure.

The hazard for this specification is

$$h_0(t)\exp\left\{\beta_{01}'\left(X_s^i, R_s^i\right) + \beta_{02}' g\left(t, X_t^i, X_t^{-i}, R_t^i, V_t^i\right)\right\}.$$

This hazard has a Cox proportional hazard form, which is widespread in the literature. What I call contagion is the set of covariates $g\left(t, X_t^i, X_t^{-i}, R_t^i, V_t^i\right)$ which are updated based on other defaults. I have written the model in some generality. Any or all of the covariates $\left(X_t^i, X_t^{-i}, R_t^i\right)$ could be removed. However, removing all the variables V_t^i causes us to lose the interpretation of contagion.

Restricting contagion to be contained in groups is restrictive. However, this is a relatively simple starting point for deriving asymptotic theory allowing for direct interactions between observations in hazard analysis. More complicated interactions are the subject of future research. The generality allowed for in the function g gives significant leeway to researchers in how they choose to specify the nature of contagion. In the simplest specification, $g\left(t, X_t^i, X_t^{-i}, R_t^i, V_t^i\right) = V_t^i$ and contagion consists of indicators that other group members have defaulted. The ability to incorporate other variables allows for extension of V_t^i. The initial impact of other defaults can be augmented by other available information $\left(t, X_t^i, X_t^{-i}, R_t^i\right)$. For example, the impact of hazard rates could be proportional to information specific to the defaulting observation $X^j(t)$. This impact could decay through time, possibly deterministically because of the inclusion of t in g. See subsection 4.3 for some specific examples of potential covariates incorporating contagion.

Implicitly, groups are of equal size in the above outlined construction. We now formalize this in the following assumption.

(B1): Observations are divided into groups of size w. Contagion is restricted to depend only on V_t^i, the other defaults in the same group. We additionally assume (A2) for the covariates X_s^i, R_s^i and $g\left(t, X_t^i, X_t^{-i}, R_t^i, V_t^i\right)$. The variables are assumed to be càglàd. The true underlying coefficients β_0 are the same for each observation.

This assumption implies that the expectation of the likelihood presented below is the same for all observations. (B1) also eases the presentation of the proofs. The most obvious way to satisfy assumption A2 on the covariates is to impose symmetry on group members. I discuss how the assumptions of symmetry and equal group sizes can be relaxed in the sequel. Dependence in the basic covariate processes

$\left(X_t^i, X_t^{-i}, R_t^i\right)$ across and within groups is allowed in the asymptotics presented below. This differs from standard results. The direct modeling of contagion is also a point of departure.

The martingale nature of default can be shown to hold in the contagion situation. The following martingale structure will be required for estimation. A proof similar to Lemma 1 will give this result. It similarly follows from Bielecki and Rutkowski (2004) Example 9.1.5. The details of the filtration and the proof are omitted. Recall the definition of the compensator:

$$\Lambda_t^i \equiv \int_0^t h_0(t) \exp\left\{\beta_{01}'\left(X_s^i, R_s^i\right) + \beta_{02}' g\left(X_t^i, X_t^{-i}, R_t^i, V_t^i\right)\right\} \mathbf{1}_{\{\tau_i \geq s\}} ds$$

Lemma 10 *There exists filtrations $\overline{\mathcal{G}}_t^n$ which satisfy the usual conditions such that $M_t^i = \mathbf{1}_{\{\tau_i \leq t\}} - \Lambda_t^i$ are $\overline{\mathcal{G}}_t^n$-martingales for all $i = 1, \ldots, n$.*

The martingale nature of M_t^i holds in great generality. In particular, it holds if common processes and block/step sampling is added to the construction in this section. In the sequel, we freely assume the martingale structure of defaults with contagion. In all cases, this is justified. The details are omitted for brevity. Conditionally independent censoring can violate the martingale structure in the contagion case. However, if all members of the groups share a censoring time the martingale structure is preserved.

4.2 Standard Estimation With Dependence

Because we have verified the martingale structure of M_t^i, we are able to apply certain standard estimation methods to this Cox proportional hazard specification. In particular, the martingale structure is a prerequisite for applying the classic results of Anderson and Gill (1982) (hereafter AG (1982)). Part of the contribution of this paper is to show the martingale structure holds, even with the type of contagious variables described above, and stress that certain standard estimation approaches are justified in this situation.

This paper differs from others in that we assume dependence in the covariates across all observations. This includes dependence in covariates across groups as the variables $\left(X_t^i, X_t^{-i}, R_t^i\right)$ are only required to satisfy an α-mixing condition. The asymptotic results in AG (1982) hold in this dependent case, provided the high level assumptions in that paper are satisfied. To the best of my knowledge, no work has been published that shows these high level assumptions hold in the case where there is dependence across all observations. In this subsection, I show the required conditions hold under appropriate assumptions.

AG (1982) derive consistency and asymptotic distributions for estimators of β_0 and $\int_0^t h_0(s)\,ds$. Their estimation uses a point process likelihood approach. In what follows, we normalize $[0,T]$ to $[0,1]$. We make the following assumption.

(B2): For all $t \in [0,1]$, $Z^i(t) = \left(X_t^i, R_t^i, g\left(t, X_t^i, X_t^{-i}, R_t^i, V_t^i\right)\right)$ has support contained in a compact rectangle normalized to be $[0,1]^q$ where q is the number of covariates. β_0 is contained in a compact rectangle $[a_1, b_1] \times \cdots \times [a_q, b_q]$.

This assumption is made to satisfy conditions in AG (1982). We make this assumption in the remainder of this section. In addition, each observation has an *i.i.d.* right censoring time C^i. This is stronger than the conditional independent censoring assumed elsewhere in the paper. Let $A^i(t) = \mathbf{1}\{\tau_i \geq t\}\mathbf{1}\{C_i \geq t\}$.

As in AG (1982), define[14]

$$
\begin{aligned}
s^{(0)}(\beta, t) &= \mathbb{E}\left[A(t)\exp\left(\beta' Z(t)\right)\right], \\
s^{(1)}(\beta, t) &= \mathbb{E}\left[A(t) Z(t)\exp\left(\beta' Z(t)\right)\right], \\
s^{(2)}(\beta, t) &= \mathbb{E}\left[A(t) Z(t)^{\otimes 2}\exp\left(\beta' Z(t)\right)\right], \\
e &= s^{(1)}/s^{(0)}, \\
v &= s^{(2)}/s^{(0)} - e^{\otimes 2}, \\
\Sigma &= \int_0^1 v(\beta_0, t)\, s^{(0)}(\beta_0, t)\, h_0(t)\, dt, \\
\overline{N}(t) &= \sum_{i=1}^n N^i(t) A^i(t)
\end{aligned}
$$

The estimator $\widehat{\beta}$ is defined as the value of β which makes the following criterion function 0:

$$
U(\beta) = \sum_{i=1}^n \left[\int_0^1 Z^i(s) A^i(s)\, dN^i(s)\right] - \int_0^1 \left[\frac{\sum_{i=1}^n A^i(s) Z^i(s)\exp\left(\beta' Z^i(s)\right)}{\sum_{i=1}^n A^i(s)\exp\left(\beta' Z^i(s)\right)}\right] d\overline{N}(s).
$$

The estimator $\widehat{\Lambda}(t)$ of $\int_0^t h_0(s)\, ds$ is defined as

$$
\widehat{\Lambda}(t) = \int_0^t \frac{1}{\sum_{i=1}^n A^i(s)\exp\left(\widehat{\beta}' Z^i(s)\right)} d\overline{N}(s).
$$

Finally, we need the following additional functions derived from the data:

$$
\begin{aligned}
\mathcal{I}(\beta) &= \int_0^1 \left\{\frac{\sum_{i=1}^n A^i(s) Z^i(s)^{\otimes 2}\exp\left(\beta' Z^i(s)\right)}{\sum_{i=1}^n A^i(s)\exp\left(\beta' Z^i(s)\right)} - \left[\frac{\sum_{i=1}^n A^i(s) Z^i(s)\exp\left(\beta' Z^i(s)\right)}{\sum_{i=1}^n A^i(s)\exp\left(\beta' Z^i(s)\right)}\right]^{\otimes 2}\right\} d\overline{N}(s). \\
J(\beta, t) &= -\int_0^t \left[\frac{\sum_{i=1}^n A^i(s) Z^i(s)\exp\left(\beta' Z^i(s)\right)}{\left\{\sum_{i=1}^n A^i(s)\exp\left(\beta' Z^i(s)\right)\right\}^2}\right] d\overline{N}(s).
\end{aligned}
$$

Proposition 11 *Assume A1-A2, B1-B2, Σ is positive definite,*

$$
\mathbb{P}\left\{A^i(t) = 1,\ \forall t \in [0,1]\right\} > 0
$$

and the covariates satisfy

$$
\mathcal{H}_l^m = \vee_{i=l}^m \sigma\left\{\eta_i, \eta_{-i}, X^i(s), X^{-i}(s), R^i(s) \middle| 0 \leq s \leq T\right\}.
$$

$$
\sum_{n>0} n^{-1}\alpha(n) < \infty.
$$

[14] For a column vector X, $X^{\otimes 2}$ represents XX'.

Then the results of AG (1982) Section 3 hold. In particular

$$n^{1/2}\left(\widehat{\beta}-\beta_0\right)\Rightarrow N\left(0,\Sigma^{-1}\right),$$

$$n^{-1}\mathcal{I}\left(\widehat{\beta}\right)\xrightarrow{p}\Sigma.$$

$n^{1/2}\left(\widehat{\beta}-\beta_0\right)$ *and the process*

$$n^{1/2}\left[\widehat{\Lambda}(t)-\Lambda_0(t)\right]+n^{1/2}\left(\widehat{\beta}-\beta_0\right)'\int_0^t e(\beta_0,s)\,h_0(s)\,ds \tag{24}$$

are asymptotically independent. (24) is asymptotically distributed as a Gaussian martingale with variance function

$$\int_0^t \frac{h_0(s)}{s^{(0)}(\beta_0,s)}ds.$$

Finally,

$$\sup_{t\in[0,1]}\left\|J\left(\widehat{\beta},t\right)-\int_0^t e(\beta_0,s)\,h_0(s)\,ds\right\|\to^p 0.$$

Proof. See Appendix C. ■

In the *i.i.d.* case, the estimates $\widehat{\beta}$ have asymptotic efficiency properties - See Anderson et al. (1994). The above result shows that, in the dependent case, the estimates have the same asymptotic variance as in the independent case. This powerful result is driven by the martingale structure of the set up. However, this result shows that previous work assuming *i.i.d.* samples has asymptotically justified confidence intervals when the dependence of covariates across observations satisfies the conditions in Proposition 11. A block/step sampling setup with no contagion gives the same results as Proposition 11 with appropriate adjustments in the assumptions. Therefore, confidence bands using AG (1982) estimates are asymptotically justified when common processes are present. It is likely that in finite samples dependence will affect the estimator's performance. This is the same issue raised above in the kernel context. Deriving different estimators of the variance is left to future research.

Proposition 11 shows that $\widehat{\Lambda}(t)$ converges to $\Lambda_0(t)$ uniformly at a $n^{1/2}$ rate. This is a strong result and it is not possible to improve on the $n^{1/2}$ rate. In a certain sense, $\widehat{\Lambda}(t)$ is the optimal estimator of $\Lambda_0(t)$ (see Johansen (1983)). It is possible to transform $\widehat{\Lambda}(t)$ to estimate $h_0(t)$. This may slow down the rate of convergence, see Anderson et al. (1994) pg. 507 for an example using kernels. $\widehat{\Lambda}(t)$ is piecewise constant, which is not optimal for frailty estimation as outlined below. The AG (1982) estimates are not justified for the frailty case outlined in Section 4.4. The estimation approach of this subsection needs to be modified to accommodate frailty. I leave this to future research.

4.3 Point Process Likelihood Specification

The estimators of the previous subsection are optimal in a number of ways. However, they estimate the integral of the baseline hazard $\int_0^t h_0(s)\,ds$ instead of $h_0(s)$. If the main interest is in an estimate of β_0, this is not an issue. When frailty is introduced in the next subsection, we will see that an estimate of

$h_0(s)$ is of substantial interest. Ad hoc transforms of the previously presented estimator $\widehat{\Lambda}(t)$ may not preserve efficiency properties. There is little theoretical research on the efficiency of such transformed estimators. In addition, the estimation results of AG (1982) can not handle the case where observations have different baseline hazards. This is also of interest when frailty is present. In this subsection, estimators are developed which directly estimate $h_0(s)$. The estimation approach and the related proofs are then modified to handle the frailty case in the next subsection. This subsection can be thought of as a prelude to frailty.

In this subsection, the proportional hazard model described in subsection 4.1 is estimated using a point process likelihood approach. The baseline hazard $h_0(t)$ is estimated directly using sieves. The results of Karr (1987) are extended to semi-nonparametric estimation where the impact of covariates is estimated instead of assumed known. Point process likelihoods have been used elsewhere in the literature. See Anderson et al. (1994) or Martinussen and Scheike (2010) for surveys. See Brémaud (1981) section IV.2 for more specifics on the following point process likelihood. Likelihood estimators often have efficiency properties. However, the efficiency properties of the following estimator is an open question.

Let $\mathcal{H}$ be the set of strictly positive continuous functions on $[0,1]$ and $\beta \in [a_1,b_1] \times \cdots \times [a_q,b_q]$. The underlying probability space $(\Omega, \mathcal{F}, P)$ supports a Poisson process with intensity 1, the covariate processes $Z^i(s)$ and the corresponding standard exponential random variables η_i. For any $(h,\beta) \in \mathcal{H} \times [a_1,b_1] \times \cdots \times [a_q,b_q]$ the following is the Radon-Nikodym derivative which changes the measure on an underlying point process which follows a standard Poisson process with $\lambda = 1$, into a single-spell point process with intensity $h(s)\exp\left(\beta' Z^i(s)\right)\mathbf{1}_{\{\tau_i \geq s\}}$.

$$\frac{d\widetilde{P}}{dP}(h,\beta) = \begin{cases} \exp\left\{\int_0^1 \left[1 - h(s)\exp\left(\beta' Z^i(s)\right)\mathbf{1}_{\{\tau_i \geq s\}}\right] ds\right\} & \tau_i > 1 \\ h(\tau_i)\exp\left(\beta' Z^i(\tau_i)\right)\exp\left\{\int_0^1 \left[1 - h(s)\exp\left(\beta' Z^i(s)\right)\mathbf{1}_{\{\tau_i \geq s\}}\right] ds\right\} & \tau_i \leq 1 \end{cases}. \tag{25}$$

The log-likelihood is

$$\log\left[\frac{d\widetilde{P}}{dP}(h,\beta)\right] = \int_0^1 \log\left[h(s)\exp\left(\beta' Z^i(s)\right)\right] dN_s^i + \int_0^1 \left[1 - h(s)\exp\left(\beta' Z^i(s)\right)\mathbf{1}_{\{\tau_i \geq s\}}\right] ds.$$

Assume no covariate in $Z^i(t)$ is a linear combination of the remaining covariates and that all $Z^i(s)$ are random. By standard arguments (see, for example, van der Vaart (1998) Lemma 5.35) the expected value of the log-likelihood is uniquely maximized at (h_0, β_0) when the observed point process has the hazard $h_0(s)\exp\left(\beta_0' Z^i(s)\right)\mathbf{1}_{\{\tau_i \geq s\}}$. This is because each choice of $(h,\beta) \in \mathcal{H} \times [a_1,b_1] \times \cdots \times [a_q,b_q]$ corresponds to a unique intensity process $h(s)\exp\left(\beta' Z^i(s)\right)\mathbf{1}_{\{\tau_i \geq s\}}$ and therefore a unique change of measure. Specifically

$$H(h,\beta) = \mathbb{E}_{(h_0,\beta_0)}\left\{\log\left[\frac{d\widetilde{P}}{dP}(h,\beta)\right]\right\},$$

$$H(h_0,\beta_0) > H(h,\beta) \qquad (h,\beta) \neq (h_0,\beta_0).$$

This identification allows us to make the log likelihood the basis of estimation. Let Θ_n be a space of functions depending on n. More specifics on the required sieve spaces Θ_n are given below. Define our estimator as

$$Q_n\left(\widehat{h},\widehat{\beta}\right) \geq \sup_{\beta,h\in\Theta_n} Q_n(h,\beta) + o_p(1)$$

where the criterion function is

$$Q_n(h,\beta) = \frac{1}{n}\sum_{i=1}^{n}\left[\int_0^1 \log\left[h(s)\exp\left(\beta' Z^i(s)\right)\right] dN_s^i + \int_0^1 \left[1 - h(s)\exp\left(\beta' Z^i(s)\right)\mathbf{1}_{\{\tau_i \geq s\}}\right] ds\right].$$

We make the high level assumption B3 on the covariate processes. This assumption is presented in Appendix C. Here, I give a number of examples of covariate processes which satisfy B3. Note that scaling of the processes may be required to fit the supports into $[0,1]^q$, but this is not a substantive issue. Any of the types of covariate processes presented below can be combined while still satisfying B3 (provided no covariate process is a linear combination of the others and all processes are random). A weak additional condition is needed when combining Examples. This is discussed in Example 16. There are many other possible covariate processes that satisfy B3.

Example 12 *Let ξ^i be a continuously distributed random variable with compact support or a random variable with finite support. Define the covariate process as $Z^i(s) = \xi^i + ct$ where $c \in \mathbb{R}$. This process satisfies assumption B3. Note that we can choose $c = 0$. In this case, the covariate processes reduce to the static covariate case. Assumption B3 can be seen as a natural extension of simple continuous or discrete distribution assumptions in the case where covariates do not change through time.*

Example 13 *Let $N^i(s)$ be a point process constructed as in Section 2 but where there may be an infinite number of ordered random times. Each random time is represented by ϕ_j. Let the covariate process have an initial distribution ξ_0^i with support contained in $[0,1]$. At each ϕ_j, a new value is drawn ξ_j^i with support contained in $[0,1]$. This distribution may be dependent on all previous draws $\xi_{j-1}^i,\ldots,\xi_0^i$ and/or the positions of the previous ϕ_j. We assume all ξ_j^i have either a continuous distribution or a finite distribution. The covariate process is*

$$Z^i(s) = \xi_0^i \mathbf{1}_{\{0\leq s\leq\phi_1\}} + \sum_{j=1}^{\infty} \xi_j^i \mathbf{1}_{\{\phi_j < s \leq \phi_{j+1}\}}.$$

If, $0 < \mathbb{P}\left\{\phi_j \in (a,b)\right\} < 1$ for all j and all intervals (a,b) contained in $[0,1]$, then Assumption B3 holds. Weaker assumptions are possible, but have more involved statements. An example of these types of covariates is movement to and from employment. This type of covariate can be used to model movement to and from any number of finite states. Another simple example which satisfies B3 is covariates which only change values at fixed times, such as weekly or quarterly.

Example 14 *We consider the same underlying point process as in Example 13. Let $N^i(s)$ be a point process constructed as in Section 2 but where there may be an infinite number of ordered random times. Each random time will be represented by ϕ_j. Let the covariate process have an initial distribution ξ_0^i with support contained in $[0,1]$. For $c \in \mathbb{R}$, define*

$$Z^i(s) = \left(\xi_0^i + cs\right)\mathbf{1}_{\{0\leq s\leq\phi_1\}} + \sum_{j=1}^{\infty}\left(cs - \phi_j\right)\mathbf{1}_{\{\phi_j < s \leq \phi_{j+1}\}}.$$

This could be used to model the length of time an observation spends in a particular state. When states change at ϕ_j*, then the length of time resets to zero. If,* $0 < \mathbb{P}\left\{\phi_j \in (a,b)\right\} < 1$ *for all* j *and all intervals* (a,b) *contained in* $[0,1]$*, then Assumption B3 holds.*

Example 15 *Assume* $X^j(s)$ *has a form taken from one of the previously given examples. Define the covariate as*

$$Z^i(s) = \exp\left(-\lambda\left(s-\tau_j\right)\right) f\left[X^j\left(\tau_j\right)\right] \mathbf{1}_{\{\tau_j < s\}},$$

where f *is a bounded continuous function and* $\lambda \in [0,\infty)$*. If* $0 < \mathbb{P}\left\{\tau_j \in (a,b)\right\} < 1$ *for all intervals* (a,b) *contained in* $[0,1]$*, then Assumption B3 holds. This form allows the affect of other defaults to decay through time. The size of the initial impact is allowed to depend on covariates of the other observation* j*. This is of interest in, for example, credit default situations. This form is used in a credit default context by Azizapour et al. (2011) and goes back to at least to Hawkes (1971). In the following theorem, we must choose a fixed* λ*. Further work must be done to allow for direct estimation of* λ*. However, we estimate a coefficient for the process* $Z^i(s)$*, which improves the flexibility of the model. Note further we can choose* $\lambda = 0$*, so the effect of another default has a fixed and permanent impact. The random time may be more complicated, such as the kth element of a group to default.*

Example 16 *Assume* $X^i(t)$ *is composed of* d *covariate processes, each of which has a form given in the previous examples. Assume further that there is a set of* k *other random times* $\mathbf{1}_{\{\tau_j < t\}}$*,* $j = 1, \ldots, k$*. Each of these additional random times has probability of default in between* 0 *and* 1 *over any subinterval of* $[0,T] = [0,1]$*. Assume that over any time interval, there is a positive probability that one of the covariate processes has a discontinuity while none of the others do. Similarly, we assume that over any time interval there is a positive probability of no discontinuity. This will cause problems for variables that only change at fixed times. We focus on the above assumptions for simplicity. If* g *is a bounded continuous function, then*

$$\left(X^i(t), g\left(X^i(t), \mathbf{1}_{\{\tau_1 < t\}}, \ldots, \mathbf{1}_{\{\tau_k < t\}}\right)\right)$$

satisfies B3.

