Semiparametric M-Estimation

Jianqiao Wang

January 28, 2022

Overview

General Scheme for Semiparametric M-Estimators

Consider a semiparametric statistical model $P_{\theta,\eta}(X)$, with i.i.d. observations $X_1, \ldots, X_n$ drawn from $P_{\theta,\eta}$, where $\theta \in \mathbb{R}^k$ and $\eta \in H$. Assume that the infinite dimensional space H has norm $\|\cdot\|$, and the true unknown parameter is (θ_0, η_0). An M-estimator $\hat{\theta}_n, \hat{\eta}_n$ of (θ, η) has the form

$$(\hat{\theta}_n, \hat{\eta}_n) = \arg\max \mathbb{P}_n m_{\theta,\eta}(X), \tag{1}$$

where m is a known, measurable function.

For simplicity, we assume the limit criterion function Pm_ψ, where $\psi = (\theta, \eta)$, has a unique and "well-separated" point of maximum ψ_0, i.e., $Pm_{\psi_0} > \sup_{\psi \in G} Pm_\psi$ for every open set G that contains ψ_0.

The Cox model with current status data

In the following paragraphs, derivatives will be denoted with superscript "()".

Let θ be the regression coefficient and Λ the baseline integrated hazard function. The MLE approach to inference for this model was discussed in Chapter 19. As an alternative estimation approach, (θ, Λ) can also be estimated by OLS:

$$(\hat{\theta}_n, \hat{\Lambda}_n) = \arg\min \mathbb{P}_n \left[1 - \delta_i - \exp\{-e^{\theta' Z_i} \Lambda(t_i)\} \right]^2$$

In this model, the nuisance parameter Λ cannot be estimated at the $\sqrt{n}$ rate, but is estimable at the $n^{1/3}$ rate.

Binary regression under misspecified link function

Suppose that we observe an i.i.d. random sample $(Y_1, Z_1, U_1), \ldots, (Y_n, Z_n, U_n)$ consisting of a binary outcome Y, a k-dimensional covariate Z, and a one-dimensional continuous covariate $U \in [0,1]$, following the additive model

$$P_{\theta,h}(Y = 1 \mid Z = z, U = u) = \phi(\theta' z + h(u)),$$

where h is a smooth function belonging to

$$\mathbb{H} = \left\{ h : [0,1] \mapsto [-1,1], \int_0^1 (h^{(s)}(u))^2 du \leq K \right\},$$

for a fixed and known $K \in (0, \infty)$ and an integer $s \geq 1$, and where $\phi : \mathbb{R} \mapsto [0,1]$ is a known continuously differentiable monotone function.

Binary regression under misspecified link function

The choices $\phi(t) = 1/(1+e^{-t})$ and cumulative normal distribution function correspond to the logit model and probit models, respectively. The maximum likelihood estimator $(\hat{\theta}_n, \hat{h}_n)$ maximizes the (conditional) log-likelihood function

$$\ell_n(\theta, h)(X) = \mathbb{P}_n \left(Y \log \phi\{\theta' Z + h(U)\} + (1-Y) \log[1 - \phi\{\theta' Z + h(U)\}] \right),$$

where $X = (Y, Z, U)$. Here, we investigate the estimation of (θ, h) under misspecification of ϕ.

Instead of maximizing the log-likelihood, we can take $(\hat{\beta}_n, \hat{h}_n)$ to be the maximizer o the penalized log-likelihood $\ell_n(\theta, h) - \lambda_n^2 J^2(h)$, where λ_n is a data-driven smoothing parameter.

Mixture models

Suppose that an observation X has a conditional density $p_\theta(x \mid z)$ given an unobservable variable $Z = z$, where p_θ is known up to the Euclidean parameter θ. If the unobservable Z possesses an unknown distribution η, then observation X has the following mixture density $p_{\theta,\eta}(x) = \int p_\theta(x \mid z) d\eta(z)$. The maximum likelihood estimator $(\hat{\theta}_n, \hat{\eta}_n)$ maximizes the log-likelihood function $\ell_n(\theta, \eta) = \mathbb{P}_n \log\{p_{\theta,\eta}(X)\}$.

Examples of mixture models include frailty models, errors-in-variable models in which the errors are modeled by a Gaussian distribution, and scale mixture models over symmetric densities.

Semiparametric M-Estimation

Analysis of the asymptotic behavior of M-estimators can be split into three main steps:

- establishing consistency (argmax theorem);
- establishing a rate of convergence;
- deriving the limiting distribution.

$\sqrt{n}$ Consistency and Asymptotic Normality

Tow approaches:

- Influence function
- Score equation

An influence function approach

For any fixed $\eta \in H$, let $\eta(t)$ be a smooth curve running through η at $t = 0$, that is $\eta(0) = \eta$. Let $a = (\partial/\partial t)\eta(t)|_{t=0}$ be a proper tangent in the tangent set $\dot{\mathcal{P}}^{(\eta)}_{P_{\theta,\eta}}$ for the nuisance parameter. For simplicity, we will use $\mathbb{A}$ to denote $\dot{\mathcal{P}}^{(\eta)}_{P_{\theta,\eta}}$ and $m(\theta, \eta)$ to denote $m(\theta, \eta; X)$. Set

$$m_1(\theta,\eta) = \frac{\partial}{\partial\theta} m(\theta,\eta), \quad m_2(\theta,\eta)[a] = \left.\frac{\partial}{\partial t}\right|_{t=0} m(\theta,\eta(t)),$$

where $a \in \mathbb{A}$. We also define

$$m_{11}(\theta,\eta) = \frac{\partial}{\partial\theta} m_1(\theta,\eta), \quad m_{12}(\theta,\eta)[a] = \left.\frac{\partial}{\partial t}\right|_{t=0} m_1(\theta,\eta(t))$$

$$m_{21}(\theta,\eta)[a] = \frac{\partial}{\partial\theta} m_2(\theta,\eta)[a], \quad m_{22}(\theta,\eta)[a_1][a_2] = \left.\frac{\partial}{\partial t}\right|_{t=0} m_1(\theta,\eta_2(t))[a_1]$$

An influence function approach

If m is a log-likelihood, one way of estimating θ is by solving the efficient score equations.

