Estimation of a High Dimensional Counting Process Without Penalty for High Frequency Events*

Luca Mucciante† and Alessio Sancetta‡

May 18, 2022

Abstract

This paper introduces a counting process for event arrivals in high frequency trading, based on high dimensional covariates. The novelty is that, under sparsity conditions on the true model, we do not need to impose any model penalty or parameters shrinkage, unlike Lasso. The procedure allows us to derive a central limit theorem to test restrictions in a two stage estimator. We achieve this by the use of a sign constraint on the intensity which necessarily needs to be positive. In particular we introduce an additive model to extract the nonlinear impact of order book variables on buy and sell trade arrivals. In the empirical application, we show that the shape and dynamics of the order book are fundamental in determining the arrival of buy and sell trades in the crude oil futures market. We establish our empirical results mapping the covariates into a higher dimensional

*We are very grateful to the Editor Peter Phillips, the Co-Editor Eric Renault, the Associate Editor, and the Referees for their detailed comments that have led to substantial improvements both in content and presentation.

†Department of Economics, Royal Holloway University of London, Egham TW20 0EX, UK. Email: lucahost@gmail.com

‡Corresponding Author. Department of Economics, Royal Holloway University of London, Egham TW20 0EX, UK. Email: asancetta@gmail.com

space. Consistently with the theoretical results, the estimated models are sparse in the number of parameters. Using this approach, we are also able to compare competing model hypotheses on the basis of an out of sample likelihood ratio type of test.

Key Words: High frequency trading, high dimensional model, Lasso, order book.

JEL Codes: C14, G10

1 Introduction

Counting processes are continuous time stochastic processes with nondecreasing trajectories, taking values in the set of positive integers. This paper is concerned with the estimation of the intensity of a counting process that depends on high dimensional covariates. In particular we are interested in modelling the intensity of high frequency trading events using a possibly large number of covariates, where the impact of each of them can be nonlinear. This further increases the dimension when the unknown nonlinearity is modelled by a series expansion or similar procedures. The motivation of this study is to use information from the order book to model the intensity of buy and sell arrivals. The importance of order book variables has already been shown by various authors with varying degrees of complexity (Hall and Hautsch 2007, Cont et al., 2014, Kercheval and Zhang, 2015, Sancetta, 2018). The intensity of the counting process can be used to predict buy or sell trade arrivals at infinitesimal time scales, using relevant microstructure variables such as order book volume imbalances, order flow, and spread as covariates.

The use of counting processes in high frequency financial modelling was pioneered by Engle and Russell (1998). Since then they have acquired an increasing popularity in the literature (Bauwens and Hautsch 2009, for a survey). A possible way to char-

acterize a counting process is via its intensity. Intuitively speaking the intensity is the instantaneous rate of occurrence of events conditional on the past history.

For definiteness, let $N := (N(t))_{t\geq 0}$ be a counting process starting at zero, and $\lambda^* := (\lambda^*(t))_{t\geq 0}$ a predictable process, both adapted to a filtration $\mathcal{F} = (\mathcal{F}_t)_{t\geq 0}$ and such that $M := (M(t))_{t\geq 0}$, where $M(t) := \left(N(t) - \int_0^t \lambda^*(r)\, dr\right)$, is an $\mathcal{F}_t$-martingale. The process λ^* is an $\mathcal{F}_t$-intensity of N. We assume that there is a predictable process $X = (X(t))_{t\geq 0}$ that takes values in $[0,1]^K$ where K is large relatively to the sample period and

$$\lambda^*(t) = X(t)' b^* \tag{1}$$

where $b^* \in [0,\infty)^K$ is an unknown sparse parameter. All vector valued quantities are column vectors and the prime symbol $'$ stands for transpose. By sparse, we mean that b^* has a small number of nonzero coefficients, relatively to K. We assume that the covariates are positive stochastic processes. Together with the assumption that the true coefficient vector b^* is nonnegative, this ensures that the intensity is a positive stochastic process. We shall show that the nonnegativity restriction naturally arises in some parametrization. Then, this positivity constraint inherits a regularization property similar to Lasso. The goal of the paper is to find an estimator for b^* when we observe a single trajectory/sample from $(N(t), X(t))_{t\geq 0}$ over a time interval $[0,T]$ with $T \to \infty$.

The empirical application in Section 3 further motivates the model, and Section 3.2 provides further discussion on the scope and limitations of the modelling strategy.

1.1 Remarks on the Model Restrictions

If the covariates are bounded, the restriction to $[0,1]^K$ is a mere linear transformation. Most covariates obtained from the order book satisfy this condition by construction. Examples include order book volume imbalances (see (10) in Section 3.1). On the

other hand, variables such as durations are not bounded. However, we can always find transformations of the data to map variables into a bounded range. The type of transformation is a function of the modelling objectives. We provide more concrete remarks in Section 3.2.

The assumptions that we shall use essentially imply that the covariates process X is ergodic. We do not assume any stationarity. This is the weakest possible assumption for econometric inference. While we do not do so in the empirical application, one example of the flexibility of the framework is to consider nonlinear Hawkes processes as one of the covariates, as long as they satisfy some stability conditions (Brémaud and Massoulié, 1996). In this case, one covariate could be set equal to $\varphi\left(\int_{(0,t)} h\left(t-s\right) dN\left(s\right)\right)$ where $\varphi\left(\cdot\right)$ is a Lipschitz function with range in $[0,1]$, and $h\left(\cdot\right)$ is positive and integrable. If h is unknown, we could suppose a finite set of such functions $\left\{h^{(l)} : l = 1, 2, ..., L\right\}$ and generate as many covariates as L to capture any self exiting nature of the intensity. Alternatively, the kernel function h would have to be estimated using a sample antecedent to the one used in the estimation of the model.

1.2 Contributions and Relation to Other Work

This paper contributes to the characterization of the impact of order book variables in the intensity of buy and sell trade arrivals. In order to do so in the most robust and simple to interpret way, it relies on a methodology that is elementary from an econometric point of view, but powerful.

The literature on modelling the arrival of high frequency events using order book variables is growing (inter alia, Hall and Hautsch 2007, Cont et al., 2014, Sancetta, 2018, Morariu-Patrichi and Pakkanen, 2018). This is an important practical problem. Algorithmic traders do track the order book (MacKenzie, 2017). The order book not only contains information about liquidity, but also helps to identify informed traders.

Nowadays, unlike traditional models of informed trading (Glosten and Milgrom, 1985, Kyle, 1985), sophisticated informed investors prefer to rely on passive execution. This means that they would avoid buying at the ask and paying the spread. Instead they would break their orders into small ones and patiently fill most of them joining the bid price. There are two reasons for this. First, informed traders want to reduce their cost. Second, they want to reduce the amount of signaling. It turns out that uninformed market participants are smart and do look at quotes to infer information beyond liquidity. In fact, our empirical results show that the intensity of buy arrivals is a nonlinear increasing function of quoted buy orders over quoted sell sizes. The empirical application focuses on crude oil futures as main instrument together with information from other auxiliary instruments.

The model in (1) is similar to Aalen (1980) multiplicative intensity model. There, we observe i.i.d. copies in the presence of censoring. For each copy, the covariates are determined at time zero, and so they are usually referred to as marks. Here we observe a single trajectory of the data. The covariates are continuous time stochastic processes and change as new information becomes available during trading. The absence of censoring leads to a simpler estimation procedure that does not require the use of U-statistics. U-statistics would be impractical for the sample sizes considered in this paper. The high dimensional version of Aalen multiplicative intensity model has been considered by several authors (e.g. Gaïffas and Guilloux, 2012, and references therein).

From a technical point of view, our results are derived minimizing a least square criterion for count processes. In our problem, we impose a nonnegativity constraint without a penalty term, unlike what is usually done in the literature (Gaïffas and Guilloux, 2012, Alaya et al., 2015, an references therein). Then, we can directly estimate the model using standard quadratic programming with no need to use a link function that ensures positivity. Hence, this work is related to results for nonnegative least

squares. Under a sparsity assumption on the vector of regression coefficients, the sign constraint imposes restrictions that lead to a regularization as effective as Lasso with no need to tune a penalty parameter. We obtain results equivalent to nonnegative least square with i.i.d. Gaussian errors (Meinshausen, 2013, Slawsky and Hein, 2013). However, in our problem, we need to control a dependent continuous time process.

The use of quadratic programming is important in our context. In practice, models can be calibrated on months of data and the size of high frequency data be so large that cannot be held in ready access memory. However, the problem is written in terms of sufficient statistics whose dimension does not grow with the longitude of the sample, hence they can be held in memory making estimation very simple.

The plan for the paper is as follows. The next section describes the estimation procedure and the assumptions for validity of the statistical procedure. Then, we show that we can obtain consistency for a high dimensional counting process with no need to use a penalty or additional tuning parameters. We conclude the section with a central limit theorem for a two-stage estimator for the parameters. Section 3 applies the results to the estimation of a point process for buy and sell arrivals where the intensity depends on order book variables in a nonlinear but additive way. In Section 3.2 we discuss the scope and limitation of the specific model we choose. Section 4 contains some final remarks. Proofs are in the online Supplementary Material to this paper in Section A.1. Additional details, which we may refer to, can also be found in the same online supplement. This includes a finite sample study using simulations.

2 Assumptions and Results

2.1 The Estimation Problem

Given (1), our estimator $\hat{b}$ for b^* is the solution of the constrained problem

$$\min_{b\geq 0}\left\{-2\int_0^T X(t)' b dN(t) + \int_0^T \left(X(t)' b\right)^2 dt\right\}. \tag{2}$$

This is a standard quadratic programming problem. Throughout, all vector inequalities are meant elementwise. Replacing the constraint with a penalty, this objective function has been used for high dimensional problems by several authors (e.g. Gaïffas and Guilloux, 2012, Alaya et al., 2015).

Note that the population version of the objective function is proportional to

$$\lim_{T\to\infty} \mathbb{E}\left[-2\frac{1}{T}\int_0^T X(t)' b dN(t) + \frac{1}{T}\int_0^T \left(X(t)' b\right)^2 dt\right]$$

and using the definition of the intensity, this is equal to

$$\lim_{T\to\infty} \mathbb{E}\left[-2\frac{1}{T}\int_0^T X(t)' b\lambda^*(t)\, dt + \frac{1}{T}\int_0^T \left(X(t)' b\right)^2 dt\right]. \tag{3}$$

When $\lambda^*(t) = X(t)' b^*$, it is easy to see that the constrained minimizer of the above display is $b = b^*$, when the constraint $b \geq 0$ holds for b^* as well. In some situations, a positivity constraint on b is meaningful. Then, we do not need to use a penalty to find consistent estimators in high dimensions. We shall give examples in Section 2.3.

2.2 Assumptions

We introduce some additional notation and terminology. Fix an arbitrary positive T. We denote the first $n = N(T)$ jump times of N by $T_0 < T_1 < \ldots < T_n \leq T$, where $T_0 = 0$. The integral of (1) is the compensator of $N(t)$ and $M(t) := \left(N(t) - \int_0^t \lambda(r)\, dr\right)$ is an $\mathcal{F}_t$-martingale (see discussion around (1)). Throughout, we use index subscripts to denote the relevant entry in either vectors or matrices. Let $S = \{i \leq K : b_i^* > 0\}$ be the set of nonzero entries of b^*, $S^c = \{i \leq K : b_i^* = 0\}$ be its complement, and $s = |S|$ be

the cardinality of S. For an arbitrary vector $a \in \mathbb{R}^K$ and $U \subset \{1, 2, ..., K\}$ we denote by $a_U \in \mathbb{R}^{|U|}$ the $|U|$-dimensional sub-vector of a obtained by removing all the entries with index not in the set U. Define $\hat{\Sigma} := \frac{1}{T}\int_0^T X(t) X(t)' dt$ and $\hat{\Sigma}_S := \frac{1}{T}\int_0^T X_S(t) X_S(t)' dt$. We use $\|\cdot\|_p$ to denote the ℓ_p norm $p \in (0, \infty]$. Finally, we write w.p.1. to mean with probability going to one.

Assumption 1 *(Model Assumption) The point process admits the intensity (1) which is supposed to be uniformly bounded by a constant $\bar{\lambda}$ (possibly going to infinity), $X := (X(t))_{t\geq 0}$ is a predictable process with values in $[0,1]^K$ for each $t \geq 0$, and $b^* \in [0, \infty)^K$.*

One main restriction is that the intensity is bounded by a constant $\bar{\lambda}$. It could be relaxed to a moment condition, but at the cost of additional technical complications. In this case, it would not be possible to obtain an error bound that is logarithmic in K.

Assumption 2 *(Eigenvalues Condition) There is a constant $\phi_{\min} > 0$ such that the eigenvalues of $\hat{\Sigma}_S$ are all greater than $\phi_{\min}$, w.p.1.*

Let $L > 0$ and $\mathcal{R}(L, S) := \{b : \|b_{S^c}\|_1 \leq L \|b_S\|_1\}$, where S the index set of active variables. The (L, S) restricted ℓ_1-eigenvalue of a matrix A is defined as

$$\phi^2_{comp}(A, L, S) := \min\left\{ s\frac{b'Ab}{\|b\|_1^2} : b \in \mathcal{R}(L, S) \right\}. \tag{4}$$

A lower bound on (4) is the weakest assumption used to derive oracle inequalities for Lasso (van de Geer and Bühlmann, 2009). We shall use the following.

Assumption 3 *(Compatibility Condition). There is a strictly positive constant ϕ such that $\phi^2_{comp}(\hat{\Sigma}, \frac{3}{\sqrt{\nu}}, S) \geq \phi$ w.p.1, where ν is as in Assumption 4.*

The positively constrained minimal ℓ_1-eigenvalue of a matrix A is defined as

$$\phi^2_{pos}(A) := \min\left\{ \frac{b'Ab}{\|b\|_1^2} : \min_i b_i \geq 0 \right\}. \tag{5}$$

A lower bound on (5) has been used by Meinshausen (2013) in the context of non-negative least squares.

Assumption 4 *(Positive Eigenvalue Condition). There is a* $\nu > 0$ *such that* $\phi_{pos}^2\left(\hat{\Sigma}\right) \geq \nu$ *w.p.1.*

We shall refer to Assumptions 1, 2, 3 and 4 simply as the Assumptions.