Proposition 17 *We make Assumptions (B1)-(B3). Choose a sequence of sieve spaces* Θ_n *with the number of basis functions used being* J_n *for* n *observations. Let there exist a sequence* $h_n \in \Theta_n$ *such that* $h_n \to^{L^1} h_0$*.* $h_0 \subset \overline{\mathcal{H}}$ *where all functions in* $\overline{\mathcal{H}}$ *are continuous and bounded above and below by known fixed constants* $C_{\min}$*,* $C_{\max}$*. Assume further that for* $h \in \Theta_n$

$$C_{\min} \leq h \leq C_{\max}, \tag{26}$$

$$\left|h'\right| \leq K_n, \tag{27}$$

where $K_n = O\left(n^{1/4-\eta}\right)$ *for a small* $\eta > 0$*. For the system of* σ*-fields*

$$\mathcal{K}_l^m = \vee_{i=l}^m \sigma\left\{\eta_i, \eta_{-i}, X_t^i, X_t^{-i}, R_t^i \middle| 0 \leq t \leq T\right\}, \tag{28}$$

assume the mixing condition

$$\sum_{n>0} n^{-1}\alpha(n) < \infty. \tag{29}$$

holds. Then

$$\widehat{\beta} \to \beta_0$$

$$\widehat{h} \to^{L^1} h_0$$

$\mathbb{P}_{\alpha_0} - a.s.$

Proof. See appendix C. ■

The sieve space containing the largest number of potential functions for approximation is the set of all functions satisfying (26)-(27). However, these functions do not have an obvious set of basis functions. In applications, a sieve with know basis functions whose coefficient can be constrained to satisfy (26)-(27) would be used. Theoretically, we would simply take the intersection of a sieve space with the set of functions satisfying (26)-(27). However, in applications there will be issues of implementation depending on the chosen basis functions. Whatever sieves are chosen, the ability to control the first derivative will be important in implementation. For example, Cardinal B-Splines can do this easily. See de Boor (2001) or Chui (1992) for more on Cardinal B-Splines and Chen (2007) for a comprehensive account of sieve estimation.

For any contagious variables in the set up of subsection 4.1, the following relationship holds

$$\mathcal{H}_l^m = \vee_{i=l}^m \sigma \left\{ \eta_i, Z^i(s) \middle| 0 \le s \le T \right\} \subset \vee_{i=l}^m \sigma \left\{ \eta_i, \eta_{-i}, X_t^i, X_t^{-i}, R_t^i \middle| 0 \le t \le T \right\}. \tag{30}$$

This is because, if we know the variables $\eta_i, \eta_{-i}, X_t^i, X_t^{-i}, R_t^i$, we can derive default times and therefore $V^i(s)$. A consequence of (30) is that mixing conditions may be put on the covariates X_t^i, X_t^{-i}, R_t^i directly. We avoid the need to assume mixing conditions on $V^i(s)$.

The relationship (30) is only useful because contagion is restricted to be within groups. If the hazard rate of observation i is affected by the defaults of all other observations, a similar upper bound on information would need to contain all covariate process from all observations. This will prevent any mixing condition from holding. In cases like these, mixing conditions have to be placed on $\mathcal{H}_l^m$ in (30). The mixing properties of $\left\{V^i(s)\right\}_{i=1}^{\infty}$ will be a direct consequence of more primitive underlying variables. What type of mixing conditions $\left\{V^i(s)\right\}_{i=1}^{\infty}$ inherits when contagion is not confined to be within groups is beyond the scope of this paper.

Proposition 17 assumes both groups of equal size and covariates with equivalent distributions across observations. Symmetry is also implicitly assumed by requiring the function g and the true underlying β_0 to be equivalent across observations. All of these conditions can be relaxed. For non-symmetric groups, the values of the true coefficients for each member and the contagious variables can be different across group members. This situation violates assumption A2, but consistency will still hold because corresponding members of different groups satisfy A2. If the members of the groups are not symmetric, one must be able to tell who the corresponding members are across groups. For example, if the groups correspond to an employer and employee pair, one must be able to tell which is which in each group when conducting estimation. The groups may also be of different sizes. This can result in a violation of assumption A2 as well. Consistency still holds provided the number of potential group sizes is finite. Each group size must constitute a fixed proportion of the sampled data asymptotically. Proofs of these extensions are omitted for brevity.

4.4 Frailty and Likelihood Estimation

The final model we examine contains our notion of frailty. This model was initially proposed as a way to better explain the observed clustering of corporate defaults using hazard models. Corporate defaults are clustered around periods of financial crisis and recession. This clustering is difficult to explain with standard hazard models. Das et al. (2007) provide some statistical evidence that hazard models with common baseline hazard functions are rejected when using data on US corporate default. A solution proposed in Duffie et al. (2009) is the frailty model described below. In this model, observations at risk of default over different calendar time intervals have different baseline hazard functions. As a result, the observed clustering can be captured. One interpretation of this model is that there is time dependent model misspecification. Another, more interesting interpretation is that a global unobserved risk factor is impacting hazard rates. More on this interpretation below. There is no reason to restrict the use of this model to corporate default. Any random economic event situation where changing macroeconomic conditions are important could potentially use this notion of frailty.

The sampling set up in this situation is an extension of block/step sampling. Recall that in block/step sampling observations begin to be at risk at calendar times $k\delta$ where $\delta > 0$ and $k \in \mathbb{N}_0$. Again, we allow adjacent blocks to overlap. Previously, we assumed a finite number of observations correspond to each calendar time and therefore each $k \in \mathbb{N}_0$. This allows for a natural ordering on observations. In the frailty case, we need to assume an increasingly large number of observations for each $k \in \mathbb{N}_0$ in the sampling. We now index each observation by ji: j corresponds to the calendar time the observation starts at $j\delta$; i is the number of the observation starting at $j\delta$. The situation is analogous to a panel. We assume that both j and i approach infinity in the sampling scheme.

I will write G^j for $j\delta$ to keep consistent with previous notation. The sampling is indexed by n. We assume n observations per G^j where we only observe the first $k(n)$ calendar times. $k(n) \to \infty$ as $n \to \infty$. The relevant covariates include observation specific covariates $X^{ji}(t)$ and common covariates $Y(t)$. We write $Z^{ji}(t) = \left(X^{ji}(t), Y\left(G^j + t\right)\right)$.

This sampling scheme is stringent, requiring a large number of observations per block. Situations where observations start to be at risk of default at regular intervals are more likely to satisfy the required conditions. Examples might include weakly, monthly or quarterly reported start times for observations. The assumption that there are the same number of observations per block is made only for notational simplicity. The somewhat rigid requirement that a large number of observations must start at precisely the same time can likely be relaxed. A minimal requirement is that a large number of observations are at risk of default over any fixed calendar time, but observations might not start at the same time. I leave extensions of the sampling scheme to future research.

Now, I define frailty and incorporate it into a Cox proportional hazard model. Frailty is a strictly positive continuous real valued function $h_0(t)$ defined on $[0, \infty)$. Below, we will assume this path is the realization of an independent stochastic process. This assumed stochastic process structure is not required for estimation, but is needed for forecasting. The frailty function takes the place of the baseline hazard in the block/step sampling case. For an observation ji, that observation's baseline hazard is the realization of the frailty process over the calendar time interval $\left[G^j, G^j + T\right]$. As a result, observations at risk over different calendar time intervals have different baseline hazards. This setup rules out standard estimators, such as those in AG (1982), because they require all observations to have the same baseline hazard. We write the baseline hazard corresponding to the jth block as h_0^j. The hazard rate for observation ji can now be written as

$$h_0\left(G^j + t\right)\exp\left\{\beta_0' Z^{ji}\left(t\right)\right\}. \tag{31}$$

Random times with this hazard rate can easily be constructed as in Section 2 above. It can be shown, with a proof almost identical to Lemma 1, that this set up has a martingale structure defined on $[0,\infty]$. The processes M_t^{ji} are defined to be zero before G^j. They have the hazard rate (31) on the interval $\left[G^j, G^j + T\right]$. After $G^j + T$, $M_t^{ji} = M_{G^j+T}^{ji}$.

The path $h_0\left(t\right)$ can be interpreted as a global latent risk factor. When the path has large values, defaults are more prevalent. When $h_0\left(t\right)$ is low, defaults are suppressed. With this interpretation, the path of $h_0\left(t\right)$ can be thought of as randomly propagating. Duffie et al. (2009) give the model additional structure by assuming $h_0\left(t\right)$ is the realization of a mean-reverting diffusion. We first present the specific Duffie et al. (2009) model. Then we show it can easily be generalized.

Assume the frailty path $h_0\left(t\right)$ is the realization of the following mean-reverting diffusion,

$$\begin{aligned} dS_t &= -\kappa S_t dt + \sigma dB_t, \\ h_0\left(t\right) &= \exp\left(S_t\right). \end{aligned}$$

[15]This is the specification in Duffie et al. (2009). The diffusion is initialized to be stationary. This specific choice of diffusion is not needed, others are possible. As a result, the idea is more general than the specification given above. Duffie et al. (2009) take a Bayesian approach to estimation with this type of frailty. However, the consistency result in Duffie et al. (2009) depends on an *i.i.d.* assumption which is not satisfied in the cases we are interested in.

In our estimation, we consider $h_0\left(t\right)$ as a realized path from some diffusion specification. We estimate $h_0\left(t\right)$ directly using point process likelihoods and sieves. This estimate recovers the entire path of $h_0\left(t\right)$ asymptotically. Consistency is achieved despite the fact that $h_0\left(t\right)$ is nowhere differentiable. Once a first stage estimate of $h_0\left(t\right)$ is obtained, it can be used in a second stage to estimate parameters characterizing the underlying frailty diffusion. There is an extensive literature on estimation of diffusions from a continuous record. See, for example, Prakasa Rao (1999) or Kutoyants (2004). Estimates of the frailty process can be used to project default probabilities in the future.

We define our estimator similarly to the one in Subsection 4.3. Again, Θ_n are spaces of functions increasing in size with n:

$$Q_n\left(\widehat{h}, \widehat{\beta}\right) \geq \sup_{\beta, h^j \in \Theta_n, j=1,\ldots,k(n)} Q_n\left(h, \beta\right) + o\left(1\right)$$

[15]Below, we will assume the frailty path is bounded above and below. This diffusion specification will violate that assumption. The reason for the boundedness assumption is to alleviate a technical condition. A more general result which does not require boundedness is possible at the expense of stronger conditions on the covariates. Another possibility is to assume a boundary on this diffusion specification at extremely small and large values. When the diffusion hits the boundary, it is reflected back. Specifications of this type are shown to exist in the literature. See Stroock and Varadhan (1979).

where the criterion function is

$$Q_n(h,\beta) = \sum_{j=1}^{k(n)} \frac{1}{n} \sum_{i=1}^{n} \left[\int_0^1 \log\left[h^j(s)\exp\left(\beta' Z^{ji}(s)\right)\right] dN_s^{ji} + \int_0^1 \left[1 - h^j(s)\exp\left(\beta' Z^{ji}(s)\right) \mathbf{1}_{\{\tau_{ji} \geq s\}}\right] ds \right].$$

Portions of the realized frailty h_0 which overlap in different calendar time blocks $\left[G^i, G^i + T\right]$ are restricted to be the same function in estimation. For example, if $[0,T]$ and $\left[\frac{1}{2}T, \frac{3}{2}T\right]$ are calendar time blocks, then the estimate corresponding to the first block over $\left[\frac{1}{2}T, T\right]$ must be the same function as the estimate corresponding to the second block over the same interval. The ability to easily implement this restriction depends on the chosen Θ_n. Some choices of Θ_n may not allow for this. We must rule these choices out. Cardinal B-Splines can achieve this restriction easily. The estimator of h_0 from the criterion function is essentially an estimator of the entire frailty path over the observed blocks.

Corollary 18 *We make Assumptions (B1)-(B3). Choose a sequence of sieve spaces Θ_n with the number of basis functions used being J_n for n observations. For all $j \in \mathbb{N}$, let there exist a sequence $h_n^j \in \Theta_n$ such that $h_n^j \to^{L^1} h_0\left(G^j + t\right)$ where the relevant interval is $[0,T]$ normalized to $[0,1]$. For all $j \in \mathbb{N}$, $h_0 \subset \overline{\mathcal{H}}$ where all functions in $\overline{\mathcal{H}}$ are continuous and bounded above and below by known fixed constants $C_{\min}$, $C_{\max}$. Assume further that for $h \in \Theta_n$*

$$C_{\min} \leq h \leq C_{\max}, \tag{32}$$

$$\left|h'\right| \leq K_n. \tag{33}$$

Let $k(n)$ and K_n be chosen such that the conditions (74)-(76) in Appendix C are satisfied and $K_n \to \infty$ at a rate slower than or equal to $o\left(\left(k(n)/n\right)^{1/2}\right)$. Let there exist $\widetilde{h}_n^j \in \Theta_n$ for each j (where functions with overlapping calendar times must agree on the overlapping intervals) such that

$$\sum_{j=1}^{k(n)} \left\{ H\left(h_0^j, \beta_0\right) - H\left(\widetilde{h}_n^j, \beta_0\right) \right\} \to 0. \tag{34}$$

Then

$$\widehat{\beta} \to \beta_0,$$

$$\widehat{h}^j \to^{L^1} h_0^j,$$

for all j, $\mathbb{P}_{\alpha_0} - a.s.$

Proof. See Appendix C. ■

Note that any sieve used must contain once differentiable functions. This may seem like a contradiction in the frailty case, where we are trying to estimate a nowhere differentiable continuous function $h_0(t)$. However, in the limit, functions in the sieve space may have arbitrarily large derivatives. Paths of diffusions may be thought of as continuous functions with infinite derivatives. So we are able to achieve L^1 convergence. We may assume $h_0(t) = S(t)$ is itself the path of a diffusion or that the observed path in the hazard is $h_0(t) = \exp(S(t))$. This may facilitate some parametric specifications of the unobserved diffusion. We restrict the path $h_0(t)$ to be bounded. The underlying diffusion must likewise be restricted.

The assumptions (74)-(76) are high level. To satisfy these conditions, we must consider the dependence between X^{ij} within blocks and across blocks. In addition, the temporal dependence of $Y(t)$ and its dependence with the processes $X^{ij}(t)$ must be accounted for. All of these interactions must be considered because it is not possible to recover $h_0^j(t)$ by sampling only from block j. When we only sample within one block $\left[G^j, G^j + T\right]$, there is no sampling of the common processes $Y(t)$. The effect of the common processes $Y(t)$ obscures estimation of $h_0^j(t)$. Only by sampling across different blocks can we effectively sample from the common processes $Y(t)$ and recover their coefficients. In their current form, the methods of AG (1982) do not apply to this situation.

In the point process likelihood approach, the needed conditions (74)-(76) are a natural extension of those used in Theorem 32 from Appendix C. However, exact specification of the types of dependence between all processes that facilitate these conditions is not obvious. One issue is that there is no simple ordering of the sample. Dependence has to be controlled in two dimensions, within blocks and across blocks. As almost sure convergence is required and we are dealing with an array structure, almost sure convergence of arrays will likely be necessary. See Liebscher (1996). I leave exact characterization of the needed dependence to future research.

A mixed proportional hazard model with this type of frailty is not identified. This is because a scaling of $h_0(s)$ is required for identification (see Van den Berg (2001)). As $h_0(s)$ is an unknown path of a diffusion, it is impossible to have such a scaling.

We do not derive a rate of convergence for the estimator in Proposition 17. Other assumptions on the path, such as Hölder continuity, are required for a rate to be derived. We note that many diffusion paths are Hölder continuous (see Röckner and Zhang (1996)).

4.5 Specification Testing

All of the hazard models given above satisfy the condition that M_t^i or $M_t^{ji}$are martingales with filtrations equal to (or similar to) those derived in Section 2. As a result of this, all of these specifications can potentially satisfy the conditions of Proposition 3 required for a martingale FCLT. In the following, we choose $g_i^{(n)}(s) = 1/\sqrt{n}$ in all cases.

Corollary 19 *If, for all $t \in [0, T]$,*

$$\frac{1}{n}\sum_{i=1}^{n}\Lambda_t^i(t) \to^p \sigma^2(t), \tag{35}$$

then

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}M^i(t) \Rightarrow^{D[0,T]} M(t), \tag{36}$$

where $M(t)$ is a continuous Gaussian martingale with variance structure $\sigma^2(t)$.

Lemma 5 holds for the processes $\Lambda_t^i(t)$. Therefore, mixing conditions as outlined in Section 3 above can be transferred to $\Lambda_t^i(t)$ and used to prove (35). This is further facilitated by the fact that, as t decreases, the same σ-field used to characterize dependence when $t = T$ can be used to describe the mixing conditions on $\Lambda_t^i(t)$. The result only has to be proven for $t = T$. In the situation of Subsection 4.4 the result is further complicated by the fact that the $\Lambda_t^{ji}(t)$ are constructed using different baseline hazards across blocks.

This result can be used as the basis of specification testing. An approach would be first to estimate the model. Then, (35) can be constructed from the data and used as the basis of confidence intervals for (36). Refinements on this implementation are beyond the scope of this paper. See Martinussen and Scheike (2010) for more discussion.

5 Conclusion

In this paper, I present hazard models with a number of dependence properties between observations. These models are designed to capture the reality of dependence between random economic events which manifest in countless different economic relationships. I first propose a model with dependent covariates determining the hazard rate of observations. These covariates can be specific to observations, for example housing prices in a model of mortgage default. The covariates can also be macroeconomic, and therefore impact all observations at risk of default. This type of macroeconomic covariate captures dependence between observations resulting from changing global economic conditions. Examples include the GDP growth rate, unemployment rate or three-month US treasury rate.

I propose a form of sampling where observations are at risk of default starting at different calendar times. I call this sampling "block/step" sampling. Assuming block/step sampling, conditions are derived under which the underlying hazard function can be consistently estimated nonparametrically using a kernel approach. A uniform rate of convergence and CLT are derived for the nonparametric estimator. Under further assumptions, the curse of dimensionality is circumvented and estimates converge nearly at a $\sqrt{n}$ rate. These results show that, given appropriate sampling, flexible estimates of the impact of macroeconomic conditions on hazard rates can be achieved under relatively weak conditions.

A second model proposed incorporates contagion into a hazard model. The covariates of one observation are allowed to depend on the defaults of other, related observations. This is intended to capture direct economic relationships between observations. Variables can be more general than simple indicators of other observations defaulting. Other observed covariates can augment indicators of default to provide a better description of how hazard rates are affected. These types of variables are used in a Cox proportional hazard model. One goal of estimation is to recover the parametric impact of the variables constituting contagion. This characterizes the magnitude that a default of one observation has on hazard rates of observations that are still at risk.

Under suitable α-mixing assumptions, standard semi-nonparametric estimation techniques for the Cox model are shown to hold. These results include a CLT for the coefficients on the covariates, including those covariates which define contagion. I additionally show how to estimate the model using a point process likelihood and sieves. This gives a direct estimator of the baseline hazard. The approach developed here is the building block of the more complicated frailty situation.

A final model incorporates a global unobserved risk factor into a Cox proportional hazard model. The baseline hazard function is no longer the same for observations at risk over different intervals of calendar time. Instead, there is a strictly positive function of $[0, \infty)$ called frailty. The interval of calendar time an observation is at risk corresponds to an interval of the frailty path. The portion of the frailty path from the calendar time an observation is at risk of default corresponds to the baseline hazard of that observation. This frailty path can be interpreted as a global unobserved risk factor. Frailty is incorporated into hazard models to improve the ability to achieve significant clustering of defaults. This type of clustering is observed in applications.

By extending block/step sampling to allow for an increasingly large number of observations in cross sections, I show how the frailty path can be estimated nonparametrically. This is done using point process likelihoods and sieve estimation. In addition, the parametric impact of the covariates can also be consistently estimated. These results hold in the presence of macroeconomic covariates described above. Again, only a general α-mixing condition is needed to control dependence in covariates.

Finally, I assume that the frailty path is a realization of a diffusion process. Once the frailty path is estimated, methods which use a continuous record to estimate diffusions can be used to characterize the underlying stochastic process that frailty follows. Estimate of the frailty diffusion can then be used in forecasting. This will increase the accuracy of out-of-sample predictions.

