SEMIPARAMETRIC ESTIMATION OF A SELF-EXCITING REGRESSION MODEL WITH AN APPLICATION IN RECURRENT EVENT DATA ANALYSIS

Fangfang Bai, Feng Chen, and Kani Chen

University of International Business and Economics,
University of New South Wales, and
Hong Kong University of Science and Technology

Supplementary Material: This supplementary file contains the proofs of the theoretical results in Section 3 of the paper.

S.1: Two technical lemmas Let $||a||$ denote the Euclidean norm for a vector a and $||f||_\infty$ the supremum norm for a function f. Let $\xrightarrow{P}$ denote convergence in probability. Furthermore, we use $\lesssim$ to indicate that the function on its left-hand side is bounded by a positive constant times the the function on its right-hand side. For any $\theta_1 = (\beta_1^\top, g_1)^\top \in \Theta$ and $\theta_2 = (\beta_2^\top, g_2)^\top \in \Theta$, define a semi-metric $\rho(\theta_1, \theta_2)$ by,

$$\rho(\theta_1, \theta_2) = ||\beta_1 - \beta_2|| + \int_0^\tau |g_1(t) - g_2(t)| \, \mathrm{d}\, t.$$

Let $N_{[\,]}(\epsilon, \mathcal{F}, \rho)$ and $N(\epsilon, \mathcal{F}, \rho)$ be the bracketing number and covering number with respect to $\rho(\cdot, \cdot)$ of a function class $\mathcal{F}$, which is defined, e.g. in van der Vaart and Wellner (1996); van der Vaart (1998).

Lemma 1. *Assume $\mathcal{F}$ is the set of all monotone polynomial splines with order d and is a q-dimensional linear space. Then for any $\eta > 0$ and $\epsilon < \eta$,*

$$\log N_{[\,]}(\epsilon, \mathcal{F}, \rho) \lesssim q \log(\frac{\eta}{\epsilon}).$$

Proof. See Lemma A1 of Lu et al. (2009). □

Lemma 2. *Suppose f is a monotone nonincreasing function with bounded r-th derivative. Then there exists a monotone nonincreasing spline function f_n with order $d \geq r+1$*

and knot sequence $0 = \xi_1 = \ldots = \xi_d < \xi_{d+1} < \cdots < \xi_{\kappa_n} < \xi_{\kappa_n+1} = \ldots = \xi_{\kappa_n+d} = \tau$*, such that*

$$||f - f_n|| = O(\kappa_n^{-r}),$$

Proof. Similar to Lemma A1 of Lu et al. (2007). □

S.2: Proof of Theorem 1 First, we show that

$$\sup_{\theta\in\Theta} |\ell_n(\theta) - E\ell(\theta, W)| \xrightarrow{P} 0.$$

Define a function class $\mathcal{G} = \{g(t) : g(t) \text{ is a nonincreasing and bounded function}\}$. By Theorem 2.7.5 in van der Vaart and Wellner (1996), we have

$$\log N_{[\,]}(\epsilon, \mathcal{G}, \rho_1) \lesssim \frac{1}{\epsilon}, \tag{S2.1}$$

where $\rho_1(g_1, g_2) = \int_0^\tau |g_1(s) - g_2(s)| \,\mathrm{d}\, s$, for any $g_1, g_2 \in \mathcal{G}$. Define two function classes

$$\begin{aligned}\mathcal{F}_1 &= \{\ell(\beta, g, W);\ \beta \in \mathcal{B},\ \text{for any fixed } g \in \mathcal{G}\},\\ \mathcal{F}_2 &= \{\ell(\beta, g, W);\ g \in \mathcal{G},\ \text{for any fixed } \beta \in \mathcal{B}\}.\end{aligned}$$

By conditions C2 and C3, both $\mathcal{F}_1$ and $\mathcal{F}_2$ have integrable envelope functions. Since for any fixed g, $\ell(\beta, g)$ is Lipschitz with respect to β, we have $N_{[\,]}(\epsilon, \mathcal{F}_1, ||\cdot||) \lesssim (\frac{1}{\epsilon})^p$. Similarly, together with (S2.1), we have $\log N_{[\,]}(\epsilon, \mathcal{F}_2, \rho_1) \lesssim \frac{1}{\epsilon}$.

