Efficient Inference for Infinite-Dimensional Parameters

Xinjie Qian

December 2, 2021

Outline

We now consider the special case that both θ and η are $\sqrt{n}$ consistent in the semiparametric model $\{P_{\theta,\eta} : \theta \in \Theta, \eta \in H\}$, where $\Theta \subset \mathbb{R}^k$. Often in this setting, η may be of some interest to the data analyst. Hence, in this chapter, η will not be considered a nuisance parameter.

Semiparametric Maximum Likelihood Estimation

Corollary 3.2

Suppose that $\dot{\ell}_{\theta,\eta}$ and $B_{\theta,\eta}h$, with h ranging over $\mathcal{H}$ and with (θ, η) ranging over a neighborhood of (θ_0, η_0), are contained in a P_{θ_0,η_0}-Donsker class, and that both $P_{\theta_0,\eta_0}\left\|\dot{\ell}_{\theta,\eta} - \dot{\ell}_{\theta_0,\eta_0}\right\|^2 \xrightarrow{P} 0$ and $sup_{h\in\mathcal{H}} P_{\theta_0,\eta_0}|B_{\theta,\eta}h - B_{\theta_0,\eta_0}h|^2 \xrightarrow{P} 0$, as $(\theta, \eta) \to (\theta_0, \eta_0)$. Also assume that Ψ is Frechet-differentiable at (θ_0, η_0) with derivative $\dot{\Psi}_0 : \mathbb{R}^k \times linH \mapsto \mathbb{R}^k \times \ell^\infty(\mathcal{H})$ that is continuously-invertible and onto its range, with inverse $\dot{\Psi}_0^{-1} : \mathbb{R}^k \times \ell^\infty(\mathcal{H}) \mapsto \mathbb{R}^k \times linH$. Then, provided $(\hat{\theta}_n, \hat{\eta}_n)$ is consistent for (θ_0, η_0) and $\Psi_n(\hat{\theta}_n, \hat{\eta}_n) = o_P(n^{-1/2})$ (uniformly over $\mathbb{R}^k \times \ell^\infty(\mathcal{H})$), $(\hat{\theta}_n, \hat{\eta}_n)$ is efficient at (θ_0, η_0) and $\sqrt{n}(\hat{\theta}_n - \theta_0, \hat{\eta}_n - \eta_0) \rightsquigarrow -\dot{\Psi}_0^{-1} Z$, where Z is the Gaussian limiting distribution of $\sqrt{n}\Psi_n(\theta_0, \eta_0)$.

Proof of Corollary 3.2

Proof

By Lemma 13.3, we can get that

$$\sqrt{n}(\Psi_n - \Psi)(\hat{\theta}_n, \hat{\eta}_n) - \sqrt{n}(\Psi_n - \Psi)(\theta_0, \eta_0) = o_P(1),$$

where the convergence is uniform. Since the Donsker assumption on the score equation ensures $\sqrt{n}\Psi_n(\theta_0, \eta_0) \rightsquigarrow Z$, for some tight, mean zero Gaussian process Z, we have satisfied all of the conditions of Theorem 2.11, and thus $\sqrt{n}(\hat{\theta}_n - \theta_0, \hat{\eta}_n - \eta_0) \rightsquigarrow -\dot{\Psi}_0^{-1} Z$.

The remaining challenge is to establish efficiency. Recall that the differentiation used to obtain the score and information operators involves a smooth function $t \mapsto \eta_t(\theta, \eta)$ for which $\eta_0(\theta, \eta) = \eta$, t is a scalar, and

$$B_{\theta,\eta} h(x) - P_{\theta,\eta} B_{\theta,\eta} h = \partial \ell_{\theta, \eta_t(\theta,\eta)}(x) / (\partial t) \Big|_{t=0},$$

and where $\ell(\theta, \eta)(x)$ is the log-likelihood for a single observation.