As we shall discuss next, there are reasons to make $\bar{\lambda}$, $\phi_{\min}$, ϕ, and ν depend on T. Hence, we shall allow $\bar{\lambda} = \bar{\lambda}(T) \to \infty$, and $\phi_{\min} = \phi_{\min}(T)$, $\phi = \phi(T)$, $\nu = \nu(T)$ to go to zero as $T \to \infty$, if needed. For ease of notation, we drop the dependence on T in what follows.

2.3 Remarks on the Assumptions

Assumption 1. We view the sign constraint as a hypothesis and estimation is carried out under this hypothesis. Specifically for the problem of high frequency trading, we have a priori knowledge whether the marginal impact of a high frequency order book covariate is increasing or decreasing. Such information can be obtained from other studies (e.g., Kercheval and Zhang, 2015, Sancetta, 2018). Hence, given such information on the direction of the impact, we show how to improve estimation in high dimensional problems.

For the sake of clarity, we now consider two examples. Consider the intensity $\lambda(Z(t)) = a_0 + a_1 Z_1(t) - a_2 Z_2(t)$ that depends on the covariate $Z(t) = [Z_1(t), Z_2(t)]'$ with values in $[0,1]^2$, where $a_i \geq 0$, $i = 0, 1, 2$, $a_0 - a_2 \geq 0$. The parameters' restriction ensures that this intensity is always nonnegative. Then, λ can be written as (1) where $X_1(t) = 1$, $X_2(t) = Z_1(t)$, $X_3(t) = 1 - Z_2(t)$, and $b_1^* = a_0 - a_2$, $b_2^* = a_1$, $b_3^* = a_2$ where $b^* \geq 0$. Hence, in our framework we are able to control the direction of the impact by the linear transformation $x \mapsto 1 - x$. From a computational point of view, this is

equivalent to changing the sign of the covariate and imposing an additional inequality constraint.

Furthermore, possible nonlinearity of the impact of a covariate can lead to high dimensional problems. For example, suppose that $Z(t)$ takes values in $[0,1]$ and $\lambda(Z(t)) = g(Z(t))$ where $g : [0,1] \to [0,\infty)$ is an unknown continuous function. Define

$$\lambda(Z(t)) = \sum_{i=0}^{K} a_i B_{i,K}(Z(t)) \tag{6}$$

where $B_{i,K}(z) = \binom{K}{i} z^i (1-z)^{K-i}$ is the i^{th}term in a K^{th} order Bernstein polynomial. By linearity, this is in the form of (1) when $X_i(t) := B_{i-1,K}(Z(t))$ and $b_i^* = a_{i-1}$, $i = 1, 2, ..., K+1$. The coefficients of Bernstein polynomials have a clear physical interpretation for $K \to \infty$:

$$\sup_{z\in[0,1]} \left| g(z) - \sum_{i=0}^{K} a_i B_{i,K}(z) \right| \to 0$$

when we define $a_i = g(i/K)$ (Lorentz, 1986, Theorem 1.1.1). Hence, the above display suggests that if $g \geq 0$, the assumption that $a_i \geq 0$ is almost necessary when K is large. The argument can be extended to a dimension greater than one. However, we note that the assumptions on $\hat{\Sigma}$ will fail as $K \to \infty$, unless we allow ϕ and ν to go to zero slowly enough. This is allowed in our results (Theorems 1 and 2).

In the empirical section of this paper, we view the sign constraint as a hypothesis on the direction of the impact of the covariates. Relying on several hypotheses we estimate models and compare their performance out of sample (see Section 3 for details). This remark would suggest that impact/sign misspecification would lead to a zero coefficient. Consider the first example in the above discussion. By impact misspecification we mean using $X_3(t) = 1 - Z_2(t)$ when in fact the impact of $Z_2(t)$ on the intensity is positive. Because of the influence from other variables, the procedure may select a covariate

with the wrong impact even when solving the population objective function (3). We carried out a number of numerical examples to find the solution to (3). We found that the constraint tends to be binding for a true negative coefficient when both s and the number of wrongly signed variables is small. However, as either s or the degree of misspecification increases, the constrained population estimator may have a positive sign even when the true one is negative. For the designs we considered in the Supplementary Material, selection of a misspecified variable when solving (3) was relatively infrequent. Details can be found in Section A.4 of the Supplementary Material.

We allow $\bar{\lambda}$, the constant upper bound on the intensity λ^*, to possibly grow to infinity. This is relevant in practice, as we may only have a crude upper bound for λ^* that depends on s, the number of active variables. To see this, note that by assumption, $\lambda^* = X'b^* \leq \|b_S\|_1 = O(s)$. Even if crude, with no further information on either the covariates or the coefficients, this is the best possible upper bound on λ^*. Hence, it is relevant to allow $\bar{\lambda} = O(s)$ and s to diverge to infinity with T. To see how information can be used to find a tighter bound on $\bar{\lambda}$, consider the intensity in (6). Then, we know that the intensity is bounded by $\max_{i \leq K} a_i$, which is the largest of the coefficients to be estimated.

Assumption 2. Unlike Meinshausen (2013) to make the proofs simpler, we use Assumption 2. This is just slightly stronger than the Compatibility Condition $\phi^2_{comp}(\hat{\Sigma}, 0, S) \geq \rho$ for some $\rho > 0$ w.p.1, which is needed in the proofs. To see this, note that, by the nonnegativity of the coefficients, the latter implies that

$$s \sum_{i,j \in S} b_i b_j \hat{\Sigma}_{i,j} \geq \rho \left(\sum_{i \in S} |b_i| \right)^2$$

w.p.1, where $\hat{\Sigma}_{i,j}$ is the i,j entry in $\hat{\Sigma}$. Suppose that ρ is a lower bound on the smallest eigenvalue of $\hat{\Sigma}_S$ w.p.1. Then,

$$s \sum_{i,j \in S} b_i b_j \hat{\Sigma}_{i,j} \geq \rho s \sum_{i \in S} b_i^2$$

w.p.1. Using the fact that the ℓ_1 norm of an s-dimensional vector is bounded by $\sqrt{s}$ times its ℓ_2 norm, this means that $\phi^2_{comp}(\hat{\Sigma}, 0, S) \geq \rho$ holds w.p.1. In practice this means that the covariates with index in the active set S need to be linearly independent w.p.1, when $T \to \infty$. This is plausible, as s should be relatively small, i.e. b^* is sparse.

It is instructive to consider conditions on X that imply this assumption. Recall that $\hat{\Sigma}_{i,j} := T^{-1} \int_0^T X_i(t) X_j(t)$. Suppose that $(X_i(t) X_j(t))_{t \geq 0}$ is ergodic, for all $i,j \in S$, in the sense that $\hat{\Sigma}_{i,j} \to \Sigma_{i,j}$ a.s., where $\Sigma_{i,j}$ is a constant, $i,j \in S$. By boundedness of the covariates, the convergence must also hold in L_2. Therefore, if S has bounded cardinality, $\hat{\Sigma}_S - \Sigma_S$ converges to zero in expected Frobenius norm. This implies convergence of the eigenvalues of $\hat{\Sigma}_S$ to the ones of Σ_S (Bosq, 2000, Theorem 4.4). Incidentally, boundedness implies that all the moments of $\hat{\Sigma}_{i,j}$ converge to $\Sigma_{i,j}$, $i,j \in S$.

The above definition of ergodic is closely linked to the existence of an asymptotic mean stationary measure such that Σ_S is the expectation of $\hat{\Sigma}_S$ with respect to that measure (Gray and Kieffer, 1980, Theorem 1, and Gray, 2009, Ch.6 for an extensive treatment). The argument does not require stationarity of $(X_i(t) X_j(t))_{t \geq 0}$. As a simple example of why this can be relevant, consider $\mathbb{E} X_i^2(t) = \sigma_i^2(t)$. It is common in high frequency data to have time varying (unconditional) volatility of certain quantities, due to intraday nonstationarity. However, by ergodicity and boundedness, there is a constant $\sigma_i^2 = \lim_T \frac{1}{T} \int_0^T \sigma_i^2(t)\, dt$. In practice, this is plausible when T spans multiple days.

Finally, we note that if s increases, the dependence among the active variables may

increase, making a lower bound $\phi_{\min}$ on the smallest eigenvalue smaller. For this reason, we allow the bounds in our results to explicitly depend on $\phi_{\min}$.

Assumption 3. We use the same Compatibility Condition as in Meinshausen (2013). Meinshausen (2013) also uses $\phi^2_{comp}(\hat{\Sigma}, 0, S)$. As discussed above, by Assumption 2, we do not need this additional condition. To verify the Compatibility Condition using population quantities, we can use approximations. Suppose that there is a sequence $\epsilon_T \to 0$ such that $\max_{i,j} \left|\hat{\Sigma}_{i,j} - \Sigma_{i,j}\right| = O_P(\epsilon_T)$, where $\Sigma_{i,j}$ is understood to be the limit (in probability) of $\hat{\Sigma}_{i,j}$, $i, j \in \{1, 2, ..., K\}$. If the number of active variables satisfies $s = o_P\left(\nu \epsilon_T^{-1}\right)$ with ν as in Assumption 3, then we can deduce that the Compatibility Condition is satisfied if $\nu > 0$ and the smallest eigenvalue of Σ is strictly positive (van de Geer and Bühlmann, 2009, Corollary 10.1 and discussion). A bound on the minimal eigenvalue of Σ requires that the covariates are not asymptotically dependent. To establish such a bound we will need rates of convergence. Ergodicity alone, as discussed in the previous remark, is not sufficient even when S has finite cardinality. To see this, note that when s is bounded we still need $\max_{i,j} \left|\hat{\Sigma}_{i,j} - \Sigma_{i,j}\right| = o_P(1)$ for $i, j \in \{1, 2, ..., K\}$ and not just for $i, j \in S$. Given that K is not necessarily bounded, this requires convergence rates.

We also note that as K increases, the smallest eigenvalue of Σ may tend to zero. For this reason, our bounds are in terms of ϕ to show how quickly the convergence rate may deteriorate if $\phi \to 0$.

Assumption 4. This assumption is the same as in Meinshausen (2013). For Assumption 4 to hold, it is sufficient that $\min_{i,j} \hat{\Sigma}_{i,j} > 0$ a.s., $i, j \in \{1, 2, ..., K\}$. This is satisfied if the covariates $X_i(t)$ do not have a disjoint support for all $t \geq 0$. Then, the average of $X_i(t) X_j(t)$ is greater than zero because $X_i(t) \in [0, 1]$, $t \geq 0$, unless some of the covariate are exactly zero. This condition is satisfied in our empirical study. However,

there are parametrizations that lead to disjoint subsets. Notable examples are splines and one-hot encoding (Alaya et al., 2019). One-hot encoding essentially builds mutually disjoint bins for each covariate and constructs dummy variables for each bin. This leads to mutually disjoint covariates. As discussed in Meinshausen (2013, Example III), the following setup covers splines and one-hot encoding, and satisfies Assumption 4. Fix a positive integer L and suppose that $\{P_l : l = 1, ..., L\}$ is a partition of $\{1, 2, ..., K\}$ such that $\hat{\Sigma}_{i,j} > \nu L$ a.s., if $i, j \in P_l$ for some l and zero if $i \in P_l$,$j \in P_k$ when $k \neq l$. This means that covariates that have an index in different partitions have disjoint support. In this case

$$b'\hat{\Sigma}b = \sum_{l=1}^{L} \sum_{i,j \in P_l} b_i b_j \hat{\Sigma}_{i,j} \geq \nu L \sum_{l=1}^{L} \sum_{i,j \in P_l} b_i b_j = \nu L \sum_{l=1}^{L} \left(\sum_{i \in P_l} b_i \right)^2 .$$

Then, using the fact that, by Jensen's inequality, for any constants $a_1, a_2, ..., a_L$, we know that $\left(\sum_{l=1}^{L} a_l\right)^2 \leq L \sum_{l=1}^{L} a_l^2$, we deduce that

$$b'\hat{\Sigma}b \geq \nu \left[\sum_{l=1}^{L} \left(\sum_{i \in P_l} b_i \right) \right]^2 = \nu \left\| b \right\|_1^2$$

when $\min_i b_i \geq 0$. Then, Assumption 4 is satisfied.

As mentioned above, we require $\hat{\Sigma}_{i,j} > \nu L$. Given that $\hat{\Sigma}_{i,j} \in [0, 1]$, we need $\nu \to 0$ if the number of partitions L diverges to infinity. This is of interest for certain nonparametric estimators like one-hot encoding and splines with an increasing number of knots.

We conclude this section giving an intuition on why Assumption 4 leads to results comparable to ℓ_1 penalization. By Assumption 4, $b'\hat{\Sigma}b \geq \nu \left\| b \right\|_1^2$ w.p.1 as long as $b \geq 0$. Hence, bounds on $b'\hat{\Sigma}b$ will translate into bounds for $\left\| b \right\|_1^2$. However, as soon as $\nu \to 0$, the control of $\left\| b \right\|_1^2$ using $b'\hat{\Sigma}b$ becomes loose and the procedure will underperform

standard estimation with an ℓ_1 penalty. Similar remarks, but in the context of regression analysis, are made in Meinshausen (2013).

2.4 Asymptotic Results

2.4.1 Consistency

We shall keep track of all the constants to see how the bound is affected. This is useful for example if $\phi_{\min} \to 0$ slowly enough. Similarly, we can allow $\bar{\lambda} \to \infty$. This is important, because $T/\bar{\lambda}$ is one of the main quantities affecting the convergence rate. Define $c(s) := \max\left\{\frac{s^2}{\phi^2}, \frac{1}{\nu}\right\}$ and $\mu_T := \frac{1}{T}\int_0^T \mathbb{E}\lambda^*(t)\,dt$. Note that $\mu_T \leq \bar{\lambda}$. We have consistency of the estimator $\hat{b}$ for b^*, under the ℓ_1 norm. Throughout, to simplify the notation we assume that $K \geq 2$ in all the results that follow.

Theorem 1 *Under the Assumptions, if* $\log K = O\left(T\bar{\lambda}\right)$*, then,*

$$\left\|\hat{b} - b^*\right\|_1 = O_P\left(\sqrt{\frac{c(s)\left(s^2\mu_T\phi_{\min}^{-2} + \bar{\lambda}\log K\right)}{T}}\right).$$

The second result is an estimation of the prediction error.