Hence, for the function class $\mathcal{F}^* = \{\ell(\theta, W) : \theta \in \Theta\}$, the bracketing number satisfies $\log N_{[\,]}(\epsilon, \mathcal{F}^*, \rho) \lesssim \frac{1}{\epsilon}$. By the Glivenko-Cantelli theorem, we have

$$\sup_{\theta\in\Theta} |\ell_n(\theta) - E\ell(\theta, W)| \xrightarrow{P} 0. \tag{S2.2}$$

Since $g_0 \in \mathcal{F}_r$, by Lemma 2 there exists a $g_{0n} \in \mathcal{F}_r^n$ such that $\sup_{t\in[0,\tau]} |g_0(t) - g_{0n}(t)| = O(\kappa_n^{-r})$. Let $\theta_{0n} = (\beta_0^\top, g_{0n})^\top$. Then clearly,

$$\rho(\theta_0, \theta_{0n}) \longrightarrow 0. \tag{S2.3}$$

Note that $\Theta_n \subset \Theta$, by (S2.2), we obtain

$$\sup_{\theta\in\Theta_n} |\ell_n(\theta) - E\ell(\theta, W)| \xrightarrow{P} 0. \tag{S2.4}$$

Clearly, Θ_n is compact with respect to $\rho(\cdot,\cdot)$, and $\ell_n(\theta)$ is continuous in $\theta \in \Theta_n$. Since θ_0 is the unique maximizer of $E\ell(\theta, W)$ on Θ, we have that θ_{0n} is the unique maximizer of $E\ell(\theta, W)$ on Θ_n. These facts and (S2.4) yield

$$\rho(\widehat{\theta}_n, \theta_{0n}) \xrightarrow{P} 0. \tag{S2.5}$$

Finally, combining (S2.3) and (S2.5), we have $\rho(\widehat{\theta}_n, \theta_0) \xrightarrow{P} 0$. □

S.3: Proof of Theorem 2 Let $g_{0n} = \arg\min_{g \in \mathcal{F}_r^n} \|g_0 - g\|_\infty$ and $\theta_{0n} = (\beta_0^\top, g_{0n})^\top$. Then by Lemma 2, we have $\|g_0 - g_{0n}\| = O(\kappa_n^{-r})$, and therefore

$$\rho(\theta_{0n}, \theta_0) = \rho_1(g_0, g_{0n}) = O(\kappa_n^{-r}). \tag{S3.1}$$

We next show that $\rho(\widehat{\theta}_n, \theta_{0n}) = O_p\big((\frac{\kappa_n}{n})^{\frac{1}{2}}\big)$. Let δ be a fixed positive constant. Consider a class of functions

$$\mathcal{M}_{n,\delta} = \{\ell(\theta, W) - \ell(\theta_{0n}, W);\ \rho(\theta, \theta_{0n}) \le \delta, \theta \in \Theta_n\}.$$

Define two function classes:

$$\mathcal{M}_1 = \{\ell(\beta, g_{0n}, W) - \ell(\beta_0, g_{0n}, W);\ ||\beta - \beta_0|| \le \delta,\ \beta \in \mathcal{B}\},$$
$$\mathcal{M}_2 = \{\ell(\beta_0, g, W) - \ell(\beta_0, g_{0n}, W);\ \int_0^\tau |g(t) - g_{0n}(t)|\,\mathrm{d}\,t \le \delta,\ g \in \mathcal{F}_r^n\}.$$