6 Appendix

A Removing the Upper Bound T

Because we are conditioning on the entire path of $Y(t)$ in the underlying filtration $\overline{\mathcal{G}}_t^n$ used in Proposition 3, the processes $g_i^{(n)}$ used to satisfy (1) above must be chosen carefully if we are to use the result in estimation. However, it is possible to satisfy (1) in a block/step sampling context. The key is that each $g_i^{(n)}$ and M_t^i are only affected by the fixed portion of the $Y(t)$ process $\left\{Y\left(G^i+t\right)\middle| t\in[0,T]\right\}$. As a result, even though the underlying filtration contains all the information from $Y(t)$, the affect of $Y(t)$ still washes out in the limit because different observations use different blocks of $Y(t)$.

It can be shown that (1) is equivalent to a similar conditional variance assumption used in Hall and Heyde (1980). These preliminary results are used in the next subsection. We assume the same notation as Section 3. Recall that we write $Z^i(t)=\left(t,X^i(t),Y\left(G^i+t\right)\right)$ and let $z=(t,x,y)\in[0,T]\times\mathcal{X}\times\mathcal{Y}$ be an interior point of the support. Define

$$g_i^{(n)}(s)=\frac{b^{(d+j+1)/2}}{n^{1/2}}K_b\left(z-Z^i(s)\right)$$

In order to use discrete time martingale methods as in Hall and Heyde (1980), we must define more filtrations. We leave out the needed completions of the filtrations for simplicity. These can easily be added. Define

$$\begin{aligned}\mathcal{G}_T^{nk} &= \mathcal{F}^n\vee\mathcal{H}_T^1\vee\cdots\vee\mathcal{H}_T^k \quad k=1,\ldots,n,\\ \mathcal{G}_t^{nk} &= \mathcal{F}^n\vee\mathcal{H}_T^1\vee\cdots\vee\mathcal{H}_T^{k-1}\vee\mathcal{H}_t^k \quad k=1,\ldots,n,\end{aligned}$$

and the random variables

$$S_k^n = \sum_{i=1}^{k} \int_0^T \frac{b^{(d+j+1)/2}}{n^{1/2}} K_b \left(z - Z^i(s)\right) dM_i(s) \qquad k = 1, \ldots, n.$$

Lemma 20 *For all $n \in \mathbb{N}$, for any $i = 1, \ldots, n$, $\int_0^t \frac{b^{(d+j+1)/2}}{n^{1/2}} K_b \left(z - Z^i(s)\right) dM_i(s)$ is a continuous-time martingale w.r.t. the right-continuous filtration $\mathcal{G}_t^{nk}$ over the interval $[0, T]$. This martingale has right-continuous paths.*

Proof. This follows from an application of Lemma 1 and Fleming and Harrington (1991) Theorem 2.4.4. ■

Proposition 21 *For each $n \in \mathbb{N}$, S_k^n is a mean-zero discrete time martingale w.r.t. the filtration $\mathcal{G}_T^{nk}$. In addition*

$$\begin{aligned} & \mathbb{E}\left[\left(\int_0^T \frac{b^{(d+j+1)/2}}{n^{1/2}} K_b \left(z - Z^k(s)\right) dM_k(s)\right)^2 \middle| \mathcal{G}_T^{n(k-1)}\right] \\ = & \mathbb{E}\left[\int_0^T \left[\frac{b^{(d+j+1)/2}}{n^{1/2}} K_b \left(z - Z^k(s)\right)\right]^2 d\Lambda^i(s) \middle| \mathcal{G}_T^{n(k-1)}\right] \\ = & \int_0^T \left[\frac{b^{(d+j+1)/2}}{n^{1/2}} K_b \left(z - Z^k(s)\right)\right]^2 d\Lambda^i(s) \end{aligned}$$

Proof. It follows from Lemma 20 that

$$\mathbb{E}\left[\int_0^T \frac{b^{(d+j+1)/2}}{n^{1/2}} K_b \left(z - Z^k(s)\right) dM_k(s) \middle| \mathcal{G}_T^{n(k-1)}\right] = 0.$$

A similar result holds for an arbitrary number of lags. As a result, S_k^n is a mean-zero martingale with respect to the given filtration. The second result holds by Lemmas 1 and 20 and Fleming and Harrington (1991) Corollary 1.4.2 and Theorem 2.4.3. ■

As a result of Proposition 21,

$$\sum_{k=1}^{n} \mathbb{E}\left[\left(\int_0^T \frac{b^{(d+j+1)/2}}{n^{1/2}} K_b \left(z - Z^k(s)\right) dM_k(s)\right)^2 \middle| \mathcal{G}_T^{n(k-1)}\right] \to^p \sigma$$

is equivalent to

$$\sum_{k=1}^{n} \int_0^T \left[\frac{b^{(d+j+1)/2}}{n^{1/2}} K_b \left(z - Z^k(s)\right)\right]^2 d\Lambda^i(s) \to^p \sigma$$

Therefore, condition (1) in Proposition 3 is equivalent to a Hall and Heyde (1980) type restriction on the conditional variance in this situation.

A.1 Another formulation

There is a second possible formulation of these ideas where we dispense with the restriction of default to intervals of length T. Assume, as before, that each observation i corresponds to a calendar time H^i at which it becomes at risk of default. Again, assume H^i is deterministic. We define its covariate processes on the interval $[0,\infty)$ as

$$\widetilde{X}^i = \begin{cases} \partial & t < H^i \\ X^i\left(t - H^i\right) & t \geq H^i \end{cases}.$$

Here, ∂ is an isolated point attached to $\mathbb{R}$. The common processes $Y(t)$ are defined as before on the interval $[0,\infty)$. We assume the hazard rate at ∂ satisfies $\alpha(\partial, y) = 0$ for all values y. $\widetilde{X}^i$ being at ∂ represents an observation not being "alive" before time H^i. The covariates $X^i(t)$ are defined on $t \in [0,\infty)$, so there is no upper or lower bound on time.

I correspondingly define random times $\widetilde{\tau}_i$ as

$$\begin{aligned} \widetilde{\Gamma}^i_t &\equiv \int_0^t \alpha\left(\widetilde{X}^i_s, Y_s\right) ds \\ \widetilde{\tau}_i &\equiv \inf\left\{ t \in \mathbb{R}_+ \middle| \, \widetilde{\Gamma}^i_t \geq \eta_i \right\} \end{aligned}$$

where η_i is an independent standard exponentially distributed random variable. Define analogous filtrations to Section 2:

$$\widetilde{\mathcal{F}}^n_t = \widetilde{\mathcal{F}}^n = \sigma\left\{ \widetilde{X}^i(u), Y(v) \middle| u, v \in [0,\infty), i = 1,\ldots,n \right\},$$

$$\widetilde{\mathcal{H}}^i_t = \sigma\left\{ \mathbf{1}_{\{\widetilde{\tau}_i \leq u\}} \middle| 0 \leq u \leq t \right\},$$

$$\widetilde{\mathcal{G}}^n_t = \widetilde{\mathcal{F}}^n \vee \widetilde{\mathcal{H}}^1_t \vee \cdots \vee \widetilde{\mathcal{H}}^n_t.$$

Finally, define

$$\widetilde{\Lambda}^i_t = \int_0^t \alpha\left(\widetilde{X}^i_s, Y_s\right) \mathbf{1}_{\{\widetilde{\tau}_i \geq s\}} ds.$$

Lemma 22 *(1) The analogous results to Lemma 1 follow in this case. In particular,* $\widetilde{M}^i_t = \mathbf{1}_{\{\widetilde{\tau}_i \leq t\}} - \widetilde{\Lambda}^i_t$ *is a martingale on* $[0,\infty)$ *with respect to the filtration* $\overline{\widetilde{\mathcal{G}}}^n_t$ *for all* $i = 1,\ldots,n$*. Here,* $\overline{\widetilde{\mathcal{G}}}^n_t$ *is the completion of the filtration* $\widetilde{\mathcal{G}}^n_t$ *as done in Lemma 1.*

Proof. Simple extension of Lemma 1. ■

Lemma 23 *(1) Let* $g^{(n)}_i(s)$ *be a locally bounded predictable process with respect to the filtration* $\overline{\widetilde{\mathcal{G}}}^n_t$*. Assume, for all* $t \in [0,\infty)$

$$\mathbb{E}\left\{ \int_0^t \left[g^{(n)}_i(s)\right]^2 \alpha\left(\widetilde{X}^i_s\right) \mathbf{1}_{\{\widetilde{\tau}_i > s\}} ds \right\} < \infty.$$

Then

$$\int_0^t g^{(n)}_i(s)\, d\widetilde{M}^i(s)$$

is a mean-zero continuous time martingale in t *over the interval* $t \in [0,\infty)$ *with respect to the filtration*

$\overline{\widetilde{\mathcal{G}}}_t^n$. *(2) Assume* $\int_0^t g_i^{(n)}(s)\, d\widetilde{M}^i(s)$ *is uniformly integrable*[16]*. Then,* $\lim_{t\to\infty}\int_0^t g_i^{(n)}(s)\, d\widetilde{M}^i(s)$ *exists a.s. We will write this random variable as* $\int_0^\infty g_i^{(n)}(s)\, d\widetilde{M}^i(s)$ *and it is measurable with respect to* $\widetilde{\mathcal{G}}_\infty^n \equiv \mathcal{F}$*, the entire underlying* σ*-field. Now,* $\int_0^t g_i^{(n)}(s)\, d\widetilde{M}^i(s)$ *is a martingale on* $t \in [0,\infty]$ *with respect to the filtration* $\widetilde{\mathcal{G}}_t^n$*.*

Proof. (1) Consequence of Fleming and Harrington (1991) Theorem 2.4.4. (2) Consequence of Protter (2005) Chapter 1 Theorem 12. ■

Lemma 23 allows us to define variables at infinity. Thus, under certain conditions, we will be able to dispense with restricting the time intervals over which default is possible. Again, we make analogous definitions for our situation based on the previous subsection in this appendix. As before, we leave off the needed completions for simplicity.

$$\widetilde{\mathcal{G}}_T^{nk} = \widetilde{\mathcal{F}}^n \vee \widetilde{\mathcal{H}}_\infty^1 \vee \cdots \vee \widetilde{\mathcal{H}}_\infty^k \quad k = 1, \ldots, n,$$

and the random variables

$$S_k^n = \sum_{i=1}^k \int_0^\infty \frac{b^{(d+j+1)/2}}{n^{1/2}} K_b\left(z - \widetilde{Z}^i(s)\right) d\widetilde{M}^i(s) \qquad k = 1, \ldots, n.$$

Proposition 24 *For each* $n \in \mathbb{N}$*,* S_k^n *is a mean-zero discrete time martingale w.r.t. the filtration* $\widetilde{\mathcal{G}}_T^{nk}$*. In addition*

$$\begin{aligned} & \mathbb{E}\left[\left.\left(\int_0^\infty \frac{b^{(d+j+1)/2}}{n^{1/2}} K_b\left(z - \widetilde{Z}^k(s)\right) d\widetilde{M}_k(s)\right)^2 \right| \widetilde{\mathcal{G}}_T^{n(k-1)}\right] \\ = \ & \mathbb{E}\left[\left.\int_0^\infty \left[\frac{b^{(d+j+1)/2}}{n^{1/2}} K_b\left(z - \widetilde{Z}^k(s)\right)\right]^2 d\widetilde{\Lambda}^i(s) \right| \widetilde{\mathcal{G}}_T^{n(k-1)}\right] \\ = \ & \int_0^\infty \left[\frac{b^{(d+j+1)/2}}{n^{1/2}} K_b\left(z - \widetilde{Z}^k(s)\right)\right]^2 d\widetilde{\Lambda}^i(s) \end{aligned}$$

Proof. This follows from almost the same proof as Proposition 21. ■

Now, a standard discrete time martingale CLT result from Hall and Heyde (1980) can be used to derive a CLT for the array S_k^n. In particular, we use Chapter 3, Corollary 3.1 from Hall and Heyde (1980). In a sense, we have turned a continuous time martingale problem into a discrete time martingale problem.

B Proofs for Sections 2 and 3

Proof (Lemma 1). Note that $\mathbf{1}_{\{\tau_i \le u\}}$ are right continuous processes. That $\mathcal{G}_t^n$ is a right continuous filtration follows from a modification of Brémaud (1981) A2, T26. In Brémaud's proof of that theorem,

[16] A stochastic process X_t is uniformly integrable if $\lim_{n\to\infty} \sup_{t\in[0,\infty)} \int_{\{|X_t|\ge n\}} |X_t|\, dP = 0$. Here, P is the underlying probability measure.

with each application of A1, T7, if the additional σ-fields

$$\mathcal{D}_s = \left\{ \sigma \left\{ Y\left(s\right) \right\}, \sigma \left\{ R^i\left(v\right) \right\},\ \sigma \left\{ X^i\left(u\right) \right\} \middle| \, 0 \le u, v \le T, s \in [0, \infty), i = 1, \ldots, n \right\}$$

are added to the class of σ-fields already used, the same proof holds exactly as before. That the completion of $\mathcal{G}_t^n$ is right continuous follows from Brémaud (1981) A2, T35.

That $\mathbf{1}_{\{\tau_i \le t\}} - \Lambda_t^i$ are $\mathcal{G}_t^n$-martingales for all $i = 1, \ldots, n$ is a straightforward application of Bielecki and Rutkowski (2004) Lemma 9.1.1. Completion of the filtration preserves the martingale property, see Dellacherie and Meyer (1980) VI.3. ■

Proof (Proposition 3). Because we have established a filtration that satisfies the usual conditions, we can apply the results of Liptser and Shiryayev (1980) using arguments similar to Ramlau-Hansen (1983) Proposition 4.2.1. Specifically, we need Corollary 2 and Remark 1 from Liptser and Shiryayev (1980). These results imply that, if for all $\epsilon > 0$,

$$\sum_{i=1}^{n} \int_0^T \left[g_i^{(n)}\left(s\right) \right]^2 \mathbf{1}_{\left\{ \left| g_i^{(n)}(s) \right| > \epsilon \right\}} d \left\langle M_i \right\rangle \left(s\right) \to^p 0$$

and if for some constant σ^2, with

$$X^n \equiv \sum_{i=1}^{n} \int_0^T g_i^{(n)}\left(s\right) dM_i\left(s\right)$$

we have

$$\left\langle X^n \right\rangle_T \to^p \sigma^2,$$

Then

$$X_T^n \Rightarrow N\left(0, \sigma^2\right).$$

For a definition of the angle bracket process $\left\langle M^i, M^j \right\rangle_t$ in this situation, see Appendix D, Fleming and Harrington (1991) or Dellacherie and Meyer (1980) VII.39. By Fleming and Harrington (1991) Theorem 2.4.3 and Lemma 2, for $i \neq j$,

$$\left\langle \int_0^{\cdot} g_i^{(n)}\left(s\right) dM_i\left(s\right), \int_0^{\cdot} g_j^{(n)}\left(s\right) dM_j\left(s\right) \right\rangle_t = \int_0^t g_i^{(n)}\left(s\right) g_j^{(n)}\left(s\right) d \left\langle M_i, M_j \right\rangle_s,$$

and for $i = j$,

$$\left\langle \int_0^{\cdot} g_i^{(n)}\left(s\right) dM_i\left(s\right), \int_0^{\cdot} g_j^{(n)}\left(s\right) dM_j\left(s\right) \right\rangle_t = \left\langle \int_0^{\cdot} g_i^{(n)}\left(s\right) dM_i\left(s\right) \right\rangle_t = \int_0^t \left[g_i^{(n)}\left(s\right) \right]^2 d \left\langle M_i \right\rangle_s.$$

Recall the polarization identity for $\left\langle \cdot, \cdot \right\rangle_t$ (Protter (2005) pg. 125),

$$\left\langle X + Y, X + Y \right\rangle = \left\langle X, X \right\rangle + \left\langle Y, Y \right\rangle + 2 \left\langle X, Y \right\rangle.$$

Also, by linearity and the definition of the angle bracket process $\left\langle \cdot, \cdot \right\rangle_t$, we have

$$\left\langle X_1, \sum_{i=1}^{n} Y_i \right\rangle = \left\langle X_1, Y_1 \right\rangle + \cdots + \left\langle X_1, Y_n \right\rangle$$

Therefore, using the above equalities,

$$\langle X^n \rangle_T = \sum_{i=1}^{n} \int_0^T \left[g_i^{(n)}(s) \right]^2 d \langle M_i \rangle (s).$$

This follows because $\langle M^i, M^j \rangle_t = 0$ for all $i \neq j$ by Fleming and Harrington (1991) Theorem 2.5.2 (see also their proof of Theorem 2.5.1). This implies

$$\int_0^t g_i^{(n)}(s)\, g_j^{(n)}(s)\, d \langle M_i, M_j \rangle_s = 0,$$

which is used above. By Fleming and Harrington (1991) section 2.5

$$\langle M_i \rangle_t = \Lambda_t^i.$$

This gives the result. The mixed normal case follows from the same results in Liptser and Shiryayev (1980). The FCLT result follows exactly the same arguments as the $t = T$ case and uses Liptser and Shiryayev (1980) corollary 2. ■

Proof (Theorem 4). Once Proposition 3 is verified and the assumptions (5)-(8) hold, the restriction of NL (1995) Theorem 1 to *i.i.d.* observations may be relaxed. Under (A2)-(A3), the expectations in the assumptions (5)-(8) are the same as in the *i.i.d.* case. Therefore, these expectations converge to the same values as those derived in NL (1995) Theorem 1. NL (1995) use the standard Bernstein inequality to show (5)-(8). However, once Proposition 3 is available, as long as (5)-(8) hold the results (9)-(11) continue to hold using the same proof as NL (1995) Theorem 1. In addition, the fact that

$$\sum_{i=1}^{n} \int_0^t \frac{b^{d+j+1}}{n} K_b^2 \left(x - Z^i(s) \right) dM_s^i$$

is a continuous time martingale in t is used to justify the use of Lenglart's inequality in the proof (see Shorack and Wellner (1986) pages 892-893). This holds in the dependent case using the information structure outlined in section 2. Finally, the result (10) is modified from the NL (1995) result by allowing for r times continuous differentiability of α and e. This allows for Taylor expansions with more terms. The term b^{-r} replaces the term b^{-2} which is found in the result from NL (1995). This type of modification is done in LNV (2003). Details are omitted. ■

Proof (Lemma 5). We show this only for the W^2 case. The W^1 case is simpler and the proofs are omitted. We assume $[0, T] = [0, 1]$ for notational convenience. The following set is equivalent to $\left\{ W^2 < M \right\}$ for any $M \in \mathbb{R}$.

$$\cup_{k=1}^{\infty} \cup_{q=1}^{k} \cup_{\{C_j\}} \cap_{j=1}^{q-1} \left[\begin{array}{c} \left\{ f\left(t, X(t), Y\left(G^i + t\right)\right) < C_j \middle| \, t \in \left(\frac{j-1}{k}, \frac{j}{k}\right] \right\} \\ \cap \left\{ f\left(t, X(t), Y\left(G^i + t\right)\right) < C_q \middle| \, t \in \left(\frac{q-1}{k}, \frac{q}{k}\right] \right\} \\ \cap \left\{ \int \sum_{j=1}^{q} C_i \mathbf{1}_{\left\{\left(\frac{j-1}{k}, \frac{j}{k}\right]\right\}}(s)\, ds < \eta_i \right\} \\ \cap \left\{ \int \sum_{j=1}^{q+1} C_i \mathbf{1}_{\left\{\left(\frac{j-1}{k}, \frac{j}{k}\right]\right\}}(s)\, ds \geq \eta_i \right\} \end{array} \right].$$

Here, $\cup_{\{C_j\}}$is the union over all sets of q elements $C_j \in \mathbb{Q}$ such that

$$\int \sum_{j=1}^{q+1} C_i \mathbf{1}_{\left\{\left(\frac{j-1}{k}, \frac{j}{k}\right]\right\}} (s)\, ds < M.$$

Because the sets $\{W^2 < M\}$ generate $\sigma(W^2)$, the result follows. ■

Proof (Proposition 6). This holds by Lemma 5 and a simple application of the Bernstein-type inequality in Bosq (1996) pg. 27. I prove the result for polynomial mixing rates. The result for exponential mixing rates holds with a similar argument. Let W_{ni} be the observations corresponding to Lemma 5 and the relevant array from conditions (5)-(8). Define

$$X_{ni} = W_{ni} - \mathbb{E}[W_{ni}].$$

This array is α-mixing with bounded elements given our assumptions. Applying the Bernstein inequality in Bosq (1998) for polynomial alpha mixing gives

$$\begin{aligned} & \mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n} X_{in} \geq \lambda\right) \\ \leq\ & C_1 \exp\left[-C_2 \lambda^2 n^{\gamma} b^{2(d+j+1)}\right] + C_3 n^{(1-\gamma)\xi - \gamma} b^{(d+j+1)/2} \end{aligned}$$

By assumption, $n^{\gamma} b 2^{(d+j+1)} \to \infty$ and $n^{(1-\gamma)\xi-\gamma} b^{(d+j+1)/2} \to 0$. This implies

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n} X_{in} \geq \lambda\right) \to 0$$

and the result follows. ■

In this appendix I first provide preliminary results similar to those in LNV (2003). I follow their approach to the proofs closely. The following lemma is a direct result of Dzhaparidze and van Zanten (2001) Theorem 3.3.