Theorem 2 *Under the Assumptions, if* $\log K = O\left(T\bar{\lambda}\right)$*, then,*

$$\frac{1}{T}\int_0^T \left(X(t)'\hat{b} - X(t)'b^*\right)^2 dt = O_P\left(\frac{c^{1/2}(s)\left(s^2\mu_T\phi_{\min}^{-2} + \bar{\lambda}\log K\right)}{T}\right).$$

Theorems 1 and 2 make explicit the dependence on the parameters ϕ, $\phi_{\min}$, ν, and $\bar{\lambda}$ defined in the Assumptions. Hence, we can have consistency even when $\phi_{\min}, \phi, \nu \to 0$ and $\bar{\lambda} \to \infty$ at suitable rates as $T \to \infty$. When $\nu, \phi, \phi_{\min}$ are fixed, $c(s) = O(s^2)$. Then, we have the following corollary.

Corollary 1 *Under the Assumptions, if* $\bar{\lambda}$ *is fixed and bounded away from infinity,* $\phi_{\min}, \phi, \nu$ *are fixed and bounded away from zero, and* $\log K = O\left(T\bar{\lambda}\right)$*, then*

$$\left\|\hat{b} - b^*\right\|_1 = O_P\left(\sqrt{\frac{s^4 + s^2 \log K}{T}}\right) \tag{7}$$

and

$$\frac{1}{T}\int_0^T \left(X(t)'\hat{b} - X(t)'b^*\right)^2 dt = O_P\left(\frac{s^3 + s\log K}{T}\right) \tag{8}$$

hold true.

The convergence rate derived here is typical of high dimensional estimation problems under some form of regularization. For example, the convergence rate in our error bounds is similar to the one derived for the nonnegative least square regression problem with i.i.d. Gaussian errors (Meinshausen, 2013, Theorems 1-2, and Slawsky and Hein, 2013, Theorem 2). We can relate to those results, assuming that the conditions of Corollary 1 hold. Then, we have ℓ_1 consistency of the estimator if $T^{-1}\left(s^4 + s^2 \ln K\right) \to 0$ (see (7)). On the other hand Meinshausen (2013, first part in his Theorem 1) say that in the regression case we have ℓ_1 consistency if $n^{-1}s^4 \ln K \to 0$, where n is the sample size.

Suppose that the regressors are bounded. Under a lower bound on the smallest nonzero elements in the regressor coefficients, results in Meinshausen (2013, Theorem 2) imply (empirical) L_2 consistency of the regression function when $n^{-1}s\ln K \to 0$. With no such condition on the regression coefficients, in the present context, we can obtain L_2 consistency of the intensity estimator when $T^{-1}\left(s^3 + s\ln K\right) \to 0$ (see (8)). We can also consider the L_2 consistency of the estimated intensity in Gaïffas and Guilloux (2012) for the Aalen intensity model, using a data driven Lasso penalty. Using a restricted eigenvalue condition, Gaïffas and Guilloux (2012, Theorem 2) achieve consistency if $n^{-1}s\ln K \to 0$, where, again, n is the sample size. Given that the model setup differs,

and we cannot rely on a penalty, the method of proof in Gaïffas and Guilloux (2012) is different. It can be difficult to discuss convergence rates beyond the aforementioned remarks.

The next result is about set identification.

Corollary 2 *Using the notation in Theorem 1, define* $\rho := \sqrt{c(s)\,\bar{\lambda}\left(s^2\phi_{\min}^{-2} + \log K\right)/T}$. *Suppose that the conditions of Theorem 1 hold. Furthermore, suppose that* $\min_{i\in S} b_i^* > \kappa$ *for some* κ *such that* $\kappa/\rho \to \infty$. *Let* $\hat{S} = \left\{i \leq K : \hat{b}_i > 0\right\}$. *Then,* $\Pr\left(\hat{S} \subset S\right) \to 0$.

Let $\hat{S}_\epsilon = \left\{i \leq K : \hat{b}_i > \epsilon\right\}$. *Then, under the above conditions, for any* ϵ *such that* $\epsilon/\rho \to \infty$ *and* $\kappa/\epsilon \to c > 1$, $\Pr\left(\hat{S}_\epsilon \neq S\right) \to 0$.

Corollary 2 says that $\hat{S}$, the estimated support of b^*, is a superset of S, the true support, w.p.1. Using a threshold on the coefficients we can achieve set identification, w.p.1. Note that we can have $\kappa \to 0$ slowly enough and similarly for ϵ.

2.4.2 Convergence in Distribution

We obtain a central limit theorem for the OLS estimator for b^*. At first we estimate $\hat{b}$ in (2) and obtain $\hat{S}_\epsilon$ as in Corollary 2. Then we compute b_ϵ^{OLS} which is the K-dimensional vector such that its entries with index in $\hat{S}_\epsilon$ are equal to

$$\left(\int_0^T X_{\hat{S}_\epsilon}(t)\, X_{\hat{S}_\epsilon}(t)'\, dt\right)^{-1} \left(\int_0^T X_{\hat{S}_\epsilon}(t)\, dN(t)\right)$$

while all the other entries are zero. Note that the cardinality of $\hat{S}_\epsilon$ can still grow with T. Given a fixed K-dimensional vector α satisfying $\alpha'\alpha = 1$ and $\alpha'_{\hat{S}_\epsilon}\alpha_{\hat{S}_\epsilon} > 0$, we are interested in the asymptotic distribution of $\sqrt{T}\alpha'\left(b_\epsilon^{OLS} - b^*\right)$ to conduct inference. Define $\hat{\Sigma}^N := \frac{1}{T}\int_0^T X(t)\, X(t)'\, dN(t)$ and $\hat{\Sigma}_S^N$ to be the submatrix that includes only the entries with row and columns indices in S. We state an additional assumption.

Assumption 5 *We have that* $\lim_T \sum_{i,j\in S}\left[Var\left(\hat{\Sigma}_{i,j}\right)+Var\left(\hat{\Sigma}^N_{i,j}\right)\right]=o\left(\phi_{\min}\right)$*, where* $\phi_{\min}$ *is as in Assumption 2, and* $\mathbb{E}\hat{\Sigma}_S$ *and* $\mathbb{E}\hat{\Sigma}^N_S$ *both converge to full rank constant matrices.*

Note that $Var\left(\hat{\Sigma}_{i,j}\right)=o\left(1\right)$ under ergodicity assumptions on X_iX_j. As already mentioned, ergodicity and boundedness imply that $\hat{\Sigma}_{i,j}$ converges in L_2 to a constant $\Sigma_{i,j}$. By boundedness we also know that $\mathbb{E}\hat{\Sigma}_{i,j}$ converges to a constant limit $\Sigma_{i,j}$. However, we need the rate of convergence to be fast enough to ensure convergence of all the entries in $\hat{\Sigma}_S$. For example, this is trivially satisfied if s is bounded or if we assume that $\max_{i,j\in S} Var\left(\hat{\Sigma}_{i,j}\right)=O\left(\epsilon_T^2\right)$ such that $s=o_P\left(\epsilon_T^{-1}\right)$. The same argument applies to $\hat{\Sigma}^N_S$. In this case, write

$$\hat{\Sigma}^N_{i,j}=\frac{1}{T}\int_0^T X_i\left(t\right)X_j\left(t\right)'\left(X\left(t\right)'b^*\right)dt+\frac{1}{T}\int_0^T X_i\left(t\right)X_j\left(t\right)'dM\left(t\right)$$

using the definition of M and intensity (see the discussion around (1)). Being a martingale, the second term on the right hand side converges to zero in probability $i,j\in S$. Under ergodicity assumptions on $\left(X_i\left(t\right)X_j\left(t\right)X_k\left(t\right)\right)_{t\geq 0}$ for $i,j,k\in S$, we also have that the first term on the right hand side (r.h.s.) converges to a constant and the same applies to its moments because of dominated convergence. We can now state a central limit theorem for our estimator.

Theorem 3 *Suppose that the assumptions of Corollary 2, Assumption 5,* $\inf_{t>0}\lambda^*\left(t\right)>0$ *and* $\lim_T \alpha_S'\left(\mathbb{E}\hat{\Sigma}_S\right)^{-1}\alpha_S>0$ *hold. If* $s^2=o\left(T\phi^2_{\min}/\bar{\lambda}\right)$*, then,* $\sqrt{T/\sigma^2_\alpha}\alpha'\left(\hat{b}^{OLS}_\epsilon-b^*\right)\to\mathcal{N}\left(0,1\right)$ *in distribution, where* $\mathcal{N}\left(0,1\right)$ *is the standard normal distribution and*

$$\sigma^2_\alpha=\lim_T \alpha_S'\left(\mathbb{E}\hat{\Sigma}_S\right)^{-1}\left(\mathbb{E}\hat{\Sigma}^N_S\right)\left(\mathbb{E}\hat{\Sigma}_S\right)^{-1}\alpha_S.$$

Moreover, $\hat{\sigma}^2_\alpha=\alpha'_{\hat{S}_\epsilon}\hat{\Sigma}^{-1}_{\hat{S}_\epsilon}\hat{\Sigma}^N_{\hat{S}_\epsilon}\hat{\Sigma}^{-1}_{\hat{S}_\epsilon}\alpha_{\hat{S}_\epsilon}$ *is a consistent estimator for* σ^2_α.

Note that the conditions are for S rather than $\hat{S}_\epsilon$. By Corollary 2, the two sets are the same, w.p.1. By assumption, $\min_{i \in S} b_i^* > \kappa$ for κ as in Corollary 2. When κ is fixed, the post selection asymptotics in Theorem 3 are valid as they are uniform in b^* (Leeb and Pötscher, 2005). When κ is allowed to go to zero as in Corollary 2, post selection asymptotics are still valid, but the convergence to a normal can be arbitrarily slow (mutatis mutandis, see Leeb and Pötscher, 2005, p.29ff.). Intuitively, this follows from the fact that as $\kappa \to 0$, it becomes harder to distinguish very small coefficients from zero coefficients.

Next we shall define a nonlinear additive model for buy and sell trades based on order book and trade variables.

3 Empirical Application: Order Book Determinants of Crude Oil Buy and Sell Trade Arrivals

We estimate the intensity of buy and sell trade arrivals, separately. We investigate the impact of features constructed from quote and trade data on these buy and sell events. The features include order book imbalance, trade imbalance, spread, and durations. Information from quote data appears to contain much information (Hall and Hautsch 2007, Cont et al., 2014 , Kercheval and Zhang, 2015, Sancetta, 2018). For example, order book imbalances are a quantity that has been used by practitioners for more than two decades (MacKenzie, 2017). However, most of the attention in the econometric literature has been on trade data.

We use data from the front month of crude oil futures traded on the Chicago Mercantile Exchange (CME). The CME ticker is CL and the sample period is 1/May/2013-5/June/2013 each day from 13:30 to 18:00 GMT. As auxiliary instruments we use information from heating oil (HG), natural gas (NG) and the S&P500 (ES) futures,

where the CME ticker is in parenthesis. The data was collected by a high frequency proprietary trading firm collocated in the Aurora data center in Chicago. The data is at the highest level of granularity, comprising all quotes and trades, time stamped at nanosecond resolution.

3.1 The Model

We use different model specifications to capture the nonlinearities in the impact of the covariates on the intensity. We separately consider buy and sell trade arrivals. We want to assess the extent to which these variables impact the intensity in a nonlinear way. We consider different hypotheses. To evaluate the hypotheses, we carry out a test for model performance. The results for the estimated models show that the impact is nonlinear as expected and allow us to characterize the shape of the impact. In order to formulate our hypotheses, we first define the covariates.

The covariates. The covariates are reported in Table 1. We apply exponential moving average (EWMA) filters to some of the covariates. The EWMA of a variable $X(t_i)$ with smoothing parameter α is

$$EWMA(X(t_i)) = \alpha EWMA(X(t_{i-1})) + (1-\alpha) X(t_i) \tag{9}$$

where $EWMA(X(t_1)) = X(t_1)$. Here, t_1 is the time of the first update in the variable X at the start of each day, while t_i is the time of the i^{th} update. EWMA's are computed for each day. Note that the covariates are updated at discrete times for each instrument and the update times are different from the trade update times T_j. We then sample the data at times that are the union of each covariate update and the times T_j for the traded instrument only, i.e. CL. This reduces the total number of updates within

each day. Finally, to ensure that the covariates are predictable, we make them left continuous by lagging them after sampling.

Variables are mapped, by linear transformation, into $[0, 1]$. However, spread and durations are first capped and then scaled by the cap so that they take values in $[0, 1]$. The book volume imbalance at level $j \in \{1, 2, ..., 5\}$, VolImb_j, is defined as

$$\text{VolImb}_j = \frac{\text{BidSize}_j - \text{AskSize}_j}{\text{BidSize}_j + \text{AskSize}_j} \tag{10}$$

where BidSize_j is the bid size (quantity) of the j^{th} level bid, and similarly for AskSize_j. This variable takes values in $[-1, 1]$. We map it to $[0, 1]$ by standard linear transformation: multiply by two and subtract one. Hence a value of 0.5 corresponds to a book volume imbalance equal to zero. The trade imbalance is computed from the EWMA of the signed traded volume every time there is a trade. We then divide it by the EWMA of the unsigned volumes. The EWMA's parameter is $\alpha = 0.98$ for both denominator and numerator. Durations are in seconds with nanosecond decimals, capped to one second. They are then passed to EWMA filters with parameter $\alpha = 0.98$ and 0.90. The spread is capped to 4 ticks and standardized by 4. Hence, the minimum value it can take is 0.25. Additional details regarding the data and the calculations are in Section A.2 in the online Supplementary Material.

Table 1: Covariates Used for Estimation. The column "Smoothing" reports the smoothing parameter used if an EWMA had been applied to the original variable.

Variables	Short Name		Smoothing
Volume Imbalance Level j	VolImb$j$, $j = 1, 2..., 5$		
Spread	Spread		
Trade Imbalance	TrdImb98		$\alpha = 0.98$
Durations	Dur98	Dur90	$\alpha = 0.98$ and 0.90

All variables are assumed to be mapped to $[0, 1]$ as described above. Then, if we

hypothesize that the impact of a covariate is negative, we apply the transformation $x \mapsto 1-x$ as discussed in Section 2.3. For example we would apply this transformation to durations: it is natural to assume that longer past durations are associated to a lower current intensity. We then apply a set of transformations $x \mapsto g_l(x)$, $l = 1, 2, 3$, where $g_1(x) = x$, $g_2(x) = (x - 0.5)^2$, $g_3(x) = x^3$. The functional forms can account for different types of impact: $g_1(x)$ has constant marginal impact as opposed to $g_3(x)$ which is close to zero for most values of x not in the proximity of one. The latter transform seems more appropriate for volume imbalances and most covariates. The function $g_2(x)$ is just convex with a minimum at $x = 0.5$. Combining these functions we can approximate different types of impact and still be able to interpret the results. We now state the hypotheses for model restriction.