Because $\ell(\beta, g, W)$ is Lipschitz with respect to β, the bracketing number of the function class $\mathcal{M}_1$ with any fixed g_{0n} satisfies $N_{[\,]}(\varepsilon, \mathcal{M}_1, \|\cdot\|) \lesssim \big(\frac{\delta}{\varepsilon}\big)^p$. Similarly, with any fixed $\beta \in \mathcal{B}$, for the function class $\mathcal{M}_2$ we have $N_{[\,]}(\varepsilon, \mathcal{M}_2, \rho_1) \lesssim \big(\frac{\delta}{\varepsilon}\big)^{\kappa_n}$. Then it follows that

$$N_{[\,]}(\varepsilon, \mathcal{M}_{n,\delta}, \rho) \le N_{[\,]}\big(\frac{\varepsilon}{2}, \mathcal{M}_1, \|\cdot\|\big) N_{[\,]}\big(\frac{\varepsilon}{2}, \mathcal{M}_2, \rho_1\big).$$

Therefore, the entropy of the function class $\mathcal{M}_{n,\delta}$ satisfies

$$\log N_{[\,]}(\varepsilon, \mathcal{M}_{n,\delta}, \rho) \lesssim \kappa_n \log\big(\frac{\delta}{\varepsilon}\big).$$

Hence, the bracketing integral $J_{[\,]}(\delta, \mathcal{M}_{n,\delta}, \rho)$ (defined e.g. in van der Vaart and Wellner, 1996, p. 324) of the function class $\mathcal{M}_{n,\delta}$ satisfies

$$\begin{aligned} J_{[\,]}(\delta, \mathcal{M}_{n,\delta}, \rho) &= \int_0^\delta [1 + \log N_{[\,]}(\varepsilon, \mathcal{M}_{n,\delta}, \rho)]^{1/2}\,\mathrm{d}\,\varepsilon \\ &\le \int_0^\delta [1 + A\kappa_n \log\frac{\delta}{\varepsilon}]^{1/2}\,\mathrm{d}\,\varepsilon \\ &\lesssim {\kappa_n}^{1/2}\delta. \end{aligned}$$

By Lemma 3.4.2 in van der Vaart and Wellner (1996), we have that

$$\begin{aligned}
&E\Big(\sup_{\frac{\delta}{2}<\rho(\theta,\theta_{0n})\le\delta}\big|(\ell_n(\theta)-\ell_n(\theta_{0n}))-E(\ell_n(\theta)-\ell_n(\theta_{0n}))\big|\Big)\\
&\le \frac{1}{\sqrt{n}}J_{[\,]}(\delta,\mathcal{M}_{n,\delta},\rho)(1+\frac{J_{[\,]}(\delta,\mathcal{M}_{n,\delta},\rho)}{\delta^2\sqrt{n}}A_3)\\
&\lesssim \frac{1}{\sqrt{n}}\kappa_n{}^{\frac{1}{2}}\delta(1+\kappa_n{}^{\frac{1}{2}}\delta/\delta^2\sqrt{n}A_3)\\
&=(\kappa_n/n)^{\frac{1}{2}}\delta(1+(\kappa_n/n)^{\frac{1}{2}}/\delta A_3)=O\big((\frac{\kappa_n}{n})^{\frac{1}{2}}\delta\big).
\end{aligned}$$

Also note by Taylor's expansion that,

$$\sup_{\delta/2<\rho(\theta,\theta_{0n})\le\delta,\theta\in\Theta_n} E(\ell(\theta,W))-E(\ell(\theta_{0n},W))\lesssim-\delta^2.$$

Now, apply Theorem 3.4.1 in van der Vaart and Wellner (1996) with $\phi_n(\delta)=\delta\cdot\kappa_n^{1/2}$, $\delta_n\equiv 0$ and $r_n=(n/\kappa_n)^{1/2}$, and we have

$$\rho(\widehat{\theta}_n,\theta_{0n})=O\big((\frac{\kappa_n}{n})^{1/2}\big).$$

This, together with (S3.1), yields that $\rho(\widehat{\theta}_n,\theta_0)=O_p\big((\frac{\kappa_n}{n})^{1/2}+\kappa_n^{-r}\big)$, which concludes the proof of Theorem 2. □

S.4: Proof of Theorem 3 The proof of the asymptotic normality proceeds as follows. The least-favorable direction for β is first obtained, and then we use a Taylor expansion for the score function of β and g along an approximately least-favorable direction.