Proof of Corollary 3.2

Proof (cont).
Note that this η_t is not necessarily an approximately least-favorable submodel. The purpose of this η_t is to incorporate the effect of a perturbation of η in the direction $h \in \mathcal{H}$. The resulting one-dimensional submodel is $t \mapsto \psi_t \equiv (\theta + ta, \eta_t(\theta, \eta))$, with derivative

$$\frac{\partial}{\partial t}\psi_t\Big|_{t=0} \equiv \dot{\psi}(a, h),$$

where $c \equiv (a, h) \in \mathbb{R}^k \times \mathcal{H}$ and $\dot{\psi} : \mathbb{R}^k \times \mathcal{H} \mapsto \mathbb{R}^k \times lin\ H \equiv \mathcal{C}$ is a linear operator that may depend on the composite (joint) parameter $\psi \equiv (\theta, \eta)$. To be explicit about which tangent in $\mathcal{C}$ is being applied to the one-dimensional submodel, we will use the notation $\psi_{t,c}$, i.e., $\partial/(\partial t)|_{t=0}\psi_{t,c} = \dot{\psi}(c)$.

Proof of Corollary 3.2

Proof (cont).

Define the abbreviated notation $U_\psi(c) \equiv a^{'}\dot{\ell}_\psi + B_\psi h - P_\psi B_\psi h$, for $c = (a, h) \in \mathcal{C}$. Our construction now gives us that for any $c_1, c_2 \in \mathcal{C}$,

$$\dot{\Psi}(\dot{\psi}_0(c_2))(c_1) = \frac{\partial}{\partial t} P_{\psi_0}[U_{\psi_{t,c_2}}(c_1)]\Big|_{t=0,\psi=\psi_0} = -P_{\psi_0}[U_{\psi_0}(c_1)U_{\psi_0}(c_2)], \tag{20.1}$$

where $\psi_0 \equiv (\theta_0, \eta_0)$.

Proof of Corollary 3.2

Proof (cont).

We know from the previous proof that the influence function for $\hat{\psi}_n \equiv (\hat{\theta}_n, \hat{\eta}_n)$ is $\tilde{\psi} \equiv -\dot{\Psi}^{-1}[U_{\psi_0}(\cdot)]$. Thus, for any $c \in \mathcal{C}$,

$$\begin{aligned} P_{\psi_0}\left[\tilde{\psi} U_{\psi_0}(c)\right] &= P_{\psi_0}\left[(-\dot{\Psi}^{-1}[U_{\psi_0}(\cdot)])U_{\psi_0}(c)\right] \\ &= -\dot{\Psi}^{-1} P_{\psi_0}[U_{\psi_0}(\cdot)U_{\psi_0}(c)] \\ &= -\dot{\Psi}^{-1}\Big[-\dot{\Psi}(\dot{\psi}_0(c))(\cdot)\Big] \\ &= \dot{\psi}_0(c). \end{aligned}$$

This means by the definition given in Section 18.1 that $\tilde{\psi}_0$ is the efficient influence function.

Since $\sqrt{n}(\hat{\psi}_n - \psi_0)$ is asymptotically tight and Gaussian with covariance that equals the covariance of the efficient influence function, we have by Theorem 18.3 that $\hat{\psi}_n$ is efficient.

For many semiparametric models where the joint parameter is regular, we have that $\eta = A$, where $t \mapsto A(t)$ is restricted to a subset $H \in D[0, \tau]$ of functions bounded in total variation, where $\tau < \infty$. The composite parameter is thus $\psi = (\theta, A)$. ψ can be viewed as an element of $\ell^\infty(\mathcal{C}_p)$ if we define

$$\psi(c) \equiv a^{'}\theta + \int_0^\tau h(s)dA(s), c \in \mathcal{C}_p, \psi \in \Omega \equiv \Theta \times H$$

As described in Section15.3.4, Ω thus becomes a subset of $\ell^\infty(\mathcal{C}_p)$.

