Estimation for the Prediction of Point Processes with Many Covariates

Alessio Sancetta*

April 13, 2017

Abstract

Estimation of the intensity of a point process is considered within a nonparametric framework. The intensity measure is unknown and depends on covariates, possibly many more than the observed number of jumps. Only a single trajectory of the counting process is observed. Interest lies in estimating the intensity conditional on the covariates. The impact of the covariates is modelled by an additive model where each component can be written as a linear combination of possibly unknown functions. The focus is on prediction as opposed to variable screening. Conditions are imposed on the coefficients of this linear combination in order to control the estimation error. The rates of convergence are optimal when the number of active covariates is large. As an application, the intensity of the buy and sell trades of the New Zealand Dollar futures is estimated and a test for forecast evaluation is presented. A simulation is included to provide some finite sample intuition on the model and asymptotic properties.

Keywords: Cox process; counting process; curse of dimensionality; forecast evaluation; greedy algorithm; Hawkes process, high frequency trading; martingale; trade arrival; variable selection.

JEL Codes: C13; C32; C55.

1 Introduction

Suppose that you want to estimate and then predict the likelihood of a trade arrival for a financial instrument that trades relatively frequently. The reason for doing so could be market making or optimal execution. These problems are quite common in the financial industry. In an application to be considered here, the instrument is the futures on the New Zealand Dollar. A trade arrival for such an instrument may depend on the state of the order book,

*Acknowledgement: I would like to thank the Editor, the Co-Editor, the referees, and Luca Mucciante for comments that led to substantial improvements both in content and presentation. E-mail: <asancetta@gmail.com>, URL: <http://sites.google.com/site/wwwsancetta/>. Address for correspondence: Department of Economics, Royal Holloway University of London, Egham TW20 0EX, UK.

which contains 5 levels on the bid and the offer. It may also depend on what happens with other related instruments, their past prices and quoted volumes dynamics, as well as on past trades. The number of possible covariates can grow quickly and become relatively large, even for high frequency data.

Problems such as the one just described can be addressed considering trade arrivals as the jump of a counting process whose intensity (the mean over an infinitesimal time period) depends on a set of covariates. This paper considers the estimation of such counting processes for problems where the data are time series, the number of covariates is large, and the functional form of the intensity does not need to be parametric.

Let $(N(t))_{t\geq 0}$ be a counting process with intensity measure

$$\Lambda(A) = \int_A \exp\{g_0(X(t))\}\, dt, \tag{1}$$

for any Borel set $A \subseteq [0,\infty)$, where g_0 is an unknown function, $X(t)$ are K dimensional covariates that can depend on t. Often, the intensity in (1) is intuitively understood to mean

$$\lim_{s\downarrow 0} \frac{\Pr(N(t+s) - N(t) = 1|\mathcal{F}_t)}{s} = \exp\{g_0(X(t))\},$$

where $\mathcal{F}_t$ is the sigma algebra generated by $(N(s), X(s))_{s\leq t}$. Given that the covariates are time dependent, the intensity may depend on the time elapsed from the last jump of $N(t)$. The covariates are predictable, for example, adapted left continuous processes. If the process is Poisson when conditioning on the covariates $X(t)$, then the counting process is usually referred to as a Cox or doubly stochastic process.

Define the stopping times $T_i := \inf\{s > 0 : N(s) \geq i\}$, $T_0 = 0$, i.e., T_i is the time of the i^{th} jump. In the empirical financial microstructure application to be considered in this paper, the jump time T_i is the time of the i^{th} trade arrival for a specific security, and the covariates will be information extracted from the order book, among other quantities. The statistical problem is where one observes $(N(t), X(t))$ up to time T. By definition of the stopping times, waiting until $T = T_n$ means that one observes n jumps. The goal is to estimate g_0. This function g_0 is only known to lie in some class of additive functions, which will be introduced in due course. The covariates and the durations between jumps are supposed to be stationary, but neither independent nor Markovian.

The time series problem where only one trajectory of the process is observed and g_0 in (1) is possibly nonlinear with a large number of covariates has not been previously discussed in the literature. The framework allows us to handle ultra-high dimensional problems where the number of covariates is exponentially larger than the sample size (n when $T = T_n$). The covariates could be time series and lagged variables. This setup is motivated by many applied problems, such as the previously mentioned trading arrival estimation problem (Bauwens and

Hautsch, 2009, for a survey and references for counting models applied to finance). A traded instrument may depend on updates and information from other instruments. This leads to a proliferation of the possible number of variables, though one might expect that either a handful of them might be relevant or many covariates could explain the intensity with a decreasing degree of importance. In the modelling application in Section 4, one ends up with more than one thousand variables with the number of trades n of about a thousand.

The main technical features of the present study are: 1. estimation of g_0 in (1), when g_0 is only known to lie in some large set of functions; 2. the number of covariates is allowed to be larger than the number of observed durations n; 3. a class of additive functions is defined and it is shown that within this class one can obtain convergence rates that are optimal in the high dimensional case; 4. the estimation problem can be solved by the Frank-Wolfe algorithm and rates of convergence are given; 5. an empirical study provides applicability of the methodology and a test for forecast superiority between counting models, showing that suitably constrained large models can perform better out of sample.

From a theoretical point of view, restrictions on the absolute summability of linear coefficients (the l_1 norm of the coefficients) in the additive model are imposed. Such a Lasso type of constraint tends to produce models that are sparse. This means that if all the coefficients are nonzero but small, tightening the constraint leads to few nonzero coefficients. It is well known that tightening a constraint on the sum of the squared coefficients (i.e., l_2 norm as in ridge regression) leads to all coefficients being small, but none being zero.

From an empirical point of view, the paper considers an estimation of the intensity for the arrival of buy and sell trades on the New Zealand Dollar futures contract. The intensity is modelled using many covariates of the same order of magnitude as the number of durations. Estimation of the intensity for buy and sell orders has been considered in the literature (e.g., Hall and Hautsch 2007). However, no study appears to consider market information (e.g., the order book) on the traded instrument as well as other related instruments. The out of sample results show that information provided by additional instruments is relevant. To evaluate the out of sample performance of competing models, an out of sample test based on the likelihood ratio is used.

Details concerning the proofs and in text derivations are provided online at Cambridge Core in supplementary material to this article. Readers may refer to the supplementary material associated with this article, available at Cambridge Core (www.cambridge.org/core/journals/econometric-theory).

1.1 Relation to the Literature

In the regression context, high dimensional additive modelling has been considered in the literature (e.g., Bühlmann and van de Geer, 2011, and references therein). This paper seems

to be the first to consider estimation with many covariates, allowing for a nonlinear link function in a time series context. Here, time series means that only one single realization of the process is observed over a window expanding in the future. This framework differs from the Cox proportional hazard model and Aalen multiplicative and additive model. In that context, estimation with many variables has been considered by various authors (e.g., Bradic et al., 2011, Gaiffas and Guilloux, 2012, amongst others) and the focus is often in recovering the true subset of active variables. This often results in stringent restrictions on the covariates' design and cross-dependence. In this paper, the focus is on prediction and on weak conditions that can lead to consistency even when the number of non-negligible covariates grows with the sample size. Beyond additivity, the estimation considered here is very general. Section 3.6 provides an overview of the applications. These include linear models, Hawkes processes with covariates, threshold models, and additive monotone functions among other possibilities.

The analysis of estimators of the intensity function usually relies on martingale methods (Andersen and Gill, 1982, van de Geer, 1995). In the context of a fixed number of covariates, nonparametric estimators are not uncommon (Nielsen and Linton, 1995, Fan et al., 1997).

The results derived here apply to parametric as well as to certain nonparametric classes of functions. In the financial econometrics literature, interest often lies in parametric modelling of a single point process (Bauwens and Hautsch, 2009, for a survey). Hence, the current paper considers the time series problem as in the financial econometrics literature, but allows for a possibly nonparametric estimation and for a large number of covariates as done in high dimensional statistics.

In a time series context, the intensity is often modelled by Hawkes processes and can be written as a predictable function of durations (e.g., Bauwens and Hautsch, 2009). The framework of this paper allows the aforementioned variables to be covariates.

1.2 Likelihood Estimation

It is well known (e.g., Brémaud, 1981, Ch.II, Theorem 16) that $\{\Lambda((T_{i-1}, T_i]) : i \in \mathbb{N}\}$ (Λ as in (1)) is i.i.d. exponentially distributed with mean 1. The likelihood is easily derived from here, assuming that Λ has density λ with respect to the Lebesgue measure (e.g., Ogata, 1978, eq.1.3).

Define the population log-likelihood

$$L(g) := \mathbb{E}g(X(0)) \exp\{g_0(X(0))\} - \mathbb{E}\exp\{g(X(0))\}, \tag{2}$$

assuming the expectations are well defined (see Section A.1.2 in the supplementary material). Suppose that g_0 in (1) lies in a set $\mathcal{G}$, momentarily assumed to be countable to avoid distracting technicalities. Then, $g_0 = \arg\sup_{g \in \mathcal{G}} L(g)$ using concavity of the log-likelihood. Given that expectations are unknown, the above is replaced by the empirical estimator $g_T := \arg\sup L_T(g)$,

where the sup is over a class of functions to be defined in the next section and the sample likelihood is

$$L_T(g) := \int_0^T g(X(t))\, dN(t) - \int_0^T \exp\{g(X(t))\}\, dt, \tag{3}$$

where $L(g) = \lim_T L_T(g)/T$ almost surely (see Section A.1.2 in the supplementary material for the proof of this statement). Supposing that one waits until a time T_n such that $N(T_n) = n$, the above can be written as

$$L_{T_n}(g) := \sum_{i=1}^{n} \left[g(X(T_i)) - \int_{T_{i-1}}^{T_i} \exp\{g(X(t))\}\, dt \right].$$

The representation in the last display is useful for actual computations.

1.3 Outline of the Paper

The plan for the paper is as follows. Section 2 defines the model for the estimator and states the goal of the paper. Section 3 states the consistency result and its optimality. A greedy algorithm is discussed as a method to carry out the estimation in practice. Section 3.6 shows applications of the main result to a variety of estimation problems and derives the convergence rates. Additional details are also given and an out of sample test based on the likelihood ratio is suggested for forecast evaluation. Section 4 applies the estimation procedure to the intensity of buy and sell trades. Section 5 provides finite sample evidence to better understand the role of the different parameters in the estimation. Section 6 contains further remarks. Proofs of the results are in Section A.1 of the supplementary material.

2 The Model

The goal is to allow for a simple interpretation of the impact of the covariates on the intensity. A good level of interpretability is gained by letting $g(x)$ be linear in x. However, the impact of each of the covariates might be nonlinear. Nonlinearities are documented in many applications, including high frequency financial data (e.g., Hasbrouck, 1991, Lillo et al., 2003). Whether these nonlinearities affect the intensity depends on the application. An additive nonlinear model is considered a reasonable trade off between interpretability and the possibility of nonlinear relations. In this case, $g(x) = \sum_{k=1}^{K} g^{(k)}(x)$, where the $g^{(k)}$'s are bounded functions, possibly zero, and for each k, $g^{(k)}(x)$ only depends on x_k, the k^{th} coordinate of $x = (x_1, x_2, ..., x_K)$ (i.e., with abuse of notation, $g^{(k)}(x) = g^{(k)}(x_k)$). This is done to reduce the notational burden.

2.1 Representation for Additive Functions

For the purpose of controlling the estimation error, it is necessary to impose structure on the set that the additive functions are supposed to lie within. Functions with the following structure are considered

$$g^{(k)}(x) = \sum_{\theta \in \Theta_k} b_\theta \theta(x) \tag{4}$$

where Θ_k is a set of functions that depends only on x_k, Θ_k is a possibly uncountable set, and the b_θ's are real valued coefficients. Given that Θ_k can be uncountable, the above representation is more general than a standard series expansion. The sum is understood to mean

$$\sum_{\theta \in \Theta_k} b_\theta := \sup \left\{ \sum_{\theta \in F} b_\theta : \mathcal{H} \subseteq \Theta_k, \, \mathcal{H} \text{ is finite} \right\}.$$

For example, we could have $g_k = b_\theta \theta$ for $\theta \in \Theta_k$, where Θ_k is a model, possibly infinite dimensional. In consequence of the additive structure of g,

$$g(x) = \sum_{k=1}^{K} \left(\sum_{\theta \in \Theta_k} b_\theta \theta(x) \right), \tag{5}$$

where the terms in the parenthesis are $g^{(k)}$ in (4), which is a function that depends on the k^{th} covariate only. This structure is suitable for estimation. Estimation within this framework requires choice of the b_θ's as well as the θ's. For practical purposes the latter might be simple parametric functions or fixed functions rather than general infinite dimensional models. Details and examples are postponed to Section 3.6. The interested reader can skim through that section for an overview. In order to impose general restrictions, suppose that the user fixes a set of weights $\mathcal{W} := \{w_\theta \in (0, \infty) : \theta \in \Theta\}$, where $\Theta := \bigcup_{k=1}^{K} \Theta_k$. This means that the weights w_θ can be different for each function θ of the k^{th} explanatory variable. Then, define $\mathcal{L}(B) = \mathcal{L}(B, \Theta, \mathcal{W}) := \left\{ \sum_{\theta \in \Theta} b_\theta \theta : \sum_{\theta \in \Theta} w_\theta |b_\theta| \leq B, w_\theta \in \mathcal{W} \right\}$. This is a subset of the functions in (5) such that the weighted absolute sum of the coefficients is bounded by a finite constant $B > 0$. The weights are often used to control the importance of each θ. For example, one can let $w_\theta^2 = Var(\theta(X(t)))$ so that intuitively, all functions have the same importance. A bound on the weighted absolute sum of the regression coefficients is common in Lasso estimation (e.g., Bühlmann and van de Geer, 2011).

Example 1 *Let $g(x) = \sum_{k=1}^{K} b_k X_k$ and π_k be the map such that $\pi_k x = x_k$ for any $x \in \mathbb{R}^K$ and x_k is the k^{th} element in x. Then, $\Theta_k := \{\pi_k\}$ contains a single function that maps $x \in \mathbb{R}^K$ into its k^{th} co-ordinate x_k. Also, let $w_\theta^2 = Var(X_k(0))$ when $\theta \in \Theta_k$ and X_k is the k^{th} co-ordinate of X. The constraint is $\sum_{k=1}^{K} |b_k| \sqrt{Var(X_k(0))} \leq B$.*

In other circumstances, the weight can serve the purpose of shrinkage within each function $g^{(k)}$, which is important in infinite dimensional spaces.

Example 2 *To avoid distracting notation, suppose that* $g(x) = \sum_{j=1}^{\infty} b_j x_k^j$, *a polynomial which depends on* $x_k \in [0,1]$ *only. Also suppose that* $\sum_{j=1}^{\infty} (j!)\,|b_j| \leq B$, *so that the weights force the coefficients to decay faster than* $j!$. *An infinite differentiable function with derivatives of all orders bounded by one can be written as the polynomial above where* $|b_j| \leq (j!)^{-1}$. *Hence, the weights allow us to account for this and the constraint induces an additional shrinkage effect on the coefficients because of the summability constraint.*

From now on, dependence on Θ and $\mathcal{W}$ will be implicit when writing $\mathcal{L}(B)$. The approximation error of functions in $\mathcal{L}(B)$ for $B < \infty$ can be related to the bound on the absolute sum of the coefficients. This is useful if one supposes that $g_0 \in \mathcal{L}(B_0)$ for some unknown but finite B_0. If the user estimates the model with $B < B_0$, an approximation error will occur. However, note that the results of the paper will allow for more general forms of misspecification. Let $P(x)$ be the marginal distribution of $X(t)$, which by stationarity does not depend on t. For any function $g : \mathbb{R}^K \to \mathbb{R}$, let $Pg = \int g(x)\, dP(x)$. The $L_r(P)$ norm is $|\cdot|_r = \left(\int |\cdot|^r\, dP\right)^{1/r}$ for $r \in [1,\infty)$, with the standard modification when $r = \infty$. The following is a re-adaptation of a result in Sancetta (2015) and can be used to control the approximation error of the estimator.

Lemma 1 *Let* $g_0 \in \mathcal{L}(B_0)$ *for* $B_0 < \infty$ *and* $\bar{\theta}_r := \sup_{\theta\in\Theta} |\theta|_r < \infty$. *Then, for any* $B < \infty$, *and* $r \in [1,\infty]$, $\min_{g\in\mathcal{L}(B)} |g_0 - g|_r \leq \underline{w}^{-1}\bar{\theta}_r \max\{B_0 - B, 0\}$.

When $g_0 \notin \mathcal{L}(B)$, define the best uniform approximation $g_B = \arg\inf |g - g_0|_\infty$ where the infimum is over $\mathcal{L}(B)$. We shall define

$$B_0 = \arg \inf_{B<\infty} |g_B - g_0|_\infty . \tag{6}$$

This means that g_{B_0} is the best uniform approximation of g_0 for any $g \in \bigcup_{B>0} \mathcal{L}(B)$.

2.2 The Goal

The user supposes that $g_0 \in \mathcal{L}(B_0)$, but ignores the value of B_0. They guess a value $\bar{B} < \infty$. If it is the case that $g_0 \in \mathcal{L}(B_0)$, and $\bar{B} \geq B_0$, there will be no approximation error. The estimation error could be high, especially if $\bar{B}$ is much larger than B_0. Once $\bar{B}$ is chosen, the log-likelihood in (3) is maximized over $\mathcal{L}\left(\bar{B}\right)$.