Lemma 25 (Dzhaparidze and van Zanten) *Let $M(t)$ be a locally square integrable martingale. Assume $|\Delta M(s)| \leq K$ for all $s \in [0, T]$. Then, for all $c, d > 0$,*

$$\mathbb{P}\left(|M_T| \geq c, \langle M \rangle_T \leq d^2\right) \leq 4 \exp\left[-\frac{c^2}{2(cK + d^2)}\right]$$

Lemma 25 is almost identical to Lemma 2 from LNV. However, dependence between the processes is accounted for. Note Λ_T^i are bounded by a nonrandom constant $\overline{\Lambda}$. This follows from our assumptions.

Lemma 26 *Let Θ be a bounded subset of $\mathbb{R}^{d+1}$ and, for each $\theta \in \Theta$, consider predictable functions $g_{1,\theta}, \ldots, g_{n,\theta}$. Suppose that for some constants L_n, K_n and $\rho_n \geq 1$, we have*

$$\begin{aligned} \left|g_{i,\theta}(t) - g_{i,\widetilde{\theta}}(t)\right| &\leq L_n \left|\theta - \widetilde{\theta}\right| \quad \text{for all } \theta, \widetilde{\theta} \in \Theta \text{ and all } i \geq 1 \text{ and } t \geq 0, \\ |g_{i,\theta}(t)| &\leq K_n \quad \text{for all } \theta \in \Theta \text{ and all } i \geq 1 \text{ and } t \geq 0, \end{aligned}$$

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}\int_{0}^{T}\left|g_{i,\theta}\left(t\right)\right|^{2}dt &\leq \rho_{n}^{2} \qquad \textit{for all } \theta\in\Theta \textit{ and all } n>1,\\ L_{n} &\leq n^{\nu} \qquad \textit{for all } \theta\in\Theta \textit{ and some } \nu<\infty,\end{aligned}$$

and

$$K_{n}\leq\sqrt{\frac{n}{\log n}}\rho_{n} \qquad \textit{for all } n>1.$$

Then, for some constants C_1, C_2, C_3 *and* $n>1$,

$$\mathbb{P}\left(\sup_{\theta\in\Theta}\frac{1}{\sqrt{n}}\left|\sum_{i=1}^{n}\int_{0}^{T}g_{i,\theta}^{(n)}dM_{i}\right|\geq C_{1}\rho_{n}\sqrt{\log n}\right)\leq C_{2}\exp\left(-C_{3}\log n\right).$$

Proof. From Lemma 25, we know that for each $\theta\in\Theta$, $a>0$ and $R>0$,

$$\begin{aligned}&\mathbb{P}\left(\frac{1}{\sqrt{n}}\left|\sum_{i=1}^{n}\int_{0}^{T}g_{i,\theta}^{(n)}dM_{i}\right|\geq a,\frac{1}{n}\sum_{i=1}^{n}\int_{0}^{T}\left[g_{i,\theta}^{(n)}\right]^{2}d\Lambda_{i}\leq R^{2}\right)\\ \leq\ & 4\exp\left[-\frac{a^{2}}{2\left(aK_{n}n^{-1/2}+R^{2}\right)}\right].\end{aligned}$$

This is exactly the same inequality as inequality (26) used in the proof of LNV Lemma 2 except the 2 is replaced by a 4. The rest of the proof holds exactly as in LNV except the constants are adjusted because of the 4. ■

Proof (Theorem 7). Define $o(z)\equiv\alpha(z)e(z)$. Following LNV (2003), we divide $\widehat{e}(z)-e(z)=\widehat{e}(z)-\mathbb{E}\widehat{e}(z)+\mathbb{E}\widehat{e}(z)-e(z)$ where $\widehat{e}(z)-\mathbb{E}\widehat{e}(z)$ is the "stochastic" part and $\mathbb{E}\widehat{e}(z)-e(z)$ is the "bias" part. This is similarly done for $\widehat{o}(z)$. Because of (A2)-(A3), linearity and the integral nature of the random variables, the bias term is the same as in the *i.i.d.* case. LNV (2003) show that $\sup_{x\in I}\left|\mathbb{E}\widehat{e}(z)-e(z)\right|$ and $\sup_{x\in I}\left|\mathbb{E}\widehat{o}(z)-o(z)\right|$ are $O(b^{r})$ using Taylor expansions and r times continuous differentiability of α and e.

Again following LNV (2003), we write $\widehat{e}(z)-\mathbb{E}\widehat{e}(z)=\sum_{i=1}^{n}\zeta_{n,i}^{c}(z)$, where $\zeta_{n,i}^{c}(z)=\zeta_{n,i}(z)-\mathbb{E}\zeta_{n,i}(z)$ and

$$\zeta_{n,i}(z)=\frac{1}{n}\int_{0}^{T}K_{b}\left(z-Z_{j}(s)\right)\mathbf{1}_{\{\tau_{j}\geq s\}}ds.$$

Across rows, $\zeta_{n,i}^{c}(z)$ are bounded mean zero random variables with an α-mixing rate such that $\alpha(n)\leq\alpha(n)\leq\overline{\alpha}\exp(-cn)$. Let $\{B(z_{1},\epsilon_{1}),\ldots,B(z_{L},\epsilon_{L})\}$ be an open cover of I. Here, $B(z_{l},\epsilon_{l})$ is an open ball of radius ϵ_{l} centered at z_{l}. We can choose a constant c_{1} such that $\epsilon_{L}\leq c_{1}/L$. Using differentiability of the kernel k, LNV (2003) show that

$$\begin{aligned}\sup_{z\in I}\left|\sum_{i=1}^{n}\zeta_{n,i}^{c}(z)\right| &\leq \max_{1\leq l\leq L}\left|\sum_{i=1}^{n}\zeta_{n,i}^{c}(z_{l})\right|+\max_{1\leq l\leq L}\sup_{x\in B(x_{L},\epsilon_{l})}\sum_{i=1}^{n}\left|\zeta_{n,i}^{c}(z_{l})-\zeta_{n,i}^{c}(z)\right|\\ &\leq \max_{1\leq l\leq L}\left|\sum_{i=1}^{n}\zeta_{n,i}^{c}(z_{l})\right|+\frac{c_{2}\epsilon_{L}}{b^{2(d+j+1)}}\end{aligned}$$

where c_2 is a constant.

$$\mathbb{P}\left(\sqrt{\frac{n^{\gamma}b^{2(d+j+1)}}{\log n}}\max_{1\leq l\leq L}\left|\sum_{i=1}^{n}\zeta_{n,i}^{c}\left(z_l\right)\right|>\lambda\right)$$
$$\leq \sum_{l=1}^{L}\mathbb{P}\left(\left|\sum_{i=1}^{n}\zeta_{n,i}^{c}\left(z_l\right)\right|>\lambda\sqrt{\frac{\log n}{n^{\gamma}b^{2(d+j+1)}}}\right)$$

Notice that $\left|n\zeta_{n,i}^{c}\left(z\right)\right|\leq c_3 b^{-(d+j+1)}$ for some constant c_3. Notice also, as in the proof of Theorem 9,

$$\mathbb{E}\left|n\zeta_{n,i}^{c}\left(z\right)\right|^2=\frac{1}{b^{d+j+1}}O\left(1\right)$$

This follows from change of variable and dominated convergence arguments similar to the proof of NL (1995) Theorem 1. We can now apply the Bernstein-type inequality for bounded α-mixing random variables from Bosq (1998).

$$\sum_{l=1}^{L}\mathbb{P}\left(\left|\sum_{i=1}^{n}\zeta_{n,i}^{c}\left(z_l\right)\right|>\lambda\sqrt{\frac{\log n}{n^{\gamma}b^{2(d+j+1)}}}\right)$$
$$\leq \sum_{l=1}^{L}\left\{C_1\exp\left[-C_2\lambda^2\log n\right]+C_3\frac{n^{5/4\gamma}\exp\left(-\frac{c}{2}n^{(1-\gamma)}\right)}{\log n}\right\},$$
.

If we choose $L=n^{\gamma}$ and λ large enough the above goes to zero. In addition

$$\sqrt{\frac{n^{\gamma}b^{2(d+j+1)}}{\log n}}\frac{c_2\epsilon_L}{n^{\gamma}b^{2(d+j+1)}}\leq\frac{c_2c_1}{\sqrt{\log n}n^{\gamma/2}b^{(d+j+1)}}\to 0.$$

This implies uniform convergence of $\widehat{e}\left(z\right)$.

LNV (2003) show that, once we establish uniform convergence for $\widehat{e}\left(z\right)$, we have

$$\sup_{z\in I}\left|\widehat{\alpha}\left(z\right)-\alpha\left(z\right)\right|\leq\frac{\kappa}{C+o_p\left(1\right)}\sup_{z\in I}\left|\widehat{o}\left(z\right)-o\left(z\right)\right|$$

for some positive constants κ, C. Recall that $\sup_{z\in I}\left|\mathbb{E}\widehat{o}\left(z\right)-o\left(z\right)\right|=O\left(b^r\right)$.

$$\sup_{z\in I}\left|\widehat{o}\left(z\right)-\mathbb{E}\widehat{o}\left(z\right)\right|$$
$$\leq \sup_{z\in I}\left|\widehat{o}\left(z\right)-\frac{1}{n}\sum_{i=1}^{n}\int_0^T K_b\left(z-Z_i\left(s\right)\right)d\Lambda_i\left(s\right)\right|$$
$$+\sup_{z\in I}\left|\frac{1}{n}\sum_{i=1}^{n}\int_0^T K_b\left(z-Z_i\left(s\right)\right)d\Lambda_i\left(s\right)-\mathbb{E}\widehat{o}\left(x\right)\right|.$$

Note that, by martingale arguments given above

$$\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\int_{0}^{T}K_{b}\left(z-Z_{i}\left(s\right)\right)d\Lambda_{i}\left(s\right)\right]=\mathbb{E}\left[\widehat{o}\left(z\right)\right].$$

We apply to $\widehat{o}(z)$ the same arguments given above to prove the uniform rate of convergence for $\widehat{e}(z)$. By boundedness and continuous differentiability of α, we have

$$\begin{aligned}
& \sup_{z\in I}\left|\frac{1}{n}\sum_{i=1}^{n}\int_{0}^{T}K_{b}\left(z-Z_{i}\left(s\right)\right)d\Lambda_{i}\left(s\right)-\mathbb{E}\left[\widehat{o}\left(z\right)\right]\right| \\
= & \sup_{z\in I}\left|\frac{1}{n}\sum_{i=1}^{n}\int_{0}^{T}K_{b}\left(z-Z_{i}\left(s\right)\right)d\Lambda_{i}\left(s\right)-\mathbb{E}\left[\frac{1}{n}\int_{0}^{T}K_{b}\left(z-Z_{i}\left(s\right)\right)d\Lambda_{i}\left(s\right)\right]\right| \\
= & O_{p}\left\{\sqrt{\frac{\log n}{n^{\gamma}b^{2(d+j+1)}}}\right\}.
\end{aligned}$$

Therefore

With the above considerations, similarly to as shown in LNV (2003) Lemma 3, if for

$$V_{n}=n^{-1}\sum_{i=1}^{n}\int_{0}^{T}K_{b}\left(z-Z_{i}\left(s\right)\right)d\left(N_{i}\left(s\right)-\Lambda_{i}\left(s\right)\right)=\widehat{o}\left(z\right)-n^{-1}\sum_{i=1}^{n}\int_{0}^{T}K_{b}\left(z-Z_{i}\left(s\right)\right)d\Lambda_{i}\left(s\right)$$

we have

$$\sup_{z\in I}\left|V_{n}\right|=O_{p}\left\{\sqrt{\frac{\log n}{n^{\gamma}b^{2(d+j+1)}}}\right\}$$

then the result holds. The exact same argument as in LNV (2003) holds (with different constants) because we have updated their Lemma 1 and 2 above in this appendix. Therefore,

$$\sup_{z\in I}\left|V_{n}\right|=O_{p}\left\{\sqrt{\frac{\log n}{nb^{d+j+1}}}\right\}$$

which is faster than necessary. Because of this,

$$\sup_{z\in I}\left|\widehat{\alpha}\left(z\right)-\alpha\left(z\right)\right|=O\left(b^{r}\right)+O_{p}\left\{\sqrt{\frac{\log n}{n^{\beta/(\beta+1)}b^{d+j+1}}}\right\}.$$

■

Next, I present the proof of Theorem 8. I closely follow the proof of Theorem 1 from LNV (2003), focusing on the modifications needed for dependence. I present here all the (extensive) notation, which is taken from LNV(2003),

$$\begin{aligned}(\widehat{\alpha}-\alpha)(z) &= \frac{(V_n(z)+B_n(z))}{\widehat{e}(z)} \\ V_n(z) &= \frac{1}{n}\sum_{i=1}^{n}\int_0^T K_b(z-Z_i(s))\,dM_i(s) \\ B_n(z) &= \frac{1}{n}\sum_{i=1}^{n}\int_0^T K_b(z-Z_i(s))\left[\alpha(Z_i(s))-\alpha(z)\right]Y_i(s)\,ds\end{aligned}$$

$$\begin{aligned}\left(\widehat{\alpha}_{Q_{-j}}-\alpha_{Q_{-j}}\right)(z_j) &= V_{Q_{-j}}(z_j)+B_{Q_{-j}}(z_j) \\ V_{Q_{-j}}(z_j) &= \frac{1}{n}\sum_{i=1}^{n}\int_0^T H_i^{(n)}(z_j,s)\,dM_i(s) \\ H_i^{(n)}(z_j,s) &= \int_{I_{-j}}\frac{K_b(z-Z_i(s))}{\widehat{e}(z)}dQ_{-j}(z_{-j}) \\ B_{Q_{-j}}(z_j) &= \int_{I_{-j}}\frac{B_n(z)}{\widehat{e}(z)}dQ_{-j}(z_{-j})\end{aligned}$$

$$\begin{aligned}\widetilde{h}_i^{(n)}(z_j,s) &= \int_{I_{-j}}\frac{W_{ni}(z,s)}{e(z)}dQ_{-j}(z_{-j}) \\ \widehat{h}_i^{(n)}(z_j,s) &= \int_{I_{-j}}\frac{W_{ni}(z,s)}{\widehat{e}(z)}dQ_{-j}(z_{-j}) \\ \ddot{h}_i^{(n)}(z_j,s) &= \int_{I_{-j}}\frac{W_{ni}(z,s)}{\widehat{e}_{-i}(z)}dQ_{-j}(z_{-j}) \\ \overline{h}_i^{(n)}(z_j,s) &= \frac{1}{(nb)^{1/2}}k\left(\frac{z_j-Z_{ji}(s)}{b}\right)\frac{q_{-j}^2(Z_{-ji}(s))}{e(z_j,Z_{-ji}(s))} \\ W_{ni}(z,s) &= \frac{b^{1/2}}{n^{1/2}}K_b(z-Z_i(s))\end{aligned}$$

$$(nb)^{1/2}\,\widetilde{V}_{Q_{-j}}(z_j)=\sum_{i=1}^{n}\int_0^T \widetilde{h}_i^{(n)}(z_j,s)\,dM_i(s)$$

By the assumed form of the bandwidth, $(nb)^{1/2}=n^{\theta}/C$. We wish to show

$$(nb)^{1/2}\left(\widehat{\alpha}_{Q_{-j}}(x)-\alpha_{Q_{-j}}(x)\right)(x_j)\Rightarrow N\left[m_j(z_j),v_j(z_j)\right].$$

Note the decomposition

$$(nb)^{1/2}\left(\widehat{\alpha}_{Q_{-j}}(x)-\alpha_{Q_{-j}}(x)\right)(x_j)=(nb)^{1/2}\left(V_{Q_{-j}}(z_j)+B_{Q_{-j}}(z_j)\right).$$

I show the result in three steps. First, prove

$$(nb)^{1/2}\,\widetilde{V}_{Q_{-j}}(z_j)\Rightarrow N\left(0,v_j(z_j)\right).$$

Second, prove

$$(nb)^{1/2}\left\{\widetilde{V}_{Q_{-j}}(z_j)-V_{Q_{-j}}(z_j)\right\}\to^p 0.$$

Finally, prove

$$n^{\theta}B_{Q_{-j}}(z_j)\to^p m_j(z_j).$$

First step:

Lemma 27 $(nb)^{1/2}\widetilde{V}_{Q_{-j}}(z_j)\Rightarrow N(0,v_j(z_j))$.

Proof. $\widetilde{h}_i^{(n)}(z_j,s)$ are left continuous and therefore predictable with respect to the filtration outlined in lemma 1 above. Therefore the martingale CLT derived in Proposition 3 holds in this situation. The same proof as in LNV Lemma 4 now holds using the Bernstein-type inequality of Bosq (1998) because the bandwidth satisfies $n^{\gamma}b^{2(d+j+1)}\to\infty$. ■

Still following LNV, second step:

$$(nb)^{1/2}\left\{\widetilde{V}_{Q_{-j}}(z_j)-V_{Q_{-j}}(z_j)\right\}\to^p 0.$$

$$\begin{aligned}
&(nb)^{1/2}\left|\widetilde{V}_{Q_{-j}}(z_j)-V_{Q_{-j}}(z_j)\right|\\
\leq\;&\left|\sum_{i=1}^{n}\int_0^T\widehat{h}_i^{(n)}(z_j,s)\,dM_i(s)-\sum_{i=1}^{n}\int_0^T\ddot{h}_i^{(n)}(z_j,s)\,dM_i(s)\right|\\
&+\left|\sum_{i=1}^{n}\int_0^T\ddot{h}_i^{(n)}(z_j,s)\,dM_i(s)-\sum_{i=1}^{n}\int_0^T\widetilde{h}_i^{(n)}(z_j,s)\,dM_i(s)\right|
\end{aligned}$$

Lemma 28 $\sum_{i=1}^n\int_0^T\widehat{h}_i^{(n)}(z_j,s)\,dM_i(s)-\sum_{i=1}^n\int_0^T\ddot{h}_i^{(n)}(z_j,s)\,dM_i(s)\to^p 0$.