We now modify the score notation slightly. For any $c \in \mathcal{C}$, let

$$
\begin{aligned}
U[\psi](c) &= \frac{\partial}{\partial t}\ell\Big(\theta + ta, A(\cdot) + t\int_0^{(\cdot)} h(s)dA(s)\Big)\Big|_{t=0} \\
&= \frac{\partial}{\partial t}\ell(\theta + ta, A(\cdot))\Big|_{t=0} + \frac{\partial}{\partial t}\ell\Big(\theta, A(\cdot) + t\int_0^{(\cdot)} h(s)dA(s)\Big)\Big|_{t=0} \\
&\equiv U_1[\psi](a) + U_2[\psi](h).
\end{aligned}
$$

It is important to note that the map $\psi \mapsto U[\psi](\cdot)$ actually has domain $lin\ \Omega$ and range contained in $\ell^\infty(\mathcal{C})$.

We then consider properties of the second derivative of the log-likelihood. Let $\bar{a} \in \mathbb{R}^k$ and $\bar{h} \in \mathcal{H}$. Denote $c = (a, h) \equiv (c_1, c_2)$. We assume the following derivative structure exists and is valid for $j = 1, 2$ and all $c \in \mathcal{C}$:

$$\begin{aligned}
&\frac{\partial}{\partial s} U_j[\theta + s\bar{a}, A + s\bar{h}](c_j)\Big|_{s=0} \\
&= \frac{\partial}{\partial s} U_j[\theta + s\bar{a}, A](c_j)\Big|_{s=0} + \frac{\partial}{\partial s} U_j[\theta, A + s\bar{h}](c_j)\Big|_{s=0} \\
&\equiv \bar{a}' \hat{\sigma}_{1j}[\psi](c_j) + \int_0^\tau \hat{\sigma}_{2j}[\psi](c_j)(u) d\bar{h}(u),
\end{aligned}$$

where $\hat{\sigma}_{1j}[\psi](c_j)$ is a random k-vector and $u \mapsto \hat{\sigma}_{2j}[\psi](c_j)(u)$ is a random function contained in $\mathcal{H}$.

In this set-up, we will need the following conditions for some $p > 0$ in order to apply Corollary 3.2:

$$\{U[\psi](c) : \|\psi - \psi_0\| \leq \epsilon, c \in \mathcal{C}_p\} \ is \ Donsker \ for \ some \ \epsilon > 0, \tag{20.2}$$

$$\sup_{c \in \mathcal{C}_p} P_0 |U[\psi](c) - U[\psi_0](c)|^2 \to 0, \ as \ \psi \to \psi_0, \tag{20.3}$$

$$\sup_{c \in \mathcal{C}_p} \|\sigma[\psi](c) - \sigma[\psi_0](c)\|_{(p)} \to 0, \ as \ \|\psi - \psi_0\|_{(p)} \to 0. \tag{20.4}$$

By Exercise 20.3.1, (20.4) implies Ψ is Frechet-differentiable in $\ell^\infty(\mathcal{C}_p)$. It is also not hard to verify that if Conditions (20.2)–(20.4) hold for some $p > 0$, then they hold for all $0 < p < \infty$ (Exercise 20.3.2).

Corollary 20.1

Assume Conditions (20.2)–(20.4) hold for some $p > 0$, that $\sigma : \mathcal{C} \mapsto \mathcal{C}$ is continuously invertible and onto, and that $\hat{\psi}_n$ is uniformly consistent for ψ_0 with

$$\sup_{c \in \mathcal{C}_1} \left| \mathbb{P}_n \Psi_n(\hat{\psi}_n)(c) \right| = o_{P_0}(n^{-1/2}).$$

Then $\hat{\psi}_n$ is efficient with

$$\sqrt{n}(\hat{\psi}_n - \psi_0)(\cdot) \rightsquigarrow Z(\sigma^{-1}(\cdot))$$

in $\ell^\infty(\mathcal{C}_1)$, where Z is the tight limiting distribution of $\sqrt{n}\mathbb{P}_n U[\psi_0](\cdot)$.