Let $\lambda = d\Lambda/d\mu$, where Λ is the intensity measure (1) and μ is the Lebesgue measure. Then, $\lambda(X(t)) = \exp\{g_0(X(t))\}$ with the right hand side as in (1). Suppose that g is fixed and bounded. Define the random norm

$$|g - g_0|_{\lambda,T} := \sqrt{\frac{1}{T}\int_0^T \left(g(X(t)) - g_0(X(t))\right)^2 d\Lambda(t)}.$$

By stationarity and ergodicity (e.g., Lemma 2 in Ogata, 1978),

$$|g - g_0|^2_{\lambda,T} \to P(g - g_0)^2 \lambda = \int (g(x) - g_0(x))^2 \lambda(x) \, dP(x), \tag{7}$$

almost surely. The goal is to define an estimator g_T in $\mathcal{L}(\bar{B})$ and obtain rates of convergence to zero of $|g_T - g_0|_{\lambda,T}$. By (7), this convergence also implies convergence of $P(g_T - g_0)^2 \lambda$, though the rate of convergence for the latter cannot be derived unless we impose dependence conditions on the covariates. If $|g_0|$ is bounded - as will be assumed here - the right hand side (r.h.s.) of (7) is proportional to $P(g - g_0)^2 = |g - g_0|^2_2$, hence the results to be derived also hold in P-integrated square error. The proofs show that the convergence results hold for the Hellinger distance between $\exp\{g_T\}$ and $\exp\{g_0\}$. To minimise the notational burden, this is not explicitly stated in the text. Details can be found in Section A.1 of the supplementary material. Note that elements $g, g' \in \mathcal{L}(\bar{B})$ will be considered the same if $P(g - g')^2 \lambda = 0$.

2.2.1 Connection to Lasso

Given the constraint on the coefficients b_θ's, minimization over $\mathcal{L}(\bar{B})$ is just the primal of an l_1 penalized likelihood estimator, i.e., Lasso. Conditioning on the sample, for each $\bar{B}$, there is a constant $\pi_{\bar{B}}$ (the Lagrange multiplier, which increases with T but at a possibly different rate than L_T), such that the left side of the following two displays are the same:

$$\arg\sup_{\theta, b_\theta} L_T \left(\sum_{\theta \in \Theta} b_\theta \theta \right),$$

where the supremum is taken over those θ's and b_θ's such that $\sum_{\theta \in \Theta} b_\theta \theta \in \mathcal{L}(\bar{B})$;

$$\arg\sup_{\theta, b_\theta} L_T \left(\sum_{\theta \in \Theta} b_\theta \theta \right) - \pi_{\bar{B}} \sum_{\theta \in \Theta} w_\theta |b_\theta|, \tag{8}$$

where the supremum is taken over those θ's and b_θ's such that $\theta \in \Theta$ and b_θ is a real number. If $\mathcal{L}(\bar{B})$ is a finite dimensional space, $\pi_{\bar{B}}/T \to 0$ (in probability) when the estimator is consistent for g_0 inside $\mathcal{L}(\bar{B})$. However, when $\mathcal{L}(\bar{B})$ is infinite dimensional, norms are not equivalent and consistency under the norm we consider in this paper does not mean consistency under the norm implied by the constraint. Hence, for infinite dimensional $\mathcal{L}(\bar{B})$, $\pi_{\bar{B}}/T$ may converge to a constant even when the estimator is consistent and g_0 lies inside $\mathcal{L}(\bar{B})$.

Estimation of the primal or dual problem gives the same solution when we are able to map the constraint into the Lagrange multiplier $\pi_{\bar{B}}$. In general, this is not straightforward. A solution for the Lasso problem is often via co-ordinate descent, though rates of convergence are usually not derived (e.g., Bühlmann and van de Geer, and references therein). Here, we

solve the constrained optimization and suggest an algorithm to do so in practice and derive the convergence rates of the algorithm (Section 3.5).

3 Consistency of the Estimator

3.1 Conditions

The following conditions are imposed. Remarks on these are in Section 3.4. To aid intuition, the conditions can be divided into three groups: stochastic restrictions, parameter space restrictions, and estimator restrictions. The conditions use the notation defined around (1) and in Section 2.1.

Condition 1 *Stochastic Restrictions.*

1. $(X(t))_{t\geq 0}$ *is a stationary, ergodic, predictable* K *dimensional process with values in a set* $\mathcal{X} \subseteq \mathbb{R}^K$ *(*$K > 1$*);*

2. *The cumulative intensity* Λ *has a density* λ *with respect to the Lebesgue measure (as in (1));*

3. $T_0 = 0$ *is the time of the last jump before the jump at time* T_1*.*

Condition 2 *Parameter Space Restrictions.*

1. *The functions in* $\Theta = \bigcup_{k=1}^{K} \Theta_k$ *are measurable, and uniformly bounded by a finite constant* $\bar{\theta} := \sup_{\theta\in\Theta} \sup_{x\in\mathcal{X}} |\theta(x)|$*. The set* Θ_k *has an* $L_\infty(P)$ ϵ*-bracketing number* $N(\epsilon, \Theta_k)$ *such that the entropy integral* $\int_0^1 \sqrt{\ln(1 + N(\epsilon, \Theta_k))}d\epsilon$ *is finite for every* k *(not bounded and can grow with the sample size); the weights in* $\mathcal{L}(B, \Theta, \mathcal{W})$ *satisfy* $\underline{w} := \inf_{\theta\in\Theta} w_\theta > 0$*;*

2. *In (1),* $\bar{g}_0 := |g_0|_\infty < \infty$ *and if* $g_0 \neq g_{B_0}$*, then* $B_0 < \infty$ *(see (6)).*

Condition 3 *Estimator Restrictions. The estimator* g_T *satisfies:*

1. $\Pr\left(g_T \notin \mathcal{L}\left(\bar{B}, \Theta, \mathcal{W}\right)\right) = o(1)$*;*

2. $L_T(g_T) \geq \sup_{g\in\mathcal{L}\left(\bar{B},\Theta,\mathcal{W}\right)} L_T(g) - O_p\left(\frac{T}{r_T^2}\right)$*, where* r_T *is as in (9) in Section 3.2.*

In general, from (9) one can deduce that $r_T^2 \lesssim T^{1/2}$, where throughout, $\lesssim$ is inequality up to a multiplicative universal finite absolute constant.

3.2 Consistency Results

It will be shown that the overall complexity in the statistical estimation depends on three factors: the logarithm of the number of variables K, $\bar{B}$ (in $\mathcal{L}\left(\bar{B}\right)$), and the entropy integral of the largest of the sets Θ_k . To ease notation, dependence on $\bar{\theta}$ and $\underline{w}$ is suppressed in what follows. More explicit bounds can be found in the proof of the results.

Theorem 1 *Suppose that there is a nondecreasing sequence r_T such that*

$$r_T^2 \lesssim \min\left\{\frac{\bar{B}^{-1}T^{1/2}}{\sqrt{\ln K}+\max_{k\leq K}\int_0^1\sqrt{\ln\left(1+N\left(\epsilon,\Theta_k\right)\right)}d\epsilon}, \frac{1}{\inf_{g\in\mathcal{L}\left(\bar{B}\right)}\left|g-g_0\right|_\infty^2}\right\}. \tag{9}$$

Under Conditions 1, 2, and 3, $\left|g_T-g_0\right|_{\lambda,T}=O_p\left(r_T^{-1}\right)$.

Note that the condition that r_T is nondecreasing implicitly imposes restrictions on $\bar{B}$, K and $N\left(\epsilon,\Theta_k\right)$. The daunting expression (9) does simplify, but it is stated in this form for flexibility. Section 3.6 considers applications of this result to a variety of problems so that the bound becomes considerably simple. To provide a sense for the sharpness of the bound, it might be convenient to suppose that the approximation error $\inf_{g\in\mathcal{L}\left(\bar{B}\right)}\left|g-g_0\right|_\infty$ is zero. Also suppose that the entropy integral is bounded by a finite constant. In this case, the rate of convergence of $\left|g_T-g_0\right|_{\lambda,T}$ is $O\left(\left(\ln\left(K\right)/T\right)^{1/4}\right)$. By stationarity and ergodicity, it is easy to see that for $T=T_n$ (T_n is the time of the n^{th} jump), $T_n\asymp n$ where $\asymp$ means equality up to a multiplicative finite absolute constant. In consequence, the bound becomes the more familiar $O\left(\left(\ln\left(K\right)/n\right)^{1/4}\right)$ for $K>1$. Results in Tsybakov (2003) show that in a regression context with Gaussian errors, no linear estimator of the convex combination of K bounded terms can achieve a rate faster than $O_p\left(\left(\ln\left(K\right)/n\right)^{1/4}\right)$ when K is of a larger order of magnitude than $n^{1/2}$ (see Theorem 2 in Tsybakov, 2003). Hence, without further assumptions, one can suppose that the convergence rate derived here is optimal in this context. Theorem 2 in Section 3.3 lends some rigor to this supposition.

In order to show the effect of the approximation error when $g_0\in\mathcal{L}\left(B_0\right)$ for an unknown but finite B_0, consider the following scenario. Let $\bar{B}\to\infty$ so that eventually $\bar{B}\geq B_0$. By Lemma 1 we deduce that the approximation error is eventually exactly zero for a finite $\bar{B}$. In consequence, the following holds true.

Corollary 1 *Suppose that $g_0\in\mathcal{L}\left(B_0\right)$. Under the conditions of Theorem 1, for any $\bar{B}\to\infty$,*

$$\left|g_T-g_0\right|_{\lambda,T}^2=O_p\left(\frac{\bar{B}\left[\sqrt{\ln K}+\max_{k\leq K}\int_0^1\sqrt{\ln\left(1+N\left(\epsilon,\Theta_k\right)\right)}d\epsilon\right]}{T^{1/2}}\right)$$

When $g_0\notin\mathcal{L}\left(B\right)$ for any B, the approximation error in Theorem 1 can be bounded using the following, which follows from the triangle inequality and Lemma 1.

Lemma 2 *Under Condition 2,*

$$\inf_{g\in\mathcal{L}\left(\bar{B}\right)}\left|g_0-g\right|_\infty \lesssim \inf_{B>0}\left\{\max\left\{B-\bar{B},0\right\}+\inf_{g\in\mathcal{L}(B)}\left|g_0-g\right|_\infty\right\}.$$

The reader interested in the scope of the possible applications can go directly to Section 3.6. The next sections provide remarks on optimality, conditions, and details for the solution of the estimation problem.

3.3 Optimality

From the previous remarks, it is reasonable to infer that the rates of convergence in Theorem 1 are optimal for K large. To avoid technicalities, consider the following simplified scenario. One may argue that less stringent conditions should make the estimation problem harder and as such, if the lower bound holds under restrictive conditions it should hold under more general conditions. Recall that $X_k(t)$ is the k^{th} element in the vector of covariates $X(t)$.

Theorem 2 *Suppose $\mathcal{L}(1) := \mathcal{L}(1,\Theta,\mathcal{W})$, where the sets Θ_k's contain bounded functions, and the weights in $\mathcal{W}$ have been set to one. Suppose that $X(t+T_{i-1}) = X(T_{i-1})$ for $t\in(0,T_i]$, i.e., $X(t)$ is constant between jumps of the point process N, and $T_0 = 0$. Also suppose that $(X(T_i))_{i\geq 0}$ forms a sequence of i.i.d. random variables and that the $X_k(T_i)$'s are independent across k, with continuous distribution function. For $K > T_n^{1/2}$ with $K = O(T_n^p)$ for any $p<\infty$, and $n\to\infty$,*

$$\inf_{g_T}\sup_{g_0\in\mathcal{L}(1)}\int_0^{T_n}\left|g_{T_n}(X(t))-g_0(X(t))\right|^2\exp\left\{g_0(X(t))\right\}dt \gtrsim \sqrt{T_n\ln\left(1+KT_n^{-1/2}\right)}$$

in probability, where the infimum is taken over all possible estimators g_{T_n} of the intensity.

Theorem 2 says that even under rather restrictive conditions, as long as the number of variables K is of order of magnitude greater than $T_n^{1/2}$ the convergence rate under $|\cdot|_\lambda$ cannot be faster than $((\ln K)/T_n)^{1/4}$.

3.4 Remarks on Conditions

It is worth emphasizing that the conditions do not restrict $g_0\in\mathcal{L}(B)$ for $B<\infty$.

Condition 1 is mild. For all practical cases, one usually restricts X to be an adapted process that is left continuous. This implies predictability (e.g., Brémaud, 1981). In consequence, the time from last jump $R(t) := \inf\{t-T_i : t-T_i>0, i=0,1,2...\}$ can be used as a covariate, as it is a predictable process. This will be the case when estimating certain nonlinear Hawkes processes in Section 3.6. $T_0=0$ is used to keep notation simple. Similarly, the condition $K>1$ is used to avoid writing $\ln(1+K)$ instead of $\ln K$ in various places.

In Condition 2, the entropy integral restriction on the class of functions is standard. It is needed as the framework is quite general, hence it requires some control of the complexity of the functions in Θ_k. The entropy integral is finite, but can grow with the sample size even though this is not made explicit in the notation (see Section 3.6.6 and the proof of Lemma 5 in the supplementary material). The $L_\infty(P)$ ϵ-bracketing number of a set Θ_k is the number of pairs of elements in a set $\mathcal{V}$ such that for each $\theta \in \Theta_k$, there is a bracket $[\theta_L, \theta_U]$ satisfying $\theta_L \leq \theta \leq \theta_U$, and $|\theta_L - \theta_U|_\infty \leq \epsilon$. The uniform norm can be replaced by the random norm $T^{-1}\int_0^T |\theta_L - \theta_U|^2 \, d\Lambda$, which is actually the norm used in the proofs. This is difficult to control and in the applications considered in this paper, the (stronger) uniform norm is used instead. To cover the case of sieve estimation and/or misspecification, g_0 is not restricted to lie in $\mathcal{L}$, but needs to be uniformly bounded.

Condition 3 only requires that asymptotically, the estimators satisfy the complexity restrictions discussed in this paper. This is weaker than assuming that the absolute sum of the coefficients is bounded by $\bar{B}$ for any sample size and that the estimators of the functions θ_k's are always in Θ_k. This setup allows us to cover different approaches for estimation without restricting attention to a specific one. Moreover, the estimator g_T only needs to maximize the sample likelihood L_T asymptotically, rather than exactly. Section 3.5 provides details on a computationally feasible estimation method.

In some circumstances we do not observe the true covariates and can only estimate the intensity using approximate data, which may not be stationary. A typical example is in the context of Hawkes processes (see Section 3.6) or when a covariate is a moving average of the past values. In the aforementioned cases, the true covariates are a causal filter of some quantity, but we can only construct the filter using an initial condition rather than observations prior to time $T_0 = 0$. Note that the true covariates still satisfy Condition 1. However, we perform optimization on surrogate data so that the last point in Condition 3 does not directly hold. The following allows us to consider such cases.

Corollary 2 *Suppose Conditions 1 and 2 hold and let r_T be as in Theorem 1. Define $\bar{B}_w := \bar{B}/\underline{w}$. Let $\tilde{X}(t)$ be arbitrary covariates, but such that*

$$\mathbb{E}\sup_{\theta\in\Theta}\int_0^T \left|\theta\left(\tilde{X}(t)\right) - \theta\left(X(t)\right)\right| dt = O\left(e^{-\bar{B}_w\bar{\theta}}\sqrt{T\ln K}\right). \tag{10}$$

Suppose that $\tilde{g}_T$ satisfies $\Pr\left(\tilde{g}_T \notin \mathcal{L}\left(\bar{B},\Theta,\mathcal{W}\right)\right) = o(1)$, and

$$\tilde{L}_T(\tilde{g}_T) \geq \sup_{g\in\mathcal{L}\left(\bar{B},\Theta,\mathcal{W}\right)} \tilde{L}_T(g) - O_p\left(\frac{T}{r_T^2}\right) \tag{11}$$

where $\tilde{L}_T$ is the log-likelihood L_T when we use covariates $\tilde{X}(t)$ instead of $X(t)$, as data. Then, $\tilde{g}_T$ is also an approximate minimiser of L_T, i.e., it satisfies Condition 3 (with error

$O_p\left(T/r_T^2\right)$). *Hence,*

$$|\tilde{g}_T - g_0|^2_{\lambda,T} = \frac{1}{T}\int_0^T |\tilde{g}_T\left(X\left(t\right)\right) - g_0\left(X\left(t\right)\right)|^2 \exp\left\{g_0\left(X\left(t\right)\right)\right\} = O_p\left(r_T^{-2}\right).$$

Moreover,

$$\frac{1}{T}\int_0^T \left|\tilde{g}_T\left(\tilde{X}\left(t\right)\right) - g_0\left(X\left(t\right)\right)\right|^2 \exp\left\{g_0\left(X\left(t\right)\right)\right\} dt = O_p\left(r_T^{-2}\right).$$

Corollary 2 says that we obtain the same rates of convergence even when the estimator is computed from the log-likelihood $\tilde{L}_T$ based on surrogate covariates, as long as the surrogate covariates satisfy (10). The last display in Corollary 2 says that $\tilde{g}_T\left(\tilde{X}\left(t\right)\right)$ is close to $g_0\left(X\left(t\right)\right)$ even though they are evaluated at different data.

3.5 Estimation Algorithm

Maximization of the log-likelihood over $\mathcal{L}\left(\bar{B}\right)$ leads to a unique maximum (within an equivalence class) because of concavity of the objective function and the convex and closed constraint. However, while suitable for theoretical derivations it is too abstract for practical implementation. The algorithm in Figure 1 can be used to solve the constrained minimization. For real valued functions g and h on $\mathbb{R}^K$, the following derivative of the log-likelihood in the direction of a function h is used

$$D_T\left(g,h\right) := \int_0^T h\left(X\left(t\right)\right) dN\left(t\right) - \int_0^T h\left(X\left(t\right)\right) \exp\left\{g\left(X\left(t\right)\right)\right\} dt.$$

There is a line search to find the coefficient ρ_j. To speed up the computations, this can be set to the deterministic value $\rho_j = 2/\left(j+1\right)$. The updated approximation to the constrained maximum at step j is denoted by F_j. The bound to be given in Theorem 3 holds in this case as well.