Proof. Because we proved uniform convergence of $\widehat{e}(z)$ in Theorem 7, the proof of Lemma 5 from LNV (2003) holds here with the same proof. Here, we provide more details for the interested reader. See also LNV (2003).

$$\begin{aligned}
&\left|\widehat{h}_i^{(n)}(z_j,s)-\ddot{h}_i^{(n)}\right|\\
=\;&\left|\int_{I_{-j}}W_{ni}(z,s)\frac{\widehat{e}_{-i}(z)-\widehat{e}(z)}{\inf_{x\in I}\left|\widehat{e}(z)\widehat{e}_{-i}(z)\right|}dQ_{-j}(z_{-j})\right|\\
\leq\;&\frac{\left[\int_{I_{-j}}W_{ni}^2(z,s)\,dQ_{-j}(z_{-j})\cdot\int_{I_{-j}}\left\{\widehat{e}_{-i}(z)-\widehat{e}(z)\right\}^2dQ_{-j}(z_{-j})\right]^{1/2}}{\inf_{z\in I}\left|\widehat{e}(z)\widehat{e}_{-i}(z)\right|}
\end{aligned}$$

where the last inequality follows from the Cauchy-Schwarz inequality. We assume Q is Lebesgue measure with constant weight C_Q. By change of variables in z_{-j}

$$\begin{aligned}\int_{I_{-j}} W_{ni}^2(z,s)\,dQ_{-j}(z_{-j}) &= \int_{I_{-j}} \frac{b}{n}\frac{1}{b^{2(d+j+1)}}k^2\left(\frac{z_j - Z_{ji}(s)}{b}\right)K^2\left(\frac{z_j - Z_{-ji}(s)}{b}\right)dQ_{-j}(z_{-j}) \\ &= \frac{1}{n}\frac{1}{b^{(d+j+1)}}k^2\left(\frac{z_j - Z_{ji}(s)}{b}\right)\int_{I_{-j}} \frac{1}{b^{(d+j)}}K^2\left(\frac{z_j - Z_{-ji}(s)}{b}\right)dQ_{-j}(z_{-j}) \\ &= \frac{1}{n}\frac{1}{b^{(d+j+1)}}k^2\left(\frac{z_j - Z_{ji}(s)}{b}\right)\int_{\left(I_{-j}^{\min}-Z_{-j}(s)\right)/b}^{\left(I_{-j}^{\max}-Z_{-j}(s)\right)/b} K^2(z_j)\,C_Q dz_{-j}.\end{aligned}$$

This implies

$$\sup_s\left[\int_{I_{-j}} W_{ni}^2(z,s)\,dQ_{-j}(z_{-j})\right] = O\left(n^{-1}b^{-(d+j+1)}\right).$$

A similar change of variables argument implies

$$\int_{I_{-j}}\left\{\widehat{e}_{-i}(z) - \widehat{e}(z)\right\}^2 dQ_{-j}(z_{-j}) = O\left(n^{-2}b^{-2(d+j+1)}\right).$$

As shown in the proof of Theorem 7, $\sup_{x\in I}|\widehat{e}(z) - e(z)| \to^p 0$. This implies, for some $\epsilon > 0$,

$$\inf_{z\in I}|\widehat{e}(z)\,\widehat{e}_{-i}(z)| \geq \epsilon + o_p(1).$$

This gives the same result as in LNV:

$$\begin{aligned}&\left|\sum_{i=1}^n\int_0^T \widehat{h}_i^{(n)}(z_j,s)\,dM_i(s) - \sum_{i=1}^n\int_0^T \ddot{h}_i^{(n)}(z_j,s)\,dM_i(s)\right| \\ \leq\ & nO_p\left(n^{-1}b^{-(d+j+1)/2}\right)O_p\left(n^{-1/2}b^{-(d+j+1)/2}\right).\end{aligned}$$

This is $o_p(1)$ because $nb^{2(d+j+1)} \to \infty$. ■

Lemma 29

$$\overline{M}_t = \sum_{i=1}^n\int_0^T \ddot{h}_i^{(n)}(z_j,s)\,dM_i(s) - \sum_{i=1}^n\int_0^T \widetilde{h}_i^{(n)}(z_j,s)\,dM_i(s) \to^p 0. \tag{37}$$

Proof. Again following LNV,

$$\begin{aligned}\overline{M}_t &= \overline{M}_t^1 + \overline{M}_t^2 + \overline{M}_t^3 \\ &= \sum_{i=1}^n\int_0^1\left\{\int_{I_{-j}} W_{ni}(z,s)\,\frac{e(z) - \mathbb{E}\left[\widehat{e}_{-i}(z)\right]}{e^2(z)}dQ_{-j}(z_{-j})\right\}dM^i(s) \\ &\quad + \sum_{i=1}^n\int_0^1\left\{\int_{I_{-j}} W_{ni}(z,s)\,\frac{\mathbb{E}\left[\widehat{e}_{-i}(z)\right] - \widehat{e}_{-i}(z)}{e^2(z)}dQ_{-j}(z_{-j})\right\}dM^i(s) \\ &\quad + \sum_{i=1}^n\int_0^1\left\{\int_{I_{-j}} W_{ni}(z,s)\,\frac{\left\{e(z) - \widehat{e}_{-i}(z)\right\}^2}{e^2(z)\,\widehat{e}_{-i}(z)}dQ_{-j}(z_{-j})\right\}dM^i(s).\end{aligned}$$

We deal with each piece individually. First consider $\overline{M}_t^1$. This proof is similar to LNV but incorporates the Bernstein-type inequality used above. Details follow.

$$\overline{M}_t^1 \leq \sum_{i=1}^n \int_0^1 \int_{I_{-j}} \left| W_{ni}(z,s) \frac{e(z) - \mathbb{E}[\widehat{e}_{-i}(z)]}{e^2(z)} \right| dQ_{-j}(z_{-j})\, dM^i(s)$$

Now consider

$$\begin{aligned}
& \int_{I_{-j}} \left| W_{ni}(z,s) \frac{e(z) - \mathbb{E}[\widehat{e}_{-i}(z)]}{e^2(z)} \right| dQ_{-j}(z_{-j}) \\
\leq & \int_{I_{-j}} W_{ni}(z,s) \frac{|e(z) - \mathbb{E}[\widehat{e}_{-i}(z)]|}{e^2(z)} dQ_{-j}(z_{-j}) \\
\leq & \frac{\sup_{x\in I} |e(z) - \mathbb{E}[\widehat{e}_{-i}(z)]|}{\inf_{x\in I} e^2(z)} \int_{I_{-j}} W_{ni}(z,s)\, dQ_{-j}(z_{-j}) \\
= & O(b^r) \int_{I_{-j}} W_{ni}(z,s)\, dQ_{-j}(z_{-j}).
\end{aligned}$$

By change of variables

$$\begin{aligned}
& \sup_s \int_{I_{-j}} W_{ni}(z,s)\, dQ_{-j}(z_{-j}) \\
= & \sup_s \frac{1}{n^{1/2} b^{1/2}} k\left(\frac{z_j - Z_j(s)}{b}\right) \int_{\left(I_{-j}^{\min} - Z_{-j}(s)\right)/b}^{\left(I_{-j}^{\max} - Z_{-j}(s)\right)/b} K^2(z_j)\, C_Q dz_{-j} \\
\leq & \frac{1}{n^{1/2} b^{1/2}} C.
\end{aligned}$$

$$\begin{aligned}
& \sum_{i=1}^n \int_0^1 \int_{I_{-j}} \left| W_{ni}(z,s) \frac{e(z) - \mathbb{E}[\widehat{e}_{-i}(z)]}{e^2(z)} \right| dQ_{-j}(z_{-j})\, dM^i(s) \\
\leq & O(b^r) \sum_{i=1}^n \int_0^1 \frac{1}{n^{1/2} b^{1/2}} C dM^i(s) \qquad (38)
\end{aligned}$$

Arguments using Proposition 3 show that (38) is a value that follows a martingale CLT multiplied by $O(b^r) = o(1)$. Therefore $\overline{M}_t^1 = o_p(1)$.

Now consider $\overline{M}_t^3$.

$$\begin{aligned}
& \int_{I_{-j}} W_{ni}(z,s) \frac{\{e(z) - \widehat{e}_{-i}(z)\}^2}{\inf_{x\in I} |e^2(z)\, \widehat{e}_{-i}(z)|} dQ_{-j}(z_{-j}) \\
\leq & \frac{\left[\int_{I_{-j}} W_{ni}^2(z,s)\, dQ_{-j}(z_{-j}) \cdot \int_{I_{-j}} \{e(z) - \widehat{e}_{-i}(z)\}^4 dQ_{-j}(z_{-j})\right]^{1/2}}{\inf_{x\in I} |e^2(z)\, \widehat{e}_{-i}(z)|}
\end{aligned}$$

Note that we previously showed

$$\sup_s \left[\int_{I_{-j}} W_{ni}^2(z,s)\, dQ_{-j}(z_{-j})\right] = O\left(n^{-1} b^{-(d+j+1)}\right).$$

By the uniform rate of convergence for $\widehat{e}_{-i}(z)$ derived in Theorem 7, we have

$$\begin{aligned} & \sup_z \int_{I_{-j}} \{e(x) - \widehat{e}_{-i}(x)\}^4 dQ_{-j}(x_{-j}) \\ \leq & \left\{ O(b^r) + O_p \left\{ \sqrt{\frac{\log n}{n^{\beta/(\beta+1)} b^{d+j+1}}} \right\} \right\}^4 \end{aligned}$$

$$\begin{aligned} \overline{M}_t^3 &\leq nO_p\left(n^{-1/2}b^{-(d+j+1)/2}\right)\left\{O(b^r) + O_p\left\{\sqrt{\frac{\log n}{n^{\gamma}b^{2(d+j+1)}}}\right\}\right\}^2. \\ &= O_p\left(n^{1/2}b^{2r-(d+j+1)/2}\right) \\ &\quad + O_p\left(n^{1/2-(\beta/2(\beta+1))}\log^{1/2}(n)\, b^{r-(d+j+1)}\right) \\ &\quad + O_p\left(n^{1/2-(\beta/(\beta+1))}\log(n)\, b^{-3(d+j+1)/2}\right) \end{aligned}$$

Now consider $\overline{M}_t^2$.

$$\begin{aligned} h_i^{(n)}(u) &= \int_{I_{-j}} \left| W_{ni}(z,u) \frac{\widehat{e}_{-i}(z) - \mathbb{E}[\widehat{e}_{-i}(z)]}{e^2(z)} \right| dQ_{-j}(z_{-j}) \\ &\leq \int_{I_{-j}} W_{ni}(z,u) \frac{|\widehat{e}_{-i}(z) - \mathbb{E}[\widehat{e}_{-i}(z)]|}{e^2(z)} dQ_{-j}(z_{-j}) \\ &\leq \frac{\sup_{z\in I}|\widehat{e}_{-i}(z) - \mathbb{E}[\widehat{e}_{-i}(z)]|}{\inf_{z\in I} e^2(z)} \sup_u \int_{I_{-j}} W_{ni}(z,u)\, dQ_{-j}(z_{-j}) \end{aligned}$$

First, recall

$$\sup_u \int_{I_{-j}} W_{ni}(z,u)\, dQ_{-j}(z_{-j}) \leq \frac{1}{n^{1/2}b^{1/2}} C.$$

Next, recall we show in the proof of Theorem 7 that

$$\mathbb{P}\left(\sqrt{\frac{n^{\gamma}b^{2(d+j+1)}}{\log n}} \sup_{z\in I} |\widehat{e}_{-i}(z) - \mathbb{E}[\widehat{e}_{-i}(z)]| > \lambda\right) \leq \frac{n^{\gamma}}{n^{\lambda^2}} O(1)$$

or

$$\mathbb{P}\left(\frac{n^{\gamma}b^{2(d+j+1)}}{\log n}\left(\sup_{z\in I}|\widehat{e}_{-i}(z) - \mathbb{E}[\widehat{e}_{-i}(z)]|\right)^2 > \lambda^2\right) \leq \frac{n^{\gamma}}{n^{\lambda^2}} O(1) \tag{39}$$

We follow Mammen and Nielsen (2007) Lemma (A1). Note that the cádlág assumptions made earlier in Mammen and Nielsen (2007) are not needed. This is important for piecewise constant covariate processes. Recall that $\alpha(z)$ is bounded above and below.

$$\begin{aligned} & \sum_{i=1}^n \mathbb{E}\int_0^T \left\{h_i^{(n)}(u)\right\}^2 \alpha\left(s, X_u^i, Y_{G^i+u}\right) \mathbf{1}_{\{\tau_i>u\}} du \\ \leq & \frac{\log n}{n^{\gamma}b^{2(d+j+1)}} \frac{C'}{nb} \sum_{i=1}^n \mathbb{E}\left[\frac{n^{\gamma}b^{2(d+j+1)}}{\log n}\left\{\sup_{z\in I}|\widehat{e}_{-i}(z) - \mathbb{E}[\widehat{e}_{-i}(z)]|\right\}^2\right] \end{aligned}$$

$\sup_{z\in I}\left|\widehat{e}_{-i}(z)-\mathbb{E}\left[\widehat{e}_{-i}(z)\right]\right|$ is bounded by C_1/b^{d+j+1}. Therefore

$$\frac{n^{\gamma}b^{2(d+j+1)}}{\log n}\left(\sup_{z\in I}\left|\widehat{e}_{-i}(z)-\mathbb{E}\left[\widehat{e}_{-i}(z)\right]\right|\right)^{2}\leq\frac{n^{\gamma}}{\log n}C_1^2$$

Using (39), we know

$$\mathbb{E}\left[\frac{n^{\gamma}b^{2(d+j+1)}}{\log n}\left\{\sup_{z\in I}\left|\widehat{e}_{-i}(z)-\mathbb{E}\left[\widehat{e}_{-i}(z)\right]\right|\right\}^{2}\right]\leq\lambda^2+\frac{n^{\gamma}}{\log n}C_1^2\times\frac{n^{\gamma}}{n^{\lambda^2}}O(1)$$

If we choose λ large enough,

$$\frac{n^{\gamma}}{\log n}C_1^2\times\frac{n^{\gamma}}{n^{\lambda^2}}O(1)=o(1).$$

this implies

$$\begin{aligned}
&\frac{\log n}{n^{\gamma}b^{2(d+j+1)}}\frac{C'}{nb}\sum_{i=1}^{n}\mathbb{E}\left[\frac{n^{\gamma}b^{2(d+j+1)}}{\log n}\left\{\sup_{z\in I}\left|\widehat{e}_{-i}(z)-\mathbb{E}\left[\widehat{e}_{-i}(z)\right]\right|\right\}^{2}\right]\\
\leq\;&\frac{\log n}{n^{\gamma}b^{2(d+j+1)}}\frac{C'}{nb}\sum_{i=1}^{n}\lambda^2+o(1)\\
=\;&\frac{\log n}{n^{\gamma}b^{2(d+j+1)+1}}\left(C'\lambda^2+o(1)\right).
\end{aligned}$$

Therefore, by the assumptions of the theorem

$$\sum_{i=1}^{n}\mathbb{E}\int_0^T\left\{h_i^{(n)}(u)\right\}^2\alpha\left(s,X_u^i,Y_{G^i+u}\right)\mathbf{1}_{\{\tau_i>u\}}du\to 0.$$

Define

$$h_{i,j}^{(n)}(u)=\int_{I_{-j}}W_{ni}(z,u)\frac{\widehat{e}_{-i,j}(z)-\mathbb{E}\left[\widehat{e}_{-i,j}(z)\right]}{e^2(z)}dQ_{-j}(z_{-j})$$

Therefore

$$\begin{aligned}
h_{i,j}^{(n)}(u)-h_i^{(n)}(u)&=\int_{I_{-j}}W_{ni}(z,u)\frac{-\frac{1}{n}\int_0^T K_b\left(z-Z^i(s)\right)\mathbf{1}_{\{\tau_i>s\}}ds}{e^2(z)}dQ_{-j}(z_{-j})\\
&\leq\frac{1}{nb^{d+j+1}}\frac{\sup_I\left|\int_0^T K\left(z-Z^i(s)\right)\mathbf{1}_{\{\tau_i>s\}}ds\right|}{\inf_I\left|e^2(z)\right|}\frac{1}{n^{1/2}b^{1/2}}C\\
&\leq\frac{1}{nb^{d+j+1}}\frac{1}{n^{1/2}b^{1/2}}C'.
\end{aligned}$$

$$\mathbb{E}\int_0^T\left\{h_{i,j}^{(n)}(u)-h_i^{(n)}(u)\right\}^2\alpha\left(s,X_u^i,Y_{G^i+u}\right)\mathbf{1}_{\{\tau_i>u\}}du\leq\frac{1}{n^2b^{2(d+j+1)}}\frac{1}{nb}C'$$

Clearly, by the assumptions of Theorem 8,

$$n\sum_{i=1}^{n}\frac{1}{n^2b^{2(d+j+1)}}\frac{1}{nb}C'=\frac{1}{nb^{2(d+j+1)+1}}C'\to 0.$$

As a result of Mammen and Nielson (2007) Lemma (A1), we conclude $\mathbb{E}\left[\overline{M}_t^2\right]^2 = o(1)$. Therefore, $\overline{M}_t^2 = o_p(1)$. Therefore, Lemma 29 is proven. ■

Putting the last three Lemmas together, we have proven $(nb)^{1/2} V_{Q_{-j}}(z_j) \Rightarrow N(0, v_j(z))$. Now we need to verify the third step: $n^\theta B_{Q_{-j}}(z_j) \to^p m_j(z_j)$. As usual, we follow LNV.

$$B_{Q_{-j}}(z_j) = \int_{I_{-j}} \frac{B_n(z)}{e(z)} dQ_{-j}(z_{-j}) + \int_{I_{-j}} B_n(z) \frac{\widehat{e}(z) - e(z)}{\widehat{e}(z) e(z)} dQ_{-j}(z_{-j})$$

$$\begin{aligned} & \left| \int_{I_{-j}} B_n(z) \frac{\widehat{e}(z) - e(z)}{\widehat{e}(z) e(z)} dQ_{-j}(z_{-j}) \right| \\ \le & \frac{\sup_{z\in I} |B_n(z)| \sup_{z \in I} |\widehat{e}(z) - e(z)|}{\inf_{z\in I} |\widehat{e}(z) e(z)|} \\ = & O_p(b^r) \left\{ O_p(b^r) + O_p\left(\frac{\log n}{n^{\beta/(\beta+1)} b^{d+j+1}} \right) \right\} \\ = & b^r \left\{ O_p(b^r) + O_p\left(\frac{\log n}{n^{\beta/(\beta+1)} b^{d+j+1}} \right) \right\} \end{aligned}$$

Note $\sup_{z\in I} |B_n(z)| = O_p(b^r)$ by (K2), a Taylor expansion and the mixing conditions. See Nielsen and Linton (1995) proof of Theorem 1 for more details. The same argument as in LNV holds and is reproduced here:

$$\mathbb{E}\left[\int_{I_{-j}} \frac{B_n(z)}{e(z)} dQ_{-j}(z_{-j}) \right] = \frac{\mu_r(k)}{r!} b^r \sum_{j=1}^d \int_{I_{-j}} \beta_j^{(r)}(z) \, dQ_{-j}(z_{-j}) \{1 + o(1)\}.$$

Notice that we can write things as

$$\int_{I_{-j}} \frac{B_n(z)}{e(z)} dQ_{-j}(z_{-j}) = \frac{1}{n} \sum_{i=1}^n \int_0^T \int_{I_{-j}} \frac{K_b(z - Z_i(s))}{e(z)} \left[\alpha(Z_i(s)) - \alpha(z) \right] Y_i(s) \, dQ_{-j}(z_{-j}) \, ds$$

This representation, Lemma 5 and the assumed mixing conditions imply that.

$$n^\theta \left\{ \int_{I_{-j}} \frac{B_n(z)}{e(z)} dQ_{-j}(z_{-j}) - \mathbb{E}\left[\int_{I_{-j}} \frac{B_n(z)}{e(z)} dQ_{-j}(z_{-j}) \right] \right\} \to^p 0.$$

C Proofs for Section 4

Proof (Proposition 11). We follow Anderson and Gill (1982) Appendix 3. Here, only the needed alterations for dependence are sketched. Note that in the proof of Theorem III.1, the deterministic times $t_1^{(n)}, \ldots, t_n^{(n)}$ need not be in increasing order. The entire proof goes through as long as there exists a distribution function y such that

$$\sup_{t\in[0,1]} \left| \frac{1}{n} \sum_{i=1}^n \mathbf{1}_{\left[t_i^{(n)}, 1\right]}(t) - y(t) \right| \to 0. \tag{40}$$

We will modify the proofs of Appendix 3 to show Condition B from AG (1982) is satisfied in this situation for $S^{(0)}(\beta,t)$. Similar arguments will give the result for $S^{(1)}(\beta,t)$ and $S^{(2)}(\beta,t)$. $s^{(0)}(\beta,t) = \mathbb{E}\left\{A(t)\exp\left(\beta' Z(t)\right)\right\} = y(t)\,\mathbb{E}\left\{\mathbf{1}\{\tau\geq t\}\exp\left(\beta' Z(t)\right)\right\}$ when the independent censoring times C^i have distribution function y. By the mixing assumptions, for any fixed $t\in[0,1]$ the random variables $\mathbf{1}\{\tau_i\geq t\}\exp\left(\beta' Z^i(t)\right)$ have α-mixing coefficients such that

$$\sum_{n>0} n^{-1}\alpha(n) < \infty. \tag{41}$$

Therefore, by Rio (1995) a SLLN holds for $\mathbf{1}\{\tau_i\geq t\}\exp\left(\beta' Z^i(t)\right)$ for any fixed β. By Pollard (1984) Theorem II.2, for any fixed t, we have

$$\sup_{\beta}\left|\frac{1}{n}\sum_{i=1}^{n}\mathbf{1}\{\tau_i\geq t\}\exp\left(\beta' Z^i(t)\right) - \mathbb{E}\left\{\mathbf{1}\{\tau\geq t\}\exp\left(\beta' Z(t)\right)\right\}\right| \to 0,$$

almost surely. In addition, the random variables $\left\|\mathbf{1}\{\tau_i\geq t\}\exp\left(\beta' Z^i(t)\right)\right\|\mathbf{1}\left\{\mathbf{1}\{\tau_i\geq t\}\exp\left(\beta' Z^i(t)\right)\notin K\right\}$ for any K compact set as specified in AG (1982) also satisfy the mixing condition (41). Therefore, they satisfy a SLLN by Rio (1995). With these extensions, the same basic argument as in AG (1982) Theorem III.1 holds. Specifically

$$\sup_{t\in[0,1],\beta}\left|\frac{1}{n}\sum_{i=1}^{n}\mathbf{1}\{\tau_i\geq t\}\exp\left(\beta' Z^i(t)\right)\mathbf{1}_{\left[t_i^{(n)},1\right]} - \mathbb{E}\left\{\mathbf{1}\{\tau\geq t\}\exp\left(\beta' Z(t)\right)\right\}y(t)\right| \to 0,$$

almost surely. Using the same idea as AG (1982) Corrollary III.2, we conclude

$$\sup_{t\in[0,1],\beta}\left|\frac{1}{n}\sum_{i=1}^{n}\mathbf{1}\{\tau_i\geq t\}\exp\left(\beta' Z^i(t)\right)\mathbf{1}_{[C^i,1]} - \mathbb{E}\left\{\mathbf{1}\{\tau\geq t\}\exp\left(\beta' Z(t)\right)\right\}y(t)\right| \to^p 0.$$

This is the required result. The remainder of the needed conditions follow from the same arguments as AG (1982) Theorem 4.1. ■

The required conditions on the covariates $Z^i(t)$ are presented here.