Note that we actually need Z to be a tight element in $\ell^\infty(\sigma^{-1}(\mathcal{C}_1))$, but the linearity of $U[\psi](\cdot)$ ensures that if $\sqrt{n}\mathbb{P}_n U[\psi_0](\cdot)$ converges to Z in $\ell^\infty(\mathcal{C}_1)$, then it will also converge weakly in $\ell^\infty(\mathcal{C}_p)$ for any $p < \infty$.

The Cox model for right-censored data

We will let θ be the regression effect and A the baseline hazard, with the observed data $X = (U, \delta, Z)$. We make the usual assumptions for this model as done in Section 4.2.2, including requiring the baseline hazard to be continuous, except that we will use (θ, A) to denote the model parameters (β, Λ). It is not hard to verify that $U_1[\psi](a) = \int_0^\tau Z' a dM_\psi(s)$ and $U_2[\psi](h) = \int_0^\tau h(s) dM_\psi(s)$, where $M_\psi(t) \equiv N(t) - \int_0^t Y(s) e^{\theta' Z} dA(s)$ and N and Y are the usual counting and at-risk processes.

The Cox model for right-censored data

It is also easy to show that the components of σ are defined by

$$\sigma_{11}a = \int_0^\tau P_0[ZZ'Y(s)e^{\theta_0'Z}]dA_0(s)a,$$
$$\sigma_{12}h = \int_0^\tau P_0[ZY(s)e^{\theta_0'Z}]h(s)dA_0(s),$$
$$\sigma_{21}a = P_0[Z'Y(\cdot)e^{\theta_0'Z}]a, \; and$$
$$\sigma_{22}h = P_0[Y(\cdot)e^{\theta_0'Z}]h(\cdot).$$

The maximum likelihood estimator is $\hat{\psi}_n = (\hat{\theta}_n, \hat{A}_n)$, where $\hat{\theta}_n$ is the maximizer of the well-known partial likelihood and $\hat{A}_n$ is the Breslow estimator. The conditions of Corollary 20.1 hold for this example.

Weighted and Nonparametric Bootstraps

Recall the nonparametric and weighted bootstrap methods for Z-estimators described in Section 13.2.3. Let $\mathbb{P}_n^\circ$ and $\mathbb{G}_n^\circ$ be the bootstrapped empirical measure and process based on either kind of bootstrap, and let $\underset{\circ}{\overset{P}{\rightsquigarrow}}$ denote either $\underset{W}{\overset{P}{\rightsquigarrow}}$ for the nonparametric version or $\underset{\xi}{\overset{P}{\rightsquigarrow}}$ for the weighted version. We will use $\hat{\psi}_n^\circ$ to denote an approximate maximizer of the bootstrapped empirical log-likelihood $\psi \mapsto \mathbb{P}_n^\circ \ell(\psi)(X)$, and we will denote $\Psi_n^\circ(\psi)(c) \equiv \mathbb{P}_n^\circ U[\psi](c)$ for all $\psi \in \Omega$ and $c \in \mathcal{C}$. We now have the following simple corollary, where $\mathcal{X}_n$ is the σ-field of the observations $X_1, ..., X_n$:

Weighted and Nonparametric Bootstraps

Corollary 20.2

Assume the conditions of Corollary 20.1, and, in addition, that $\hat{\psi}_n^\circ \overset{as*}{\to} \psi_0$ unconditionally and

$$P\Big(\sqrt{n} \sup_{c \in \mathcal{C}_1} |\Psi_n(\hat{\psi}_n^\circ)(c)| \Big| \mathcal{X}_n\Big) = o_P(1). \tag{20.7}$$

Then the conclusions of Corollary 20.1 hold and $\sqrt{n}(\hat{\psi}_n - \hat{\psi}_n) \underset{\circ}{\overset{P}{\rightsquigarrow}} Z(\sigma^{-1}(\cdot))$ in $\ell^\infty(\mathcal{C}_1)$, i.e., the limiting distribution of $\sqrt{n}(\hat{\psi}_n - \psi_0)$ and the conditional limiting distribution of $\sqrt{n}(\hat{\psi}_n - \hat{\psi}_n)$ given $\mathcal{X}_n$ are the same.