Figure 1. Log-Likelihood Optimization

Set:
$m \in \mathbb{N}$
$F_0 := 0$
$\bar{B} < \infty$
For: $j = 1, 2, ..., m$
$\theta_j := \arg\sup_{\theta\in\Theta} \left|D_T\left(F_{j-1},\theta\right)\right| / w_\theta$
$b_j := \frac{\bar{B}}{w_\theta} sign\left(D_T\left(F_{j-1},\theta_j\right)\right)$
$\rho_j := \arg\max_{\rho\in[0,1]} L_T\left(\left(1-\rho\right)F_{j-1} + \rho b_j\theta_j\right)$ or $\rho_j := 2/\left(j+1\right)$
$F_j\left(X\right) := \left(1-\rho_j\right)F_{j-1}\left(X\right) + \rho_j b_j \theta_j\left(X\right)$

Theorem 3 *Let F_m be the resulting estimator from Figure 1. Define $\bar{B}_w := \bar{B}/\underline{w}$. Then,*

$$L_T(F_m) \geq \sup_{g \in \mathcal{L}(\bar{B})} L_T(g) - \frac{8Te^{\bar{B}_w\bar{\theta}}\left(\bar{B}_w\bar{\theta}\right)^2}{m+2}$$

where the notation is from Condition 2.

The algorithm in Figure 1 belongs to the family of Frank-Wolfe algorithms (e.g., Jaggi, 2013, for the general proof of the convergence towards the optimum point, and Sancetta, 2016, for its statistical properties for linear models). The following identifies a suitable number of iterations for the purpose of consistent estimation.

Corollary 3 *If $m^{-1} = o\left(T^{-1/2}e^{-\bar{B}_w\bar{\theta}}\left(\bar{B}_w\bar{\theta}\right)^{-2}\right)$, then F_m in Figure 1 satisfies Condition 3. Hence, if $\bar{B}$ is bounded, $m^{-1} = o\left(T^{-1/2}\right)$.*

3.6 Application to Various Estimation Methods and Model Specifications

The class of functions is general and can accommodate various estimation methods and model specifications. Below, different models, function classes, and estimators are discussed. There is overlap for some of the applications, but the variations in terms of approximation error make them different enough to justify their individual treatment.

To avoid some oddities in the discussion, define the map $(x_1, x_2, ..., x_K) = x \mapsto \pi_k(x) = x_k$ so that by composition, for any f on $\mathbb{R}$, $f \circ \pi_k(x) = f(x_k)$. In all the examples, it is tacitly assumed that the support of each covariate is $[0,1]$. This is done for simplicity to avoid distracting technicalities even when not necessary. In various occasions, we may have a nontrivial approximation error. In this case, the following will be used to indicate a set that contains the true g_0,

$$\mathcal{G}(B) := \left\{ g = \sum_{k=1}^{K} b_k f_k \circ \pi_k : f_k \in \mathcal{H}, \sum_{k=1}^{K} |b_k| \leq B \right\}, \tag{12}$$

where $\mathcal{H}$ is a class of univariate functions which will be defined within each section below, depending on the application. In all the examples of this section, all the weights w_θ's in $\mathcal{W}$ are supposed to be equal to one without further mention. Then, when $\Theta_k = \{f \circ \pi_k : f \in \mathcal{H}\}$, $\mathcal{L}(B) = \mathcal{G}(B)$. Suppose that $f_{V,k}$ is an approximation to a function $f_k \in \mathcal{H}$, then

$$\left| \sum_{k=1}^{K} b_k f_k - \sum_{k=1}^{K} b_k f_{V,k} \right|_\infty \leq B \max_{k \leq K} |f_k - f_{V,k}|_\infty \tag{13}$$

when $\sum_{k=1}^{K} |b_k| \leq B$. This will be used in some of the examples in order to estimate the approximation error. In this case, (13) will be used in conjunction with Lemma 2 where B is

just a bounded constant (e.g., $B = B_0$). Finally, to avoid trivialities $K > 1$ in all the bounds below. The bounds are of particular interest when $K \gtrsim T^{1/2}$. Note that in the examples, we can have bounds such as $|g_T - g_0|^2_{\lambda,T} \lesssim \bar{B}\sqrt{(\ln K)/T}$. It is tacitly assumed that we require the r.h.s. to be $O(1)$. Proofs of the following corollaries to Theorem 1 can be found in Section A.1.5 of the supplementary material.

3.6.1 Linear Model with Many Variables

Let $\Theta_k := \{\pi_k\}$ which maps $x \in \mathbb{R}^K$ into its k^{th} co-ordinate x_k. Then, $g(x) = \sum_{k=1}^K b_k x_k$. The following holds true.

Corollary 4 *Suppose that $g_0 \in \mathcal{L}(\bar{B})$. Under Conditions 1 and 3, $|g_T - g_0|^2_{\lambda,T} \lesssim \left(\frac{\ln K}{T}\right)^{1/2}$ in probability.*

The corollary implies that the estimator is consistent even in the ultra high dimensional case $K = O\left(e^{T^c}\right)$ for $c \in [0, 1)$.

3.6.2 Hawkes Process with Many Covariates

There are many versions of the Hawkes process. For the sake of illustration, consider a nonlinear function of the standard exponential decay case (e.g., Brémaud and Massoulié, 1996). Define the family of processes $\left\{\left(\tilde{f}_a(t)\right)_{t\geq 0} : a \in [\underline{a}, \bar{a}] \subset (0,\infty)\right\}$, where for each a, $\tilde{f}_a(t) := f\left(\int_{[0,t)} e^{-a(t-s)} dN(s)\right)$ and f is a bounded Lipschitz function. The process $\tilde{f}_a(t)$ is not stationary because it is initiated at $t = 0$. In consequence, it fails Condition 1 and cannot be used as one of the covariates. Define the family $\left\{(f_a(t))_{t\geq 0} : a \in [\underline{a}, \bar{a}] \subset (0,\infty)\right\}$, where $f_a(t) = f\left(\int_{(-\infty,t)} e^{-a(t-s)} dN(s)\right)$ and f is as before. The processes f_a's are stationary, but not observable. Despite the notational difference, one can verify the conditions of Corollary 2 to see that Theorem 1 still holds. We also need to verify that using $f_a(t)$ the counting process is stationary.

Corollary 5 *Under Condition 1, the point process with intensity density $\lambda(t) = \exp\{f_{a_0}(t) + g_0(X(t))\}$ (for any $a_0 \in (\underline{a}, \bar{a})$) has a stationary distribution. Moreover, suppose that the log-likelihood with intensity $\exp\left\{\tilde{f}_a(t) + g(X(t))\right\}$ is maximized w.r.t. $g \in \mathcal{L}(\bar{B})$ and $a \in [\underline{a}, \bar{a}]$ by g_T and a_T (even approximately with the same error as in Condition 3). Suppose that $\bar{B}$ is fixed, and $g_0 \in \mathcal{L}(\bar{B})$, then, in probability,*

$$\left|\left(g_T + \tilde{f}_T\right) - (g_0 + f_0)\right|^2_{\lambda,T} \lesssim \frac{\sqrt{\ln K} + \sqrt{\ln T} + \max_{k\leq K} \int_0^1 \sqrt{\ln(1 + N(\epsilon, \Theta_k))} d\epsilon}{\sqrt{T}}. \quad (14)$$

Also suppose that $\Theta_k := \{\pi_k\}$, then $\left|\left(g_T + \tilde{f}_T\right) - (g_0 + f_0)\right|^2_{\lambda,T} \lesssim \left(\frac{\ln KT}{T}\right)^{1/2}$ in probability.

Note that to ease notation, we use $\ln KT = \ln(KT)$, throughout the paper.

3.6.3 Threshold Model with Many Variables

Suppose that $\varphi : \mathbb{R} \to [0,1]$ is Holder's continuous with parameter $\alpha \in (0,1]$, i.e., $|\varphi(x) - \varphi(y)| \lesssim |x-y|^{\alpha}$. Consider the class of linear threshold functions $f(x,z) := a_1 x + a_2 x \varphi(c_1 z - c_2)$, $x, z \in \mathbb{R}$, where a_1, a_2, c_1, c_2 are unknown real coefficients, with $a_1, a_2, c_1, c_2 \in [-1,1]$. Denote the set of such functions by $\mathcal{H}$.

Let $(Z(t))_{t\geq 0}$ be a predictable stationary and ergodic real valued process taking values in $[0,1]$ as for the X_k's. Refer to it as a threshold variable. Then, $f(X_k(t), Z(t))$ is a transition process, for the k^{th} covariate: the impact of X_k depends on the threshold variable Z. Hence, $f(x,z)$ is a smooth transition function (see van Dijk et al., 2002, for a survey of smooth regression models based on this functional specification).

The class of functions with elements $\varphi(c_1 z - c_2)$ with bounded z has finite entropy integral (e.g., deduce this from Theorem 2.7.11 in van der Vaart and Wellner, 2000). Given that $a_1, a_2 \in [-1,1]$, it follows that $\mathcal{H}$ has finite entropy integral. Let $\Theta_k := \{f \circ (\pi_k, \iota) : f \in \mathcal{H}\}$, where ι is the identity map $\iota(z) = z$ (i.e., $f \circ (\pi_k x, \iota z) = f(x_k, z)$).

Corollary 6 *Let Z be as described before. Suppose that $g_0 \in \mathcal{L}(B_0)$. Under Conditions 1 and 3, for the estimator $g_T \in \mathcal{L}(\bar{B})$, for any $\bar{B} \to \infty$, $|g_T - g_0|^2_{\lambda,T} \lesssim \bar{B}\left(\frac{\ln K}{T}\right)^{1/2}$, eventually, in probability.*

3.6.4 Expansion in Terms of a Fixed Dictionary under l_1 Constraint

Consider the case of univariate functions with representation $f = \sum_{v=1}^{\infty} a_v e_v$ where $\{e_v : v = 1,2,...\}$ is a dictionary and $\sum_{v=1}^{\infty} |a_v| < \infty$. Subspaces of such functions are considered in Barron et al. (2008). A typical example is when f is a polynomial. Then, let $\sum_{v=1}^{V} a_v e_v(x_k)$ be the (truncated) representation for the functions of the k^{th} covariate for some finite V. Then, suppose that g_0 can be written as

$$g(x) = \sum_{k=1}^{K} b_k \sum_{v=1}^{V} a_{v_k,k} e_{v_k}(x_k) \tag{15}$$

so that $\Theta_k = \{e_v \circ \pi_k : v = 1,2,...,V\}$ and $\sum_{k=1}^{K}\sum_{v=1}^{V} |b_k a_{v_k,k}| \leq B_0$. In this case, one can directly estimate the coefficients $b_k a_{v_k,k}$ and reduce the optimization over Θ to the selection of an element $e_v \circ \pi_k$ in Θ. There are V fixed elements in each Θ_k. Hence, the entropy integral for each Θ_k is a constant multiple of $\sqrt{\ln V}$. If no approximation error is incurred (i.e. g_0 can be written as (15)), then $|g_T - g_0|^2_{\lambda,T} \lesssim \left(\frac{\ln KV}{T}\right)^{1/2}$, as in the linear case (Section 3.6.1), but with KV variables instead of K.

This framework adapts to sieve estimation of smooth functions, in which case an approximation error is incurred. For definiteness suppose that $\{e_v : v = 1, 2, ...\}$ are trigonometric polynomials with period one, rather than a general dictionary. Let $\mathcal{H}$ be the class of Holder continuous functions on $[0,1]$ with exponent $\alpha > 1/2$, constant one and uniformly bounded by one, i.e., $|f(x) - f(y)| \leq |x-y|^{\alpha}$ and $|f|_{\infty} \leq 1$, if $f \in \mathcal{H}$. By Bernstein Theorem (e.g., Katznelson, 2002, p. 33), if $f \in \mathcal{H}$, there is a finite absolute constant c_{α} depending only on $\alpha > 1/2$ such that $f = \sum_{v=1}^{\infty} a_v e_v$ and $\sum_{v=1}^{\infty} |a_v| \leq c_{\alpha}$, where the equality holds in the sup norm. Hence, in what follows we can take $\mathcal{H}$ to be equivalent to the class of functions with such series expansion. Let $\mathcal{H}_V$ be the set of trigonometric polynomials up to order V. By Jackson Theorem (e.g., Katznelson, 2002, p.49), for any $f \in \mathcal{H}$, there is a trigonometric polynomial of order V, say $f_V \in \mathcal{H}_V$, such that $|f_V - f|_{\infty} \lesssim V^{-\alpha}$. Suppose that $g_0 \in \mathcal{G}(1)$ (in (12)), then using subscript 0 to denote the coefficients of g_0,

$$g_0 = \sum_{k=1}^{K} b_{0k} \left(\sum_{v=1}^{\infty} a_{0vk} e_v \right) = \sum_{k=1}^{K} (\bar{a}_{0k} b_{0k}) \left(\sum_{v=1}^{\infty} \left(\frac{a_{0vk}}{\bar{a}_{0k}} \right) e_v \right)$$

setting $\bar{a}_{0k} := \sum_{v=1}^{\infty} |a_{0vk}|$. By the aforementioned remarks concerning Bernstein Theorem, there is a finite constant c_{α} such that $\bar{a}_{0k} \leq c_{\alpha}$. Hence, $\sum_{k=1}^{K} (\bar{a}_{0k} b_{0k}) \leq c_{\alpha}$, using the constraint on the b_{0k}'s implied by restricting attention to $g_0 \in \mathcal{G}(1)$. Let $\Theta_k := \left\{ \sum_{v=1}^{V} a_v e_v \circ \pi_k : \sum_{v=1}^{V} |a_v| \leq 1 \right\}$. Using (13) we can derive the approximation error for this problem and deduce the following consistency rates.

Corollary 7 *Let $g_0 \in \mathcal{G}(1)$ (as in (12)) with $\mathcal{H}$ Holder continuous with exponent $\alpha > 1/2$. Under Conditions 1 and 3, for $g_T \in \mathcal{L}\left(\bar{B}\right)$, there is a finite constant c_{α} such that $|g_T - g_0|_{\lambda,T}^2 \lesssim \bar{B}\left(\frac{\ln KV}{T}\right)^{1/2} + V^{-2\alpha} + \max\left\{c_{\alpha} - \bar{B}, 0\right\}^2$ in probability. Hence, for any $\bar{B} \to \infty$, choosing $V \asymp (T/\ln T)^{1/(4\alpha)}$, $|g_T - g_0|_{\lambda,T}^2 \lesssim \bar{B} T^{-1/2} (\ln KT)^{1/2}$, in probability.*

3.6.5 Neural Networks

Suppose $f(x) = \int_{\mathbb{R}} \int_{\mathbb{R}} \varphi(a_1 x + a_0) \, d\nu(a_0, a_1)$ for $x \in [0,1]$, where ν is a signed measure of finite variation equal to $1/2$, and φ is as in Section 3.6.3. Up to a scaling constant, any continuous bounded function on $[0,1]$ admits this representation (e.g., Yukich et al., 1995, Section II). Denote such class of univariate functions by $\mathcal{H}$. Usually, φ is a sigmoidal function, a monotone function such that $\lim_{x \to \infty} \varphi(x) = 1$ and $\lim_{x \to -\infty} \varphi(x) = 0$. Consider the truncated series expansion $\sum_{v=1}^{V} a_{1v} \varphi(a_{2v} x - a_{3v})$ for some finite V. Denote the set of such series expansions with V terms by

$$\mathcal{H}_V := \left\{ f(x) = \sum_{v=1}^{V} a_{1v} \varphi(a_{2v} x - a_{3v}) : \sum_{v=1}^{V} |a_{1v}| \leq 1, \, a_{2v}, a_{3v} \in \mathbb{R} \right\}.$$

Let $\Theta_k := \{f \circ \pi_k : f \in \mathcal{H}_V\}$. Suppose that $g_0 \in \mathcal{G}(B_0)$ (in (12)). The uniform error incurred by the best approximation in $\mathcal{H}_V$ for $\mathcal{H}$ is $V^{-1/2}$ P-almost surely (Theorem 2.1 in Yukich et al., 1995). Hence, using (13), the sieve with $V^{-1} = O\left(T^{-1/2}\right)$ leads to an approximation error for g_0 that is $O\left(T^{-1/4}\right)$. By the arguments in Section 3.6.4 and the fact that φ is Holder's continuous as in Section 3.6.3, the following is deduced.

Corollary 8 *Suppose that $g_0 \in \mathcal{G}(B_0)$. Under Conditions 1 and 3, for the estimator $g_T \in \mathcal{L}\left(\bar{B}\right)$, for any $V \geq 1$,*

$$|g_T - g_0|^2_{\lambda,T} \lesssim \bar{B}\left(\frac{\ln KV}{T}\right)^{1/2} + \max\left\{B_0 - \bar{B}, 0\right\}^2 + V^{-1}$$

in probability. Hence, choosing $V \asymp T^{1/2}$, for any $\bar{B} \to \infty$, $|g_T - g_0|^2_{\lambda,T} \lesssim \bar{B}\left(\frac{\ln KT}{T}\right)^{1/2}$ in probability.

3.6.6 Shape Constrained Estimator: Many Monotone Lipschitz Functions

Consider estimation under monotone function constraints. Suppose $\mathcal{H}$ is the class of monotone increasing Lipschitz functions with domain $[0,1]$ and bounded by one. Let the Lipschitz constant be known and equal to α. Let $\mathcal{H}_V$ be the class of univariate Bernstein polynomials of order V. Recall that f_V is a Bernstein polynomial of order V if $f_V(x) = \sum_{v=0}^{V}\binom{V}{v}a_v x^v (1-x)^{V-v}$, $x \in [0,1]$, for any real a_v. If $a_v \geq a_{v-1}$ for all v's, the polynomial is monotonically increasing. If also $a_v - a_{v-1} \leq \alpha/V$ for all v's, it is Lipschitz with constant α (e.g., Lorentz, 1986, Ch.1.4). Hence, under these constraints on the coefficients of the polynomial, $\mathcal{H}_V$ is a subset of functions with Lipschitz constant bounded by α. Moreover, for each $f \in \mathcal{H}$ there is an $f_V \in \mathcal{H}_V$ such that $|f_V - f|_\infty \lesssim \alpha V^{-1/2}$ (e.g., Lorentz, 1986, Theorem 1.6.1). Let $\Theta_k := \{f \circ \pi_k : f \in \mathcal{H}_V\}$. Estimation of monotone functions with known Lipschitz constraint can be conveniently performed by Bernstein polynomials, using the algorithm in Section 3.5. The estimation problem becomes a linear programming problem at each step. To see this, define $q_v(x) := \binom{V}{v}x^v(1-x)^{V-v}$. In particular, $D_T(g,\theta)$ in Section 3.5 is linear in θ. Hence, maximization of $D_T(F_{j-1},\theta)$ w.r.t. $\theta \in \Theta_k$ is equivalent to

$$\max_{\{a_v : v \leq V\}} \sum_{v=0}^{V} a_v \left[\int_0^T q_v(X_k(t))\, dN(t) - \int_0^T q_v(X_k(t)) \exp\{g(X(t))\}\, dt\right]$$

such that $0 \leq a_{v-1} \leq a_v \leq 1$, and $a_v - a_{v-1} \leq \alpha/V$, $v = 1, 2, ..., V$. This is routinely solved by the simplex method for each k. From Corollary 2.7.2 in van der Vaart and Wellner (2000), deduce that the entropy integral for functions in $\mathcal{H}_V$ is a constant multiple of $\alpha^{1/2}$. The following uses this observation when applying Theorem 1.