From the expression of the logliklihood (eq. (2.4) in the paper), the score function for β and the score operator for g are respectively,

$$\begin{aligned}
\ell_\beta &= -XC+\int_0^C\frac{X}{X^\top\beta+\int_0^t g(t-s)\,\mathrm{d}N(s)}\,\mathrm{d}N(t)\\
\ell_g[h] &= -\int_0^C\int_0^t h(t-s)\,\mathrm{d}N(s)\,\mathrm{d}t+\int_0^C\frac{\int_0^t h(t-s)\,\mathrm{d}N(s)}{X^\top\beta+\int_0^t g(t-s)\,\mathrm{d}N(s)}\,\mathrm{d}N(t).
\end{aligned}$$

Step 1.

For semiparametric models, the least-favorable submodel is the submodel that achieves the infimum of the information over all submodels (Bickel et al., 1993; van der Vaart, 1998). First, we show that the least-favorable submodel exists.

Note that $\ell_g[\cdot]$ is a linear operator from $\mathcal{F}_r$ to $L_2(P_{\theta_0})$, where P_{θ_0} is the true probability distribution of W, and that the closed linear space spanned by the score functions for g is,

$$\Big\{-\int_0^C\int_0^t h(t-s)\,\mathrm{d}\,N(s)\,\mathrm{d}\,t+\int_0^C\frac{\int_0^t h(t-s)\,\mathrm{d}\,N(s)}{X^\top\beta+\int_0^t g(t-s)\,\mathrm{d}\,N(s)}\;\mathrm{d}\,N(t);\;h\in\mathcal{F}_r\Big\}.$$

The dual operator $\ell_g^*: L_2(P_{\theta_0})\to\mathcal{F}_r$, satisfies that for any $h\in\mathcal{F}_r$ and measurable function $u(W)$,

$$E[\ell_g[h](W)u(W)]=\int_0^\tau \ell_g^*[u](t)h(t)\,\mathrm{d}\,t.$$

To find the least-favorable direction for β is equivalent to solve the following equation,

$$\ell_g^*[\ell_g[h]]=\ell_g^*\ell_\beta. \tag{S4.1}$$

Since equation (S4.1) is a Fredholm-type equation, the existence of the solution is equivalent to showing that the equation $\ell_g^*[\ell_g[h]]=0$ has a trivial solution. Note that if $\ell_g^*\big[\ell_g[h]\big]=0$, then $E[\ell_g[h]^2]=0$, that is $\ell_g[h]=0$. It is clear that $h=0$. Therefore, the least-favorable direction for β exists. Actually, the least-favorable direction for β is the projection of the score function ℓ_β on the linear closed space spanned by the score function $\ell_g[h]$.

Step 2.

Denote the least-favorable direction for β as $h^*(t)$. We choose an approximate submodel $(\widehat{\beta}_n+\epsilon b,\widehat{g}_n+\epsilon\widehat{h}_n)$, where $\widehat{h}_n$ is the spline approximation for the least-favorable function $h^*(t)\in\mathcal{F}_r$ (so $||\widehat{h}_n(t)-h^*(t)||=O(\kappa_n^{-r})$).

Since the estimator $(\widehat{\beta}_n,\widehat{g}_n)$ maximizes the log likelihood function along this submodel, then

$$P_n\big[\ell_\beta(\widehat{\beta}_n,\widehat{g}_n)+\ell_g(\widehat{\beta}_n,\widehat{g}_n)[\widehat{h}_n]\big]=0.$$

