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