Model hypothesis 1. The impact of all volume imbalances and trade imbalance is positive for buy trade intensity and negative for sell intensity. The impact of the spread and durations is always negative. Finally, we apply, separately, the following two transformations: g_1 and g_3, i.e. linear and cubic. Hence, this doubles the final number of covariates to estimate.

Model hypothesis 2. The impact of all covariates is as in model hypothesis 1. However, we finally apply, separately, the following two transformations: g_2 and g_3, i.e. quadratic and cubic. Again, this doubles the final number of covariates to estimate.

Model hypothesis 3. The impact of all covariates is as in model hypothesis 1, except for spread, for which we now suppose a positive impact. Finally, we apply, separately the same transformations as in model hypothesis 2.

Remarks on the model hypotheses. We briefly comment on the hypotheses. First, the direction of the impact is a maintained hypothesis when estimating the model. If this is wrong, it is likely that the estimated coefficient will be zero. Second, regarding the transformations, we allow for the possibility of combining two functions (out of three) in order to derive a more flexible model. At the same time, we also rely on the estimation procedure to select the best submodel in an automated fashion. For example, assume that the direction of the impact of the covariates is as in model hypothesis 1 and that the impact is linear. Then, model hypothesis 1 is true, while model hypotheses 2 is false. We have included model hypothesis 3 to illustrate the point that using the wrong direction of the impact for the spread can lead to a zero coefficient.

For this problem, we have high confidence in the direction of the impact based on the existing literature (recall the discussion in Section 2.3).

After the application of EWMA's filters and the transformations, each hypothesis has 18 covariates for each of the 4 instruments, which means 72 parameters. Rather than making this a pure exercise in data mining, we prefer to keep the number of covariates relatively small to simplify our discussion and focus on the significance of the results.

3.2 Scope and Limitations

The goal is to present a model in the same vein as the ones used by high frequency trading firms. The focus is on the order book as the main source of information. The EWMA's of trade durations attempt to capture the well known autocorrelation of durations (Engle and Russell, 1998). The model is geared towards high frequency predictions in live trading. It is then customary to restrict variables to finite ranges to avoid consuming corrupted/invalid messages from the exchange. This may sound problematic for some fat tailed variables like durations. However, as durations increase,

the intensity should shrink. Capping durations to a relatively high value is equivalent to saying that the decreasing marginal effect of durations beyond the cap value is zero. Given that for crude oil we can have hundreds of trade arrivals within a second, capping durations to one second seemed reasonable.

The model does not include an intraday seasonal component. However, part of this seasonality is implicitly captured by a slow moving EWMA of durations, so we did not further increase the number of variables as we want to showcase the importance of the order book.

A final comment pertains to the separate estimation of buy and sell intensities. Suppose the following feedback loop effect

$$\begin{aligned}\lambda^{buy} =& X' b^{buy} + \rho^{buy} \lambda^{sell} \\ \lambda^{sell} =& X' b^{sell} + \rho^{sell} \lambda^{buy},\end{aligned}$$

where ρ^{buy} and ρ^{sell} are constants in $[0, 1)$. This system has the reduced form

$$\begin{aligned}\lambda^{buy} =& X' \left(b^{buy} + \rho^{buy} b^{sell}\right) \left(1 - \rho^{buy}\rho^{sell}\right)^{-1} \\ \lambda^{sell} =& X' \left(b^{sell} + \rho^{sell} b^{buy}\right) \left(1 - \rho^{buy}\rho^{sell}\right)^{-1}.\end{aligned}$$

In consequence, separate estimation of the buy and sell intensities is equivalent to estimation of the above reduced form model.

3.3 Results

To carry out a test of performance measurement, we split the data in two halves. We use the data until 18/May/2013 for estimation and the subsequent sample for testing. We shall only report results for buy trades. The ones for sell trades are essentially

Table 2: Estimation Results for Buy Trades. For each of the three hypotheses, the active covariate and the estimated coefficient are reported. The superfix and suffix in each covariate name are separated by an underscore. The superfix represents the ticker and the suffix the index in the transform g_l $l = 1, 2, 3$, defined at the end of the paragraph "the covariates" in Section 3.1. The total number of estimated parameters is denoted by K and the estimated number of nonzero elements by $\hat{s}$.

Hypothesis 1		Hypothesis 2		Hypothesis 3	
active cov.	$\hat{b}$	active cov.	$\hat{b}$	active cov.	$\hat{b}$
CL_VolImb1_3	2.50	CL_VolImb1_2	1.25	CL_VolImb1_2	1.28
CL_TrdImb98_3	0.04	CL_VolImb1_3	2.04	CL_VolImb1_3	2.05
CL_Dur98_3	0.44	CL_VolImb2_2	4.61	CL_VolImb2_2	4.65
CL_Dur90_3	3.51	CL_VolImb3_2	1.48	CL_VolImb3_2	1.51
HO_Spread_3	2.48	CL_TrdImb98_2	8.08	CL_TrdImb98_2	8.16
HO_Dur98_1	0.16	CL_Dur98_3	0.29	CL_Dur98_3	0.29
HO_Dur98_3	0.52	CL_Dur90_3	3.12	CL_Dur90_3	3.12
HO_Dur90_3	1.87	HO_Spread_3	1.96	HO_Dur98_1	0.29
		HO_Dur98_1	0.28	HO_Dur98_3	0.35
		HO_Dur98_3	0.38	HO_Dur90_3	1.95
		HO_Dur90_3	1.91		
K	$\hat{s}/K$	K	$\hat{s}/K$	K	$\hat{s}/K$
73	0.11	73	0.15	73	0.14

identical.

The estimator for all models is sparse, as only 10 to 15 percent of the coefficients are nonzero (Table 2). An interesting finding is that the estimation based on hypothesis 2 leads to an impact of order book volume imbalances that is not monotonic. Moreover, all hypotheses suggest that order book volume imbalances beyond the third level are not important.

We then test the performance of the models out of sample. As already mentioned, we split the sample into estimation and test samples of roughly the same size each. Let $\hat{\lambda}^{(k)}$ be the intensity estimator from (2) using the estimation sample for model

hypothesis $k \in \{1, 2, 3\}$. The loglikelihood on the test sample is

$$L_T^{(k)} = \int_0^T \ln \hat{\lambda}^{(k)}(t)\, dN(t) - \int_0^T \hat{\lambda}^{(k)}(t)\, dt.$$

Its predictable part is

$$H_T^{(k)} := \int_0^T \ln \hat{\lambda}^{(k)}(t)\, \lambda^*(t)\, dt - \int_0^T \hat{\lambda}^{(k)}(t)\, dt.$$

Under the null hypothesis that model hypothesis 1 and hypothesis 2 perform the same, we have that $H_T^{(1)} - H_T^{(2)} = 0$, against the one sided alternative that model 2 is worse, i.e. $H_T^{(1)} - H_T^{(2)} > 0$. The standardized likelihood ratio test statistic $\mathrm{LR}_T^{(1,2)} := \left(L_T^{(1)} - L_T^{(2)}\right) / \sqrt{T\hat{\sigma}_T^2}$ is asymptotically standard normal, where

$$\hat{\sigma}_T^2 = \frac{1}{T} \sum_{j=1}^{n} \left[\ln \left(\hat{\lambda}^{(1)}(T_j) / \hat{\lambda}^{(2)}(T_j) \right) \right]^2$$

(Sancetta, 2018, Proposition 1). When carrying out the test, we standardize each intensity by $\hat{\lambda}_0 := \int_0^T \hat{\lambda}^{(k)}(t)\, dt / N(T)$ to ensure that the model comparisons are not affected by the value of the expected intensity. Recall that $\hat{\lambda}^{(k)}$ is unbiased if $\hat{\lambda}_0 = 1$.

The test is performed for different pairs of model hypotheses. We find that hypothesis 2 is favored out of sample (Table 3). From the results in Tables 2 and 3, we infer that information from auxiliary instruments is relevant and instruments are interlinked. Moreover, we find that the impact of volume imbalances is not monotonic. It is also worth mentioning that the results for the sell intensity led to similar conclusions except for one case. We did not reject the hypothesis that model hypothesis 2 (H2) is not better than model hypothesis 3 (H3). Recall that the difference in these two model hypotheses is that the impact of the spread is in the wrong expected direction for H3. In this case, the spread is not selected when estimation is carried out under H3. The

spread does not seem to be an important variable in this problem because all products are very liquid. Hence, their spread tends to be very tight most of the time. The only exception is heating oil (HO). In this case, inspection of the data did show some variation for the spread. However, overall its importance is not as crucial as the other covariates. It is reassuring that under H3, the spread for heating oil, having the wrong sign is not selected.

Table 3: Test of Model Performance for Buy Trades. The standardized out of sample likelihood ratio statistic $\mathrm{LR}_T^{(1,2)}$ between different model combinations is reported. The columns identify the null hypothesis. For example, H2-H1 is the null that the model implied by hypothesis 1 performs as well as the one from hypothesis 2. A large value rejects the null in favor of the alternative that the model from hypothesis 2 performs better.

	H2-H1	H3-H1	H2-H3
t-stat	38.59	37.92	2.65
p-value	0	0	0

4 Conclusion

This paper studies the estimation of a point process where the intensity is a function of high dimensional variables. We rely on a nonnegativity constraint for the intensity to show that the estimation problem can be solved by standard quadratic programming with no need to include a penalty and tune additional parameters. The resulting estimator is consistent and the error only grows logarithmically in the number of estimated parameters, as long as the true parameter is sparse. Our motivation is the estimation of a possibly nonlinear additive model for the intensity of buy and sell trades of crude oil futures prices. The covariates that affect the intensity are order book and trade variables on crude oil together with information from auxiliary instruments such as natural gas, heating oil and the S&P500 futures. The results show that the impact of

variables constructed from the order book and trades is nonlinear and that the instruments are interconnected. A test for model performance is used and it shows that we can compare competing hypotheses for the direction of the impact of the covariates and the functional form of the intensity.

Supplementary Material

Luca Mucciante and Alessio Sancetta (2022): Supplement to "Estimation of a High Dimensional Counting Process Without Penalty for High Frequency Events", Econometric Theory Supplementary Material. To view, please visit: [[doi will be inserted here by typesetter]]

References

[1] Aalen, O. (1980) A Model for Nonparametric Regression Analysis of Counting Processes. In W. Klonecki, A. Kozek and J. Rosiński (eds.), Mathematical Statistics and Probability Theory: Proceedings of the Sixth International Conference Wisla (Poland) 1978, Lecture Notes in Statistics, 1-25. New York: Springer.

[2] Alaya, M.Z., S. Gaïffas, A. Guilloux (2015) Learning the Intensity of Time Events With Change-Points. IEEE Transactions on Information Theory 61, 5148-5171.

[3] Alaya, M.Z., S. Bussy, S. Gaïffas, and A. Guilloux (2019) Binarsity: A Penalization for One-Hot Encoded Features in Linear Supervised Learning. Journal of Machine Learning Research 20, 1-34.

[4] 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.

[5] Bosq, D. (2000). Linear Processes in Function Spaces: Theory and Applications. Lecture Notes in Statistics 149. New York: Springer.

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

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

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

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

[9] Glosten, L.P., and P. Milgrom (1985) Bid, Ask and Transaction Prices in a Specialist Market with Heterogeneously Informed Traders. Journal of Financial Economics 14, 71-100.

[10] Gray, R.M. (2009) Probability, Random Processes, and Ergodic Properties. Boston, MA: Springer.

[11] Gray, R.M. and J.C. Kieffer (1980) Asymptotically Mean Stationary Measures. Annals of Probability 8, 962-973.

[12] Kercheval, A. and Y. Zhang (2015) Modelling High-Frequency Limit Order Book Dynamics with Support Vector Machines. Quantitative Finance 15, 1315-1329.

[13] Kyle, A.S. (1985) Continuous Auctions and Insider Trading. Econometrica 53, 1315-1336.

[14] Leeb, H. and B.M. Pötscher (2005) Model Selection and Inference: Facts and Fiction. Econometric Theory 21, 21-59.

[15] 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.

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

[17] MacKenzie, D. (2017) A Material Political Economy: Automated Trading Desk and Price Prediction in High-Frequency Trading. Social Studies of Science 47, 172-194.

[18] Meinshausen, N. (2013) Sign-Constrained Least Squares Estimation for High-Dimensional Regression. Electronic Journal of Statistics 7, 1607-1631.

[19] Morariu-Patrichi, M. and M.S. Pakkanen (2018) State-Dependent Hawkes Processes and their Application to Limit Order Book Modelling. https://arxiv.org/abs/1809.08060.

[20] Sancetta, A. (2018) Estimation for the Prediction of Point Processes with Many Covariates. Econometric Theory 34, 598-627.

[21] Slawsky, M. and M. Hein (2013) Nonnegative Least Squares for High-Dimensional Linear Models: Consistency and Sparse Recovery without Regularization. Electronic Journal of Statistics 7, 3004-3056.

[22] van de Geer, S.A. and P. Bühlmann (2009) On the Conditions Used to Prove Oracle Results for the Lasso. Electronic Journal of Statistics 3, 1360–1392.

Supplementary Material to Estimation of a High Dimensional Counting Process Without Penalty for High Frequency Events" by L. Mucciante and A. Sancetta

A.1 Proofs

Mutatis mutandis, the proofs of Theorem 1 and 2 follow the arguments in Meinshausen (2013). There are notable differences, however. These are due to the fact that we consider a continuous time process.

We introduce some additional notation. Define the oracle estimator b^{oracle} to be the solution of the following minimization problem

$$\min_{b\geq 0: b_{S^c}=0} \left\{ -2\int_0^T X(t)' b dN(t) + \int_0^T \left(X(t)' b\right)^2 dt \right\}. \tag{A.1}$$

This is an oracle estimator because it assumes knowledge of S^c, the index set of zero coefficients.