By the Lipschitz property, the function class

$$\Big\{\ell_\beta(\beta,g)+\ell_g(\beta,g)[h];\;||\beta-\beta_0||\lesssim\big(\frac{\kappa_n}{n}\big)^{-1/2},\;||g-g_0||\lesssim\kappa_n^{-r},\;||h-h^*||\lesssim\kappa_n^{-r}\Big\}, \tag{S4.2}$$

is a Donsker class. Therefore, we have

$$\sup_{||\beta-\beta_0||\lesssim\left(\frac{\kappa_n}{n}\right)^{-1/2},||g-g_0||\lesssim\kappa_n^{-r},||h-h^*||\lesssim\kappa_n^{-r}}\Big|\Big|\sqrt{n}(P_n-P)\big(\ell_\beta(\beta,g)+\ell_g(\beta,g)[h] - [\ell_\beta(\beta_0,g_0)+\ell_g(\beta_0,g_0)[h^*]]\big)\Big|\Big|=o_p(1). \tag{S4.3}$$

Combining (S4.2) and (S4.3), we have that

$$\sqrt{n}P_n\big(\ell_\beta(\beta_0,g_0)+\ell_g(\beta_0,g_0)[h^*]\big)+o_p(1)=-\sqrt{n}P\big(\ell_\beta(\widehat{\beta}_n,\widehat{g}_n)+\ell_g(\widehat{\beta}_n,\widehat{g}_n)[\widehat{h}_n]\big).$$

Then, after Taylor expansion of the right hand side of last equation, we have

$$\begin{aligned}\sqrt{n}P_n\big(\ell_\beta(\beta_0,g_0)+\ell_g(\beta_0,g_0)[h^*]\big)+o_p(1)=&-\sqrt{n}P\big(\ell_{\beta\beta}(\beta_0,g_0)+\ell_{\beta g}(\beta_0,g_0)[h^*]\big)(\widehat{\beta}_n-\beta_0)\\&-\sqrt{n}P\big\{\ell_{\beta g}(\beta_0,g_0)[\widehat{g}_n-g_0]+\ell_{gg}(\beta_0,g_0)[h^*,\widehat{g}_n-g_0]\big\}\\&+\sqrt{n}O(\|\widehat{\beta}-\beta_0\|^2+\|\widehat{g}_n-g_0\|^2+\|\widehat{h}_n-h^*\|^2),\end{aligned}\tag{S4.4}$$

where $\ell_{\beta g}(\beta_0,g_0)[\widehat{g}_n-g]$ is the derivative of ℓ_β along the path $\beta=\beta_0, g=g_0+\epsilon(\widehat{g}_n-g)$ and $\ell_{gg}(\beta_0,g_0)[h^*,\widehat{g}_n-g_0]$ is the derivative of $\ell_g[h^*]$ along the path $\beta=\beta_0, g=g_0+\epsilon(\widehat{g}_n-g)$. Since h^* is the least-favorable direction for β, we have that

$$P\big\{\ell_{\beta g}(\beta_0,g_0)[\widehat{g}_n-g]+\ell_{gg}(\beta_0,g_0)[h^*,\widehat{g}_n-g_0]\big\}=0.$$

Moreover, by the assumed conditions that $\kappa_n^2/n\longrightarrow 0$ and $n\kappa_n^{-4r}\longrightarrow 0$ as $n\to\infty$, and the convergence rate of $(\widehat{\beta}_n^\top,\widehat{g}_n)^\top$, the third term of the right hand side of (S4.4) is $o_p(1)$. Therefore, (S4.4) becomes

$$\begin{aligned}&\sqrt{n}P\big\{\ell_{\beta\beta}(\beta_0,g_0)+\ell_{\beta g}(\beta_0,g_0)[h^*]\big\}(\widehat{\beta}_n-\beta_0)\\&\qquad=-\sqrt{n}P_n\big(\ell_\beta(\beta_0,g_0)+\ell_g(\beta_0,g_0)[h^*]\big)+o_p(1).\end{aligned}\tag{S4.5}$$

Step 3.