(**B3**): Assume the following for each observation $i\in\mathbb{N}_0$. Assume for each $s\in[0,1]$, $Z^i(s)$ has support contained in a compact rectangle normalized to be $[0,1]^q$ for ease of notation. $Z^i(s)$ must be random and no process in $Z^i(s)$ can be a linear combination of the other processes. Assume

$$\mathbb{E}\left[\mathbf{1}_{\{\tau_i\geq s\}}\middle| Z^i(s), 0\leq s\leq 1\right] \geq M^{\min}(s) > 0 \qquad \forall s\in(0,1]\text{, a.s.} \tag{42}$$

for some deterministic function $M^{\min}(s)$ on $s\in[0,1]$. Let $S\subset D^q[0,1]$ and give S the relative Skorokhod topology. Assume

$$\overline{\mathbb{P}}\left\{Z^i(s+)\in S\right\} = 1.$$

For all $x_0(s)\in S$ for all $k\in\{1,\ldots,q\}$ there exists $x_0(s)+\epsilon(s)\in S$ such that , $\epsilon^i(s)=0$ for $i\neq j$, $s\in[0,1]$ and $\epsilon^k(s)>0$ or $\epsilon^k(s)<0$ on some interval with positive Lebesgue measure contained in $[0,1]$, finally $\epsilon^k(s)=0$ for s outside of that interval. Let $\beta\in[a_1,b_1]\times\cdots\times[a_q,b_q]$. Either (1.)

For all $x_0 \in S$ and all $\epsilon > 0$,

$$\overline{\mathbb{P}}\left\{\omega \middle| Z^i(s+)(\omega) \in S, d\left(Z^i(s+)(\omega), x_0\right) < \epsilon\right\} > 0,$$

$$\exists x_0' \in S,\ x_0' \neq x_0 \qquad s.t. \quad d\left(x_0, x_0'\right) < \epsilon,$$

where $d(\cdot,\cdot)$ is the Skorokhod metric. Or (2.) S consists of a finite number of paths x, each with positive probability.

Condition (42) will always be satisfied because we can choose a point process with hazard

$$\left[\inf_{v\in[0,1]} h_0(v)\right](s) \inf_{\beta,z}\left[\exp\left(\beta' z\right)\right] \mathbf{1}_{\{\tau_i \geq s\}}$$

which will correspond to an $M^{\min}(s)$ that satisfies the condition. We include this assumption for completeness.

Lemma 30 *Let f_1, f_2, $\{f_n\}$ be bounded strictly positive functions defined on $[0,1]$. If*

$$\int_0^1 \left[\frac{f_n}{f_1} - 1 - \log\left(\frac{f_n}{f_1}\right)\right] f_1 f_2 ds \to 0,$$

then

$$f_n \to^{L^1} f.$$

Proof. Note that the function $f(x) = x - 1 - \log(x)$ on the interval $(0, \infty)$ is uniquely minimized at 1 where its value is 0. The result follows easily. This results is used in Karr (1987) and Grenander (1981). ■

Lemma 31 *Under assumptions (B1)-(B3), if*

$$H(h_0, \beta_0) - H(h_n, \beta_n) \to 0,$$

then

$$h_n \to^{L^1} h_0,$$

$$\beta_n \to \beta_0.$$

Proof. By a similar manipulation as in Karr's (1987) proof of Theorem 3.3,

$$
\begin{aligned}
& H\left(h_0, \beta_0\right) - H\left(h_n, \beta_n\right) \\
= & \ \mathbb{E}\left\{\int_0^1 \left[h_n(s) \exp\left(\beta_n' Z^i\right) - h_0(s) \exp\left(\beta_0' Z^i\right)\right] \mathbf{1}_{\{\tau_i \geq s\}} ds\right\} \\
& -\mathbb{E}\left\{\int_0^1 \log\left[\frac{h_n(s)\exp\left(\beta_n' Z^i\right)}{h_0(s)\exp\left(\beta_0' Z^i\right)}\right] dN_s^i\right\} \\
= & \ \mathbb{E}\left\{\int_0^1 \left[\frac{h_n(s)\exp\left(\beta_n' Z^i\right)}{h_0(s)\exp\left(\beta_0' Z^i\right)} - 1\right] h_0(s)\exp\left(\beta_0' Z^i\right) \mathbf{1}_{\{\tau_i \geq s\}} ds\right\} \\
& -\mathbb{E}\left\{\int_0^1 \log\left[\frac{h_n(s)\exp\left(\beta_n' Z^i\right)}{h_0(s)\exp\left(\beta_0' Z^i\right)}\right] h_0(s)\exp\left(\beta_0' Z^i\right) \mathbf{1}_{\{\tau_i \geq s\}} ds\right\} \\
= & \ \mathbb{E}\left\{\int_0^1 \left[\frac{h_n(s)\exp\left(\beta_n' Z^i\right)}{h_0(s)\exp\left(\beta_0' Z^i\right)} - 1 - \log\left(\frac{h_n(s)\exp\left(\beta_n' Z^i\right)}{h_0(s)\exp\left(\beta_0' Z^i\right)}\right)\right] h_0(s)\exp\left(\beta_0' Z^i\right) \mathbf{1}_{\{\tau_i \geq s\}} ds\right\} \\
= & \ \mathbb{E}\left\{\mathbb{E}\left[\left. \begin{array}{c} \int_0^1 \left[\frac{h_n(s)\exp\left(\beta_n' Z^i\right)}{h_0(s)\exp\left(\beta_0' Z^i\right)} - 1 - \log\left(\frac{h_n(s)\exp\left(\beta_n' Z^i\right)}{h_0(s)\exp\left(\beta_0' Z^i\right)}\right)\right] \\ h_0(s)\exp\left(\beta_0' Z^i\right) \mathbf{1}_{\{\tau_i \geq s\}} ds \end{array} \right| Z^i(t),\ t \in [0,1]\right]\right\} \\
= & \ \mathbb{E}\left\{\begin{array}{c} \int_0^1 \left[\frac{h_n(s)\exp\left(\beta_n' Z^i\right)}{h_0(s)\exp\left(\beta_0' Z^i\right)} - 1 - \log\left(\frac{h_n(s)\exp\left(\beta_n' Z^i\right)}{h_0(s)\exp\left(\beta_0' Z^i\right)}\right)\right] \\ h_0(s)\exp\left(\beta_0' Z^i\right) \mathbb{E}\left[\mathbf{1}_{\{\tau_i \geq s\}} \middle| Z^i(t),\ t \in [0,1]\right] ds \end{array}\right\} \\
\geq & \ \mathbb{E}\left\{\int_0^1 \left[\frac{h_n(s)\exp\left(\beta_n' Z^i\right)}{h_0(s)\exp\left(\beta_0' Z^i\right)} - 1 - \log\left(\frac{h_n(s)\exp\left(\beta_n' Z^i\right)}{h_0(s)\exp\left(\beta_0' Z^i\right)}\right)\right] h_0(s)\exp\left(\beta_0' Z^i\right) M^{\min}(s)\, ds\right\}
\end{aligned} \tag{43}
$$

Here, we note that we can convert the covariate processes in (43) to their right continuous cádlág versions without changing the expectation. This is because of the integral in the expectation and that cáglád processes can only have a countable number of discontinuities. From now on in the proof, we have changed $Z^i(s)$ to $Z^i(s+)$ and therefore can deal with the Skorokhod space. This allows us to exploit assumption (B3). By assumption, (43) converges to zero. Lemma 30 implies that, for a fixed x, if

$$
\int_0^1 \left[\frac{h_n(s)\exp\left(\beta_n' x\right)}{h_0(s)\exp\left(\beta_0' x\right)} - 1 - \log\left(\frac{h_n(s)\exp\left(\beta_n' x\right)}{h_0(s)\exp\left(\beta_0' x\right)}\right)\right] h_0(s)\exp\left(\beta_0' x\right) M^{\min}(s)\, ds \to 0, \tag{44}
$$

then

$$
h_n(s)\exp\left(\beta_n' x\right) \to^{L^1} h_0(s)\exp\left(\beta_0' x\right). \tag{45}
$$

Assume (45) does not hold for an open ball in S around a path $x_0 \in S$. By assumption, this open ball has positive probability. This implies (44) does not hold for this set. Because of the positive probability of Z^i having a realization in this set, (43) would fail to converge to zero if the assumption is true. So (45) can not fail on an open ball in S. Therefore, (45) must hold for $x \in D$ where D is a dense set of paths in S with the relative Skorokhod topology.

A consequence is that $\sup_n \int |h_n(s)|\, ds$ is bounded. This holds because β is restricted to a compact interval. If it did not hold, (45) would fail at all paths $x(s) \in [0,1]^d$. Similarly, $\int |h_n(s)|\, ds \nrightarrow 0$ because if this happened $h_n(s)\exp\left(\beta_n' x\right) \to^{L^1} 0$ which can not happen because $h_0(s)\exp\left(\beta_0' x\right)$ is strictly positive.

Assume there exists a path $x_0 \in S$ such that (45) fails. Because (45) must hold on a dense set, by assumption (B3) and by the definition of the Skorokhod metric, for any $\eta > 0$ there exists a perturbation $\epsilon(s)$ such that (45) is satisfied for $x_0(s) + \epsilon(s) \in S$ and

$$\sup_{t \in [0,1]} \|\epsilon(s)\| < \eta$$

.

$$\begin{aligned}
& \int \left| h_n(s) \exp\left(\beta_n' x_0(s)\right) - h_0(s) \exp\left(\beta_0' x_0(s)\right) \right| ds \\
= & \int \left| \begin{array}{c} h_n(s) \exp\left(\beta_n' x_0(s)\right) - h_n(s) \exp\left(\beta_n' x_0(s)\right) \exp\left(\beta_n' \epsilon(s)\right) \\ + h_n(s) \exp\left(\beta_n' x_0(s)\right) \exp\left(\beta_n' \epsilon(s)\right) - h_0(s) \exp\left(\beta_0' x_0(s)\right) \exp\left(\beta_0' \epsilon(s)\right) \\ + h_0(s) \exp\left(\beta_0' x_0(s)\right) \exp\left(\beta_0' \epsilon(s)\right) - h_0(s) \exp\left(\beta_0' x_0(s)\right) \end{array} \right| ds \\
\leq & \int \left| h_n(s) \exp\left(\beta_n' x_0(s)\right) - h_n(s) \exp\left(\beta_n' x_0(s)\right) \exp\left(\beta_n' \epsilon(s)\right) \right| ds \\
& + \int \left| h_n(s) \exp\left(\beta_n' x_0(s)\right) \exp\left(\beta_n' \epsilon(s)\right) - h_0(s) \exp\left(\beta_0' x_0(s)\right) \exp\left(\beta_0' \epsilon(s)\right) \right| ds \\
& + \int \left| h_0(s) \exp\left(\beta_0' x_0(s)\right) \exp\left(\beta_0' \epsilon(s)\right) - h_0(s) \exp\left(\beta_0' x_0(s)\right) \right| ds \\
= & \int \left| h_n(s) \exp\left(\beta_n' x_0(s)\right) \right| \left| 1 - \exp\left(\beta_n' \epsilon(s)\right) \right| ds + o(1) \\
& + \int \left| h_0(s) \exp\left(\beta_0' x_0(s)\right) \right| \left| \exp\left(\beta_0' \epsilon(s)\right) - 1 \right| ds \\
\leq & C_{\max}^1 \sup_{\beta, s} \left| 1 - \exp\left(\beta' \epsilon(s)\right) \right| + o(1) \\
& + C_{\max}^2 \sup_{s} \left| \exp\left(\beta_0' \epsilon(s)\right) - 1 \right|.
\end{aligned}$$

Because we may choose $\epsilon(s)$ such that (45) holds for any $\eta > 0$, for any $\gamma > 0$ we can choose $\epsilon(s)$ such that these exists an N such that

$$\int \left| h_n(s) \exp\left(\beta_n' x_0\right) - h_0(s) \exp\left(\beta_0' x_0\right) \right| ds < \gamma$$

for all $n \geq N$. Therefore, (45) holds for $x_0(s)$ and therefore (45) holds for all $x \in S$.

Because (45) holds for all $x \in S$, then if $\beta_n \to \beta_0$ this implies $h_n(s) \to^{L^1} h_0(s)$. We can see this from the following

$$
\begin{aligned}
& \int \left| h_n(s) \exp\left(\beta_n' x_0(s)\right) - h_0(s) \exp\left(\beta_0' x_0(s)\right) \right| ds \\
= & \int \left| h_n(s) \left[\exp\left(\beta_0' x_0(s)\right) + \sum_{i=1}^{d} x_0^i(s) \exp\left(c' x_0(s)\right) \left(\beta_n^i - \beta_0^i\right) \right] - h_0(s) \exp\left(\beta_0' x_0(s)\right) \right| ds \\
\geq & \int \left| h_n(s) \exp\left(\beta_0' x_0(s)\right) - h_0(s) \exp\left(\beta_0' x_0(s)\right) + h_n(s) \sum_{i=1}^{d} x_0^i(s) \exp\left(c' x_0(s)\right) \left(\beta_n^i - \beta_0^i\right) \right| ds \\
\geq & \int \left| \left| \left[h_n(s) - h_0(s)\right] \exp\left(\beta_0' x_0(s)\right) \right| - \left| h_n(s) \sum_{i=1}^{d} x_0^i(s) \exp\left(c' x_0(s)\right) \left(\beta_n^i - \beta_0^i\right) \right| \right| ds \\
\geq & \left| \int \left| \left[h_n(s) - h_0(s)\right] \exp\left(\beta_0' x_0(s)\right) \right| ds - \int \left| h_n(s) \sum_{i=1}^{d} x_0^i(s) \exp\left(c' x_0(s)\right) \left(\beta_n^i - \beta_0^i\right) \right| ds \right| \qquad (46)
\end{aligned}
$$

We have proven that (46) converges to zero. If $h_n(s) \nrightarrow^{L^1} h_0(s)$ and $\beta_n \rightarrow \beta_0$ we have a contradiction because, as we showed above, $\sup_n \int |h_n(s)| \, ds$ is bounded.

Note that (46) must converge to zero for any $x_0 \in S$ and for $x_0(s) + \epsilon(s) \in S$ where $\epsilon(s)$ perturbs only one covariate as outlined in the assumptions. We can define such a perturbation for each covariate by the theorem assumptions. Above we have proven

$$
\int \left| h_n(s) \exp\left(\beta_n' x_0(s)\right) - h_0(s) \exp\left(\beta_0' x_0(s)\right) \right| ds \rightarrow 0 \qquad (47)
$$

$$
\int \left| h_n(s) \exp\left(\beta_n' x_0(s)\right) \exp\left(\beta_n' \epsilon(s)\right) - h_0(s) \exp\left(\beta_0' x_0(s)\right) \exp\left(\beta_0' \epsilon(s)\right) \right| ds \rightarrow 0 \qquad (48)
$$

We now use a Taylor expansion of the term $\exp\left(\beta_n' \epsilon(s)\right)$ around β_n in (48).

$$
\begin{aligned}
& \int \left| \begin{array}{c} h_n(s) \exp\left(\beta_n' x_0(s)\right) \left[\exp\left(\beta_0' \epsilon(s)\right) + \epsilon^i(s) \exp\left(c' \epsilon(s)\right) \left(\beta_n^i - \beta_0^i\right)\right] \\ - h_0(s) \exp\left(\beta_0' x_0(s)\right) \exp\left(\beta_0' \epsilon(s)\right) \end{array} \right| ds \\
= & \int \left| \begin{array}{c} \left[h_n(s) \exp\left(\beta_n' x_0(s)\right) - h_0(s) \exp\left(\beta_0' x_0(s)\right)\right] \exp\left(\beta_0' \epsilon(s)\right) \\ + h_n(s) \exp\left(\beta_n' x_0(s)\right) \epsilon^i(s) \exp\left(c' \epsilon(s)\right) \left(\beta_n^i - \beta_0^i\right) \end{array} \right| ds \\
\geq & \int \left| \begin{array}{c} \left| \left[h_n(s) \exp\left(\beta_n' x_0(s)\right) - h_0(s) \exp\left(\beta_0' x_0(s)\right)\right] \exp\left(\beta_0' \epsilon(s)\right) \right| \\ - \left| h_n(s) \exp\left(\beta_n' x_0(s)\right) \epsilon^i(s) \exp\left(c' \epsilon(s)\right) \left(\beta_n^i - \beta_0^i\right) \right| \end{array} \right| ds \\
\geq & \left| \begin{array}{c} \int \left| \left[h_n(s) \exp\left(\beta_n' x_0(s)\right) - h_0(s) \exp\left(\beta_0' x_0(s)\right)\right] \exp\left(\beta_0' \epsilon(s)\right) \right| ds \\ - \int \left| h_n(s) \exp\left(\beta_n' x_0(s)\right) \epsilon^i(s) \exp\left(c' \epsilon(s)\right) \left(\beta_n^i - \beta_0^i\right) \right| ds \end{array} \right| \qquad (49)
\end{aligned}
$$

(47) shows that the first term in (49) converges to zero. Therefore, because $\int |h_n(s)| \, ds \nrightarrow 0$ and by the definition of $\epsilon(s)$, $\beta_n^i \rightarrow \beta_0^i$. Because we can define an appropriate $\epsilon(s)$ for each covariate by the assumptions, we have $\beta_n \rightarrow \beta_0$. As a result, $h_n(s) \rightarrow^{L^1} h_0(s)$. ■

I prove the following theorem below. This theorem implies Proposition 17 presented in subsection 4.3. I prove this implication later in this appendix.

Theorem 32 *We make Assumptions (B1)-(B3). Assume we choose a sequence of sieve spaces* Θ_n *with* *the number of basis functions used being J_n for n observations. Let there exist a sequence $h_n \in \Theta_n$ such that $h_n \to^{L^1} h_0$. Assume further that for $h \in \Theta_n$*

$$C^n_{\min} \leq h \leq C^n_{\max} \tag{50}$$

$$\left|\frac{h'}{h}\right| \leq K_n. \tag{51}$$

In addition, the constants $C^n_{\min}$, $C^n_{\max}$ and K_n must satisfy the following $\mathbb{P}_{\alpha_0} - a.s.$:

$$K_n \int_0^1 \left| \frac{1}{n} \sum_{i=1}^n \left[\mathbb{E} \left\{ \int_0^s h_0(t) \exp\left(\beta_0' Z^i(s)\right) \mathbf{1}_{\{\tau_i \geq t\}} dt \right\} - \int_0^s h_0(t) \exp\left(\beta_0' Z^i(s)\right) \mathbf{1}_{\{\tau_i \geq t\}} dt \right] \right| ds \to 0, \tag{52}$$

$$C^n_{\max} \sup_\beta \left(\int_0^1 \left| \frac{1}{n} \sum_{i=1}^n \left\{ \mathbb{E}\left[\exp\left(\beta' Z^i(s)\right) \mathbf{1}_{\{\tau_i \geq s\}} \right] - \exp\left(\beta' Z^i(s)\right) \mathbf{1}_{\{\tau_i \geq s\}} \right\} \right| ds \right) \to 0 \tag{53}$$

$$\left|\log\left(C^n_{\min}\right)\right| \vee \left|\log\left(C^n_{\max}\right)\right| \sup_\beta \left| \begin{array}{c} \frac{1}{n} \sum_{i=1}^n \left(\beta' Z^i(1)\right) \int_0^1 h_0(t) \exp\left(\beta_0' Z^i(s)\right) \mathbf{1}_{\{\tau_i \geq t\}} dt \\ -\mathbb{E}\left[\left(\beta' Z^i(1)\right) \int_0^1 h_0(t) \exp\left(\beta_0' Z^i(s)\right) \mathbf{1}_{\{\tau_i \geq t\}} dt \right] \end{array} \right| \to 0 \tag{54}$$

Assume $\overline{C}_n = Cn^{-1/4+\eta}$ for a small $\eta > 0$, where C is an arbitrary constant. Assume

$$1 \Bigg/ \left(\left|\log\left(C^n_{\min}\right)\right| \vee \left|\log\left(C^n_{\max}\right)\right| + \sup_{\beta,x} \left|\beta' x\right| \right) = \overline{C}_n, \tag{55}$$

and

$$\frac{1}{K_n} = \overline{C}_n. \tag{56}$$

Then

$$\widehat{\beta} \to \beta_0$$

$$\widehat{h} \to^{L^1} h_0$$

$\mathbb{P}_{\alpha_0} - a.s.$

The bound facilitated by the Burkholder inequality in the following proof is crude. In one instance, we bound a martingale at time $t = 1$ by its supremum over the interval $t \in [0, 1]$, then apply the Burkholder inequality. In another, we bound the integral of a martingale over $t \in [0, 1]$ by its supremum over the same interval, and again apply the Burkholder inequality. These bounds can likely be improved upon. This could improve the choice of the sequences K_n, $C^n_{\min}$ and $C^n_{\max}$.