Throughout, all vector equalities and inequalities are meant elementwise. Finally, we use the symbol $\lesssim$ when the left hand side (l.h.s.) is bounded by an absolute constant times the right hand side.

A.1.1 Preliminary Results

The next two lemmas will be useful in the sequel. For a martingale $Z = (Z(t))_{t\geq 0}$ let $\Delta Z(t) = Z(t) - Z(t-)$ be its jump and $\langle Z, Z\rangle_t$ its predictable quadratic variation, $t \geq 0$. We recall a classical Bernstein inequality for martingales (van de Geer, 1995, Lemma 2.1).

Lemma 1 *Let Z be a real valued locally square integrable martingale such that $Z(0) = 0$ and that $\max_{t\leq T}|\Delta Z(t)| \leq a$. Then for every $\epsilon > 0$ and $\Gamma > 0$,*

$$\Pr\left(\sup_{t\in[0,T]}|Z_t| > \epsilon \text{ and } \langle Z,Z\rangle_T \leq \Gamma\right) \leq 2\exp\left(-\frac{\epsilon^2}{2(a\epsilon+\Gamma)}\right).$$

Let 1_B be the indicator function of an arbitrary but measurable set B. Bernstein inequality implies the following maximal inequality (van der Vaart and Wellner, 2000, Lemma 2.2.10).

Lemma 2 *Let $K \in \mathbb{N}$ and $Z_1, \ldots, Z_K$ be arbitrary real valued random variables. Assume that for a measurable set B and some constants $a \geq 0$ and $\Gamma > 0$*

$$\Pr(|Z_i| > \epsilon \text{ and } B) \leq 2\exp\left(-\frac{\epsilon^2}{2(a\epsilon+\Gamma)}\right)$$

for any $\epsilon > 0$ and $i = 1, 2, ..., K$. Then, we have that

$$\mathbb{E}\left(\max_{1\leq i\leq K}|Z_i|\,1_B\right) \lesssim a\log(1+K) + \sqrt{\Gamma\log(1+K)}.$$

We find a bound for $\mathbb{E}\max_{1\leq i\leq K}\left|\int_0^T X_i(t)\,dM(t)\right|$, where X_i is the i^{th} entry in X. Its proof is based on the previous two lemmas.

Lemma 3 *Suppose that the Assumptions hold. If $\log(1+K) = O\left(T\bar{\lambda}\right)$, then*

$$\mathbb{E}\max_{1\leq i\leq K}\left|\int_0^T X_i(t)\,dM(t)\right| = O\left(\sqrt{\bar{\lambda}T\log(1+K)}\right). \tag{A.2}$$

Proof. The result is proved by an application of Lemma 2. Define the set

$$B := \left\{\max_{1\leq i\leq K}\left|\int_0^T X_i^2(t)\,\lambda^*(t)\,dt\right| \leq \Gamma\right\}$$

where Γ is a positive constant to be fixed in due course. It is easy to see that B is measurable. We shall use Lemma 1 to show that

$$\Pr\left(\left|\int_0^T X_i(t)\,dM(t)\right| > \epsilon \text{ and } B\right) \leq 2\exp\left(-\frac{\epsilon^2}{2(a\epsilon+\Gamma)}\right)$$

for every $i = 1, \ldots, K$, and $\epsilon > 0$, with $a = 1$ and $\Gamma = T\bar{\lambda}$. From the assumptions we have made $\int_0^t X_i(s)\,dM(s)$ is a locally square integrable martingale. In addition

$$\left|\int_{t-}^t X_i(s)\,dM(s)\right| = \int_{t-}^t X_i(s)\,dN(s) \leq N(t) - N(t-) \leq 1 \tag{A.3}$$

for every $i = 1, \ldots, K$, because the compensator is continuous and X_i takes values in $[0, 1]$. Moreover,

$$\max_{1\leq i\leq K}\left|\int_0^T X_i^2(t)\,\lambda^*(t)\,dt\right| \leq \int_0^T \lambda^*(t)\,dt \leq T\bar{\lambda}. \tag{A.4}$$

The predictable quadratic variation of $\int_0^t X_i(s)\,dM(s)$ is $\int_0^t X_i^2(s)\,\lambda^*(s)\,ds$. Taking into account (A.3) and (A.4), the hypotheses of Lemma 1 are met and we have that

$$\Pr\left(\left|\int_0^T X_i(t)\,dM(t)\right| > \epsilon \text{ and } B\right) \leq 2\exp\left(-\frac{\epsilon^2}{2\left(\epsilon+T\bar{\lambda}\right)}\right).$$

The above display allows us to apply Lemma 2 and obtain

$$\begin{aligned}\mathbb{E}\left(\max_{1\leq i\leq K}\left|\int_0^T X_i(t)\,dM(t)\right| 1_B\right) &\lesssim \log(1+K) + \sqrt{T\bar{\lambda}\log(1+K)} \\ &\lesssim \sqrt{T\bar{\lambda}\log(1+K)}\end{aligned} \tag{A.5}$$

using the fact that $\log(1+K) = O\left(T\bar{\lambda}\right)$. By (A.4), the event B has probability one. Then, the statement of the lemma follows from the above display. ■

Next, we show that b^{oracle} in (A.1) is close to b^*. The fact that we cannot rely on Gaussian distributional assumptions leads to a bound that is $\phi_{\min}^{-1/2}$ times the equivalent bound derived in Lemma 4 in Meinshausen (2013).

Proposition 1 *Suppose that the Assumptions hold. Then,* $\left\|b^* - b^{oracle}\right\|_1 = O_P\left(\sqrt{\frac{s^2\mu_T}{\phi_{\min}^2 T}}\right)$, *where* $\mu_T := \frac{1}{T}\int_0^T \mathbb{E}\lambda^*(t)\,dt$.

Proof. Recall that X_S is obtained by selecting the columns of X having index in S. Define the ordinary least square estimator

$$b_S^{OLS} := \left(\int_0^T X_S(t)\, X_S(t)'\, dt\right)^{-1}\left(\int_0^T X_S(t)\, dN(t)\right). \tag{A.6}$$

This is the solution of

$$\min_{b\in\mathbb{R}^s}\left\{-2\int_0^T X_S(t)'\, b dN(t) + \int_0^T \left(X_S(t)'\, b\right)^2 dt\right\}.$$

By the Eigenvalues Condition, b_S^{OLS} is well defined. Let $\lambda^{OLS} := X_S(t)'\, b_S^{OLS}$, then $\lambda^{oracle} := X(t)'\, b^{oracle}$ minimizes the following functional

$$\lambda \to \left\|\lambda^{OLS} - \lambda\right\|_{L_2}^2 := \int_0^T \left(\lambda^{OLS}(t) - \lambda(t)\right)^2 dt \tag{A.7}$$

among the functions $\lambda = X(t)'\, b$, where $b \geq 0$ and $b_{S^c} = 0$. This follows from the properties of linear projections. It can also derived directly if we show that the objective function in (A.1) equals (A.7) except for the term $\int_0^T \lambda^{OLS}(t)^2\, dt$ which does not depend on λ. Then, it is sufficient to show that $-2\int_0^T X(t)'\, b dN(t) = -2\int_0^T \lambda^{OLS}(t)\,\lambda(t)\, dt$.

To this end, for b such that $b_{S^c} = 0$,

$$\begin{aligned} & -2\int_0^T \lambda^{OLS}(t)\,\lambda(t)\,dt \\ = \; & -2\int_0^T b_S' X_S(t)\,X_S(t)'\,dt\left[\left(\int_0^T X_S(t)\,X_S(t)'\,dt\right)^{-1}\int_0^T X_S(t)\,dN(t)\right] \\ = \; & -2\int_0^T b_S' X_S(t)\,dN(t) \end{aligned}$$

where in the first equality we have used the fact that $\lambda(t) = X(t)'b = X_S(t)'b_S$, the definition of λ^{OLS}, and (A.6). The above display proves our claim. By these remarks and using the fact that λ^* is a feasible vector, we have that

$$\left\|\lambda^{OLS} - \lambda^{oracle}\right\|_{L_2}^2 \leq \left\|\lambda^{OLS} - \lambda^*\right\|_{L_2}^2. \tag{A.8}$$

Using (A.8) and the triangle inequality, we deduce that

$$\left\|\lambda^{oracle} - \lambda^*\right\|_{L_2} \leq 2\left\|\lambda^{OLS} - \lambda^*\right\|_{L_2}. \tag{A.9}$$

By the Doob-Meyer decomposition and (1), $dN(t) = X_S(t)'b_S^* dt + dM(t)$. Recall that b_S^* is the population parameter obtained by deleting the zero entries in b^* so that $\lambda^*(t) = X_S(t)'b_S^*$. Then, using the definition of b_S^{OLS} in (A.6), we find that

$$\begin{aligned} b_S^{OLS} &= \left(\int_0^T X_S(t)\,X_S(t)'\,dt\right)^{-1}\int_0^T X_S(t)\left(X_S(t)'b_S^* dt + dM(t)\right) \\ &= b_S^* + \left(\int_0^T X_S(t)\,X_S(t)'\,dt\right)^{-1}\int_0^T X_S(t)\,dM(t). \end{aligned}$$

In consequence, we have that $\left\|\lambda^{OLS}-\lambda^{*}\right\|_{L_2}^2$ is equal to

$$\begin{aligned}&\left\|X_S'\left(b_S^{OLS}-b_S^{*}\right)\right\|_{L_2}^2\\&=\left(\int_0^T X_S\left(t\right)' dM\left(t\right)\right)\left(\int_0^T X_S\left(t\right)X_S\left(t\right)' dt\right)^{-1}\left(\int_0^T X_S\left(t\right)dM\left(t\right)\right).\end{aligned}\tag{A.10}$$

Here, define $Z:=\frac{1}{\sqrt{T}}\int_0^T X_S\left(t\right)dM\left(t\right)$. By the Eigenvalues Condition, $\hat{\Sigma}_S^{-1}$ has maximal eigenvalue bounded by $\phi_{\min}^{-1}$, w.p.1. Hence, we deduce that (A.10) is equal to $Z'\hat{\Sigma}_S^{-1}Z=O_P\left(\phi_{\min}^{-1}Z'Z\right)$. Then, it is sufficient to bound $\mathbb{E}Z'Z$. To this end, using the isometry property of martingales,

$$\begin{aligned}\frac{\mathbb{E}Z'Z}{s}=&\frac{1}{s}\sum_{i\in S}\mathbb{E}\left(\frac{1}{\sqrt{T}}\int_0^T X_i\left(t\right)dM\left(t\right)\right)^2\\=&\frac{1}{s}\sum_{i\in S}\mathbb{E}\frac{1}{T}\int_0^T X_i^2\left(t\right)\lambda^{*}\left(t\right)dt\leq\mu_T,\end{aligned}$$

where the last inequality follows from the fact that the covariates are in $[0,1]$ and the definition of μ_T. In consequence we can bound (A.10) accordingly and deduce that

$$\left\|\lambda^{OLS}-\lambda^{*}\right\|_{L_2}^2 \;=\; O_P\left(s\mu_T\phi_{\min}^{-1}\right).\tag{A.11}$$

Recall that Assumption 2 implies the Compatibility Condition $\phi_{comp}^2\left(\hat{\Sigma},0,S\right)\geq\phi_{\min}$ (see the remarks on Assumption 2 in Section 2.3). Since $b_{S^c}^{oracle}-b_{S^c}^{*}=0$, by the aforementioned Compatibility Condition, we find that

$$\begin{aligned}\left\|\lambda^{oracle}-\lambda^{*}\right\|_{L_2}^2 &= \left(b^{oracle}-b^{*}\right)'\int_0^T X\left(t\right)X\left(t\right)'dt\left(b^{oracle}-b^{*}\right)\\&\geq \frac{T}{s}\phi_{\min}\left\|b^{oracle}-b^{*}\right\|_1^2.\end{aligned}\tag{A.12}$$

Putting together (A.9), (A.11) and (A.12) we deduce the statement of the proposition.

■

The next step is to prove that $\hat{b}$ is close to b^{oracle}. Mutatis mutandis, this is equivalent to Meinshausen (2013, eq.(11)). However, we have the extra factor $\bar{\lambda}$ because we cannot rely on a Gaussian distributional assumption.

Proposition 2 *Suppose that the Assumptions hold. Then,*

$$\left\|\hat{b}-b^{oracle}\right\|_1 = O_P\left(\sqrt{\frac{c\left(s\right)\left(\mu_T s^2 \phi_{\min}^{-2}+\bar{\lambda}\log\left(1+K\right)\right)}{T}}\right)$$

where $c\left(s\right):=\max\left\{\frac{s^2}{\phi^2},\frac{1}{\nu}\right\}$ *and* μ_T *is as in Proposition 1.*

Proof. By definition of $\hat{b}$, $\hat{\delta}:=\hat{b}-b^{oracle}$ solves

$$\min_{\delta\in\mathbb{R}^K}\left\{-2\int_0^T X\left(t\right)'\left(b^{oracle}+\delta\right)dN\left(t\right)+\int_0^T\left(X\left(t\right)'\left(b^{oracle}+\delta\right)\right)^2 dt\right\}$$

such that $\delta+b^{oracle}\geq 0$. The above display is equivalent to Meinshausen (2013, eq.(9)). The zero vector is a feasible solution of the above problem. Then, it holds that

$$\begin{aligned}&-2\int_0^T X\left(t\right)'\left(b^{oracle}+\hat{\delta}\right)dN\left(t\right)+\int_0^T\left(X\left(t\right)'\left(b^{oracle}+\hat{\delta}\right)\right)^2 dt\\ \leq\ &-2\int_0^T X\left(t\right)'b^{oracle}dN\left(t\right)+\int_0^T\left(X\left(t\right)'b^{oracle}\right)^2 dt.\end{aligned}$$

Expanding the square, we have that

$$\begin{aligned}&-2\int_0^T X\left(t\right)'\left(b^{oracle}+\hat{\delta}\right)dN\left(t\right)\\ &+\int_0^T\left[\left(X\left(t\right)'b^{oracle}\right)^2+\left(X\left(t\right)'\hat{\delta}\right)^2+2\hat{\delta}'X\left(t\right)X\left(t\right)'b^{oracle}\right]dt\\ \leq&-2\int_0^T X\left(t\right)'b^{oracle}dN\left(t\right)+\int_0^T\left(X\left(t\right)'b^{oracle}\right)^2 dt.\end{aligned}$$

By simple algebra, the above display implies that

$$-2\int_0^T X(t)'\hat{\delta}dN(t)+\int_0^T\left[\left(X(t)'\hat{\delta}\right)^2+2\hat{\delta}'X(t)X(t)'b^{oracle}\right]dt\leq 0.$$

Adding and subtracting $2\int_0^T\hat{\delta}'X(t)X(t)'b^*dt$, and rearranging the terms, we deduce that

$$\begin{aligned}\int_0^T\left(X(t)'\hat{\delta}\right)^2dt &\leq 2\int_0^T X(t)'\hat{\delta}dM(t)\\ &\quad+2\hat{\delta}'\int_0^T X(t)X(t)'\left(b^*-b^{oracle}\right)dt. \qquad (A.13)\end{aligned}$$

We start controlling the r.h.s. of (A.13). For the first term, using Lemma 3, we have that

$$\begin{aligned}\int_0^T X(t)'\hat{\delta}dM(t) &= \sum_{i=1}^K\int_0^T X_i(t)dM(t)dt\hat{\delta}_i\\ &\leq \max_{1\leq i\leq K}\left|\int_0^T X_i(t)dM(t)dt\right|\left\|\hat{\delta}\right\|_1\\ &= O_P\left(\sqrt{\bar{\lambda}T\log(1+K)}\left\|\hat{\delta}\right\|_1\right) \qquad (A.14)\end{aligned}$$

while the second term can be bounded as follows

$$\begin{aligned}\int_0^T\hat{\delta}'X(t)X(t)'\left(b^*-b^{oracle}\right)dt &\leq T\sum_{i,j=1}^K\left|\hat{\delta}_i\right|\left|b_j^*-b_j^{oracle}\right|\\ &= T\left\|\hat{\delta}\right\|_1\left\|b^*-b^{oracle}\right\|_1 \qquad (A.15)\end{aligned}$$

because $\int_0^T X_i(t)X_j(t)dt\leq T$.