Corollary 9 *Let $g_0 \in \mathcal{G}(B_0)$. Under Conditions 1 and 3, for $g_T \in \mathcal{L}\left(\bar{B}\right)$, $V \gtrsim \alpha^{3/2} \times T^{1/2}$, and $\bar{B} \to \infty$, $|g_T - g_0|^2_{\lambda,T} \lesssim \bar{B}\left(\frac{\alpha + \ln K}{T}\right)^{1/2}$, in probability.*

If the Lipschitz constant is not known, we can let $\alpha \to \infty$ in the estimation. In this case, the entropy integral is finite, but not bounded.

3.7 Choice of $\bar{B}$

Given the relation with l_1 penalization (see (8)), the model degrees of freedom can be approximated by the resulting number of active variables (e.g. Bradic et al., 2011). Hence, the value $\bar{B}$ can be chosen by maximizing the Akaike's penalized likelihood (AIC): $AIC_T(B) := \sup_{g \in \mathcal{L}(B)} L_T(g) - K_B$ where K_B is the number of nonzero parameters in $g_T = \arg\sup_{g \in \mathcal{L}(B)}$. This is less computationally intensive than cross-validation. (In a time series context, cross-validation requires some care except for a special few cases; e.g., Burman et al., 1994).

For very large sample size, AIC will select models that are also very large. In this case cross-validation with a large validation sample (i.e., leaving out a large proportion of the data) tends to select smaller models. Hence, the method to be used depends on the context. See Sections 4.2 and 5 for further discussion and applications. Finally, note that to speed up the calculations for the choice of $\bar{B}$, the algorithm in Section 3.5 can be used without line search.

3.8 Model Fit and Out of Sample Evaluation

Model adequacy can be carried out in large samples using the log-likelihood evaluated out of sample. The out of sample log-likelihood ratio for two competing models $g_t, g_t' \in \mathcal{L}\left(\bar{B}\right)$ which are predictable at time t is

$$L_S\left(g, g'\right) = \int_0^S \left[g_t\left(X\left(t\right)\right) - g_t'\left(X\left(t\right)\right)\right] dN\left(t\right) - \int_0^S \left[\exp\left\{g_t\left(X\left(t\right)\right)\right\} - \exp\left\{g_t'\left(X\left(t\right)\right)\right\}\right] dt.$$

In practice, one may split the sample and estimate g_t and g_t' on the first half or every so often using past observations. The predictable part of the log-likelihood ratio is

$$H_S\left(g, g'\right) = \int_0^S \left[g_t\left(X\left(t\right)\right) - g_t'\left(X\left(t\right)\right)\right] d\Lambda\left(t\right) - \int_0^S \left[\exp\left\{g_t\left(X\left(t\right)\right)\right\} - \exp\left\{g_t'\left(X\left(t\right)\right)\right\}\right] dt,$$

where $\Lambda(t)$ is a short for $\Lambda([0,t])$. Model g outperforms g' if $H_S(g,g') > 0$. (If $g = g_0$, $H_S(g,g') \geq 0$, with equality only if $g' = g_0$, see Lemma 4 in the supplementary material.) The following null hypothesis can be tested: $H_S(g,g') = 0$ against a one or two sided alternative. Under the null,

$$L_S\left(g, g'\right) = \int_0^S \left[g_t\left(X\left(t\right)\right) - g_t'\left(X\left(t\right)\right)\right] d\left(N\left(t\right) - \Lambda\left(t\right)\right).$$

The following martingale result is the justification for the testing procedure.

Proposition 1 *Suppose that g_t and g'_t are predictable bounded processes and $H_S(g, g') = 0$. Suppose that as $S \to \infty$*

$$\frac{1}{S}\int_0^S \left[g_t(X(t)) - g'_t(X(t))\right]^2 d\Lambda(t) \to \sigma^2 > 0$$

in probability. Let $\hat{\sigma}_S^2 := \frac{1}{S}\int_0^S \left[g_t(X(t)) - g'_t(X(t))\right]^2 dN(t)$. Then, $L_S(g, g')/\sqrt{S\hat{\sigma}_S^2}$ converges in distribution to a standard normal random variable.

The testing framework falls within the prequential framework of Dawid (e.g., Seillier-Moiseiwitsch and Dawid, 1993, for applications).

This methodology can be applied in various ways. As an example, consider a sample of size $2T$. Use $[0, T]$ to find the estimators g_T and g'_T. Conduct the test on $(T, 2T]$ so that, mutatis mutandis, $S = T$ in the proposition. In this case, g_T and g'_T are predictable. We need to suppose that the testing sample size S increases to infinity in order to apply the result. If the size T of the testing sample is large, the asymptotic result is applicable.

4 Application to Estimation and Forecasting of Trade Arrivals of New Zealand Dollar Futures

One motivation for the estimation method discussed here was to understand the variables that affect the trade arrivals of the New Zealand Dollar futures, i.e. the futures on NZDUSD traded on the Chicago Mercantile Exchange (CME). The New Zealand Dollar is a liquid currency futures, but not as much as other currency futures (Fx futures) such as the Euro, Australian Dollar, and the Swiss Franc (against the Dollar). What are the variables that affect a trade arrival such as a buy trade? Are these variables, and relations if any, stable in the sense that one can forecast a buy trade arrival tomorrow having estimated a model with today's data? These questions are important to the understanding of market microstructures, and the general etiology of the Fx futures markets and its relation to other instruments like equity markets, commodities, etc. In fact, the New Zealand Dollar belongs to the commodity Fx group that includes the Australian Dollar and Canadian Dollar. These are the currencies of countries whose economy relies on commodity exports. Anecdotal evidence seems to suggest that the New Zealand Dollar tends to increase in value when risk appetite increases.

Below, the data are described and subsequently the model is estimated.

4.1 Data and Variables Description

The estimation of the intensity of trade arrivals is an important problem (e.g., Hall and Hautsch, 2007). New Zealand Dollar futures (the NZDUSD futures front month contract, whose ticker is 6N) are traded on the Chicago Mercantile Exchange. Two days of trading between 8am to 5pm GMT are considered in particular, 10/09/2013 and 11/09/2013. The time slot is based on liquidity considerations. Data are proprietary and were collected with high precision time stamps by a Chicago proprietary trading firm with co-located servers in the Aurora data centre in Chicago. In consequence, trades were classified as buy or sell with minimal probability of error. The data has nanosecond time stamps, and trades time stamps have been adjusted to account for delays in the CME network and reporting (these adjustments are in the order of half a millisecond). This ensures that only information prior to the trade is used to define covariates. Buy and sell intensities are estimated separately. The covariates are derived using information from 6N as well as from other contracts that are perceived as likely to have an impact.

Covariates are constructed from the following CME futures: NZDUSD (6N), AUDUSD (6A), EURUSD (6E), GBPUSD (6B), CADUSD (6C), JPYUSD (6J), CHFUSD (6S), MXNUSD (6M), Crude Oil (CL), Gold (GC), and mini S&P500 (ES). For each instrument, covariates were derived from order book and trade updates. In particular, the variables are mid-price returns, bid-ask spread, volume imbalances for the first two levels, trade imbalances, and trade duration. Variables are updated every time there is a change in their value. For example, the return is computed when there is a mid price change from the previous mid. Volume imbalances are computed as the difference of the bid and ask quantities on each level (the contracts usually quote prices for 5 levels). These differences are then standardized by the sum on the bid and ask quantity on that level. Trade imbalances are the signed traded size, positive if a buy and negative if a sell. Excluding the spread, moving averages of all variables are also computed. In particular, moving averages of order 1, 2, 4, 8, 16, 32, 64, and 128 are used. This is to allow information at slightly different frequencies to affect the intensity in a way similar to MIDAS. Overall, the total number of variables is 508 including a constant. A model that allows for squares and third powers of all the standardized variables is also estimated. In this case, the total number of variables is 1,522 including a constant. Once the feature variables are computed, in order to reduce the computational burden these are sampled only when there is an update in the NZDUSD futures. The argument is that if an instrument leads 6N, then the book for 6N would update before a trade.

4.2 Computational Details

The two-day sample is split into three parts. The first half of day one is the estimation sample. The second half of day one is the validation sample. The second day is the testing sample.

The variables are winsorized at the 95% quantile and then standardized by it so as to take values in $[-1, 1]$. If a variable takes both positive and negative values, winsorization is applied to the absolute value which is then signed. For estimation of the cubic polynomial, powers of the variables are computed after having mapped the variables into $[-1, 1]$. The quantile is computed using the data from day one only. Hence, winsorization on the testing sample is based on the previous day 95% quantile. After winsorization, the set of weights $\mathcal{W}$ is chosen equal to the sample estimator of the L_2 norm, i.e. $w_\theta = \left(\frac{1}{T}\int_0^T \theta^2\left(X\left(t\right)\right)dt\right)^{1/2}$ over day one. This ensures that all variables are given the same importance. The model is estimated for $B \in \{2, 4, 8, 16\}$ on the estimation sample. We set $\bar{B}$ equal to the B that maximizes the likelihood on the validation sample. This method is an alternative to AIC when the sample size is large. With this choice of $\bar{B}$, the model is then re-estimated using the data in the first day, i.e. both estimation and validation sample. This approach is feasible in a large sample and avoids some of the drawbacks of cross-validation for dependent observations.

4.3 Estimation Results

It is difficult to clearly and concisely report the variables that appear to be most important for the intensity. In fact, a large number of variables are included by the method described here, though they have small coefficients. For the linear model, the chosen $\bar{B}$ results in a model for buy and sell trades with 77 and 68 covariates, respectively. For the cubic case, the number was slightly larger. Including many variables with relatively small coefficients produces an averaging effect across many variables and can provide a hedge against instability and noise, in a similar way to forecast combination.

The intersection of the first ten variables in the linear model for buy and sell trades is reported in Table 1. These variables can be seen as some form of a more stable subset of variables (Meinshausen and Bühlmann, 2010, for formal methods on stability selection).

Table 1: Most important variables affecting buy and sell trade arrival in linear model.

Instrument	Variable
6N	Volume Imbalance on Level 1
6N	Volume Imbalance on Level 2
6N	Spread
6A	Duration from Last Trade

Interestingly, past durations of 6N (the New Zealand Dollar futures) do not seem to be as important so they are not included in Table 1. However, the durations of 6A (the Australian

Dollar) appear to be important. The Australian Dollar tends to correlate with the New Zealand Dollar but it is more liquid. Hence, it might provide useful information on trade arrival. Past durations have been found to be important predictors in some high frequency financial applications (e.g., Engle and Russell, 1998). However, order book information seems to have a greater impact (e.g., Cont et al., 2014). In the next section, linear and cubic models using only the variables in Table 1 will also be used for comparison and will be referred to as the restricted linear and cubic models.

4.3.1 Out of Sample Performance

Having estimated the model on the first day, it is of interest to see if the model can be used to explain a trade arrival out of sample. This is done by computing the average log-likelihood ratio $L_S(g, g')/S$, and $\hat{\sigma}_S/\sqrt{S}$ on the second day (see Proposition 1). Confidence intervals can then be constructed using Proposition 1. The goal is to assess the out of sample performance of the linear and cubic models as well as the restricted models (the ones with variables in Table 1). It is of interest to verify if restricting attention to a linear model might produce similar out of sample results. When compared to the constant intensity (Conts.), the constant is computed as the out of sample maximum likelihood estimator, i.e., the best constant intensity with hindsight.

Table 2 shows that all of the models do improve on the constant intensity with overwhelming evidence. When looking at the relative merits of the unrestricted models, it becomes unclear whether a cubic model adds value out of sample. Looking at the restricted linear model relative to the unrestricted linear model, there is overwhelming evidence that the unrestricted model should be preferable. It is interesting that when comparing the restricted models, there is overwhelming evidence that a cubic model does improve on the linear one. From these results one could infer that modelling nonlinearities does pay off when looking at small dimensional models. However, when models are linear but have many covariates, nonlinear impact of book and trade variables is less obvious. The simulation results of Section 5 support this claim.

Table 2: Out of sample performance of models: g vs. g' with g and g' as defined in the headings below.

	Lin. vs. Const.		Cubic vs. Const.	
	Buy	Sell	Buy	Sell
Avg.Log-LR.$\times 10^2$	3.77	4.48	4.02	4.56
S.E.$\times 10^2$	0.33	0.25	0.31	0.26
P-Val.	$<$0.01	$<$0.01	$<$0.01	$<$0.01
	Cubic vs. Linear			

	Buy	Sell		
Avg.Log-LR.$\times 10^2$	0.25	0.08		
S.E.$\times 10^2$	0.08	0.07		
P-Val.	$<$0.01	0.22		
	Lin. Restr. vs. Const.		Lin. Restr. vs. Lin.	
	Buy	Sell	Buy	Sell
Avg.Log-LR.$\times 10^2$	1.14	1.24	-2.63	-3.25
S.E.$\times 10^2$	0.15	0.14	0.20	0.19
P-Val.	$<$0.01	$<$0.01	$<$0.01	$<$0.01
	Cubic Restr. vs. Lin. Restr.			
	Buy	Sell		
Avg.Log-LR.$\times 10^2$	0.26	0.25		
S.E.$\times 10^2$	0.10	0.06		
P-Val.	$<$0.01	$<$0.01		

5 Numerical Examples

As remarked in Section 1.2, $\{\Lambda((T_{i-1}, T_i]) : i \in \mathbb{N}\}$ (Λ as in (1)) is i.i.d. exponentially distributed with mean 1. For simplicity in the simulations, it is assumed that the covariates only update immediately after each jump time T_i. Hence, the intervals $(T_{i-1}, T_i]$ are simulated from an exponential distribution with parameter $\exp\{g(X(T_{i-1}))\}$, i.e., with mean $\exp\{-g(X(T_{i-1}))\}$. The covariates are standard Gaussian random variables with Toeplitz covariance $Cov(X_k(t), X_l(t)) = \rho^{|k-l|}$ and uncorrelated over time. The variables have been capped to 2 in absolute value, i.e., they take values in $[-2, 2]$.

The parameters in the simulation are $K \in \{10, 50\}$ number of covariates, $T = T_{100}$ (recall $N(T_n) = n$) sample size, and $\rho \in \{0, 0.75\}$. Different choices of g_0, and Θ are considered. These are summarized as follows. For estimation simplicity, Θ is a finite set of functions.

5.1 True Unknown Model g_0

Here we describe various options for the true function g_0. The true function g_0 takes the form $g_0(x) = \sum_{k=1}^{K} g_0^{(k)}(x)$, where the functions $g_0^{(k)}$ are defined as follows.

True additive functions. Linear: $g_0^{(k)}(x) = b_{0k} x_k$; NonLinear: $g_0^{(k)}(x) = b_{0k}(|x_k| + 0.5 x_k)$.

Active variables. FewLarge $b_{0k} = 1$ for $k = 1, 2, 3$, $b_{0k} = 0$ for $k > 3$; ManySmall $b_{0k} = 1/\sqrt{10}$ for $k \leq 10$, $b_{0k} = 0$ for $k > 10$. Even when there is no model misspecification, these values are unknown to the researcher.

5.2 Estimator in $\mathcal{L}(B, \Theta, \mathcal{W})$

Here we define the parameter space $\mathcal{L}(B, \Theta, \mathcal{W})$ used by the researcher. Estimation is carried out allowing for model misspecification. Hence, depending on the design the choice of functions does not need to correspond to the true functions $g_0^{(k)}$ (Section 5.1). The estimated models are of the form $g(x) = \sum_{k=1}^{K} \sum_{\theta \in \Theta_k} b_\theta \theta(x)$. Details regarding Θ_k and the estimation of the b_θ's are as follows.

Functions in Θ. Linear (Lin): $\theta(x) = x_k$ for $\theta \in \Theta_k$; Monomials (Poly): $\theta(x) = (x_k/2)^a$ for $\theta \in \Theta_k$ with $a = 1, 2, 3$. A constant is added by default in the estimations. When the true function is linear (i.e., $g_0^{(k)}(x) = b_{0k} x_k$) there is no misspecification error. However, the coefficients still need to be estimated and many of them can be zero. When the true function is nonlinear, misspecification error will be incurred even when the estimation is carried out using a polynomial (Poly). However, in this case the degree of misspecification will be small.

Choice of $\bar{B}$ and $\mathcal{W}$ The parameter $\bar{B}$ is chosen as the $B \in \{1, 4, 8, 16\}$ that maximizes AIC_T as defined in Section 3.7. In this case, the sample size is relatively small and the performance of AIC_T and cross-validation (leaving out many variables) was similar. Hence, AIC_T is preferred for computational convenience. We applied the algorithm in Section 3.5 with $F_0 = \ln(N(T)/T)$ rather than $F_0 = 0$. In this case e^{F_0} is an estimator of $P\lambda$, the expected intensity. The main reason was to reduce fine tuning of the set of possible values of B to the different functions and simulation designs. The simulation design is such that as the number of active variables increases, $P\lambda$ increases and in consequence the optimal B.

The weights in $\mathcal{W}$ are chosen to be the sample L_2 norm as in Section 4.2. Note that no winsorization is applied to the variables, as they are already bounded.