Proof (Theorem 32). We will show that $H(h_0, \beta_0) - H\left(\widehat{h}, \widehat{\beta}\right) \to 0$ almost surely and the result follows from Lemma 31. Again we follow Karr (1987) Theorem 3.3. Define $\widetilde{\alpha} = \left(\widetilde{h}, \beta_0\right)$ where

$$\int \left|\widetilde{h} - h_0\right| ds \leq \inf_{h \in \Theta_n} \int |h - h_0| \, ds + o(1).$$

The minima need not be unique because we are dealing with L^1.

$$
\begin{aligned}
H(\alpha_0) - H(\widehat{\alpha}) &= H(\alpha_0) - H(\widetilde{\alpha}) \\
&\quad + H(\widetilde{\alpha}) - Q_n(\widetilde{\alpha}) \\
&\quad + Q_n(\widetilde{\alpha}) - Q_n(\widehat{\alpha}) \\
&\quad + Q_n(\widehat{\alpha}) - H(\widehat{\alpha}) \\
&\leq o(1) \qquad (57) \\
&\quad + H(\widetilde{\alpha}) - Q_n(\widetilde{\alpha}) \\
&\quad + o(1) \\
&\quad + Q_n(\widehat{\alpha}) - H(\widehat{\alpha})
\end{aligned}
$$

If we show the second and fourth terms in (57) converge to zero a.s., then $H(\alpha_0) - H(\widehat{\alpha}) \to 0$ a.s. and therefore $\widehat{\alpha} \to \alpha_0$. Note that the third line in (57) is $o(1)$ provided the other lines are $o(1)$. This is because $\widehat{\alpha}$ is chosen to maximize $Q_n(\alpha)$ and $0 \leq H(\alpha_0) - H(\widehat{\alpha})$. Consider the fourth term

$$
\begin{aligned}
Q_n(\widehat{\alpha}) - H(\widehat{\alpha}) &\leq \frac{1}{n}\sum_{i=1}^{n}\int_0^1 \log\left[\widehat{h}(s)\exp\left(\widehat{\beta}' Z^i(s)\right)\right] dN_s^i + \int_0^1 \left[1 - \widehat{h}(s)\exp\left(\widehat{\beta}' Z^i(s)\right)\mathbf{1}_{\{\tau_i \geq s\}}\right] ds \\
&\quad - \mathbb{E}_{(\alpha_0)}\left[\int_0^1 \log\left[\widehat{h}(s)\exp\left(\widehat{\beta}' Z^i(s)\right)\right] dN_s^i + \int_0^1 \left[1 - \widehat{h}(s)\exp\left(\widehat{\beta}' Z^i(s)\right)\mathbf{1}_{\{\tau_i \geq s\}}\right] ds\right] \\
&= \mathbb{E}\left[\int_0^1 \widehat{h}(s)\exp\left(\widehat{\beta}' Z^i(s)\right)\mathbf{1}_{\{\tau_i \geq s\}} ds\right] - \frac{1}{n}\sum_{i=1}^{n}\int_0^1 \widehat{h}(s)\exp\left(\widehat{\beta}' Z^i(s)\right)\mathbf{1}_{\{\tau_i \geq s\}} ds \qquad (58) \\
&\quad + \frac{1}{n}\sum_{i=1}^{n}\int_0^1 \log\left[\widehat{h}(s)\exp\left(\widehat{\beta}' Z^i(s)\right)\right] dN_s^i - \mathbb{E}_{(\alpha_0)}\left[\int_0^1 \log\left[\widehat{h}(s)\exp\left(\widehat{\beta}' Z^i(s)\right)\right] dN_s^i\right].
\end{aligned}
$$

We will from now on suppress the α_0 on the expectations with the understanding that expectations are taken with respect to the true underlying measure. Using integration by parts, (58) becomes

$$\begin{aligned}
(58) \quad = \quad & \mathbb{E}\left[\int_0^1 \widehat{h}(s)\exp\left(\widehat{\beta}' Z^i(s)\right)\mathbf{1}_{\{\tau_i\geq s\}}ds\right] - \frac{1}{n}\sum_{i=1}^n \int_0^1 \widehat{h}(s)\exp\left(\widehat{\beta}' Z^i(s)\right)\mathbf{1}_{\{\tau_i\geq s\}}ds \\
& -\frac{1}{n}\sum_{i=1}^n \int_0^1 N^i(s)\frac{\widehat{h}'(s)}{\widehat{h}(s)}ds + \log\left[\widehat{h}(1)\exp\left(\widehat{\beta}' Z^i(1)\right)\right]N^i(1) - 0(a.s.) \\
& -\mathbb{E}\left[-\int_0^1 N^i(s)\frac{\widehat{h}'(s)}{\widehat{h}(s)}ds\right] - \mathbb{E}\left[\log\left[\widehat{h}(1)\exp\left(\widehat{\beta}' Z^i(1)\right)\right]N^i(1)\right] \\
= \quad & \mathbb{E}\left[\int_0^1 \widehat{h}(s)\exp\left(\widehat{\beta}' Z^i(s)\right)\mathbf{1}_{\{\tau_i\geq s\}}ds\right] - \frac{1}{n}\sum_{i=1}^n \int_0^1 \widehat{h}(s)\exp\left(\widehat{\beta}' Z^i(s)\right)\mathbf{1}_{\{\tau_i\geq s\}}ds \qquad (59) \\
& -\frac{1}{n}\sum_{i=1}^n \int_0^1 N^i(s)\frac{\widehat{h}'(s)}{\widehat{h}(s)}ds + \frac{1}{n}\sum_{i=1}^n \int_0^1 \left[\frac{\widehat{h}'(s)}{\widehat{h}(s)}\int_0^s h_0(t)\exp\left(\beta_0' Z^i(t)\right)\mathbf{1}_{\{\tau_i\geq t\}}dt\right]ds \qquad (60) \\
& +\frac{1}{n}\sum_{i=1}^n \log\left[\widehat{h}(1)\exp\left(\widehat{\beta}' Z^i(1)\right)\right]N^i(1) - \mathbb{E}\left[\log\left[\widehat{h}(1)\exp\left(\widehat{\beta}' Z^i(1)\right)\right]N^i(1)\right] \qquad (61) \\
& +\mathbb{E}\left[\int_0^1 N^i(s)\frac{\widehat{h}'(s)}{\widehat{h}(s)}ds\right] - \frac{1}{n}\sum_{i=1}^n \int_0^1 \left[\frac{\widehat{h}'(s)}{\widehat{h}(s)}\int_0^s h_0(t)\exp\left(\beta_0' Z^i(t)\right)\mathbf{1}_{\{\tau_i\geq t\}}dt\right]ds. \qquad (62)
\end{aligned}$$

After this expansion, we need to consider the absolute value of (59) + (60) + (61) + (62). Therefore, we consider the absolute value of each of the lines from (59)-(62) individually as an upper bound.

$$\begin{aligned}
|(59)| \quad = \quad & \left|\frac{1}{n}\sum_{i=1}^n \left(\int_0^1 \widehat{h}(s)\left\{\mathbb{E}\left[\exp\left(\widehat{\beta}' Z^i(s)\right)\mathbf{1}_{\{\tau_i\geq s\}}\right] - \exp\left(\widehat{\beta}' Z^i(s)\right)\mathbf{1}_{\{\tau_i\geq s\}}\right\}ds\right)\right| \\
\leq \quad & \left|\int_0^1 \left(\widehat{h}(s)\frac{1}{n}\sum_{i=1}^n \left\{\mathbb{E}\left[\exp\left(\widehat{\beta}' Z^i(s)\right)\mathbf{1}_{\{\tau_i\geq s\}}\right] - \exp\left(\widehat{\beta}' Z^i(s)\right)\mathbf{1}_{\{\tau_i\geq s\}}\right\}\right)ds\right| \\
\leq \quad & C^n_{\max}\int_0^1 \left|\frac{1}{n}\sum_{i=1}^n \left\{\mathbb{E}\left[\exp\left(\widehat{\beta}' Z^i(s)\right)\mathbf{1}_{\{\tau_i\geq s\}}\right] - \exp\left(\widehat{\beta}' Z^i(s)\right)\mathbf{1}_{\{\tau_i\geq s\}}\right\}\right|ds \\
\leq \quad & C^n_{\max}\sup_{\beta}\left(\int_0^1 \left|\frac{1}{n}\sum_{i=1}^n \left\{\mathbb{E}\left[\exp\left(\beta' Z^i(s)\right)\mathbf{1}_{\{\tau_i\geq s\}}\right] - \exp\left(\beta' Z^i(s)\right)\mathbf{1}_{\{\tau_i\geq s\}}\right\}\right|ds\right).
\end{aligned}$$

$$\begin{aligned}
|(60)| \quad = \quad & \left|\frac{1}{n}\sum_{i=1}^n \int_0^1 N^i(s)\frac{\widehat{h}'(s)}{\widehat{h}(s)}ds - \frac{1}{n}\sum_{i=1}^n \int_0^1 \left[\frac{\widehat{h}'(s)}{\widehat{h}(s)}\int_0^s h_0(t)\exp\left(\beta_0' Z^i(s)\right)\mathbf{1}_{\{\tau_i\geq t\}}dt\right]ds\right| \\
= \quad & \left|\int_0^1 \left[\frac{\widehat{h}'(s)}{\widehat{h}(s)}\left\{\frac{1}{n}\sum_{i=1}^n \left(N^i(s) - \int_0^s h_0(t)\exp\left(\beta_0' Z^i(s)\right)\mathbf{1}_{\{\tau_i\geq t\}}dt\right)\right\}\right]ds\right| \\
\leq \quad & K_n\int_0^1 \left|\frac{1}{n}\sum_{i=1}^n \left(N^i(s) - \int_0^s h_0(t)\exp\left(\beta_0' Z^i(s)\right)\mathbf{1}_{\{\tau_i\geq t\}}dt\right)\right|ds \\
\leq \quad & K_n\sup_{s\in[0,1]}\left|\frac{1}{n}\sum_{i=1}^n \left(N^i(s) - \int_0^s h_0(t)\exp\left(\beta_0' Z^i(s)\right)\mathbf{1}_{\{\tau_i\geq t\}}dt\right)\right|. \qquad (63)
\end{aligned}$$

The term in the supremum of (63) is a martingale by arguments given in Section 2. As noted in Karr

(1987), this term is subject to the Burkholder inequality and therefore can be bounded in probability. A similar term will arise again, so we wait to give a specific bound.

$$
\begin{aligned}
|(61)| &= \left|\frac{1}{n}\sum_{i=1}^{n}\log\left[\widehat{h}(1)\exp\left(\widehat{\beta}'Z^i(1)\right)\right]N^i(1)-\mathbb{E}\left[\log\left[\widehat{h}(1)\exp\left(\widehat{\beta}'Z^i(1)\right)\right]N^i(1)\right]\right| \\
&\leq \left|\frac{1}{n}\sum_{i=1}^{n}\log\left[\widehat{h}(1)\exp\left(\widehat{\beta}'Z^i(1)\right)\right]\left(N^i(1)-\int_0^1 h_0(t)\exp\left(\beta_0'Z^i(t)\right)\mathbf{1}_{\{\tau_i\geq t\}}dt\right)\right| \quad (64) \\
&\quad+\left|\begin{array}{c}\frac{1}{n}\sum_{i=1}^{n}\log\left[\widehat{h}(1)\exp\left(\widehat{\beta}'Z^i(1)\right)\right]\int_0^1 h_0(t)\exp\left(\beta_0'Z^i(t)\right)\mathbf{1}_{\{\tau_i\geq t\}}dt \\ -\mathbb{E}\left[\log\left[\widehat{h}(1)\exp\left(\widehat{\beta}'Z^i(1)\right)\right]N^i(1)\right]\end{array}\right|. \quad (65)
\end{aligned}
$$

We handle the terms (64) and (65) separately

$$
\begin{aligned}
|(64)| &= \left|\frac{1}{n}\sum_{i=1}^{n}\log\left[\widehat{h}(1)\exp\left(\widehat{\beta}'Z^i(1)\right)\right]\left(N^i(1)-\int_0^1 h_0(t)\exp\left(\beta_0'Z^i(t)\right)\mathbf{1}_{\{\tau_i\geq t\}}dt\right)\right| \\
&\leq \left(\left|\log\left(C^n_{\min}\right)\right|\vee\left|\log\left(C^n_{\max}\right)\right|+\sup_{\beta,z}\left|\beta'z\right|\right)\left|\frac{1}{n}\sum_{i=1}^{n}\left(N^i(1)-\int_0^1 h_0(t)\exp\left(\beta_0'Z^i(t)\right)\mathbf{1}_{\{\tau_i\geq t\}}dt\right)\right| \\
&\leq \left(\left|\log\left(C^n_{\min}\right)\right|\vee\left|\log\left(C^n_{\max}\right)\right|+\sup_{\beta,z}\left|\beta'z\right|\right) \quad (66) \\
&\quad\times\sup_s\left|\frac{1}{n}\sum_{i=1}^{n}\left(N^i(s)-\int_0^s h_0(t)\exp\left(\beta_0'Z^i(t)\right)\mathbf{1}_{\{\tau_i\geq t\}}dt\right)\right|.
\end{aligned}
$$

$$
\begin{aligned}
|(65)| &= \left|\begin{array}{c}\frac{1}{n}\sum_{i=1}^{n}\log\left[\widehat{h}(1)\exp\left(\widehat{\beta}'Z^i(1)\right)\right]\int_0^1 h_0(t)\exp\left(\beta_0'Z^i(t)\right)\mathbf{1}_{\{\tau_i\geq t\}}dt \\ -\mathbb{E}\left[\log\left[\widehat{h}(1)\exp\left(\widehat{\beta}'Z^i(1)\right)\right]N^i(1)\right]\end{array}\right| \\
&= \left|\begin{array}{c}\frac{1}{n}\sum_{i=1}^{n}\log\left[\widehat{h}(1)\exp\left(\widehat{\beta}'Z^i(1)\right)\right]\int_0^1 h_0(t)\exp\left(\beta_0'Z^i(t)\right)\mathbf{1}_{\{\tau_i\geq t\}}dt \\ -\mathbb{E}\left[\log\left[\widehat{h}(1)\exp\left(\widehat{\beta}'Z^i(1)\right)\right]\int_0^1 h_0(t)\exp\left(\beta_0'Z^i(t)\right)\mathbf{1}_{\{\tau_i\geq t\}}dt\right]\end{array}\right| \\
&\leq \left|\log\left(C^n_{\min}\right)\right|\vee\left|\log\left(C^n_{\max}\right)\right|\left|\begin{array}{c}\frac{1}{n}\sum_{i=1}^{n}\left(\widehat{\beta}'Z^i(1)\right)\int_0^1 h_0(t)\exp\left(\beta_0'Z^i(t)\right)\mathbf{1}_{\{\tau_i\geq t\}}dt \\ -\mathbb{E}\left[\left(\widehat{\beta}'Z^i(1)\right)\int_0^1 h_0(t)\exp\left(\beta_0'Z^i(t)\right)\mathbf{1}_{\{\tau_i\geq t\}}dt\right]\end{array}\right| \\
&\leq \left|\log\left(C^n_{\min}\right)\right|\vee\left|\log\left(C^n_{\max}\right)\right|\sup_{\beta}\left|\begin{array}{c}\frac{1}{n}\sum_{i=1}^{n}\left(\beta'Z^i(1)\right)\int_0^1 h_0(t)\exp\left(\beta_0'Z^i(t)\right)\mathbf{1}_{\{\tau_i\geq t\}}dt \\ -\mathbb{E}\left[\left(\beta'Z^i(1)\right)\int_0^1 h_0(t)\exp\left(\beta_0'Z^i(t)\right)\mathbf{1}_{\{\tau_i\geq t\}}dt\right]\end{array}\right|. \quad (67)
\end{aligned}
$$

$$
\begin{aligned}
|(62)| &= \left| \mathbb{E}\left[\int_0^1 N^i(s)\frac{\widehat{h}'(s)}{\widehat{h}(s)}ds\right] - \frac{1}{n}\sum_{i=1}^n \int_0^1 \left[\frac{\widehat{h}'(s)}{\widehat{h}(s)}\int_0^s h_0(t)\exp\left(\beta_0' Z^i(t)\right)\mathbf{1}_{\{\tau_i \geq t\}}dt\right]ds \right| \\
&= \left| \begin{array}{c} \mathbb{E}\left[\int_0^1 \left[\frac{\widehat{h}'(s)}{\widehat{h}(s)}\int_0^s h_0(t)\exp\left(\beta_0' Z^i(t)\right)\mathbf{1}_{\{\tau_i \geq t\}}dt\right]ds\right] \\ -\frac{1}{n}\sum_{i=1}^n \int_0^1 \left[\frac{\widehat{h}'(s)}{\widehat{h}(s)}\int_0^s h_0(t)\exp\left(\beta_0' Z^i(t)\right)\mathbf{1}_{\{\tau_i \geq t\}}dt\right]ds \end{array} \right| \\
&= \left| \frac{1}{n}\sum_{i=1}^n \int_0^1 \frac{\widehat{h}'(s)}{\widehat{h}(s)} \left[\begin{array}{c} \mathbb{E}\left\{\int_0^s h_0(t)\exp\left(\beta_0' Z^i(t)\right)\mathbf{1}_{\{\tau_i \geq t\}}dt\right\} \\ -\int_0^s h_0(t)\exp\left(\beta_0' Z^i(t)\right)\mathbf{1}_{\{\tau_i \geq t\}}dt \end{array} \right] ds \right| \\
&\leq K_n \int_0^1 \left| \frac{1}{n}\sum_{i=1}^n \left[\begin{array}{c} \mathbb{E}\left\{\int_0^s h_0(t)\exp\left(\beta_0' Z^i(t)\right)\mathbf{1}_{\{\tau_i \geq t\}}dt\right\} \\ -\int_0^s h_0(t)\exp\left(\beta_0' Z^i(t)\right)\mathbf{1}_{\{\tau_i \geq t\}}dt \end{array} \right] ds \right|. \qquad (68)
\end{aligned}
$$

As outlined in Section 2, the terms

$$
M_t^n = \sum_{i=1}^n \left(N^i(s) - \int_0^s h_0(t)\exp\left(\beta_0' Z^i(t)\right)\mathbf{1}_{\{\tau_i \geq t\}}dt \right) \qquad (69)
$$

are martingales even when the covariates Z^i are dependent. Therefore, the Burkholder inequality can be used to bound the supremum of (69), even with dependence. See Appendix D for the relevant statement of Burkholder's inequality and Dellacherie and Meyer (1980) for a more general statement. For any $\epsilon > 0$,

$$
\begin{aligned}
\mathbb{P}_{\alpha_0}\left\{\frac{1}{n\overline{C}_n}\sup|M_t^n| > \epsilon\right\} &\leq \frac{1}{\left(n\overline{C}_n\epsilon\right)^4}\mathbb{E}\left\{\sup|M_t^n|^4\right\} \qquad (70) \\
&\leq \frac{1}{\left(n\overline{C}_n\epsilon\right)^4}C_2\mathbb{E}\left\{[M^n]_1^2\right\}
\end{aligned}
$$

The first inequality of (70) comes from Markov's inequality. The quadratic covariation map $(A, B) \to [A, B]$ is bilinear (Protter (2005) pg. 66). Therefore

$$
[M^n]_t = \sum_{j=1}^n \left(\sum_{i=1}^n \left[M^i, M^j\right]_t \right)
$$

The semimartingales M^i are "quadratic pure jump" (Protter (2005) pg. 70-71). By Protter (2005) Theorem 28 pg. 75, this implies $\left[M^i, M^j\right]_1 = 0$ a.s. because jumps in both processes only happen at the same time with probability zero. We note that $\left[M^i\right]_1 = N_1^i$ (Protter (2005) pg. 70). Therefore,

$$
[M^n]_1 = \sum_{i=1}^n N_1^i.
$$

Because $0 \leq N_1^i \leq 1$,

$$
\begin{aligned}
\mathbb{E}\left\{[M^n]_1^2\right\} &= \sum_{i=1}^n \mathbb{E}\left\{\left(N_1^i\right)^2\right\} + 2\sum_{i \neq j}\mathbb{E}\left\{N_1^i N_1^j\right\} \\
&\leq n + 2n(n-1)
\end{aligned}
$$

This implies

$$\frac{1}{\left(n\overline{C}_n\epsilon\right)^4}C_2\mathbb{E}\left\{[M^n]_1^2\right\}=O\left(n^2\overline{C}_n^4\epsilon^4\right)$$

Therefore, we need $n\overline{C}_n^2\to\infty$ and $\overline{C}_n=n^{-1/4+\eta}$. Above, $\overline{C}_n$ takes the following two values:

$$\begin{aligned}\overline{C}_n &= 1\bigg/\left(\left|\log\left(C^n_{\min}\right)\right|\vee\left|\log\left(C^n_{\max}\right)\right|+\sup_{\beta,x}\left|\beta'x\right|\right),\\ \overline{C}_n &= \frac{1}{K_n}.\end{aligned}$$

By the assumptions of the theorem, a Borel-Cantelli argument gives

$$\frac{1}{n\overline{C}_n}\sup\left|M^n_t\right|\to 0,\qquad \mathbb{P}_{\alpha_0}-a.s.$$

Therefore, by the assumptions of the theorem, $\left|Q_n\left(\widehat{\alpha}\right)-H\left(\widehat{\alpha}\right)\right|\to 0$. Notice that throughout the proof the exact value of $\widehat{\alpha}$ was irrelevant and the results hold for an arbitrary sequence $\alpha_n\in\Theta_n$ under the assumptions on Θ_n. Therefore, $\left|H\left(\widetilde{\alpha}\right)-Q_n\left(\widetilde{\alpha}\right)\right|\to 0$ and by (57), $H\left(\alpha_0\right)-H\left(\widehat{\alpha}\right)\to 0$. By Lemma 31,

$$\widehat{\beta}\to\beta_0$$

$$\widehat{h}\to^{L^1}h_0$$

$\mathbb{P}_{\alpha_0}-a.s.$ ■

Proof (Proposition 17). For η small, set

$$p=1/\left(\frac{3}{4}+\eta\right)$$

For each fixed s, let

$$B^{i,1}_s=\int_0^s h_0\left(t\right)\exp\left(\beta_0'Z^i\left(s\right)\right)\mathbf{1}_{\{\tau_i\geq t\}}dt$$

satisfy the mixing condition

$$\sum_{n>0}n^{p-2}\alpha\left(n\right)<\infty.$$

For all $\beta\in[a_1,b_1]\times\cdots\times[a_q,b_q]$ and all fixed s, let the variables

$$B^{i,2}_s=\exp\left(\beta'Z^i\left(s\right)\right)\mathbf{1}_{\{\tau_i\geq s\}}$$

$$B^{i,3}=\left(\beta'Z^i\left(1\right)\right)\int_0^1 h_0\left(t\right)\exp\left(\beta_0'Z^i\left(s\right)\right)\mathbf{1}_{\{\tau_i\geq t\}}dt$$

satisfy the mixing condition

$$\sum_{n>0}n^{-1}\alpha\left(n\right)<\infty.\tag{71}$$

The theorem holds under these assumptions and the assumptions stated in the proposition. We make mixing conditions on the underlying covariates instead of $B^{i,1}_s$, $B^{i,2}_s$ and $B^{i,3}$. A proof similar to Lemma

5 shows that mixing conditions on the covariate processes $Z^i(s)$ are transferred to mixing conditions on the relevant sequences of random variables $B_s^{i,1}$, $B_s^{i,2}$ and $B^{i,3}$. These proofs are omitted for brevity. As in Karr (1987), we apply a Marcinkiewicz-Zygmund type strong law of large numbers to show (52)-(54) are satisfied. That paper is not as specific about the argument as I would like, so I provide more details here. The Marcinkiewicz-Zygmund strong law we comes from Rio (1995). This strong law allows for dependence between and within covariate processes in our application. Note that the random variables $B_s^{i,1}$, $B_s^{i,2}$ and $B^{i,3}$ are bounded. This simplifies the results in Rio (1995) as discussed in that paper.