Hence, inserting (A.14) and (A.15) in (A.13) and using Proposition 1, we deduce

that

$$\frac{1}{T}\int_0^T \left(X(t)'\,\hat{\delta}\right)^2 dt = O_P\left(\left\|\hat{\delta}\right\|_1 \left[\sqrt{\frac{\bar{\lambda}\log(1+K)}{T}} + \sqrt{\frac{s^2\mu_T}{\phi_{\min}^2 T}}\right]\right). \tag{A.16}$$

Now, we find a lower bound for the l.h.s. of the above display. Mutatis mutandis, in the remainder of the proof, we follow Meinshausen (2013, p.1625-1626). Set $D := \left\{i \leq K : \hat{\delta}_i < 0\right\}$ and its complement $D^c := \left\{i \leq K : \hat{\delta}_i \geq 0\right\}$. (In Meinshausen, 2013, these sets are denoted by M and M^c, respectively.) By definition $D \subseteq S$ and in consequence $S^c \subseteq D^c$. To see this, note that $\hat{\delta}_i < 0$ implies $0 \leq \hat{b}_i < b_i^{oracle}$ because $\hat{b}_i, b_i^{oracle} \geq 0$. We consider the following two complementary cases: $||\hat{\delta}_{D^c}||_1 \geq \frac{3}{\sqrt{\nu}}||\hat{\delta}_D||_1$ and $||\hat{\delta}_{D^c}||_1 < \frac{3}{\sqrt{\nu}}||\hat{\delta}_D||_1$.

Case I: $||\hat{\delta}_{D^c}||_1 \geq \frac{3}{\sqrt{\nu}}||\hat{\delta}_D||_1$**.** We have that

$$\begin{aligned}\hat{\delta}'\hat{\Sigma}\hat{\delta} &= \sum_{i,j\in D}\hat{\delta}_i\hat{\Sigma}_{ij}\hat{\delta}_j + \sum_{i,j\in D^c}\hat{\delta}_i\hat{\Sigma}_{ij}\hat{\delta}_j + 2\sum_{i\in D, j\in D^c}\hat{\delta}_i\hat{\Sigma}_{ij}\hat{\delta}_j \\ &\geq \sum_{i,j\in D^c}\hat{\delta}_i\hat{\Sigma}_{ij}\hat{\delta}_j + 2\sum_{i\in D, j\in D^c}\hat{\delta}_i\hat{\Sigma}_{ij}\hat{\delta}_j\end{aligned}$$

because $\sum_{i,j\in D}\hat{\delta}_i\hat{\Sigma}_{ij}\hat{\delta}_j \geq 0$. By the Cauchy–Schwarz inequality

$$\begin{aligned}\left|\sum_{i\in D, j\in D^c}\hat{\delta}_i\hat{\Sigma}_{ij}\hat{\delta}_j\right| &\leq \left(\sum_{i,j\in D}\hat{\delta}_i\hat{\Sigma}_{ij}\hat{\delta}_j\right)^{1/2}\left(\sum_{i,j\in D^c}\hat{\delta}_i\hat{\Sigma}_{ij}\hat{\delta}_j\right)^{1/2}. \\ &\leq \left\|\hat{\delta}_D\right\|_1\left(\sum_{i,j\in D^c}\hat{\delta}_i\hat{\Sigma}_{ij}\hat{\delta}_j\right)^{1/2}\end{aligned}$$

where we used the fact that $\hat{\Sigma}_{ij} \in [0,1]$ in the last step. By the above two displays, we deduce that

$$\hat{\delta}'\hat{\Sigma}\hat{\delta} \;\geq\; \sum_{i,j\in D^c} \hat{\delta}_i\hat{\Sigma}_{ij}\hat{\delta}_j - 2\left(\sum_{i,j\in D^c} \hat{\delta}_i\hat{\Sigma}_{ij}\hat{\delta}_j\right)^{1/2} \left\|\hat{\delta}_D\right\|_1 . \tag{A.17}$$

We now use the fact that $||\hat{\delta}_{D^c}||_1 \geq \frac{3}{\sqrt{\nu}}||\hat{\delta}_D||_1$ and the Positive Eigenvalue Condition, so that (A.17) becomes

$$\hat{\delta}'\hat{\Sigma}\hat{\delta} \;\geq\; \nu\left\|\hat{\delta}_{D^c}\right\|_1^2 - 2\frac{\nu}{3}\left\|\hat{\delta}_{D^c}\right\|_1^2 \geq \frac{\nu}{3}\left\|\hat{\delta}_{D^c}\right\|_1^2.$$

Multiplying and dividing by $\left(1+\frac{\sqrt{\nu}}{3}\right)^2$ and using the assumed inequality $||\hat{\delta}_{D^c}||_1 \geq \frac{3}{\sqrt{\nu}}||\hat{\delta}_D||_1$, we find that

$$\begin{aligned}\hat{\delta}'\hat{\Sigma}\hat{\delta} \;&\geq\; \frac{\nu}{3\left(1+\frac{\sqrt{\nu}}{3}\right)^2}\left(\left(1+\frac{\sqrt{\nu}}{3}\right)\left\|\hat{\delta}_{D^c}\right\|_1\right)^2 \\ &\gtrsim\; \nu\left(\left\|\hat{\delta}_{D^c}\right\|_1 + \left\|\hat{\delta}_D\right\|_1\right)^2 = \nu\left\|\hat{\delta}\right\|_1^2.\end{aligned}$$

Note that we can assume $\nu \leq 1$. Using the above display, together with (A.16) we conclude that $||\hat{\delta}||_1 = O_P\left(\sqrt{\frac{\bar{\lambda}\log(1+K)}{\nu^2 T}} + \sqrt{\frac{s^2\mu_T}{\nu^2\phi_{\min}^2 T}}\right)$.

Case II: $||\hat{\delta}_D||_1 > \frac{\sqrt{\nu}}{3}||\hat{\delta}_{D^c}||_1$**.** Note that $S^c \subseteq D^c$, so that

$$||\hat{\delta}_{S^c}||_1 \leq ||\hat{\delta}_{D^c}||_1 \leq \frac{3}{\sqrt{\nu}}||\hat{\delta}_D||_1 \leq \frac{3}{\sqrt{\nu}}||\hat{\delta}_S||_1.$$

Hence, we have shown that $\hat{\delta} \in \mathcal{R}(\frac{3}{\sqrt{\nu}}, S)$. We apply the Compatibility Condition and deduce that $\hat{\delta}'\hat{\Sigma}\hat{\delta} \geq (\phi/s)||\hat{\delta}||_1^2$. Using again (A.16) we have that

$$||\hat{\delta}||_1 = O_P\left(\sqrt{\frac{s^2\bar{\lambda}\log(1+K)}{\phi^2 T}} + \sqrt{\frac{s^4\mu_T}{\phi^2\phi_{\min}^2 T}}\right).$$

Defining $c(s) = \max\left\{\frac{s^2}{\phi^2}, \frac{1}{\nu}\right\}$, and using the basic inequality $(x+y)^2 \leq 2(x^2+y^2)$ for any x, y, we deduce the statement of the proposition. ■

A.1.2 Proof of Theorems 1 and 2

Proof of Theorem 1. By the triangle inequality, we find that

$$\left\|\hat{b} - b^*\right\|_1 \leq \left\|\hat{b} - b^{oracle}\right\|_1 + \left\|b^{oracle} - b^*\right\|_1.$$

By Propositions 1 and 2, we see that $\left\|b^* - b^{oracle}\right\|_1 = o_P\left(\left\|\hat{b} - b^{oracle}\right\|_1\right)$. Using Proposition 2 we then obtain the bound of the theorem.

Proof of Theorem 2. By a basic inequality

$$\begin{aligned}\frac{1}{T}\int_0^T \left(X(t)'\hat{b} - X(t)'b^*\right)^2 dt \leq & \frac{2}{T}\int_0^T \left(X(t)'\hat{b} - X(t)'b^{oracle}\right)^2 dt \\ & + \frac{2}{T}\int_0^T \left(X(t)'b^{oracle} - X(t)'b^*\right)^2 dt. \qquad \text{(A.18)}\end{aligned}$$

By (A.16), and a basic inequality, we find that

$$\int_0^T \left(X(t)'\hat{b} - X(t)'b^{oracle}\right)^2 dt = O_p\left(\sqrt{\frac{\left(s^2\mu_T\phi_{\min}^{-2} + \bar{\lambda}\log(1+K)\right)}{T}}\left\|\hat{\delta}\right\|_1\right).$$

Theorem 1 gives a bound for $\left\|\hat{\delta}\right\|_1$ so that the r.h.s. of the above display is bounded above in probability by a constant multiple of

$$\frac{c^{1/2}(s)\left(s^2\mu_T\phi_{\min}^{-2}+\bar{\lambda}\log(1+K)\right)}{T}. \tag{A.19}$$

Given the fact that the covariates take values in $[0,1]$, and that $\left\|b^{oracle}-b^*\right\|_2^2 \leq \left\|b^{oracle}-b^*\right\|_1^2$, the second term on the r.h.s. of (A.18) is $O_P\left(\frac{s^2\mu_T}{\phi_{\min}^2 T}\right)$ because of Proposition 1. By these remarks, we deduce that the l.h.s. of (A.18) is bounded above by a quantity of the same order of magnitude as (A.19), and this proves the result.

A.1.2.1 Proof of Corollary 2

Under the conditions of the corollary, the set $\left\{\hat{S}\subset S\right\}$ is the same as the set difference of

$$\left\{\hat{b}_i=0 \text{ and } b_i^*>\kappa \text{ for at least one } i\leq K\right\}$$

and

$$\left\{\hat{b}_i>0 \text{ and } b_i^*\leq\kappa \text{ for at least one } i\leq K\right\}.$$

This set difference is contained in $\left\{\max_{i\leq K}\left|\hat{b}_i-b_i^*\right|>\kappa\right\}$. Bounding the maximum by the sum, the result is proved if $\Pr\left(\left\|\hat{b}-b^*\right\|_1>\kappa\right)\to 0$. Noting that $\mu_T\leq\bar{\lambda}$, this is the case by Theorem 1 and the choice of κ. This shows the inclusion.

Under the conditions on b^* and the definition of S_ϵ, the event $\left\{\hat{S}_\epsilon\neq S\right\}$ is contained in the union of the events

$$\left\{\hat{b}_i\leq\epsilon \text{ and } b_i^*>\kappa \text{ for at least one } i\leq K\right\} \tag{A.20}$$

and

$$\left\{\hat{b}_i>\epsilon \text{ and } b_i^*=0 \text{ for at least one } i\leq K\right\}. \tag{A.21}$$

By the same argument used in the proof of the first result, the event in (A.20) is contained in $\left\{\left\|\hat{b}-b^*\right\|_1 > \kappa - \epsilon\right\}$. The probability of this latter event goes to zero if $\kappa/\epsilon \to c > 1$ as required, given the conditions on κ and ϵ. The event in (A.21) is contained in $\left\{\left\|\hat{b}-b^*\right\|_1 > \epsilon\right\}$ and this also goes to zero under the condition on ϵ.

A.1.3 Proof of Theorem 3

We first need a result on convergence in distribution.

Lemma 4 *Let N be a point process with predictable intensity λ^* bounded above by a constant $\bar{\lambda} > 0$. Suppose that $Z = (Z(t))_{t\in[0,T]}$ is a predictable stochastic process such that $\mathbb{E}\frac{1}{T}\int_0^T |Z(t)|^2 \lambda^*(t)\,dt \to 1$ and $\max_{t\in[0,T]} |Z(t)|^4$ is bounded above by a quantity $z_{4,T} := o\left(T/\bar{\lambda}\right)$. Then,*

$$\frac{1}{\sqrt{T}}\int_0^T Z(t)\,dM(t) \to \mathcal{N}(0,1),$$

in distribution, where $\mathcal{N}(0,1)$ is the standard normal distribution.