5.3 Simulation Results

The following loss function is considered to assess the model fit,

$$Loss(g) := \frac{\int_T^{T+S} [g_0(X(t)) - g(X(t))]^2 dN(t)}{\int_T^{T+S} [g_0(X(t)) - \gamma_0]^2 dN(t)} \tag{16}$$

where $\gamma_0 := \frac{\int_T^{T+S} g_0(X(t)) dN(t)}{N(T+S) - N(T)}$. This loss function is justified noting that when S is large, $Loss(g) \simeq |g_0 - g|_\lambda^2 / \left[\inf_{\gamma > 0} |g_0 - \gamma|_\lambda^2\right]$. Hence, the numerator in $Loss$ is an approximation to the convergence criterion of Theorem 1, while the denominator is the error incurred by γ_0 which is the best constant approximation with hindsight. The standardization ensures that $Loss(g) \in [0, 1)$ if g improves over γ_0, if not $Loss(g) \geq 1$. The denominator in $Loss$ is the benchmark for the finite sample experiment carried out here. In the simulations, data

are generated for a sample period $[0, T_{1100}]$, and the model is estimated on $[0, T_{100}]$ and out of sample performance is evaluated on $[T_{100}, T_{1100}]$. Hence, in *Loss*, $T = T_{100}$ and $S = T_{1100} - T_{100}$. Table 3 reports the median of $Loss\,(g_{T_{100}})$ (LOSS) together with the 75% and 25% quantile.

Overall, different choices of true model (linear or convex) and basis functions allow us to gauge the main features of the estimator. The results in Table 3 can be summarized as follows. There is a clear advantage in using a nonlinear model when the true model is nonlinear, but also a considerable loss (mostly due to estimation error) when the true model is linear. For nonlinear estimators such as polynomials, a judicious choice of $\mathcal{W}$ to dump the effect of higher order coefficients can make the estimator more robust. The present choice of $\mathcal{W}$ is equivalent to standardizing the variables by their L_2 norm. This is simple, but might lead to big oscillations if the order of polynomial is not as small as it is here. The choice of $\mathcal{W}$ is an important part of the modelling and estimation procedure when dealing with polynomials. An increase in variables correlation produces better forecasts. This is in contrast with the problem of variable screening. Numerical experiments of the author - not reported here - as well as related results in the literature (e.g., Bradic et al., 2011) show that in this context, false discovery of active variables increases substantially with correlation. This is natural, as correlation confounds the merits of each single variable. The forecasting and variable screening are related but complementary problems, which require separate treatment.

Table 3: Simulation results relative to the best constant intensity with hindsight. Estimation is based on samples of size T_{100} corresponding to $N\,(T_{100}) = 100$ number of jumps. The table reports the median (Med.) and the 25 (Q25%) and 75 (Q75%) percent quantiles of $Loss \times 100$ (*Loss* as in (16)). A number below 100 means a relative improvement on the best constant intensity with hindsight.

	Loss×100			Loss×100		
	Med.	Q25%	Q75%	Med.	Q25%	Q75%
	$\rho = 0$			$\rho = 0.75$		
g_0 is Linear FewLarge $K = 10$						
Lin	3.70	2.37	5.72	1.69	1.11	2.61
Poly	5.54	3.60	8.10	2.76	1.85	4.19
g_0 is Linear FewLarge $K = 50$						
Lin	6.83	4.92	9.41	3.64	2.34	5.26
Poly	10.61	8.01	13.66	4.30	2.82	6.65
g_0 is Linear ManySmall $K = 10$						
Lin	13.42	9.90	19.16	2.72	1.93	3.87

Poly	32.88	24.93	41.65	4.05	2.63	5.90
g_0 is Linear ManySmall $K = 50$						
Lin	47.02	35.29	57.26	4.81	3.42	6.22
Poly	60.83	51.45	74.57	6.23	4.64	8.30
g_0 is Convex FewLarge $K = 10$						
Lin	81.08	75.13	92.10	70.08	63.90	77.56
Poly	19.92	15.03	26.82	9.35	7.19	12.77
g_0 is Convex FewLarge $K = 50$						
Lin	110.23	90.67	123.28	83.53	72.46	95.66
Poly	35.47	28.08	45.36	14.70	11.86	19.36
g_0 is Convex ManySmall $K = 10$						
Lin	97.95	87.20	112.47	73.59	66.67	82.64
Poly	17.16	14.42	20.58	5.49	4.60	6.90
g_0 is Convex ManySmall $K = 50$						
Lin	104.12	94.80	114.59	67.38	62.55	74.41
Poly	48.24	40.72	57.95	10.67	8.86	13.13

5.3.1 Simulations with Dynamics: Hawkes Process with Covariates

The previous simulations considered time independent covariates. Here, we make the covariates time dependent, following an autoregressive process and also allow the intensity to follow a Hawkes process. Consider the intensity

$$\lambda(t) = \exp\left\{\ln\left(c_0 + \int_{(0,t)} e^{-a_0(t-s)} dN(s)\right) + g_0(X(t))\right\} \tag{17}$$

This is in the form of Section 3.6.2, though the function $f(\cdot) = \ln(c_0 + \cdot)$ is bounded below (because its domain is positive), it is not bounded above. Here, $c_0 > 0$ is required to avoid degeneracy. To directly apply the results in Section 3.6.2 we could use $f(\cdot) = \max\{\ln(c_0 + \cdot), \bar{c}\}$ instead for some finite $\bar{c}$, in which case the process is assured to be stationary (see Corollary 5). The process simplifies to

$$\lambda(t) = \left(c_0 + \int_{(0,t)} e^{-a_0(t-s)} dN(s)\right) \exp\{g_0(X(t))\}. \tag{18}$$

Using results for marked Hawkes processes (e.g., Brémaud et al., 2002), one could surmise that (18) would be stationary if $a_0 > \mathbb{E}\exp\{g_0(X(t))\}$. To the author's knowledge, formal existing results do not fit exactly into the framework of (18). In the simulations we add a constant to the true model, i.e., $g_0(x) = \gamma + \sum_{k=1}^{K} g_0^{(k)}(x)$ where $\gamma = -\mathbb{E}\exp\left\{\sum_{k=1}^{K} g_0^{(k)}(X(t))\right\}$, so

that $\mathbb{E}\exp\{g_0(X(t))\} = 1$. This should ensure the aforementioned stationarity of (18) when $a_0 > 1$. Other than that, the true models for g_0 are as in Section 5.1. In the simulations we verified that the term in parenthesis on the r.h.s. of (18) remains bounded, thus ensuring stationarity with no need of a capping constant $\bar{c}$. This model can be simulated and estimated, and details concerning this and some of the calculations to be discussed below can be found in Section A.2 of the supplementary material.

As in the previous simulation, we let $X(t) = X(T_{i-1})$ for $t \in (T_{i-1}, T_i]$. However, the $X(T_i)$'s now follow the vector autoregression $X(T_i) = 0.95X(T_{i-1}) + \varepsilon_i$, $X(T_0) = \varepsilon_0$, where the K dimensional innovations ε_i's are generated as the i.i.d. truncated Gaussian with Toeplitz covariance exactly as the i.i.d. $X(T_i)$'s used in Section 5.3. If the $X(T_i)$'s were independent as in the previous simulation, the dependence in the Hawkes component would be confounded by the independent variability in $\exp\{g_0(X(T_i))\}$. Given the dependence structure, we use a larger sample size T_n with $n = 200$. In the simulations, we set $c_0 = 2$ and $a_0 = 1.3$.

Except for these differences, the set up is the same as in the previous simulation. However, we have c_0 and a_0 as extra parameters to be estimated. The goal of the simulations is to see how the remarks made in the case of time independent variables may hold in this case. Results are reported in Table 4. Results in Table 3 and Table 4 are not directly comparable because of the scaling required for stationarity. However, we can establish conclusions in relative terms.

Table 4 confirms the overall situation of Table 3. However, time series dependence makes the problem harder, as expected. The relative benefit of estimating a nonlinear model when the true g_0 is nonlinear decreases substantially in the present scenario. For example, in the case of Convex ManySmall with $K = 50$, the ratio of the loss for Lin and Poly in Table 3 is $104.12/48.24 = 2.16$, while in Table 4 is $52.55/42.55 = 1.23$.

Table 4: Simulation results relative to the best constant intensity with hindsight. The model is as in (17). Estimation is based on samples of size T_{200} corresponding to $N(T_{200}) = 200$ number of jumps. The table reports the median (Med.) and the 25 (Q25%) and 75 (Q75%) percent quantiles of $Loss \times 100$ ($Loss$ as in (16)). A number below 100 means a relative improvement on the best constant intensity with hindsight.

	Loss×100			Loss×100		
	Med.	Q25%	Q75%	Med.	Q25%	Q75%
	$\rho = 0$			$\rho = 0.75$		
g_0 is Linear FewLarge $K = 10$						
Lin	1.04	0.71	1.51	0.54	0.37	0.81
Poly	1.53	1.01	2.34	0.95	0.70	1.31

g_0 is Linear FewLarge $K = 50$						
Lin	3.53	1.78	7.78	2.72	1.96	3.74
Poly	4.76	2.87	8.29	4.15	3.26	5.60
g_0 is Linear ManySmall $K = 10$						
Lin	2.53	1.70	3.76	0.95	0.59	1.38
Poly	5.22	3.57	7.01	1.77	1.27	2.52
g_0 is Linear ManySmall $K = 50$						
Lin	19.70	13.32	28.09	4.66	3.13	6.96
Poly	20.49	14.23	28.27	6.22	4.28	8.88
g_0 is Convex FewLarge $K = 10$						
Lin	45.96	37.24	57.86	36.94	29.25	48.42
Poly	5.82	4.38	8.16	3.26	2.40	4.70
g_0 is Convex FewLarge $K = 50$						
Lin	72.03	56.68	91.19	44.44	36.51	55.90
Poly	26.40	20.22	36.27	14.58	11.59	20.43
g_0 is Convex ManySmall $K = 10$						
Lin	33.96	26.78	46.12	21.46	17.07	28.86
Poly	14.35	10.79	19.01	5.43	3.93	7.84
g_0 is Convex ManySmall $K = 50$						
Lin	52.55	42.57	68.41	32.56	24.28	42.81
Poly	42.55	34.79	50.73	23.48	17.94	29.86

6 Concluding Remarks

This paper introduced a general framework for estimation of high dimensional point processes with a focus on forecasting. The estimation methodology is feasible using a greedy algorithm. The rates of consistency in the case of many additive components are optimal. A set of examples for the applicability of different estimation procedures and their convergence rates are derived as corollaries of the main result. This asymptotic analysis differs from the one where only a few variables are active, which is usually addressed in the high dimensional statistical literature. In finance, because of a very low signal to noise ratio it is often found that most of the variables are cross-sectionally correlated but are weak predictors. As a consequence, no one variable dominates. Hence the asymptotic analysis carried out here is in this vein. The empirical study of the prediction of buy and sell trade arrivals for futures on the New Zealand Dollar seems to confirm that using a small subset of the variables might be suboptimal. Hence, it is beneficial to use many variables as long as they are properly aggregated.

More inferential procedures should be devised for high dimensional model estimation. In

finance, many applications require an assessment of model performance out of sample. In high frequency, the size of the dataset is large and the estimation procedures must be computationally feasible. This paper provides some solutions in this direction. For very large sample sizes, one may need to forego the use of the likelihood and work with approximations. In this case, the intensity density could be directly modelled as an additive model and the likelihood replaced with a square loss contrast estimator (e.g., Gaiffas and Guilloux, 2012). Applications in this vein will be the subject of future research.

References

[1] Andersen, P.K. and R.D. Gill (1982) Cox's Regression Model for Counting Processes: A Large Sample Study. Annals of Statistics 10, 1100-1120.

[2] Barron, A.R., A. Cohen, W. Dahmen and R.A. DeVore (2008) Approximation and Learning by Greedy Algorithms. Annals of Statistics 36, 64-94.

[3] Bauwens, L. and N. Hautsch (2009) Modelling Financial High Frequency Data Using Point Processes. In T.G. Andersen, R.A. Davis, J.-P. Kreiss and T. Mikosch (eds.), Handbook of Financial Time Series, 953-982. New York: Springer.

[4] Bradic, J., J. Fan and J. Jiang (2011) Regularization for Cox's Proportional Hazards Model with NP-Dimensionality. Annals of Statistics 39, 3092-3120.

[5] Brémaud, P. (1981) Point Processes and Queues: Martingales Dynamics. New York: Springer.

[6] Brémaud, P. and L. Massoulié (1996) Stability of Nonlinear Hawkes Processes. Annals of Probability 24, 1563-1588.

[7] Bühlmann, P. and S. van de Geer (2011) Statistics for High-Dimensional Data. London: Springer.

[8] Brémaud, P., G. Nappo and G.L. Torrisi (2002) Rate of Convergence to Equilibrium of Marked Hawkes Processes. Journal of Applied Probability 39, 123-136.

[9] Burman, P., E. Chow and D. Nolan (1994) A Cross-Validatory Method for Dependent Data. Biometrika 81, 351-358.

[10] Cont, R., A. Kukanov and S. Stoikov (2014) The Price Impact of Order Book Events. Journal of Financial Econometrics 12, 47-88.

[11] Engle, R.F. and J.R. Russell (1998) Autoregressive Conditional Duration: A New Model for Irregularly Spaced Transaction Data. Econometrica 66, 1127-1162.

[12] Fan, J., I. Gijbels and M. King (1997) Local Likelihood and Local Partial Likelihood in Hazard Regression. Annals of Statistics 25, 1661-1690.

[13] Gaiffas, S. and A. Guilloux (2012), High Dimensional Additive Hazards Models and the Lasso. Electronic Journal of Statistics 6 , 522–546.

[14] Hall, D. and N. Hautsch (2007) Modelling the Buy and Sell Intensity in a Limit Order Book Market. Journal of Financial Markets 10, 249-286.

[15] Hasbrouck, J. (1991) Measuring the Information Content of Stock Prices. Journal of Finance 46, 179-207.

[16] Jaggi, M. (2013) Revisiting Frank-Wolfe: Projection-Free Sparse Convex Optimization. Journal of Machine Learning Research (Proceedings ICML 2013). URL: <http://jmlr.org/proceedings/papers/v28/jaggi13-supp.pdf>

[17] Katznelson, Y. (2002) An Introduction to Harmonic Analysis. Cambridge: Cambridge University Press.

[18] Lorentz, G.G. (1986) Bernstein Polynomials. New York: Chelsea Publishing Company.

[19] Lillo, F., J.D. Farmer and R.N. Mantegna (2003) Master Curve for the Price-Impact Function. Nature 421,129-130.

[20] Meinshausen, N. and P. Bühlmann (2010) Stability Selection (with discussion). Journal of the Royal Statistical Society B 72, 417-473.

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

[22] Ogata, Y. (1978) The Asymptotic Behaviour of the Maximum Likelihood Estimator for Stationary Point Processes. Annal of the Institute of Statistical Mathematics 30 (A), 243-261.

[23] Sancetta A. (2015) A Nonparametric Estimator for the Covariance Function of Functional Data. Econometric Theory 31, 1359-1381.

[24] Sancetta A. (2016) Greedy Algorithms for Prediction. Bernoulli 22, 1227-1277.

[25] Seillier-Moiseiwitsch, F. and A.P. Dawid (1993). On Testing the Validity of Sequential Probability Forecasts. Journal of the American Statistical Association 88, 355-359.

[26] Tsybakov, A.B. (2003) Optimal Rates of Aggregation. Proceedings of COLT-2003, Lecture Notes in Artificial Intelligence, 303-313.

[27] van de Geer (1995) Exponential Inequalities for Martingales, with Application to Maximum Likelihood Estimation for Counting Processes. Annals of Statistics 23, 1779-1801.

[28] van Dijk, D., T. Teräsvirta and P.H. Franses (2002) Smooth Transition Autoregressive Models - A Survey of Recent Developments. Econometric Reviews 21, 1-47.

[29] van der Vaart, A. and J.A. Wellner (2000) Weak Convergence and Empirical Processes. New York: Springer.

[30] Yukich, J.E., M.B. Stinchcombe and H. White (1995) Sup-Norm Approximation Bounds for Networks Through Probabilistic Methods. IEEE Transactions on Information Theory 41, 1021-1027.

Supplementary Material to "Estimation for the Prediction of Point Processes with Many Covariates" by Alessio Sancetta

A.1 Proofs of Results

The notation is collected in the next subsection so that the reader can refer to it when needed.

A.1.1 Preliminary Lemmas and Notation

Write $\mathcal{L}_0 := \mathcal{L}(B_0, \Theta, \mathcal{W})$, $\bar{\mathcal{L}} := \mathcal{L}\left(\bar{B}, \Theta, \mathcal{W}\right)$ and $\mathcal{L} := \mathcal{L}(B, \Theta, \mathcal{W})$ for arbitrary, but fixed B. By Condition 2, the envelope function of $\bar{\mathcal{L}}$, is

$$\sup_{g \in \bar{\mathcal{L}}} \sup_{z \in \mathbb{R}} |g(z)| \leq \bar{B}\bar{\theta}/\underline{w} =: \bar{g}. \tag{A.1}$$

From the main text, recall that $\bar{B}_w := \bar{B}/\underline{w}$. Throughout, to keep notation simpler, suppose that $K > 1$.

To ease notation, write $\Lambda(t)$ for $\int_0^t d\Lambda(s) = \int_0^t \lambda(X(s))\, ds$, $\int_0^t e^g d\mu$ for $\int_0^t e^{g(X(s))} ds$ and similarly for $\int_0^t g dN$, $\int_0^t g d\Lambda$ $\int_0^t g d\mu$, etc., where μ is the Lebesgue measure. Hence, arguments $X(t)$ and t are dropped, but this should cause no confusion: all integrals here are w.r.t. $dN(t)$, $d\mu(t)$ etc., and the argument of all the functions is $X(t)$. Also, $\lambda(X(s)) = e^{g_0(X(s))}$, where $\bar{g}_0 := |g_0|_\infty$. With no loss of generality, to keep notation simple also suppose that $|g_{B_0}|_\infty \leq \bar{g}_0$. If this were not the case, we can just redefine $\bar{g}_0$ to be an upper bound for the uniform norms of g_0 and g_{B_0} (recall the definition of B_0 in (6)). It then follows from (6) that $\sup_{B>0} |g_B|_\infty \leq \bar{g}_0$ because g_B is the best uniform approximation for g_0 in $\mathcal{L}(B)$, and for $B \geq B_0$, (6) implies $g_B = g_{B_0}$. These facts will be used freely in the proofs without further mention. Define the following random Hellinger metric $d_T(g, g_0) = \sqrt{\frac{1}{2}\int_0^T \left(e^{g/2} - e^{g_0/2}\right)^2 d\mu}$. Sometimes, it will be useful to consider the identity $d_T^2(g, 0) = \frac{1}{2}\int_0^T \left|e^{g/2} - 1\right|^2 d\mu$.