As a result of the specified strong law, there exists a set Ω_0 of probability 1 such that, for a countable dense set $\overline{S} \subset [0,1]$ if $\overline{s} \in \overline{S}$ the following holds for all $\omega \in \Omega_0$:

$$K_n \left| \frac{1}{n} \sum_{i=1}^{n} \left[\mathbb{E}\left\{ \int_0^{\overline{s}} h_0(t) \exp\left(\beta_0' Z^i(s)\right) \mathbf{1}_{\{\tau_i \geq t\}} dt \right\} - \int_0^{\overline{s}} h_0(t) \exp\left(\beta_0' Z^i(s)\right) \mathbf{1}_{\{\tau_i \geq t\}} dt \right] \right| \to 0. \tag{72}$$

Let $\widetilde{s} \notin \overline{S}$ and $\widetilde{s} + \delta \in \overline{S}$ with δ arbitrarily small. This can always be done because $\overline{S}$ is dense in $[0,1]$.

$$\begin{aligned}
& K_n \left| \frac{1}{n} \sum_{i=1}^{n} \left[\mathbb{E}\left\{ \int_0^{\widetilde{s}} h_0(t) \exp\left(\beta_0' Z^i(s)\right) \mathbf{1}_{\{\tau_i \geq t\}} dt \right\} - \int_0^{\widetilde{s}} h_0(t) \exp\left(\beta_0' Z^i(s)\right) \mathbf{1}_{\{\tau_i \geq t\}} dt \right] \right| \\
\leq \; & K_n \left| \frac{1}{n} \sum_{i=1}^{n} \left[\mathbb{E}\left\{ \int_0^{\widetilde{s}} h_0(t) \exp\left(\beta_0' Z^i(s)\right) \mathbf{1}_{\{\tau_i \geq t\}} dt \right\} - \mathbb{E}\left\{ \int_0^{\widetilde{s}+\delta} h_0(t) \exp\left(\beta_0' Z^i(s)\right) \mathbf{1}_{\{\tau_i \geq t\}} dt \right\} \right] \right| \\
& + K_n \left| \frac{1}{n} \sum_{i=1}^{n} \left[\mathbb{E}\left\{ \int_0^{\widetilde{s}+\delta} h_0(t) \exp\left(\beta_0' Z^i(s)\right) \mathbf{1}_{\{\tau_i \geq t\}} dt \right\} - \int_0^{\widetilde{s}+\delta} h_0(t) \exp\left(\beta_0' Z^i(s)\right) \mathbf{1}_{\{\tau_i \geq t\}} dt \right] \right| \\
& + K_n \left| \frac{1}{n} \sum_{i=1}^{n} \left[\int_0^{\widetilde{s}+\delta} h_0(t) \exp\left(\beta_0' Z^i(s)\right) \mathbf{1}_{\{\tau_i \geq t\}} dt - \int_0^{\widetilde{s}} h_0(t) \exp\left(\beta_0' Z^i(s)\right) \mathbf{1}_{\{\tau_i \geq t\}} dt \right] \right| \\
= \; & K_n C^1(\delta) + o_{a.s.}(1) + K_n C^2(\delta).
\end{aligned} \tag{73}$$

A sequence δ can be chosen to converge to 0 fast enough so $K_n C^1(\delta) = o(1)$ and $K_n C^2(\delta) = o(1)$. This implies (72) holds for all $s \in [0,1]$. Therefore, condition (52) from Theorem 32 holds. Along with the assumptions in Proposition 17, a simple bracketing argument implies conditions (53) and (54) in Theorem 32. See Pollard (1984) Section II.2 for the required bracketing results. This needs to be coupled with an argument similar to that given above for condition (52). The details are omitted. ■

The same decomposition into terms as done in the proof of Theorem 32 can be done with the extended form of block/step sampling outlined in Subsection 4.4. The difference now is we additionally have to sum over $k(n)$. We still must show the terms of the decomposition converge to zero almost surely. The terms handled by the Burkholder inequality can still be handled the same way as the martingale structure is preserved. Provided $K_n \to \infty$ at a rate slower than or equal to $o\left((k(n)/n)^{1/2}\right)$, the martingale terms converge to zero almost surely with the same proof. The remaining terms that need to be handled are:

$$K_n \int_0^1 \left| \sum_{j=1}^{k(n)} \frac{1}{n} \sum_{i=1}^{n} \left[\begin{array}{c} \mathbb{E}\left\{ \int_0^s h_0^j(t) \exp\left(\beta_0' Z^{ji}(s)\right) \mathbf{1}_{\{\tau_{ji} \geq t\}} dt \right\} \\ - \int_0^s h_0^j(t) \exp\left(\beta_0' Z^{ji}(s)\right) \mathbf{1}_{\{\tau_{ji} \geq t\}} dt \end{array} \right] \right| ds \to 0 \;\; a.s., \tag{74}$$

$$\sup_{\beta}\left(\int_{0}^{1}\left|\sum_{j=1}^{k(n)}\frac{1}{n}\sum_{i=1}^{n}\left\{\mathbb{E}\left[\exp\left(\beta' Z^{ji}(s)\right)\mathbf{1}_{\{\tau_{ji}\geq s\}}\right]-\exp\left(\beta' Z^{ji}(s)\right)\mathbf{1}_{\{\tau_{ji}\geq s\}}\right\}\right|ds\right)\to 0 \;\; a.s., \tag{75}$$

$$\sup_{\beta}\left|\sum_{j=1}^{k(n)}\frac{1}{n}\sum_{i=1}^{n}\left[\begin{array}{c}\left(\beta' Z^{ji}(1)\right)\int_0^1 h_0^j(t)\exp\left(\beta_0' Z^{ji}(s)\right)\mathbf{1}_{\{\tau_{ji}\geq t\}}dt\\ -\mathbb{E}\left\{\left(\beta' Z^{ji}(1)\right)\int_0^1 h_0(t)\exp\left(\beta_0' Z^{ji}(s)\right)\mathbf{1}_{\{\tau_{ji}\geq t\}}dt\right\}\end{array}\right]\right|\to 0 \;\; a.s. \tag{76}$$

Proof. The result follows by a slight modification of the proof of Theorem 32. It can be shown that

$$\sum_{j=1}^{k(n)}\left\{H^j\left(h_0^j,\beta_0\right)-H^j\left(\widehat{h}^j,\widehat{\beta}\right)\right\}\to 0, \tag{77}$$

$a.s.$ and the result follows. The main difference between the proof is that in (77) there is a sum over the number of blocks $k(n)$. Note that the Burkholder inequality applies to martingales defined on $[0,\infty]$ as in this extension. See Dellacherie and Meyer (1980). ■

D Background Martingale Theory

In this appendix, I briefly present some martingale theory used in this paper. For far more comprehensive accounts see Fleming and Harrington (1991), Protter (2005) or Dellacherie and Meyer (1980). Processes are defined on a probability space $(\Omega,\mathcal{F},\mathbb{P})$ equipped with a filtration of sub-σ-fields $(\mathcal{F}_t, t\geq 0)$.

Proposition 33 (Fleming and Harrington (1991) Corollary 1.4.2.) *Let M_t be a right continuous martingale with respect to a right continuous filtration and assume $\mathbb{E}M^2(t)<\infty$ for any $t\geq 0$. Then there exists a unique increasing right-continuous predictable process $\langle M,M\rangle_t$ such that $\langle M,M\rangle_0=0$ a.s., $\mathbb{E}\langle M,M\rangle_t<\infty$ for each t and $M_t^2-\langle M,M\rangle_t$ is a right-continuous martingale.*

Proposition 34 (Fleming and Harrington (1991) Theorem 1.4.2.) *Let $M_1(t)$ and $M_2(t)$ be two right continuous martingale with respect to a right continuous filtration and assume $\mathbb{E}M_i^2(t)<\infty$ for any $t\geq 0$, $i=1,2$. Then there exists a unique right-continuous predictable process $\langle M_1,M_2\rangle_t$ such that $\langle M_1,M_2\rangle_0=0$ a.s., $\mathbb{E}\langle M_1,M_2\rangle_t<\infty$ for each t and $M_1M_2-\langle M_1,M_2\rangle$ is a martingale.*

The left-continuous version of a process $X(t)$ is given by $X(-t)=\lim_{s\uparrow t}X(s)$ and by definition $X(-0)=0$.

Definition 35 (Quadratic Variation Process, Protter (2005)) *Let $M(t)$ be a martingale with finite variation. The quadratic variation process is denoted $[M,M]_t$ and defined as*

$$[M,M]_t=M_t^2-2\int M(-t)\,dM(t).$$

This definition is equivalent to the following definition. Let $\{t_i^n\}$ denote a sequence of partitions of $[0,T]$. Each successive partition if finer than the previous one and assume

$$\sup\left|t_i^n-t_{i-1}^n\right|\to 0$$

as $n \to \infty$. *Then*

$$M_0^2 + \sum_i \left(M\left(t_i^n\right) - M\left(t_{i-1}^n\right)\right)^2 \to^p [M, M]_T .$$

This definition defines a process on $[0, T]$*, as the previous one does, by using partitions constrained to* $[0, t]$ *with* $t \leq T$*.*

The following is a version of the Burkholder inequality used in this paper. More general versions exist.

Theorem 36 (Burkholder Inequality) *There exists a universal constant* C *such that for every martingale* $M(t)$ *and every finite time* T*,*

$$\mathbb{E}\left\{\sup_{t\in[0,T]} |M_t|^4\right\} \leq C\mathbb{E}\left\{[M^n]_T^2\right\}.$$

References

[1] Aalen, O.O., Ø. Borgan & H.K. Gjessing (2010) *Survival and Event History Analysis.* Springer.

[2] Andersen, P.K., Ø. Borgan, R.D. Gill & N. Keiding (1994) *Statistical Models Based on Counting Processes.* Springer.

[3] Andersen, P.K. & R.D. Gill (1982) Cox's Regression Model for Counting Processes: A Large Sample Study. *Annals of Statistics*

[4] Azizpour, S., K. Giesecke and G. Schwenkler (2011) Exploring the Sources of Default Clustering. Unpublished.

[5] Bielecki, T.R. & M. Rutkowski (2004) *Credit Risk: Modeling, Valuation and Hedging.* Springer.

[6] Billingsley, P. (1999) *Convergence of Probability Measures, 2nd ed.* Wiley.

[7] Bosq, D. (1998) *Nonparametric Statistics for Stochastic Processes.* Springer.

[8] Brémaud, P. (1981) *Point Processes and Queues: Martingale Dynamics.* Springer-Verlag.

[9] Burda, M., M. Harding & J. Hausman (2008) A Bayesian Mixed Logit-Probit Model for Multinomial Choice. *Journal of Econometrics*, 147 (2), 232-246.

[10] Burda, M., M. Harding & J. Hausman (2011) Understanding Choice Intensity: A Poisson Mixture Model with Logit-based Random Utility Selective Mixing. Unpublished.

[11] Carrasco, M. & J-P. Florens (2000) Generalization of GMM to a Continuum of Moment Conditions. *Econometric Theory* 16, 797-834.

[12] Chen, X. (2007) Large sample sieve estimation of semi-nonparametric models, chapter 76 in Handbook of Econometrics, Vol. 6B, 2007, eds. James J. Heckman and Edward E. Lleamer, North-Holland.

[13] Chen, X., Z. Liao and Y. Sun (2011) Sieve Inference for Weakly Dependent Data. Working paper.

[14] Chen, X. & X. Shen (1998) Sieve Extremum Estimates for Weakly Dependent Data. *Econometrica* 66 (2), 289-314.

[15] Cherubini, U., E. Luciano & W. Vecchiato (2006) *Copula Methods in Finance.* Wiley Finance.

[16] Chui, C.K. (1992) *An Introduction to Wavelets.* Academic Press.

[17] Collin-Dufresne, P., R. Goldstein & J. Helwege (2009) How Large Can Jump-to-Default Risk Premia be? Modeling Contagion via the Updating of Beliefs. Unpublished manuscript.

[18] Das, S.R., D. Duffie, N. Kapadia & L. Saita (2007) Common Failings: How Corporate Defaults are Correlated. *Journal of Finance* 62 (1), 93-117.

[19] Davis, M. & V. Lo (2001) Infectious Defaults. *Quantitative Finance 1, 382-386.*

[20] de Boor, C. (2001) *A Practical Guide to Splines.* Springer.

[21] Dellacherie, C. & P.A. Meyer (1980) *Probabilities and Potential B.* North Holland.

[22] Doraszelski, U. & K.L. Judd (2010) Avoiding the Curse of Dimensionality in Dynamic Stochastic Games. Unpublished.

[23] Doukhan, P. & M.H. Neumann (2005) Probability and Moment Inequalities for Sums of Weakly Dependent Random Variables, with Applications. Unpublished.

[24] Duffie, D., A. Eckner, G. Horel & L. Saita (2009) Frailty Correlated Default. *Journal of Finance* 64 (5), 2089-2123.

[25] Duffie, D., L. Saita & K. Wang (2007) Multi-period Corporate Default Prediction with Stochastic Covariates. *Journal of Financial Economics* 83, 635-665.

[26] Dzhaparidze, K. & J.H. van Zanten (2001) On Bernstein-type Inequalities and Martingales. *Stochastic Processes and their Applications* 93, 109-117.

[27] Ethier, S.N. & T.G. Kurtz (1986) *Markov Processes: Characterization and Convergence.* Wiley.

[28] Fleming, T.R. & D.P. Harrington (1991) *Counting Processes and Survival Analysis*, Wiley.

[29] Giesecke, K. & B. Kim (2011) Systemic Risk: What Defaults Are Telling Us. Unpublished Manuscript.

[30] Grenander, U. (1981) *Abstract Inference.* Wiley.

[31] Hall, P., C.C. Heyde (1980) *Martingale Limit Theory and its Application.* Academic Press.

[32] Hawkes, A.G. (1971) Spectra of some self-exciting and mutually exciting point processes. *Biometrika* 58 (1), 83-90.

[33] Hougaard, P. (2000) *Analysis of Multivariate Survival Data.* Springer.

[34] Jacod, J. & A.N. Shiryaev (2003) *Limit Theorems for Stochastic Processes, 2nd ed.* Springer.

[35] Jarrow, R.A. & F. Yu (2001) Counterparty Risk and the Pricing of Defaultable Securities. *Journal of Finance* 56 (5), 1765-1799.

[36] Johansen, S. (1983) An Extension of Cox's Regression Model. *International Statistical Review* 51, 165-174.

[37] Karr, A.F. (1987) Multiplicitive Likelihood Estimation in the Multiplicative Intensity Model Via Sieves. *Annals of Statistics* 15 (2), 473-490.

[38] Koopman, S.J., A. Lucas & B. Schwaab (2011) Modeling Frailty-Correlated Defaults Using Many Macroeconomic Covariates. *Journal of Econometrics* 162, 312-325.

[39] Kutoyants, Y.A. (2004) *Statistical Inference for Ergodic Diffusion Processes.* Springer.

[40] Lang, S. (1993) *Real and Functional Analysis: Third Edition.* Springer.

[41] Liebscher, E. (1996) Strong Convergence of Sums of α-Mixing Random Variables with Applications to Density Estimation. *Stochastic Processes and their Applications* 65, 69-80.

[42] Linton, O.B., J.P. Nielsen & S. Van de Geer (2003) Estimating Multiplicative and Additive Hazard Functions by Kernel Methods. *Annals of Statistics* 31 (2), 464-492.

[43] Liptser, R.Sh. & A.N. Shiryayev (1980) A Functional Central Limit Theorem for Semimartingales. *Theory of Probability and its Applications* 25 (4), 667-688.

[44] Liptser, R.Sh. & A.N. Shiryayev (1989) *Theory of Martingales.* Kluwer.

[45] Mammen, E. & J.P. Nielsen (2007) A General Approach to the Predictability Issue in Survival Analysis with Applications. *Biometrika* 94 (4), 873-892.

[46] Martinussen, T. and Scheike T.H. (2010) *Dynamic Regression Models for Survival Data.* Springer.

[47] Mayer, C., K. Pence & S.M. Sherlund (2009) The Rise of Mortgage Defaults. *Journal of Economic Prospectives* 23 (1), 27-50.

[48] Merlevède, F., M. Peligrad & E. Rio (2010) Bernstein Inequality and Moderate Deviations Under Strong Mixing Condtions. *IMS Colletions, High Dimensional Probability 5, 273-292.*

[49] Merlevède, F., M. Peligrad & E. Rio (2010) A Bernstein Type Inequality and Moderate Deviations for Weakly Dependent Sequences. *Probability Theory and Related Fields.*

[50] Modha, D.S. & E. Masry (1996) Minimum Complexity Regression Estimation with Weakly Dependent Observations. *IEEE Transactions on Information Theory* 42 (6), 2133-2145.

[51] Nielsen, J.P. (1990) Kernel Estimation of Densities and Hazards: a Counting Process Approach. Ph.D dissertation, Dept. Biostatistics, Univ. California, Berkeley.

[52] Nielsen, J.P. & O.B. Linton (1995) Kernel Estimation in a Nonparametric Marker Dependent Hazard Model. *Annals of Statistics* 23 (5), 1735-1748.

[53] Pollard, D. (1984) *Convergence of Stochastic Processes.* Springer-Verlag.

[54] Prakasa Rao, B.L.S. (1999) *Statistical Inference for Diffusion Type Processes.* Arnold.

[55] Protter, P.E. (2005) *Stochastic Integration and Differential Equations.* Springer.

[56] Ramlau-Hansen, H. (1983) Smoothing Counting Process Intensities by Means of Kernel Functions. *Annals of Statistics* 11, 453-466.

[57] Röckner, M. & T.S. Zhang (1996) Finite Dimensional Approximation of Diffusion Processes on Infinite Dimensional Spaces. *Stochastics and Stochastics Reports* 57, 37-55.

[58] Roussas, G.G. (1996) Exponential Probability Inequalities with some Applications. *IMS Lecture Notes* 30, 303-319.

[59] Schönbucher, P.J. (2005) *Credit Derivatives Pricing Models: Models, Pricing and Implementation.* Wiley Finance.

[60] Shorack, G.R. & J.A. Wellner (1986) *Empirical Processes with Applications to Statistics.* SIAM.

[61] Stroock, D.W. & S.R.S. Varadhan (1979) *Multidimensional Diffusion Processes.* Springer.

[62] Van den Berg, G.J. (2001) Duration Models: Specification, Identification and Multiple Durations, chapter 55 in Handbook of Econometrics, Vol. 5, 2001, eds. James J. Heckman and Edward E. Lleamer, Elsevier Science.

[63] Van der Vaart, A.W. (1998) *Asymptotic Statistics.* Cambridge.