Proof. Let $\Delta_i = ((i-1)/\bar{\lambda}, i/\bar{\lambda}]$, $i = 1,2,...,n$ for some integer n. To avoid trivialities suppose that $n = \bar{\lambda}T$. Then,

$$\frac{1}{\sqrt{T}}\int_0^T Z(t)\,dM(t) = \frac{1}{\sqrt{n}}\sum_{i=1}^n Y_i \tag{A.22}$$

where $Y_i := |\Delta_i|^{-1/2}\int_{\Delta_i} Z(t)\,dM(t)$. By construction $\mathbb{E}_{i-1}Y_i = 0$, where $\mathbb{E}_{i-1}$ is the expectation conditioning on $(Y_j)_{j\leq i-1}$. By assumption, we have that $\mathbb{E}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^n Y_i\right)^2 \to 1$ using the standard isometry for square integrable martingales. For the martingale central limit theorem to apply to (A.22), it is sufficient that $\sum_{i=1}^n \mathbb{E}\,|Y_i/\sqrt{n}|^2\, 1\{|Y_i/\sqrt{n}| > \epsilon\} \to 0$ for any $\epsilon > 0$. By Holder's inequality and Markov inequality, this is clearly implied by $\sum_{i=1}^n \mathbb{E}\,|Y_i/\sqrt{n}|^4 \to 0$. By the Burkholder, Davis, Gundy inequality (Kallenberg,

1997, Theorem 23.12), $\mathbb{E}\left|\int_{\Delta_i} Z(t)\,dM(t)\right|^4 \lesssim \mathbb{E}\left(\int_{\Delta_i} |Z(t)|^2\,dN(t)\right)^2$. We also have that $\int_{\Delta_i} |Z(t)|^2\,dN(t) \leq \sup_{t>0}|Z(t)|^2 \int_{\Delta_i} dN(t)$. These two remarks imply that

$$\mathbb{E}\left|\frac{Y_i}{\sqrt{n}}\right|^4 \lesssim \frac{z_{4,T}}{n^2|\Delta_i|^2}\mathbb{E}\left[\int_{\Delta_i} dN(t)\right]^2.$$

We use the fact that the intensity is bounded above by $\bar{\lambda}$. Then, we see that $\mathbb{E}\left[\int_{\Delta_i} dN(t)\right]^2 \leq |\Delta_i|\bar{\lambda} + \left(|\Delta_i|\bar{\lambda}\right)^2 \leq 2$, using an upper bound in terms of a Poisson random variable with intensity $|\Delta_i|\bar{\lambda} = 1$. By assumption, $z_{4,T} = o\left(T/\bar{\lambda}\right)$. In consequence the above display is $o(n^{-1})$ because by construction, $n|\Delta_i| = T$ and $n = T\bar{\lambda}$. Hence, we have shown that $\sum_{i=1}^n \mathbb{E}|Y_i/\sqrt{n}|^4 = o(1)$ and the lemma is proved. ■

Proof of Theorem 3. By Corollary 2, the event $\left\{\hat{S}_\epsilon = S\right\}$ has probability going to one uniformly in b^*. We can then derive the result on this set only, with no further mention. Hence, we have that $\alpha'\left(b^{OLS} - b^*\right) = \alpha_S'\left(b_S^{OLS} - b_S^*\right)$. By the Doob-Meyer decomposition for N we have that the r.h.s. is equal to $\alpha_S\left(\int_0^T X_S(t)\,X_S(t)'\,dt\right)^{-1}\left(\int_0^T X_S(t)\,dM(t)\right)$. By Assumption 5, $\hat{\Sigma}_S$ converges to $\mathbb{E}\hat{\Sigma}_S$ in expected Frobenius norm. Hence, by Assumption 2, $\mathbb{E}\hat{\Sigma}_S$ has minimum eigenvalue greater than some nonzero constant multiple of $\phi_{\min}$. We shall apply Lemma 4 to $\frac{1}{\sqrt{T}}\int_0^T Z(t)\,dM(t)$, where $Z(t) := \alpha_S'\left(\mathbb{E}\hat{\Sigma}_S\right)^{-1} X_S(t)/\sigma_\alpha$. By construction, $\mathbb{E}\frac{1}{T}\int_0^T |Z(t)|^2\lambda^*(t)\,dt = 1$. Hence, we only need to check that $\max_{t\in[0,T]}|Z(t)|^4 = o\left(T/\bar{\lambda}\right)$. To ease notation, define $A := \alpha_S\left(\mathbb{E}\hat{\Sigma}_S\right)^{-1}$. By direct calculation, using the fact that $X_S(t) \leq 1_s$ elementwise, where 1_s is the s-dimensional column vector of ones,

$$\sigma_\alpha^4|Z(t)|^4 \leq \text{Trace}\left(A1_s1_s'A'\right)^2 \leq \text{Trace}\left(\left(1_s1_s'\right)^2\right)\text{Trace}\left(\left(AA'\right)^2\right),$$

as the trace of a scalar is equal to the scalar, then using the properties of traces, and the Cauchy-Schwarz inequality for traces. Clearly, $\text{Trace}\left(\left(1_s1_s'\right)^2\right) = s^2$. Given that

$\alpha'\alpha = 1$, $AA' = \alpha_S' \left(\mathbb{E}\hat{\Sigma}_S\right)^{-2} \alpha_S$ is bounded above by the reciprocal of the squared minimum eigenvalue of $\mathbb{E}\hat{\Sigma}_S$. By this remark, we know that $AA' = O\left(\phi_{\min}^{-2}\right)$. Hence, the r.h.s. of the above display is bounded above by a constant multiple of $\phi_{\min}^{-2} s^2$. It is easy to show that $\sigma_\alpha^4 > 0$, noting that $\mathbb{E}\hat{\Sigma}_S^N \geq \mathbb{E}\hat{\Sigma}_S \inf_t \lambda^*(t)$ elementwise. Then, $\sigma_\alpha^4 \gtrsim \alpha_S' \left(\mathbb{E}\hat{\Sigma}_S\right)^{-1} \alpha_S$. By assumption, the r.h.s. is asymptotically bounded away from zero. By these remarks, we conclude that $Z(t)^4 \lesssim \phi_{\min}^{-2} s^2$. By the constraint on s, the conditions of Lemma 4 are satisfied so that $T^{-1/2} \int_0^T Z(t)\, dM(t)$ converges to a standard normal random variable. The fact that $\hat{\sigma}_\alpha^2 \to \sigma_\alpha^2$ is probability essentially follows by Assumption 5.

A.2 Additional Details on Data and Covariates Definition

Here, we give additional details regarding Section 3. The trades were accurately classified as buy or sell. During busy times, when many trades are executed, CME might not send the resulting book update for some time as there is a limit in the size of each packet being sent through the network. For this reason, if a trade arrives and the book is not updated, we construct an imputed book. Again this operation is admissible (was carried out in live trading) and avoids any bias due to lack of synchronicity. Finally, we also subtract 400 microseconds from trade times in order to account for some delay on the side of CME when sending trade messages as opposed to order book messages. We do so to avoid the risk of asynchronicity and consequently spurious relations. This approach matched closely live trading. To summarize, the data processing and variables construction is the same as in live trading to ensure that we do not induce any forward looking bias.

We now provide additional details for the definition of the covariates. A signed

trade is defined to be the trade size times either one if the trade price is greater than or equal to the mid price immediately preceding the trade, or minus one otherwise. The variable TrdImb98 is computed as follows. Let

$$\text{TrdImb98}\left(t_i\right) := \begin{cases} \frac{EWMA(\text{signedTradedVolume}(t_i))}{EWMA(\text{tradedVolume}(t_i))} & \text{if } t_i \text{ is a trade update} \\ \text{TrdImb98}\left(t_{i-1}\right) & \text{otherwise} \end{cases}$$

where the EWMA's are as in (9) with parameter $\alpha = 0.98$. Both signed traded volumes and traded volumes are updated only when a trade is reported. The EWMA is computed and updated only at these event times. When using trade variables as covariates, we do not adjust their timestamp by 400 microseconds in order to ensure that they can only be used once received, as in live trading. Note that if t_i is not an update for the trade imbalance, we just report the last available value of the trade imbalance. A similar approach is applied to the durations.

The duration variables are in nanosecond resolution with nanoseconds as decimals. Hence to map durations in $[0, 1]$ we cap them at one second.

We compute the spread in ticks and cap it at 4 ticks. We also force the spread to take the minimum value of one tick. This is because a spread equal to zero is not a tradable event. We then map this spread into $[0, 1]$ dividing it by the cap, which is 4. In consequence, the transformed spread variable only takes values in $\{0.25, .5, 0.75, 1\}$.

A.3 Finite Sample Analysis via Simulations

We present simulation results to gain further understanding of the procedure in a finite sample. Recall that $\hat{b}$ and b^* are the estimated parameter and the true parameter, respectively. The goodness of fit of the estimator is measured via four statistics.

Relative ℓ_2 error (Norm2). The Monte Carlo approximation of the ℓ_2 norm of the relative error: $\left\|b^* - \hat{b}\right\|_2 / \left\|b^*\right\|_2$.

Relative ℓ_1 error (Norm1). The Monte Carlo approximation of the ℓ_1 norm of the relative error: $\left\|b^* - \hat{b}\right\|_1 / \left\|b^*\right\|_1$.

Relative ℓ_0 error (Norm0). The Monte Carlo approximation of the ℓ_0 norm of the relative error: $\left\|b^* - \hat{b}\right\|_0 / \left\|b^*\right\|_0$. The ℓ_0 norm $\|\cdot\|_0$ is the number of nonzero coefficients.

False Positives (FP). The number of coefficients estimated to be strictly positive when the true ones are zero, i.e. false positives.

False Negatives (FN). The number of coefficients estimated to be zero when the true ones are strictly positive, i.e. false negatives.

For FP and FN, a generic entry of the vector $\hat{b}$, say $\hat{b}_i$, is set equal to zero if $\hat{b}_i < 10^{-5}$.

A.3.1 The True Model

The true model is given by $\lambda^*(t) = X(t)' b^*$ where the first s entries of b^* are equal to 10 and zero otherwise. The number of active covariates is $s = 10$. The covariates are assumed to be constant between two consecutive jumps of the counting process N. The covariates process is given by

$$X(T_j) = \alpha X(T_{j-1}) + (1-\alpha) U_j \tag{A.23}$$

where α is a scalar and U_j is uniformly distributed in $[0,1]^K$, with Gaussian copula with scaling parameter R. Each of the entries in the process in (A.23) has expectation $1/2$ for any $\alpha \in [0,1)$. We set $X(T_1) = U_1$.

To simulate the durations $\{(T_j - T_{j-1}) : j = 1, 2, ..., n\}$ of the counting process N, we note that for $j = 1, 2, ..., n$,

$$\int_{T_{j-1}}^{T_j} \lambda(t)\, dt = \int_{T_{j-1}}^{T_j} X(t)' b^* dt, \tag{A.24}$$

are i.i.d. exponential random variable with mean one (Brémaud, 1981, Chapter II, Theorem 16). In our case (A.24) and (A.23) mean that $\left[X(T_{j-1})' b^*\right](T_j - T_{j-1})$ is an exponential random variable with mean equal to one.

Monte Carlo approximations are derived using 250 simulations. Table A.1 shows the results for $\alpha \in \{0, 0.9\}$, $n \in \{10^3, 10^4, 10^5\}$, $K = 1000$, $s = 10$ and three different dependence structures for the covariates. In particular we consider: $R = I$ (uncorrelated design), R having (i, j) entry equal to $\rho^{|i-j|}$ (Toeplitz design), and $R = I + \rho(1_K 1_K' - I)$ (equicorrelated design). Here I is the K-dimensional identity matrix, 1_K is the K-dimensional column vector of ones, and $\rho = 0.9$. As we expected, smaller values of K/n and s/n correspond to smaller errors (Norm2, Norm1, Norm2, FP, FN). In the uncorrelated case the results are, in general, better than either equicorrelated case or Toeplitz design, as expected. When the covariates are uncorrelated, it is less difficult to identify the active covariates. However, given that the first $s = 10$ covariates are active, a decaying correlation among the covariates (Toeplitz design) seems beneficial especially in terms of FN and FP errors. Conversely, the equicorrelated design makes prediction harder as covariates are confounded. Finally, as expected, an increase in time series dependence in (A.23) is associated to higher errors.

Table A.1: Simulation Results. Results for different designs are reported using the same notation as in the text. The different correlation structures (corr.) are none for $\rho = 0$, equi for the equicorrelated case, and toep for the Toeplitz structure. The total number of covariates and the number of active ones are fixed to $K = 1000$ and $s = 10$, respectively. For each design, the first row reports the mean and the second the standard error.

(corr, α, n)	Norm2	Norm1	Norm0	FP	FN
(none, 0.00, 1000)	1.00	1.57	4.60	35.98	4.59
	0.01	0.01	0.03	0.35	0.10
(none, 0.00, 10000)	0.14	0.56	5.16	41.58	0.00
	0.00	0.01	0.04	0.42	0.00
(none, 0.00, 100000)	0.01	0.17	5.26	42.65	0.00
	0.00	0.00	0.05	0.47	0.00
(none, 0.90, 1000)	2.37	1.97	2.28	12.80	9.63
	0.03	0.01	0.02	0.19	0.03
(none, 0.90, 10000)	1.29	1.75	4.09	30.86	7.01
	0.01	0.01	0.03	0.32	0.09
(none, 0.90, 100000)	0.29	0.83	5.42	44.22	0.08
	0.00	0.01	0.05	0.48	0.02
(toep, 0.00, 1000)	0.82	0.83	1.66	6.56	2.55
	0.02	0.01	0.02	0.15	0.07
(toep, 0.00, 10000)	0.16	0.34	1.62	6.24	0.08
	0.01	0.01	0.02	0.15	0.02
(toep, 0.00, 100000)	0.02	0.11	1.63	6.29	0.00
	0.00	0.00	0.02	0.16	0.00
(toep, 0.90, 1000)	2.33	1.65	1.74	7.42	7.79
	0.05	0.01	0.02	0.17	0.07
(toep, 0.90, 10000)	1.19	1.12	1.89	8.85	4.39
	0.03	0.01	0.02	0.21	0.07
(toep, 0.90, 100000)	0.39	0.58	2.01	10.12	0.90
	0.01	0.01	0.02	0.24	0.05
(equi, 0.00, 1000)	1.77	1.92	3.05	20.48	9.04
	0.01	0.01	0.02	0.24	0.06
(equi, 0.00, 10000)	0.71	1.30	5.04	40.36	2.32
	0.01	0.01	0.04	0.42	0.08
(equi, 0.00, 100000)	0.09	0.44	5.47	44.74	0.00
	0.00	0.00	0.05	0.47	0.00
(equi, 0.90, 1000)	4.26	1.99	2.34	13.38	9.84
	0.09	0.00	0.56	5.56	0.06
(equi, 0.90, 10000)	2.27	1.96	2.41	14.11	9.65
	0.02	0.01	0.02	0.20	0.04
(equi, 0.90, 100000)	1.24	1.72	4.34	33.41	6.63
	0.01	0.01	0.03	0.35	0.09

A.4 The Effect of Directional Misspecification

Consider the true intensity $\lambda^*(t) = X(t)' b^*$ where b^* can have negative entries. Given that the covariates take values in $[0,1]$, we can ensure that the intensity is positive as long as there is an intercept whose coefficient is at least as large as the sum of the negative coefficients (see also the remarks on Condition 1 at the start of Section 2.3). Assuming a constant limit $\Sigma := \lim_T \mathbb{E}\hat{\Sigma}$ exists, we can add and subtract $b^{*\prime}\Sigma b^*$ in (3), use the definition of λ^*, and complete the square. By this remark, we can deduce that the minimizer of (3) is the solution of

$$\inf_{b \geq 0} (b - b^*)' \Sigma (b - b^*). \tag{A.25}$$

We denote the solution by $\tilde{b}$. This is not guaranteed to satisfy $\tilde{b}_i b_i^* \geq 0$, $i = 1, 2, ..., K$. The latter condition implies that the sign constraint never results in a variable $\tilde{b}_i > 0$ when $b_i^* \leq 0$. We carried out a number of numerical examples to see under what conditions we can expect $\tilde{b}_i b_i^* \geq 0$, $i = 1, 2, ..., K$. At a high level, for $\tilde{b}_i b_i^* \geq 0$ $i = 1, 2, ..., K$ to be satisfied, we need sparsity in the sense that the cardinality of S is small relative to K and the number of negative coefficients.