Lemma 3 *Suppose that f, f' are functions on $\mathbb{R}^K$. Then,*

$$\frac{1}{8}\int_0^T \left(f - f'\right)^2 e^{f'} d\mu \leq d_T^2\left(f, f'\right). \tag{A.2}$$

Proof. Multiplying and dividing by $e^{f'}$,

$$d_T^2\left(f, f'\right) = \frac{1}{2}\int_0^T e^{f'}\left(e^{(f-f')/2} - 1\right)^2 d\mu. \tag{A.3}$$

Expand the square in the above display

$$\left(e^{(f-f')/2}-1\right)^2 = e^{(f-f')} - 2e^{(f-f')/2} + 1.$$

By Taylor expansion of the two exponentials, the above is equal to

$$\sum_{j=0}^{\infty}\frac{(f-f')^j}{j!} - 2\sum_{j=0}^{\infty}\frac{(f-f')^j}{j!}\left(\frac{1}{2}\right)^j + 1 = \sum_{j=2}^{\infty}\frac{(f-f')^j}{j!}\left(1-\frac{1}{2^{j-1}}\right) \geq \frac{(f-f')^2}{4}.$$

Inserting in (A.3) deduce (A.2). ■

Lemma 4 *Suppose that* $|g_{B_0}|_\infty \leq \bar{g}_0$. *Then,*

$$0 \leq \int_0^T \left[(g_0 - g_{B_0})\, d\Lambda - (e^{g_0} - e^{g_{B_0}})\, d\mu\right] \leq \frac{1}{2}e^{2\bar{g}_0}\int_0^T (g_0-g)^2\, d\Lambda.$$

Proof. By definition of $d\Lambda = e^{g_0}d\mu$,

$$\begin{aligned}\int_0^T \left[(g_0-g)\, d\Lambda - (e^{g_0}-e^{g})\, d\mu\right] &= \int_0^T \left[(g_0-g)\, e^{g_0} - (e^{g_0}-e^{g})\right] d\mu \\ &= \int_0^T \left[(g_0-g) + e^{-(g_0-g)} - 1\right] e^{g_0} d\mu. \qquad \text{(A.4)}\end{aligned}$$

For any fixed real x, by Taylor series with remainder, for some x_* in the convex hull of $\{0, x\}$,

$$e^{-x} - 1 + x = \frac{x^2}{2}e^{-x_*}. \qquad \text{(A.5)}$$

Apply this equality to $x = g_0 - g$ and insert it in the square brackets on the r.h.s. of (A.4) to deduce the upper bound in the lemma because $|g_0 - g_{B_0}|_\infty \leq 2\bar{g}_0$. For any $x > 0$, the following inequality holds:

$$0 \leq (x - \ln x - 1) \qquad \text{(A.6)}$$

with equality only if $x = 1$. Apply this inequality to $x = \exp\{-(g_0 - g_{B_0})\}$ and insert it in the square brackets on the r.h.s. of (A.4) to deduce the lower bound in the lemma. ■

A.1.2 Solution of the Population Likelihood

For simplicity, as in Condition 1 suppose that $T_0 = 0$. Then, by Lemma 2 in Ogata (1978),

$$L(g) = \lim_T \frac{L_T(g)}{T} = \lim_T \frac{1}{T}\int_0^T (gdN - e^g d\mu) = P\left(ge^{g_0} - e^g\right)$$

almost surely, where L_T is the log-likelihood at time T (e.g., Ogata, 1978, eq.1.3). Taking first derivatives, the first order condition is $P\left(he^{g_0} - he^g\right) = 0$ for any $h \in \bar{\mathcal{L}}$. Hence, if $g = g_0$,

the condition is satisfied. To check uniqueness, verify that the second order condition for concavity, i.e., $-Ph^2e^g < 0$, holds for any $h \neq 0$. Using the lower bound $e^{-\bar{g}} \leq e^g$, deduce that $-Ph^2e^g \leq -e^{-\bar{g}}Ph^2 < 0$ holds for any $h \neq 0$ P-almost everywhere. Given that $-L(g)$ is convex and $\bar{\mathcal{L}}$ is convex and closed, the maximizer of $L(g)$ is unique.

A.1.3 Proof of Theorem 1

The result is derived for the Hellinger distance d_T rather than the norm $|\cdot|_{\lambda,T}$.

Define $C_T^2 := C^2 \times T \max\left\{r_T^{-2}, 2e^{3\bar{g}_0}\left|g_0 - g_{\bar{B}}\right|_\infty^2\right\}$ and the martingale $M = N - \Lambda$ (Λ in (1) is the compensator of N). Here, r_T is a nondecreasing sequence which will be defined in due course. With the present notation, the last display in the proof of lemma 4.1 in van de Geer (1995) states that

$$\frac{1}{2}\int_0^T (g - g_0)\, dM \geq d_T^2(g, g_0) + \frac{1}{2}L_T(g, g_0), \tag{A.7}$$

where $L_T(g, g_0) := L_T(g) - L_T(g_0)$ for any g, so also for $g = g_T$. (The above display is only valid when g_0 is the true function, but it is not required that $g_0 \in \mathcal{L}(B)$ for some B.) By Condition 3, and the inequality $L_T(g_T, g_{\bar{B}}) \geq L_T(g_T) - \sup_{g\in\bar{\mathcal{L}}} L_T(g)$, deduce that

$$L_T(g_T, g_0) = L_T(g_T, g_{\bar{B}}) + L_T(g_{\bar{B}}, g_0) \geq -\left(C_T^2/2\right) + L_T(g_{\bar{B}}, g_0) \tag{A.8}$$

choosing C large enough, in the definition of C_T. Hence, inserting (A.8) in (A.7), deduce that

$$\begin{aligned} & \Pr\left(d_T(g_T, g_0) > C_T\right) \\ \leq\ & \Pr\left(\frac{1}{2}\left[\int_0^T (g - g_0)\, dM - L_T(g_{\bar{B}}, g_0)\right] \geq d_T^2(g, g_0) - \frac{C_T^2}{4}\right. \\ & \left. \text{and } d_T^2(g, g_0) > C_T^2 \text{ for some } g \in \bar{\mathcal{L}}\right) \end{aligned} \tag{A.9}$$

To bound the term in the square bracket, add and subtract $\int_0^T g_{\bar{B}} dM$ and note that $L_T(g_{\bar{B}}, g_0)$ can be written as $\int_0^T \left[(g_{\bar{B}} - g_0)\, dM + (g_{\bar{B}} - g_0)\, d\Lambda - (e^{g_{\bar{B}}} - e^{g_0})\, d\mu\right]$. This implies that

$$\begin{aligned} \int_0^T (g - g_0)\, dM - L_T(g_{\bar{B}}, g_0) &= \int_0^T \left[(g - g_{\bar{B}}) + (g_{\bar{B}} - g_0)\right] dM \\ &\quad - \int_0^T \left[(g_{\bar{B}} - g_0)\, dM + (g_{\bar{B}} - g_0)\, d\Lambda - (e^{g_{\bar{B}}} - e^{g_0})\, d\mu\right] \\ &= \int_0^T (g - g_{\bar{B}})\, dM + \int_0^T \left[(g_0 - g_{\bar{B}})\, d\Lambda - (e^{g_0} - e^{g_{\bar{B}}})\, d\mu\right] \\ &\leq \int_0^T (g - g_{\bar{B}})\, dM + \frac{1}{2}e^{2\bar{g}_0}\int_0^T (g_0 - g_{\bar{B}})^2\, d\Lambda \end{aligned}$$

using Lemma 4 in the inequality. From the above calculations, and the fact that $\int_0^T (g_0 - g_{\bar{B}})^2 \, d\Lambda \leq Te^{\bar{g}_0} |g_0 - g_{\bar{B}}|_\infty^2$, deduce that (A.9) is less than

$$\begin{aligned} & \Pr\left(\frac{1}{2}\int_0^T (g - g_{\bar{B}})\, dM \geq d_T^2(g, g_0) - \frac{C_T^2}{4} - \frac{1}{2}Te^{3\bar{g}_0} |g_{\bar{B}} - g_0|_\infty^2 \right. \\ & \left. \text{and } d_T^2(g, g_0) > C_T^2 \text{ for some } g \in \bar{\mathcal{L}}\right) \\ \leq \; & \Pr\left(\frac{1}{2}\int_0^T (g - g_{\bar{B}})\, dM \geq d_T^2(g, g_0) - \frac{C_T^2}{2} \text{ and } d_T^2(g, g_0) > C_T^2 \text{ for some } g \in \bar{\mathcal{L}}\right), \end{aligned}$$

using the definition of C_T. The above is bounded by $\Pr\left(\sup_{g\in\bar{\mathcal{L}}} \int_0^T (g - g_{\bar{B}})\, dM \geq C_T^2\right)$, which is further bounded by

$$\frac{1}{C_T^2}\mathbb{E}\left|\sup_{g\in\bar{\mathcal{L}}}\int_0^T (g - g_{\bar{B}})\, dM\right| \leq \frac{2}{C_T^2}\mathbb{E}\left|\sup_{g\in\bar{\mathcal{L}}}\int_0^T g dM\right|$$

using Markov inequality and then the triangle inequality because $g_{\bar{B}} \in \bar{\mathcal{L}}$. Write $g = \sum_\theta b_\theta \theta$. Note that

$$\sup_{g\in\bar{\mathcal{L}}}\left|\int_0^T g dM\right| = \sup_{b_\theta, \theta\in\Theta}\left|\int_0^T \left(\sum_\theta b_\theta \theta\right) dM\right| \leq \bar{B}_w \sup_{\theta\in\Theta}\left|\int_0^T \theta dM\right|$$

where the supremum runs over all the b_θ's such that $\sum_\theta |b_\theta| \leq \bar{B}_w$. According to these calculations, to bound (A.9) it is sufficient to bound

$$\frac{2\bar{B}_w}{C_T^2}\mathbb{E}\sup_{\theta\in\Theta}\left|\int_0^T \theta dM\right|. \tag{A.10}$$

Let $\{\Pi_l(\epsilon) : v = 1, 2, .., N_\Pi\}$ be a partition of Θ into $N_\Pi(\epsilon)$ elements such that $\sup_{\theta,\theta'\in\Pi_l(\epsilon)} |\theta - \theta'| \leq \epsilon$. By Condition 2, one can construct such partition with $N_\Pi(\epsilon) \lesssim N(\epsilon, \Theta)$ and such that

$$\sup_{\theta,\theta'\in\Pi_l(\epsilon)} \left|\theta - \theta'\right|_\infty \leq |\theta_{U,l} - \theta_{L,l}|_\infty \tag{A.11}$$

where $[\theta_{L,l}, \theta_{U,l}]$ is an ϵ-bracket for the functions in Π_l, under the uniform norm. It follows that $N_\Pi\left(2\bar{\theta}\right) = 1$ because the diameter of Θ under the uniform norm is bounded by $2\bar{\theta}$. To bound (A.10), use the following maximal inequality from Nishiyama (1998, Theorem 2.2.3), which is specialized to the present framework.

Lemma 5 *Under Conditions 1 and 2,*

$$\mathbb{E}\max_{t\in[0,T]}\max_{\theta\in\Theta}\left|\int_0^t \theta dM\right| \lesssim C_{1,T}\int_0^{2\bar{\theta}} \sqrt{\ln(1 + N_\Pi(\epsilon))} d\epsilon + \frac{C_{2,T}}{\bar{\theta}C_{1,T}} \tag{A.12}$$

for any $C_{2,T} \geq \int_0^T \bar{\theta}^2 d\Lambda$, *and* $C_{1,T} \geq |\Theta|_{\Pi,T}$, *where*

$$|\Theta|_{\Pi,T} := \sup_{\epsilon \in (0,\bar{\theta}]} \max_{l \leq N_\Pi(\epsilon)} \frac{\sqrt{\int_0^T \left(\sup_{\theta,\theta' \in \Pi_l(\epsilon)} |\theta - \theta'|\right)^2 d\Lambda}}{\epsilon}.$$

From the discussion around (A.11) replace $N_\Pi(\epsilon)$ with $N(\epsilon, \Theta)$. The application of Lemma 5 essentially requires to find a bound for $C_{1,T}$ and $C_{2,T}$. Given that $\lambda = d\Lambda/d\mu$ is bounded by $e^{\bar{g}_0}$, from the discussion around (A.11), $|\Theta|_{\Pi,T} \leq \sqrt{e^{\bar{g}_0} T}$ and we set $C_{1T} = C_1 \sqrt{e^{\bar{g}_0} T}$ for some C_1 to be chosen later. Also, deduce that we can choose $C_{2,T} = \bar{\theta} e^{\bar{g}_0} T$. This implies that $C_{2,T}/\bar{\theta} C_{1,T} = \sqrt{e^{\bar{g}_0} T / C_1}$. Hence, the first term on the r.h.s. of (A.12) is of no smaller order of magnitude than the second (i.e., not smaller than a constant multiple of $T^{1/2}$). Thus, in what follows, we can incorporate $C_{2,T}/\bar{\theta} C_{1,T}$ into it without further mention. Hence, an application of Lemma 5 bounds (A.10) by

$$\frac{2\bar{B}_w}{C_T^2} \mathbb{E} \sup_{\theta \in \Theta} \left| \int_0^T \theta dM \right| \lesssim \frac{2\bar{B}_w \sqrt{e^{\bar{g}_0} T}}{C_T^2} \int_0^{2\bar{\theta}} \sqrt{\ln(1 + N(\epsilon, \Theta))} d\epsilon. \tag{A.13}$$

Using the definition of C_T, and choosing $r_T^2 \lesssim \left[e^{3\bar{g}_0} |g_0 - g|_\infty^2 \right]^{-1}$, the above is a constant multiple of

$$r_T^2 \frac{\bar{B}_w e^{\bar{g}_0/2}}{T^{1/2}} \int_0^{2\bar{\theta}} \sqrt{\ln(1 + N(\epsilon, \Theta))} d\epsilon$$

which is required to be $O(1)$, as it is an upper bound for (A.9) . This implies

$$r_T^2 \lesssim \frac{T^{1/2}}{\bar{B}_w e^{\bar{g}_0/2} \int_0^{2\bar{\theta}} \sqrt{\ln(1 + N(\epsilon, \Theta))} d\epsilon}.$$

But, r_T is also required not to go to zero, and in fact it is supposed to diverge to infinity unless the approximation error is nonvanishing. Therefore, the r.h.s. of the above display needs to be bounded away from zero.

To bound the entropy integral, recall that $\Theta = \bigcup_{k=1}^K \Theta_k$. The bracketing number of a union of sets is bounded above by the sum of the bracketing numbers of the individual sets. Hence, $N(\epsilon, \Theta) \leq \sum_{k=1}^K N(\epsilon, \Theta_k)$. Using the inequality $\ln(1 + xy) \leq \ln x + \ln(1 + y)$ for real $x, y \geq 1$, this implies that

$$\begin{aligned} \int_0^{2\bar{\theta}} \sqrt{\ln(1 + N(\epsilon, \Theta_k))} d\epsilon &\leq \int_0^{2\bar{\theta}} \max_{k \leq K} \sqrt{\ln K + \ln(1 + N(\epsilon, \Theta_k))} d\epsilon \\ &\leq 2\bar{\theta} \sqrt{\ln K} + \max_{k \leq K} \int_0^{2\bar{\theta}} \sqrt{\ln(1 + N(\epsilon, \Theta_k))} d\epsilon. \end{aligned}$$

Also, given that $\bar{\theta}$ is bounded and the entropy above is decreasing in ϵ, the above display can be bounded by a multiple of

$$\sqrt{\ln K} + \max_{k \leq K} \int_0^1 \sqrt{\ln\left(1 + N\left(\epsilon, \Theta_k\right)\right)} d\epsilon. \tag{A.14}$$

Also, we can discard the terms that are bounded, i.e., $\bar{g}_0$ and $\bar{\theta}$, but kept so far just to highlight what their contribution might be. Similarly, $\bar{B}_w$ can be replaced by $\bar{B}$ because it enters the bound as a multiplicative constant. These calculations imply that there is a sequence r_T as in the statement in the theorem such that for C large enough,

$$\Pr\left(\frac{r_T^2}{T} d_T^2\left(g_T, g_0\right) > C\right) \leq \frac{1}{C^2}.$$

By the relation between $d_T^2\left(g_T, g_0\right)/T$ and $\left|g_T - g_0\right|_{\lambda,T}^2$ (see (A.2)), the theorem follows.

A.1.4 Proof of Theorem 2

To ease notation, $T = T_n$. We adapt the calculations in the proof of Theorem 2 in Tsybakov (2003). This requires an upper bound for the Kullback-Leibler distance between two intensity densities, and the construction of a suitable subset of $\mathcal{L}(1)$ (using the notation of our theorem). The result in Tsybakov (2003) will then provide the necessary lower bound as stated in our Theorem 2.

To this end, let $N^{(1)}$ and $N^{(2)}$ be point processes with intensities e^{g_1} and e^{g_2} such that $|g_k|_\infty \leq \bar{g}$, $k = 1, 2$. Let the sigma algebra generated by the process $X = (X(t))_{t \geq 0}$ be denoted by $\mathcal{F}^X$. The Kullback-Leibler distance between two intensity densities e^{g_1} and e^{g_2}, restricted to $[0, T]$, and conditioning on $\mathcal{F}^X$ is

$$K\left(g_1, g_2 | \mathcal{F}^X\right) = \mathbb{E}_X \int_0^T \left(g_1 - g_2\right) dN^{(1)} - \int_0^T \left(e^{g_1} - e^{g_2}\right) d\mu$$

where $\mathbb{E}_X$ is the expectation conditional on $\mathcal{F}^X$. The above follows noting that conditioning on $\mathcal{F}^X$, durations are exponentially distributed with intensity density $\exp\{g_1(X(t))\})$. Then,

$$K\left(g_1, g_1 | \mathcal{F}^X\right) = \int_0^T \left(g_1 - g_2\right) e^{g_1} d\mu - \int_0^T \left(e^{g_1} - e^{g_2}\right) d\mu \leq \frac{e^{3\bar{g}}}{2} \int_0^T \left|g_1 - g_2\right|^2 d\mu$$

using (A.5) and the fact that $|g_k|_\infty \leq \bar{g}$, $k = 1, 2$. This provides the necessary upper bound for the Kullback-Leibler distance, to be used in the proof of Theorem 2 in Tsybakov (2003).