We are going to show that the matrix $P\big\{\ell_{\beta\beta}(\beta_0,g_0)+\ell_{\beta g}(\beta_0,g_0)[h^*]\big\}$ is nonsingular. It suffices to show that for a vector a, if

$$a^\top P\big\{\ell_{\beta\beta}(\beta_0,g_0)+\ell_{\beta g}(\beta_0,g_0)[h^*]\big\}a=0,\tag{S4.6}$$

then $a=0$. Since h^* is the least-favorable direction for β, (S4.6) becomes

$$P\Big[a^\top\ell_\beta(\beta_0,g_0)+a^\top\ell_g(\beta_0,g_0)[h^*]\Big]^2=0.$$

Then, $a^\top\ell_\beta(\beta_0,g_0)+a^\top\ell_g(\beta_0,g_0)[h^*]=0$. We can conclude that $a^\top(X+\int_0^t h^*(t-s)\,\mathrm{d}\,N(s))=0$. By C4, we have $a=0$. Hence, the derivative matrix is nonsingular.

By equation (S4.5), we have that

$$\sqrt{n}(\widehat{\beta}_n-\beta_0)=-\Big\{P\big[\ell_{\beta\beta}(\beta_0,g_0)+\ell_{\beta g}(\beta_0,g_0)[h^*]\big]\Big\}^{-1}\sqrt{n}P_n\big(\ell_\beta(\beta_0,g_0)+\ell_g(\beta_0,g_0)[h^*]\big)+o_p(1).$$

Hence, $\sqrt{n}(\widehat{\beta}_n - \beta_0)$ converges to a normal distribution. The influence function is

$$-\Big\{P\big[\ell_{\beta\beta}(\beta_0, g_0) + \ell_{\beta g}(\beta_0, g_0)[h^*]\big]\Big\}^{-1}\big\{\ell_\beta(\beta_0, g_0) + \ell_g(\beta_0, g_0)[h^*]\big\},$$

and thus the estimator $\widehat{\beta}_n$ is semiparametrically efficient.

To show the consistency of the variance estimator $\hat{\Sigma}_\beta$, the key is to show the information operator $\mathcal{I}$, given by

$$\mathcal{I}\left[\begin{pmatrix} b_1 \\ h_1 \end{pmatrix}, \begin{pmatrix} b_2 \\ h_2 \end{pmatrix}\right] = -\left(b_1^\top P\ell_{\beta\beta} b_2 + b_1^\top P\ell_{\beta g}[h_2] + b_2^\top P\ell_{\beta g}[h_1] + P\ell_{g,g}[h_1, h_2]\right),$$

is uniformly consistently estimated by $\left(\iota(\cdot)^\top, \pi_{B_n}(\cdot)^\top\right)\mathcal{I}_n\left(\iota(\cdot)^\top, \pi_{B_n}(\cdot)^\top\right)^\top$, where ι is the identity map, π_{B_n} is the operator that determines the vector of coefficients of the MBS approximation to a function, so that $g_n(t) = B_n(t)^\top \pi_{B_n}(g)$, and $\mathcal{I}_n = -P_n \begin{pmatrix} \ell_{\beta\beta} & \ell_{\beta g}[B_n] \\ \ell_{\beta g}[B_n]^\top & \ell_{gg}[B_n, B_n], \end{pmatrix}$ is the observed information matrix. The details are omitted due to similarity to those of the consistency proof for the variance estimator in Zeng and Lin (2006). □

References

Bickel, P., Klaassen, C. A. J., Ritov, Y., and Wellner, J. A. (1993). *Efficient and Adaptive Estimation for Semiparametric Models*. The John Hopkins University Press, New York.

Lu, M., Zhang, Y., and Huang, J. (2007). Estimation of the mean function with panel count data using monotone polynomial splines. *Biometrika*, 94(3):705–718.

Lu, M., Zhang, Y., and Huang, J. (2009). Semiparametric estimation methods for panel count data using monotone B-splines. *Journal of the American Statistical Association*, 104(487):1060–1070.

van der Vaart, A. (1998). *Asymptotic Statistics*. Cambridge university press, New York.

van der Vaart, A. W. and Wellner, J. A. (1996). *Weak Convergence and Empirical Processes*. Springer, New York.

Zeng, D. and Lin, D. Y. (2006). Efficient estimation of semiparametric transformation models for counting processes. *Biometrika*, 93(3):627–640.