We consider $\Sigma = T_n^{-1} \int_0^{T_n} X(t) X(t)' dt$. Recall that T_n is the time such that $N(T_n) = n$. Note this is just a method to construct the matrix Σ. Hence, Σ is regarded as a population quantity for the purpose of this section. Here, X is as in (A.23) with $\alpha = 0$. We are not interested in ancillary quantities such as α. We are using X as a way to construct different designs for Σ. A small n allows us to assess results when Σ is nearly singular. Except for restricting $\alpha = 0$, X is constructed as in Section A.3. We use different values for b^*. Let K_N denote the cardinality of $\{i \leq K : b_i^* < 0\}$. We set the first s entries in b^* to be positive. Entries $s+1$ to $s+K_N$ are set to negative numbers, while the remaining entries are set to zero. The absolute values of the en-

tries in b^* are chosen to satisfy different designs. We consider three designs: random values equal to the absolute value of standard normal random variables (gauss); fixed values equal to 1 (equal); fixed values equal to 1 for positive and 10 for negative values (skewed). We shall refer to these designs with the name given in the parenthesis.

To ensure that $X(t)' b^*$ results in a bona fide intensity for all $t \geq 0$, under possibly negative b^* entries, we impose some additional constraints in the construction of X, as well as b^*. We let the first entry in X be a constant. This is tantamount to ensuring that there is an intercept. Given that the first entry in X is a constant and the covariates take values in the unit interval, we let $b_1^* = 10^{-5} - \sum_{i=s+1}^{s+K_N} b_i^*$. Recall that $b_i^* < 0$, $i = s+1, s+2, ..., s+K_N$. This means that the intensity is uniformly bounded below by 10^{-5}. This argument follows from the remarks at the start of this section.

Given that Σ is randomly generated and for some designs also b^*, we carry out 1000 simulation for each design. Each time we compute the following statistics.

True discovery rate (TDR). We define this to be $\frac{\left|\left\{i \leq K : b_i^* > 0 \text{ and } \tilde{b}_i > 0)\right\}\right|}{\left|\left\{i \leq K : b_i^* > 0\right\}\right|}$. Recall that for a set A, $|A|$ is its cardinality. In population, the true discovery rate is always 1. However, the effect of the constraint under misspecification can lead to a lower true discovery rate in population.

Average Sign Coherence (ASC). We define this to be $\left|\left\{i \leq K : \tilde{b}_i b_i^* \geq 0)\right\}\right| / K$. Note that $\tilde{b}_i b_i^* < 0$ only if $\tilde{b}_i > 0$ and $b_i^* < 0$ because $\tilde{b}_i \geq 0$ due to the constraint. This is a weaker requirement than TDR. However, it is an important one. We would like variables that have negative coefficient not to be selected in the population.

The results show that we may expect lower ASC as we increase either s or the number of misspecified signed variables. Increasing the dependence reduces ASC. Finally, we also note that when the entries in b^* are random, despite the regularity in the entries of Σ, it is more likely to obtain an ASC less than one. These remarks apply to the case

when Σ can be nearly singular, i.e. $K = n = 100$. When $n = 1000$, the ASC is equal to one for most designs. Table A.2 reports the results.

A.4.1 Challenges Beyond Numerical Illustration

We look at the Karush-Kuhn-Tucker conditions for (A.25) to improve our understanding of the problem and relate to the results in Table A.2. With no loss of generality, suppose that the coefficients in b^* are ordered as follows: $b^* = (b_S^{*\prime}, b_{S^c}^{*\prime})'$. When we allow for misspecification, $S^c := \{i \leq K : b_i^* \leq 0\}$. Then,

$$\Sigma = \begin{pmatrix} \Sigma_{SS} & \Sigma_{SS^c} \\ \Sigma_{S^cS} & \Sigma_{SS} \end{pmatrix}. \tag{A.26}$$

To ensure a unique solution, assume that Σ is strictly positive definite. By the Karush-Kuhn-Tucker conditions,

$$\Sigma_{SS^c}\left(b_{S^c} - b_{S^c}^*\right) + \Sigma_{SS}\left(b_S - b_S^*\right) = \tau_S \tag{A.27}$$

$$\Sigma_{S^cS}\left(b_S - b_S^*\right) + \Sigma_{S^cS^c}\left(b_{S^c} - b_{S^c}^*\right) = \tau_{S^c} \tag{A.28}$$

where $\tau = (\tau_S', \tau_{S^c}')'$ is the Lagrange multiplier. The Lagrange multiplier satisfies $\tau \in [0, \infty)^K$. By strict positive definiteness of Σ, the constraint is not binding for the i^{th} variable if and only if $\tau_i = 0$.

From (A.27), we deduce that

$$\left(b_S - b_S^*\right) = -\Sigma_{SS}^{-1}\Sigma_{SS^c}\left(b_{S^c} - b_{S^c}^*\right) + \Sigma_{SS}^{-1}\tau_S. \tag{A.29}$$

Substituting in (A.28) and rearranging, we have that

$$\left(\Sigma_{S^cS^c} - \Sigma_{S^cS}\Sigma_{SS}^{-1}\Sigma_{SS^c}\right)\left(b_{S^c} - b^*_{S^c}\right) + \Sigma_{S^cS}\Sigma_{SS}^{-1}\tau_S = \tau_{S^c}. \tag{A.30}$$

We use $\tilde{\tau}$ to denote the value of the Lagrange multiplier at the constrained optimal solution.

We ask under what conditions ASC is equal to one, i.e. $\tilde{b}_{S^c} = 0$. Recall that $\tilde{b}$ denotes the unique optimal solution. For the moment suppose that the TDR is also equal to one, i.e. $\tilde{\tau}_S = 0$, i.e. the constraint is not binding for $\tilde{b}_S$. Then, from (A.30) we must have

$$-\left(\Sigma_{S^cS^c} - \Sigma_{S^cS}\Sigma_{SS}^{-1}\Sigma_{SS^c}\right) b^*_{S^c} = \tau_{S^c} \tag{A.31}$$

where $\left[\tau_{S^c}\right]_i > 0$ if $\left[\tilde{b}_{S^c}\right]_i < 0$. We use $\left[\tau_{S^c}\right]_i$ to denote the i^{th} entry in τ_{S^c} and similarly for $\left[\tilde{b}_{S^c}\right]_i$. Define $\Sigma_{S^cS^c|S} := \left(\Sigma_{S^cS^c} - \Sigma_{S^cS}\Sigma_{SS}^{-1}\Sigma_{SS^c}\right)$. Let $\left[\Sigma_{S^cS^c|S}\right]_{i,j}$ denote the i,j entry in $\Sigma_{S^cS^c|S}$. Using positive definiteness, it is not difficult to show that there is an $\epsilon > 0$ such that

$$\left[\Sigma_{S^cS^c|S}\right]_{i,i} \geq \left|\left[\Sigma_{S^cS^c|S}\right]_{i,j}\right| + \epsilon \text{ and } \sum_j \left[\Sigma_{S^cS^c|S}\right]_{i,j} \geq \epsilon, \ \forall i. \tag{A.32}$$

From (A.31) and using this notation, $\left[\tau_{S^c}\right]_i > 0$ if and only if $-\sum_j \left[\Sigma_{S^cS^c|S}\right]_{i,j}\left[\tilde{b}_{S^c}\right]_j > 0$. Using (A.32), this is the case if the entries in $\tilde{b}_{S^c}$ are either zero or have the same negative entries. In Table A.2, this remark applies to the designs "equal" and "skewed", but not to "gauss".

As shown in Table A.2 the assumption that TDR is equal to one is a strong one. Rewrite (A.29) as

$$\tilde{b}_S = b^*_S + \Sigma_{SS}^{-1}\Sigma_{SS^c}b^*_{S^c} + \Sigma_{SS}^{-1}\tilde{\tau}_S$$

under the assumption that the ASC equals one. If the constraint is not binding for $\tilde{b}_S$, i.e. $\tilde{b}_S > 0$, we have that $\tilde{\tau}_S = 0$. This happens when the entries in b^*_S dominate the ones in $\Sigma^{-1}_{SS}\Sigma_{SS^c}b^*_{S^c}$. For example, it tends to occur when the entries in $b^*_{S^c}$ are mostly zero or small relatively to b^*_S. In Table A.2 this corresponds to the design "gauss" and to some extent "equal", but not "skewed". However, the structure of Σ also plays a crucial role. For the design "equi", which applies to the construction of Σ, we find little difference on whether the coefficients in b^* are restricted to "equal" or "skewed". Finally, the value of the smallest eigenvalue of Σ does matter, as can be seen when we construct a nearly singular matrix using $n = K = 100$. In this case, the ϵ in (A.32) can be arbitrarily close to zero for some of the 1000 simulations of Σ.

References

[1] Brémaud, P. (1981) Point Processes and Queues: Martingales Dynamics. Berlin: Springer.

[2] Kallenberg, O. (1997) Foundations of Modern Probability. New York: Springer.

[3] Meinshausen, N. (2013) Sign-Constrained Least Squares Estimation for High-Dimensional Regression. Electronic Journal of Statistics 7, 1607-1631.

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

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

Table A.2: The Effect of the Sign Constraint under Misspecification. Results for the solution of (A.25) under different designs are reported using the same notation as in the text. The different correlation structures (corr.) are none for $\rho = 0$, equi for the equicorrelated case, and toep for the Toeplitz structure. The total number of covariates used to generate Σ is equal to $K = 100$. Results are averaged across 1000 different random designs.

$(\text{corr}, b^*, s, K_N)$	TDR	ASC	TDR	ASC
	$n = 100$		$n = 10000$	
(none, gauss, 5, 5)	0.6145	0.9986	0.9920	0.9996
(none, gauss, 5, 50)	0.7278	0.8624	0.9768	0.9830
(none, gauss, 50, 5)	0.8490	0.9954	0.9931	0.9994
(none, gauss, 50, 50)	0.7254	0.8545	0.9750	0.9808
(none, equal, 5, 5)	0.7470	1.0000	1.0000	1.0000
(none, equal, 5, 50)	0.8530	0.8976	1.0000	1.0000
(none, equal, 50, 5)	0.9663	0.9996	1.0000	1.0000
(none, equal, 50, 50)	0.8027	0.8908	1.0000	1.0000
(none, skewed, 5, 5)	0.3003	1.0000	0.9990	1.0000
(none, skewed, 5, 50)	0.6543	0.8984	0.8980	1.0000
(none, skewed, 50, 5)	0.4861	1.0000	0.9993	1.0000
(none, skewed, 50, 50)	0.5471	0.8992	0.8559	1.0000
(toep, gauss, 5, 5)	0.4245	1.0000	0.4228	1.0000
(toep, gauss, 5, 50)	0.3003	0.9998	0.2533	1.0000
(toep, gauss, 50, 5)	0.6697	1.0000	0.9223	1.0000
(toep, gauss, 50, 50)	0.2652	0.9997	0.6712	1.0000
(toep, equal, 5, 5)	0.3423	1.0000	0.2515	1.0000
(toep, equal, 5, 50)	0.2960	0.9998	0.2500	1.0000
(toep, equal, 50, 5)	0.7164	1.0000	0.9412	1.0000
(toep, equal, 50, 50)	0.2711	0.9998	0.7289	1.0000
(toep, skewed, 5, 5)	0.2238	1.0000	0.2500	1.0000
(toep, skewed, 5, 50)	0.2623	0.9999	0.2500	1.0000
(toep, skewed, 50, 5)	0.1856	1.0000	0.5102	1.0000
(toep, skewed, 50, 50)	0.1091	0.9999	0.1815	1.0000
(equi, gauss, 5, 5)	0.3405	1.0000	0.3350	1.0000
(equi, gauss, 5, 50)	0.2500	1.0000	0.2500	1.0000
(equi, gauss, 50, 5)	0.8277	0.9944	0.9604	0.9999
(equi, gauss, 50, 50)	0.0740	0.9972	0.1087	1.0000
(equi, equal, 5, 5)	0.2500	1.0000	0.2500	1.0000
(equi, equal, 5, 50)	0.2500	1.0000	0.2500	1.0000
(equi, equal, 50, 5)	0.9651	0.9979	1.0000	1.0000
(equi, equal, 50, 50)	0.0205	1.0000	0.0204	1.0000
(equi, skewed, 5, 5)	0.2500	1.0000	0.2500	1.0000
(equi, skewed, 5, 50)	0.2500	1.0000	0.2500	1.0000
(equi, skewed, 50, 5)	0.0246	1.0000	0.0204	1.0000
(equi, skewed, 50, 50)	0.0204	1.0000	0.0204	1.0000