Now, follow Bunea et al. (2007, p. 1693) with minor adjustments. For each k, we shall construct a function, say f_k, in Θ_k. Let $A_j = \sum_{i=1}^j 1\{T_i - T_{i-1} \geq a\}$, i.e. the number of durations greater than a amongst the first j durations. Throughout, $1\{\cdot\}$ is the indicator

function. Clearly, $A_n \leq n$ with equality only if $a = 0$. Define

$$f_k(x) = \gamma \sum_{j=1}^{n} \phi_k\left(\frac{A_j}{A_n}\right) \frac{1\{x_k = X_k(T_{j-1})\}\, 1\{T_j - T_{j-1} \geq a\}}{\sqrt{T_j - T_{j-1}}}$$

where $\gamma > 0$ is a constant to be chosen in due course, and $\{\phi_k(s) : k = 1, 2, ..., K\}$ are bounded functions w.r.t. $s \in [0,1]$, and such that $\frac{1}{A_n}\sum_{j=1}^{A_n} \phi_k\left(\frac{j}{A_n}\right)\phi_l\left(\frac{j}{A_n}\right) = \delta_{kl}$, where $\delta_{kl} = 1$ if $k = l$, zero otherwise (e.g., mutatis mutandis, as in Bunea et al., 2007, p. 1693). The functions f_k's are uniformly bounded in absolute value by a constant multiple of $\gamma/\sqrt{a}$. Hence $f_k \in \Theta_k$, for each k, choosing γ small enough. It follows that

$$\begin{aligned}\int_0^T f_k(X(t))\, f_l(X(t))\, dt &= \sum_{j=1}^{n} f_k(X(T_{j-1}))\, f_l(X(T_{j-1}))\,(T_j - T_{j-1}) \\ &= \gamma^2 \sum_{j=1}^{A_n} \phi_k\left(\frac{j}{A_n}\right)\phi_l\left(\frac{j}{A_n}\right) = \gamma^2 A_n \delta_{kl}.\end{aligned}$$

The first step follows because $X(t)$ is predictable and only changes after a jump. The second step follows by the definition of the f_k's because by continuity of the distribution of $X(0)$ and stationarity, $\Pr(X(T_i) = X(T_j)) = 0$ for $i \neq j$. Also, note that unless $\{T_j - T_{j-1} \geq a\}$ is true, the j^{th} term in the definition of f_k will be zero.

Let $\mathcal{C}$ be the subset of $\mathcal{L}(1)$ which consists of arbitrary convex combinations of $m \leq K/6$ of the f_k's with weight $1/m$ so that the weights sum to one. In consequence, for any $g_1, g_2 \in \mathcal{C}$,

$$\int_0^T (g_1 - g_2)^2\, d\mu \asymp A_n \gamma^2 / m.$$

Let $p_a := \Pr(T_j - T_{j-1} \geq a)$. We claim that $\Pr(A_n < np_a/2) \to 0$ exponentially fast. Hence, the r.h.s. of the above display is proportional to $n\gamma^2/m$ with probability going to one. This claim will be verified at the end of the proof.

Now, by suitable choice of small γ, it is possible to follow line by line the argument after eq. (10) in Tsybakov (2003, proof of Theorem 2). This would give us a result for $\int_0^T (g_T - g_0)^2\, d\mu$ rather than $\int_0^T (g_T - g_0)^2\, \lambda d\mu$ and in terms of n rather than $T = T_n$. To replace n with T_n as in the statement of the theorem, note that T_n/n converges almost surely to $(P\lambda)^{-1}$, which is bounded. Finally, $\int_0^T (g_T - g_0)^2\, \lambda d\mu \gtrsim \int_0^T (g_T - g_0)^2\, d\mu$ by the conditions of the theorem.

It remains to show that the claim on A_n holds true. For any positive decreasing function h on the reals, the sets $\{A_n < cn\}$ and $\{h(A_n) > h(cn)\}$ are the same; here $c \in (0,1)$ is a constant to be chosen in due course. Hence, by Markov inequality $\Pr(A_n < cn) \leq$

$\mathbb{E}h\left(n^{-1/2}A_n\right)/h\left(cn^{1/2}\right)$, which implies the following lower bound

$$\Pr\left(A_n \geq cn\right) \geq 1 - \frac{\mathbb{E}h\left(A_n/\sqrt{n}\right)}{h\left(c\sqrt{n}\right)}.$$

It remains to show that the second term on the r.h.s. goes to zero. To this end, let $h\left(s\right) = e^{-ts}$, for some fixed $t > 0$. For p_a as previously defined in the proof, write

$$\frac{A_n}{\sqrt{n}} = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(1\left\{T_i - T_{i-1} \geq a\right\} - p_a\right) + \sqrt{n}p_a.$$

The first term on the r.h.s. is a root-n standardized sum of i.i.d. centered Bernoulli random variables. Hence, it has a moment generating function which is bounded (use the proof of the central limit theorem for Bernoulli random variables). By this remark,

$$\frac{\mathbb{E}h\left(A_n/\sqrt{n}\right)}{h\left(c\sqrt{n}\right)} = \frac{\mathbb{E}\exp\left\{-tA_n/\sqrt{n}\right\}}{\exp\left\{-tc\sqrt{n}\right\}} \lesssim e^{-t(p_a - c)\sqrt{n}}.$$

Choose $c = p_a/2$ to see that the r.h.s. goes to zero exponentially fast for any $t > 0$, as previously claimed.

A.1.5 Proof of Lemma 1 and Corollaries

Proof. [Lemma 1] The proof is a minor re-adaptation of Lemma 4 in Sancetta (2015). Note that if $B \geq B_0$, the lemma is clearly true because in this case, $\mathcal{L}_0 \subseteq \mathcal{L} := \mathcal{L}\left(B, \Theta, \mathcal{W}\right)$. Hence, assume $B < B_0$ and w.n.l.g. $B = \rho B_0$ for $\rho \in (0,1)$. Write

$$g_0 = \sum_{\theta\in\Theta} b_\theta\theta = \sum_{\theta\in\Theta}\lambda_\theta\bar{b}\theta$$

where the λ_θ's are nonnegative and add to one, and $\bar{b} = \sum_{\theta\in\Theta}|b_\theta|$. Note that the constraint $\sum_{\theta\in\Theta} w_\theta\left|b_\theta\right| \leq B_0$ for functions in $\mathcal{L}_0$ implies $\bar{b} \leq B_0/\underline{w}$. Define $g'\left(x\right) = \rho g_0\left(x\right)$ for ρ such that $B = \rho B_0$ so that $g' \in \mathcal{L}$. Using this choice of g', by standard inequalities,

$$\left|g_0 - g'\right|_r \leq \left|\sum_{\theta\in\Theta}\lambda_\theta\bar{b}\theta - \sum_{\theta\in\Theta}\lambda_\theta\rho\bar{b}\theta\right|_r \leq \left|\bar{b}\left(1-\rho\right)\right|\sum_{\theta\in\Theta}\lambda_\theta\left|\theta\right|_r \leq \bar{b}\left(1-\rho\right)\max_{\theta\in\Theta}\left|\theta\right|_r \leq \frac{\bar{\theta}_r}{\underline{w}}\left(B_0 - B\right)$$

using the definition of ρ. This proves the result, because for g' above, $\inf_{g\in\mathcal{L}}\left|g_0 - g\right|_r \leq \left|g_0 - g'\right|_r$. ■

Proof. [Corollary 2] We need to show that $L_T\left(\tilde{g}_T, g_B\right) \geq -\left(C_T^2/2\right)$ with C_T as in the proof of Theorem 1 and r_T as in (9), e.g., $C_T^2 \gtrsim \bar{B}\sqrt{T\ln K}$. To this end, recall that $\tilde{L}_T\left(g\right) = \int_0^T g\left(\tilde{X}\left(t\right)\right)dN\left(t\right) - \int_0^T \exp\left\{g\left(\tilde{X}\left(t\right)\right)\right\}dt$, which is the log-likelihood when we use $\tilde{X}$ instead

of X. Note that the counting process N is still the same whether we use X or $\tilde{X}$, as jumps are observable. By definition, $\tilde{g}$ is the approximate maximizer of $\tilde{L}_T(g)$, but not necessarily the maximizer of $L_T(g)$. It would be enough to show that $L_T(\tilde{g}_T, g_B) \gtrsim -C_T^2$ in probability, as by a re-definition of the constant in C_T, the proof in Theorem 1 would go through. Given these remarks, write

$$L_T(\tilde{g}_T, g_B) \geq \tilde{L}_T(\tilde{g}_T, g_B) - \left| L_T(\tilde{g}_T, g_B) - \tilde{L}_T(\tilde{g}_T, g_B) \right|.$$

Using (11) we have that $\tilde{L}_T(\tilde{g}_T, g_B) \gtrsim -C_T^2$ as in (A.8). To bound the second term on the r.h.s. of the above display, it is sufficient to bound a constant multiple of

$$\begin{aligned}
& \sup_{g \in \bar{\mathcal{L}}} \left| L_T(g) - \tilde{L}_T(g) \right| \\
= & \sup_{g \in \bar{\mathcal{L}}} \left| \int_0^T \left[g(X(t)) - g\left(\tilde{X}(t)\right) \right] dN(t) - \int_0^T \left[\exp\{g(X(t))\} - \exp\left\{ g\left(\tilde{X}(t)\right) \right\} \right] dt \right| \\
\leq & \sup_{g \in \bar{\mathcal{L}}} \left| \int_0^T \left[g(X(t)) - g\left(\tilde{X}(t)\right) \right] dN(t) \right| + \sup_{g \in \bar{\mathcal{L}}} \left| \int_0^T \left[\exp\{g(X(t))\} - \exp\left\{ g\left(\tilde{X}(t)\right) \right\} \right] dt \right| \\
=: & \; I + II.
\end{aligned}$$

First, find a bound for II. By the mean value theorem in Banach spaces,

$$II \leq \sup_{g \in \bar{\mathcal{L}}} e^{\bar{g}} \int_0^T \left| g(X(t)) - g\left(\tilde{X}(t)\right) \right| dt. \tag{A.15}$$

Now,

$$\begin{aligned}
\sup_{g \in \bar{\mathcal{L}}} \int_0^T \left| g(X(t)) - g\left(\tilde{X}(t)\right) \right| dt & \leq \sup_{\left\{ b_\theta : \sum_{\theta \in \Theta} |b_\theta| \leq \bar{B}_w \right\}} \int_0^T \sum_{\theta \in \Theta} |b_\theta| \left| \theta\left(\tilde{X}(t)\right) - \theta(X(t)) \right| dt \\
& \leq \bar{B}_w \max_{\theta \in \Theta} \int_0^T \left| \theta\left(\tilde{X}(t)\right) - \theta(X(t)) \right| dt
\end{aligned}$$

because the supremum over the simplex is achieved at one of its edges. By the conditions of the corollary, the above display is $O_p\left(\bar{B} e^{-\bar{g}} \sqrt{T \ln K}\right)$. Hence, deduce that (A.15) is $O_p\left(\bar{B}\sqrt{T \ln K}\right) = O_p(C_T)$ (recall the notation in (A.1)).

It remains to bound I. Adding and subtracting $\int_0^T \left[g(X(t)) - g\left(\tilde{X}(t)\right) \right] d\Lambda(t)$, and using the triangle inequality,

$$I \leq \sup_{g \in \bar{\mathcal{L}}} \left| \int_0^T \left[g(X(t)) - g\left(\tilde{X}(t)\right) \right] dM(t) \right| + \sup_{g \in \bar{\mathcal{L}}} \left| \int_0^T \left[g(X(t)) - g\left(\tilde{X}(t)\right) \right] d\Lambda(t) \right|.$$

The first term in the above display can be incorporated in the l.h.s. of (A.7) and bounded as

in the proof of Theorem 1. To bound the second term on the above display by definition of $d\Lambda$,

$$\sup_{g\in\bar{\mathcal{L}}}\left|\int_0^T\left[g\left(X\left(t\right)\right)-g\left(\tilde{X}\left(t\right)\right)\right]\exp\left\{g_0\left(X\left(t\right)\right)\right\}dt\right|\leq\sup_{g\in\bar{\mathcal{L}}}e^{\bar{g}_0}\int_0^T\left|g\left(X\left(t\right)\right)-g\left(\tilde{X}\left(t\right)\right)\right|dt.$$

From the derived bound for II deduce that the r.h.s. is $O_p\left(C_T^2\right)$. This completes the proof of the first statement in the corollary, as all the conditions of Theorem 1 are satisfied. To show the last statement of the corollary, use the inequality $\left|g\left(X\left(t\right)\right)-g\left(\tilde{X}\left(t\right)\right)\right|^2\leq 2\bar{g}\left|g\left(X\left(t\right)\right)-g\left(\tilde{X}\left(t\right)\right)\right|$ together with a trivial modification of the previous display. ■

Proof. [Corollary 4] The approximation error is zero by assumption. Given that Θ_k has one single element, the entropy integral is trivially finite. Hence, (9) simplifies as in the statement of the corollary. ■

Proof. [Corollary 5] Define the set

$$\mathcal{B}:=\left\{\sup_{t>0}\left|\int_0^t\left(t-s\right)e^{-\underline{a}(t-s)}dN\left(s\right)\right|\leq\beta\right\}$$

for some $\beta<\infty$. In the proof of Theorem 1 write

$$\Pr\left(d_T\left(g_T,g_0\right)>C_T,\right)\leq\Pr\left(d_T\left(g_T,g_0\right)>C_T,\text{ and }\mathcal{B}\right)+\Pr\left(\mathcal{B}^c\right)$$

where $\mathcal{B}^c$ is the complement of $\mathcal{B}$. We shall apply Corollary 2 to the first term on the r.h.s., and then show that the last term in the above display is negligible.

At first, show that the process with intensity density $\lambda\left(t\right)=\exp\left\{f_{a_0}\left(t\right)+g_0\left(X\left(t\right)\right)\right\}$ is stationary. To this end, we apply Theorem 2 in Brémaud and Massoulié (1996). Using their notation, their nonlinear function $\phi\left(\cdot\right)$ in their eq.(1) is here defined as $\exp\left\{f\left(\cdot\right)\right\}\exp\left\{g_0\left(X\left(t\right)\right)\right\}$, which is random, unlike their case. However, in the proof of their Theorem 2 they only use the fact that $\left|\phi\left(y\right)-\phi\left(y'\right)\right|\leq\alpha\left|y-y'\right|$ for some finite constant α (see their eq.(23) and first display on p.1580). This is the case here as well. To see this, recall the definition of f (see Section 3.6.2) which is bounded and Lipschitz. Then,

$$\left|\exp\left\{f\left(y\right)\right\}\exp\left\{g_0\left(X\left(t\right)\right)\right\}-\exp\left\{f\left(y'\right)\right\}\exp\left\{g_0\left(X\left(t\right)\right)\right\}\right|\leq\exp\left\{\bar{g}_0\right\}\left|f\left(y\right)-f\left(y'\right)\right|$$

(recall $\bar{g}_0$ is the uniform norm of g_0). We also need to note that $\exp\left\{g_0\left(X\left(t\right)\right)\right\}$ is stationary, bounded and predictable. This ensures that the intensity $\lambda\left(t\right)$ is bounded and predictable, which is required in the lemmas used in Brémaud and Massoulié (1996). Hence Condition 1 is satisfied.

To verify Condition 2, we verify that the entropy integral of the process $\tilde{f}_a$ is finite in a sense to be made clear below. We shall postpone this to the end of the proof.

Hence, mutatis mutandis, we now verify (10) in Corollary 2. To this end, we bound $c_T := \mathbb{E}\max_{a\in[\underline{a},\bar{a}]}\int_0^T \left|f_a(t) - \tilde{f}_a(t)\right| dt$. Corollary 2 requires c_T to be $O\left(e^{-\bar{B}_w\bar{\theta}}\sqrt{T\ln K}\right)$. By the Lipschitz condition and $a \in [\underline{a},\bar{a}]$,

$$\int_0^T \left|f_a(t) - \tilde{f}_a(t)\right| dt \lesssim \int_0^T e^{-\underline{a}t}\left(\int_{(-\infty,0)} e^{\underline{a}s} dN(s)\right) dt.$$

Using the fact that Λ is the compensator of N, and that Λ has bounded density $\exp\{f_{a_0}(t) + g_0(X(t))\}$, deduce that

$$\begin{aligned}\mathbb{E}\max_{a\in[\underline{a},\bar{a}]}\int_0^T \left|f_a(t) - \tilde{f}_a(t)\right| dt &\leq \mathbb{E}\left[\left(\int_{(-\infty,0)} e^{\underline{a}s} dN(s)\right)\left(\int_0^T e^{-\underline{a}t} dt\right)\right] \\ &\lesssim \frac{1}{\underline{a}}\mathbb{E}\int_{(-\infty,0)} e^{\underline{a}s} d\Lambda(s) \lesssim \frac{1}{\underline{a}^2} < \infty.\end{aligned}$$

This verifies (10) in Corollary 2.