Weighted and Nonparametric Bootstraps

Proof of Corollary 20.2.

By Theorem 2.6, the conditional bootstrapped distribution of a Donsker class is automatically consistent. Since conditional weak convergence implies unconditional weak convergence (as argued in the proof of Theorem 10.4), both Lemma 13.3 and Theorem 2.11 apply to Ψ_n°, and thus

$$\sup_{c \in \mathcal{C}_p} \left| \sqrt{n}(\hat{\psi}_n - \psi_0)(\sigma(c)) - \sqrt{n}(\Psi_n^\circ - \Psi)(c) \right| = o_{P_0}(1),$$

unconditionally, for any $0 < p < \infty$. Combining this with previous results for $\hat{\psi}_n$, we obtain for any $0 < p < \infty$

$$\sup_{c \in \mathcal{C}_p} \left| \sqrt{n}(\hat{\psi}_n - \hat{\psi}_n)(\sigma(c)) - \sqrt{n}(\Psi_n^\circ - \Psi_n)(c) \right| = o_{P_0}(1).$$

Since $\{U[\psi_0](c) : c \in \mathcal{C}_p\}$ is Donsker for any $0 < p < \infty$, we have the desired conclusion by reapplication of Theorem 2.6 and the continuous invertibility of σ.

The Piggyback Bootstrap

The "profile sampler": generating random realizations θ_n such that $\sqrt{n}(\theta_n - \hat{\theta}_n)$ given the data has the same limiting distribution as $\sqrt{n}(\hat{\theta}_n - \theta_0)$ does unconditionally.

The piggyback bootstrap will utilize these θ_n realizations to improve computational efficiency.
Notation: For any $\theta \in \Theta$, let $\hat{A}^{\circ}_{\theta} = argmax_A \mathbb{P}^{\circ}_n \ell(\theta, A)(X)$, where $\mathbb{P}^{\circ}_n$ is the weighted bootstrap empirical measure.

The Piggyback Bootstrap

The main idea of the piggyback bootstrap is to generate a realization of θ_n, then generate the random weights $\xi_1, ..., \xi_n$ in $\mathbb{P}_n^{\circ}$ independent of both the data and θ_n, and then compute $\hat{A}^{\circ}_{\theta_n}$.

This generates a joint realization $\hat{\psi}_n^{\circ} \equiv (\theta_n, \hat{A}_n^{\circ})$.

For instance, one can generate a sequence of θ_ns, $\theta_n^{(1)}, ..., \theta_n^{(m)}$, using the profile sampler.

The Piggyback Bootstrap

Under some regularity conditions, the conditional distribution of $\sqrt{n}(\hat{\psi}_n^\circ - \hat{\psi}_n)$ converges to the same limiting distribution as $\sqrt{n}(\hat{\psi}_n - \psi_0)$ does unconditionally.

Hence the realizations $\hat{\psi}_{(1)}^\circ, \ldots, \hat{\psi}_{(m)}^\circ$ can be used to construct joint confidence bands for Hadamard-differentiable functions of $\psi_0 = (\theta_0, A_0)$ (Theorem 12.1).

For example, this could be used to construct confidence bands for estimated survival curves from a proportional odds model for a given covariate value.

The Piggyback Bootstrap

Corollary 20.3

Assume some conditions in addition to the conditions of Corollary 20.1. Then the conclusions of Corollary 20.1 hold and

$$\sqrt{n}\begin{pmatrix}\theta_n - \hat{\theta}_n \\ \hat{A}^{\circ}_{\theta_n} - \hat{A}_n\end{pmatrix} \underset{M,\xi}{\overset{P}{\rightsquigarrow}} Z(\sigma^{-1}(\cdot)),\ in\ \ell^{\infty}(\mathcal{C}_1).$$