To verify Condition 2 for $\tilde{f}_a$, we need an estimate of the entropy integral for the family of stochastic processes $\mathcal{A} := \left\{\left(\tilde{f}_a(t)\right)_{t\geq 0} : a \in [\underline{a},\bar{a}]\right\}$. This means that we need to bound

$$\begin{aligned}\sup_{t>0}\left|\tilde{f}_a(t) - \tilde{f}_{a'}(t)\right| &\lesssim \sup_{t>0}\left|\int_0^t \left(e^{-a(t-s)} - e^{-a'(t-s)}\right) dN(s)\right| \\ &\leq \sup_{t>0}\left|\int_0^t (t-s)\, e^{-\underline{a}(t-s)} dN(s)\right| dt\left|a - a'\right|\end{aligned}$$

using a first order Taylor expansion, and the lower bound on a, a'. On $\mathcal{B}$, the above is $\beta|a - a'|$. It is then easy to see that the entropy integral is a constant multiple of $\beta^{1/2}$ because the uniform ϵ-bracketing number of $[\underline{a}\beta, \bar{a}\beta]$ has size $\beta(\bar{a} - \underline{a})/\epsilon$.

In consequence, we can apply Corollary 2. Let $\beta = O(\ln T)$). There is no approximation error, so that r_T^{-2} (r_T as in (9)) becomes as in (14). The term $\sqrt{\ln T}$, in the numerator of (14), is proportional to the entropy integral of $\mathcal{A}$.

To conclude, we show that $\mathcal{B}^c$, the complement of $\mathcal{B}$, is such that $\Pr(\mathcal{B}^c) \to 0$ as $\beta \to \infty$. By Markov inequality,

$$\Pr(\mathcal{B}^c) \leq \frac{\mathbb{E}\sup_{t>0}\left|\int_0^t (t-s)\, e^{-\underline{a}(t-s)} dN(s)\right|}{\beta}.$$

Recalling that $M = N - \Lambda$, by the triangle inequality, the numerator on the r.h.s. can be bounded by

$$\mathbb{E}\sup_{t>0}\left|\int_0^t (t-s)\, e^{-\underline{a}(t-s)} dM(s)\right| + \mathbb{E}\sup_{t>0}\left|\int_0^t (t-s)\, e^{-\underline{a}(t-s)} d\Lambda(s)\right| =: I + II.$$

The first integral inside the square is a bounded predictable function w.r.t. a martingale, and is a martingale. By the Burkholder-Davis-Gundy inequality,

$$I^2 \lesssim \sup_{t>0} \mathbb{E} \int_0^t \left| (t-s)\, e^{-\underline{a}(t-s)} \right|^2 d\Lambda(s) \leq e^{\bar{g}_0} \sup_{t>0} \int_0^t \left| (t-s)\, e^{-\underline{a}(t-s)} \right|^2 ds = O(1).$$

By a similar argument $II = O(1)$. These bounds imply that $\Pr(\mathcal{B}^c) \to 0$. The last statement in the corollary is deduced from the proof of Corollary 4. ■

Proof. [Corollary 6] By Lemma 1, the approximation error will be zero as soon as $\bar{B} \geq B_0$, which will be eventually the case as $\bar{B} \to \infty$ and B_0 is finite. By the remarks in Section 3.6.3 the entropy integral is finite. Hence, the bound follows from (9). ■

Proof. [Corollary 7] By Lemma 2 and (13) the approximation error is a constant multiple of $V^{-2\alpha} + \max\left\{c_\alpha - \bar{B}, 0\right\}^2$. The univariate square uniform approximation rate $V^{-2\alpha}$ follows by the remarks in Section 3.6.4. Given that there are V elements in each Θ_k the entropy integral is a constant multiple of $\sqrt{\ln(1+V)}$. Inserting in (9), the bound is deduced as long as $V > 1$. In particular for $V \gtrsim (T/\ln T)^{1/(4\alpha)}$ the bound simplifies further. ■

Proof. [Corollary 8] The proof is the same as for Corollary 7. ■

Proof. [Corollary 9] As stated in Section 3.6.6, the approximation rate of Bernstein polynomials under the squared uniform loss is a constant multiple of $\alpha^2 V^{-1}$. Hence, by Lemma 2 and (13), the approximation error is a constant multiple of $\alpha^2 V^{-1} + \max\left\{B_0 - \bar{B}, 0\right\}^2$. In consequence, as $\bar{B} \to \infty$, the approximation error is eventually $O\left(\sqrt{\alpha/T}\right)$ when $V \gtrsim T^{1/2}\alpha^{3/2}$. By the remarks in Section 3.6.6, the entropy integral is $\alpha^{1/2}$. Inserting in (9) the bound follows. ■

A.1.6 Proof of Theorem 3

Define $h := b\theta$, and let $t \in [0,1]$. Let

$$h_m := \arg\sup_{h \in \bar{\mathcal{L}}} D_T(F_{m-1}, h - F_{m-1}).$$

By linearity, the maximum is obtained by a function $h = b\theta$ with $\theta \in \Theta_k$ for some k and $|b| \leq \bar{B}$. Hence, it is sufficient to maximize the absolute value of D_T w.r.t. θ as the coefficient b is not constrained in sign. Define,

$$G(F_{m-1}) := D_T(F_{m-1}, h_m - F_{m-1}),$$

so that for any $g \in \bar{\mathcal{L}}$,

$$L_T(g) - L_T(F_{m-1}) \leq G(F_{m-1}) \tag{A.16}$$

by concavity. For $m \geq 0$, define $\bar{\rho}_m = 2/(m+2)$. By concavity, again,

$$L_T(F_m) = \max_{\rho \in [0,1]} L_T(F_{m-1} + \rho(h - F_{m-1})) \geq L_T(F_{m-1}) + D_T(F_{m-1}, h - F_{m-1})\bar{\rho}_m + \frac{\bar{C}}{2}\bar{\rho}_m^2$$

where

$$\bar{C} := \min_{h,g \in \bar{\mathcal{L}}, t \in [0,1]} \frac{2}{t^2}\left[L_T(g + t(h-g)) - L_T(g) - D_T(g, t(h-g))\right] < 0.$$

The above two displays together with (A.16), imply

$$\begin{aligned} L_T(F_m) - L_T(g) &\geq L_T(F_{m-1}) - L_T(g) + \bar{\rho}_m G(F_{m-1}) + \frac{\bar{C}}{2}\bar{\rho}_m^2 \\ &\geq L_T(F_{m-1}) - L_T(g) + \bar{\rho}_m(L_T(g) - L_T(F_{m-1})) + \frac{\bar{C}}{2}\bar{\rho}_m^2 \\ &= (1 - \bar{\rho}_m)(L_T(F_{m-1}) - L_T(g)) + \frac{\bar{C}}{2}\bar{\rho}_m^2 \\ &\geq \frac{2\bar{C}}{m+2} \end{aligned} \tag{A.17}$$

for the given choice of $\bar{\rho}_m$ (mutatis mutandis, as in the proof of Theorem 1 in Jaggi (2013)). It remains to bound $\bar{C}$. By Taylor's expansion in Banach spaces,

$$L_T(g + t(h-g)) = L_T(g) + D_T(g, t(h-g)) + \frac{1}{2}H_T\left(g_*, t^2(h-g)^2\right),$$

for $g_* = t_* g + (1 - t_*)h$, and some $t_* \in [0,1]$, where

$$H_T\left(g, t^2(h-g)\right) = -\int_0^T t^2(h-g)^2 e^g ds.$$

This means that

$$\bar{C} \geq \min_{h,g \in \bar{\mathcal{L}}, t \in [0,1]} \frac{2}{t^2}\left[-\frac{1}{2}\int_0^T t^2(h(X(s)) - g(X(s)))^2 e^{\bar{g}} ds\right] \geq -4Te^{\bar{g}}\bar{g}^2 \geq -4Te^{\bar{B}\bar{\theta}/\underline{w}}\left(\bar{B}\bar{\theta}/\underline{w}\right)^2$$

using (A.1). Substituting in (A.17) gives the result.

A.1.7 Proof of Proposition 1

Let $M := N - \Lambda$ and $h_t := g_t - g_t'$. To ease notation, suppose for the moment that S is an integer. Then, under the conditions of the proposition (the null hypothesis),

$$L_S(g, g') = \sum_{s=1}^{S} \int_{s-1}^{s} h_t(X(t))\, dM(t) = \sum_{s=1}^{S} Y_s.$$

Then, $\{Y_s : s = 1, 2, ...\}$ is a sequence of martingale differences. This follows from the law of iterated expectations and the fact that h_t is a predictable process. Denote the expectation conditioning on $\{Y_i : i \leq s\}$ by $\mathbb{E}_s$. The result will follow by an application of Theorem 2.3 in McLeish (1974). To this end, it is sufficient to show that (i.) $\mathbb{E}\left|\frac{1}{S}\sum_{s=1}^{S} Y_s^2\right| \to \sigma^2$, (ii.) $\lim_{S\to\infty} \mathbb{E}\max_{s\leq S} Y_s^2/S < \infty$ and (iii.) $\max_{s\leq S}\left|Y_s/\sqrt{S}\right| \to 0$ in probability. Note that

$$\mathbb{E}\left|\frac{1}{S}\sum_{s=1}^{S} Y_s^2\right| = \mathbb{E}\left|\frac{1}{S}\sum_{s=1}^{S} \mathbb{E}_{s-1} Y_s^2\right| \tag{A.18}$$

using iterated expectations and the fact that the elements in the sum are positive. Note that

$$\begin{aligned}
\mathbb{E}_{s-1} Y_s^2 &= \mathbb{E}_{s-1}\left[\int_{s-1}^{s} h_t\left(X\left(t\right)\right) dM\left(t\right)\right]^2 \\
&= \mathbb{E}_{s-1}\left[\int_{s-1}^{s} h_t^2\left(X\left(t\right)\right) d\Lambda\left(t\right)\right]
\end{aligned}$$

(e.g., Ogata, 1978, e.q. 2.1). Hence,

$$\frac{1}{S}\sum_{s=1}^{S} \mathbb{E}_{s-1} Y_s^2 = \left[\frac{1}{S}\sum_{s=1}^{S} \mathbb{E}_{s-1}\int_{s-1}^{s} h_t^2\left(X\left(t\right)\right) d\Lambda\left(t\right)\right].$$

By these remarks, (A.18) is equal to

$$\begin{aligned}
\mathbb{E}\left|\frac{1}{S}\sum_{s=1}^{S} \mathbb{E}_{s-1}\int_{s-1}^{s} h_t^2\left(X\left(t\right)\right) d\Lambda\left(t\right)\right| &= \frac{1}{S}\sum_{s=1}^{S} \mathbb{E}\int_{s-1}^{s} h_t^2\left(X\left(t\right)\right) d\Lambda\left(t\right) \\
&= \mathbb{E}\frac{1}{S}\int_{0}^{S} h_t^2\left(X\left(t\right)\right) d\Lambda\left(t\right),
\end{aligned}$$

using the fact that the terms in the sum are positive. By the conditions of the proposition

$$\sigma_S^2 := \frac{1}{S}\int_{0}^{S} h_t^2\left(X\left(t\right)\right) d\Lambda\left(t\right) \to \sigma^2 > 0$$

in probability. The sequence $\left(\sigma_S^2\right)_{S\geq 1}$ is uniformly bounded. In consequence, convergence in probability implies convergence in L_1, i.e. $\mathbb{E}\sigma_S^2 \to \sigma^2$. This verifies the first condition (i.). Now,

$$\mathbb{E}\max_{s\leq S}\frac{Y_s^2}{S} \leq \frac{1}{S}\mathbb{E}\sum_{s=1}^{S} Y_s^2$$

bounding the maximum by the sum. By the previous calculations deduce that the above is bounded, which then verifies the second condition (ii.). Finally,

$$\begin{aligned}\max_{s\leq S}\left|Y_s\right|/\sqrt{S} &= \frac{1}{\sqrt{S}}\max_{s\leq S}\left|\int_{s-1}^{s} h_t\left(X\left(t\right)\right)dM\left(t\right)\right| \\ &\lesssim \frac{1}{\sqrt{S}}\max_{s\leq S}\left|\int_{s-1}^{s} dN\left(t\right)\right| + \frac{1}{\sqrt{S}}\max_{s\leq S}\left|\int_{s-1}^{s} d\Lambda\left(t\right)\right| \\ &= \frac{1}{\sqrt{S}}\max_{s\leq S}\left[N\left(s\right)-N\left(s-1\right)\right] + \frac{1}{\sqrt{S}}\max_{s\leq S}\Lambda\left(\left[s-1,s\right]\right)\end{aligned}$$

where the inequality uses the fact that h_t is bounded. The last term on the r.h.s. is $O_p\left(S^{-1/2}\right)$. A counting process N is increasing with the intensity. Since $\lambda\left(X\left(s\right)\right) \leq e^{\bar{g}_0}$ uniformly in s, there is a counting process N' with intensity density $e^{\bar{g}_0}$ such $\Pr\left(N\left(s\right)>n\right) \leq \Pr\left(N'\left(s\right)>n\right)$. In consequence, for any s, $\mathbb{E}\left[N\left(s\right)-N\left(s-1\right)\right]^4 \leq \mathbb{E}\left[N'\left(s\right)-N'\left(s-1\right)\right]^4 \leq C$ for some absolute constant C that depends on $\bar{g}_0$ only. The last inequality follows because N' is Poisson with intensity $e^{\bar{g}_0}$. By these remarks,

$$\begin{aligned}\mathbb{E}\frac{1}{\sqrt{S}}\max_{s\leq S}\left[N\left(s\right)-N\left(s-1\right)\right] &\leq \frac{1}{\sqrt{S}}\left(\mathbb{E}\max_{s\leq S}\left|N\left(s\right)-N\left(s-1\right)\right|^4\right)^{1/4} \\ &\leq \frac{1}{\sqrt{S}}\left(\sum_{s=1}^{S}\mathbb{E}\left|N\left(s\right)-N\left(s-1\right)\right|^4\right)^{1/4}\end{aligned}$$

bounding the maximum by the sum. Deduce that the above is $\left(C/S\right)^{1/4} = o\left(1\right)$. This verifies the third condition (iii.) required for the application of Theorem 2.3 in McLeish (1974).

If S is not an integer, write $\lfloor S\rfloor$ for its integer part. Then,

$$\frac{1}{\sqrt{S}}L_S\left(g,g'\right) = \left(\frac{\lfloor S\rfloor}{S}\right)^{1/2}\frac{1}{\sqrt{\lfloor S\rfloor}}\sum_{s=1}^{\lfloor S\rfloor}Y_s + \frac{1}{\sqrt{S}}\int_{\lfloor S\rfloor}^{S} h_t\left(X\left(t\right)\right)dM\left(t\right).$$

Clearly, $\lfloor S\rfloor/S \to 1$. Moreover, by arguments similar to the ones used to verify the third condition (iii.) above, we deduce that the last term on the r.h.s. is $o_p\left(1\right)$. This shows the result using σ_S as scaling sequence rather than $\hat{\sigma}_S$. However, $\left|\hat{\sigma}_S^2-\sigma_S^2\right| = \left|\frac{1}{S}\int_0^S h_t^2\left(X\left(t\right)\right)dM\left(t\right)\right| \to 0$ a.s., and we can use $\hat{\sigma}_S^2$ to define the t-statistic. This completes the proof.

A.2 Details Regarding Section 5.3.1

Define $Y_i := \exp\{g_0(X(T_i))\}$ and $Z_i := \sum_{T_j \leq T_i} e^{-a_0(T_i - T_j)}$, and recall $R(T_{i+1}) = T_{i+1} - T_i$. Note that for $t \in (T_i, T_{i+1}]$, $\lambda(t) = \left(c_0 + Z_i e^{-a_0(t-T_i)}\right) Y_i$. In consequence,

$$\Lambda((T_i, T_{i+1}]) = \int_{T_i}^{T_{i+1}} \lambda(t)\, dt = \left[c_0 R(T_{i+1}) + \frac{Z_i}{a_0}\left(1 - e^{-a_0 R(T_{i+1})}\right)\right] Y_i$$

is exponentially distributed with mean one, conditioning on $\mathcal{F}_i := (T_i, Z_i, Y_i)$. Moreover, $Z_i = Z_{i-1} e^{-a_0(T_i - T_{i-1})} + 1$ with $Z_0 = 1$. Hence, define $c_1 = c_0 Y_i$, $c_2 = Y_i Z_i$, and simulate i.i.d. $[0,1]$ uniform random variables U_i's. We simulate $R(T_i)$ setting it equal to the s that solves $c_1 s + \frac{c_2}{a_0}(1 - e^{-a_0 s}) = -\ln U_i$. Given an initial guess $(2, 1.5)$ of of the true $(c_0, a_0) = (2, 1.3)$ we estimate $\exp\{g_T(X(t))\}$. Given $\exp\{g_T(X(t))\}$ we estimate c and a in $\left(c + \sum_{T_i < t} e^{-a(t-T_i)}\right) \exp\{g_T(X(t))\}$. We perform a second iteration.

Estimation of g is done using the algorithm in Section 3.5. In this case, the relevant part of the likelihood is

$$\sum_{i=1}^{n} g(T_{i-1}) - \sum_{i=1}^{n} \exp\{g(T_{i-1})\}\, \Delta_i$$

where

$$\Delta_i = cR(T_i) + \frac{Z_{i-1}}{a}\left(1 - e^{-aR(T_i)}\right)$$

and c and a are set to their guess/estimated values. Estimation of c and a is via maximum likelihood given $\exp\{g_T(X(t))\}$.

References

[1] Brémaud, P. and L. Massoulié (1996) Stability of Nonlinear Hawkes Processes. Annals of Probability 24, 1563-1588.

[2] Bunea, F., A. Tsybakov and M. Wegkamp (2007) Aggregation for Gaussian Regression. Annals of Statistics 35, 1674-1697.

[3] McLeish, D.L. (1974) Dependent Central Limit Theorems and Invariance Principles. Annals of Probability 2, 620-628.

[4] Ogata, Y. (1978) The Asymptotic Behaviour of the Maximum Likelihood Estimator for Stationary Point Processes. Annal of the Institute of Statistical Mathematics 30 (A), 243-261.

[5] Nishiyama, Y. (1998) Entropy Methods for Martingales. Ph.D. thesis, University of Utrecht. URL: http://oai.cwi.nl/oai/asset/14076/14076D.pdf.

[6] Sancetta A. (2015) A Nonparametric Estimator for the Covariance Function of Functional Data. Econometric Theory 31, 1359-1381.

[7] Tsybakov, A.B. (2003) Optimal Rates of Aggregation. Proceedings of COLT-2003, Lecture Notes in Artificial Intelligence, 303-313.