For general M-estimators, define $m_2(\theta,\eta)[A] = (m_2(\theta,\eta)[a_1], \ldots, m_2(\theta,\eta)[a_k])$, where $A = (a_1, \ldots, a_k)$ and $a_1, \ldots, a_k \in \mathbb{A}$. We define $m_{12}[A_1]$ and $m_{22}[A_1][A_2]$ accordingly, where $A_1 = (a_{11}, \ldots, a_{1k}), A_2 = (a_{21}, \ldots, a_{2k})$ and $a_{ij} \in \mathbb{A}$. Assume there exists an $A^* = (a_1^*, \ldots, a_k^*)$, where $\{a_i^*\} \in \mathbb{A}$, such that for any $A = (a_1, \ldots, a_k), \{a_i\} \in \mathbb{A}$,

$$P(m_{12}(\theta_0, \eta_0)[A] - m_{22}(\theta_0 \eta_0)[A^*][A]) = 0$$

An influence function approach

Define $\tilde{m}(\theta, \eta) = m_1(\theta, \eta) - m_2(\theta, \eta)[A^*]$. θ is then estimated by solving $\mathbb{P}_n \tilde{m}(\theta, \hat{\eta}_n; X) = 0$, where we substitute an estimator $\hat{\eta}_n$ for the unknown nuisance parameter.

A variation of this approach is to obtain an estimator $\hat{\eta}_n(\theta)$ of η for each given value of θ and then solve θ from

$$\mathbb{P}_n \tilde{m}(\theta, \hat{\eta}_n(\theta); X) = 0.$$

In some cases, estimators satisfying the above equation may not exist. Hence we weaken it to the following "nearly-maximizing" condition:

$$\mathbb{P}_n \tilde{m}(\hat{\theta}_n, \hat{\eta}_n) = o_P(n^{-1/2}).$$

An influence function approach

A1 (Consistency and rate of convergence) Assume

$$|\hat{\theta}_n - \theta_0| = o_P(1), \quad \|\hat{\eta}_n - \eta_0\| = O_P(n^{-c_1}),$$

for some $c_1 > 0$, where $|\cdot|$ will be used in this chapter to denote the Euclidean norm.

A2 (Finite variance) $0 < det(I^*) < \infty$, where *det* denotes the determinant of a matrix and

$$\begin{aligned} I^* =& \{P(m_{11}(\theta_0, \eta_0) - m_{21}(\theta_0, \eta_0)[A^*])\}^{-1} \\ & \times P[m_1(\theta_0, \eta_0) - m_2(\theta_0, \eta_0)[A^*]]^{\otimes 2} \\ & \times \{P(m_{11}(\theta_0, \eta_0) - m_{21}(\theta_0, \eta_0)[A^*])\}^{-1} \end{aligned}$$

An influence function approach

A3 (Stochastic equicontinuity) For any $\delta_n \downarrow 0$ and $C > 0$,

$$\sup_{|\theta-\theta_0|\le\delta_n, \|\eta-\eta_0\|\le Cn^{-c_1}} |\sqrt{n}(\mathbb{P}_n - P)(\tilde{m}(\theta,\eta) - \tilde{m}(\theta_0,\eta_0))| = o_P(1).$$

A4 (Smoothness of the model) For some $c_2 > 1$ satisfying $c_1 c_2 > 1/2$ and for all (θ,η) satisfying $\{(\theta,\eta) : |\theta-\theta_0| \le \delta_n, \|\eta-\eta_0\| \le Cn^{-c_1}\}$,

$$\begin{aligned} &\Big|P\Big\{(\tilde{m}(\theta,\eta) - \tilde{m}(\theta_0,\eta_0)) - (m_{11}(\theta_0,\eta_0) - m_{21}(\theta_0,\eta_0)[A^*])(\theta-\theta_0) \\ &- \left(m_{12}(\theta_0,\eta_0)[\frac{\eta-\eta_0}{\|\eta-\eta_0\|}] - m_{22}(\theta_0,\eta_0)[A^*][\frac{\eta-\eta_0}{\|\eta-\eta_0\|}]\right)\|\eta-\eta_0\|\Big\}\Big| \\ =&o(|\theta-\theta_0|) + O(\|\eta-\eta_0\|^{c_2}). \end{aligned}$$

An influence function approach

- Condition A2 corresponds to the nonsingular information condition for the MLE.
- Condition A3 can be verified via entropy calculations and certain maximal inequalities.
- Condition A4 can be checked via Taylor expansion techniques for functionals.

An influence function approach

Theorem

Suppose that $(\hat{\theta}_n, \hat{\eta}_n)$ satisfies $\mathbb{P}_n \tilde{m}(\hat{\theta}_n, \hat{\eta}_n) = o_P(n^{-1/2})$, and that Conditions A1-A4 hold, then

$$\sqrt{n}(\hat{\theta}_n - \theta_0) = -\sqrt{n}\{P(m_{11}(\theta_0, \eta_0) - m_{21}(\theta_0, \eta_0)[A^*])\}^{-1} \times \mathbb{P}_n(m_1(\theta_0, \eta_0) - m_2(\theta_0, \eta_0)[A^*]) + o_P(1).$$

Hence $\sqrt{n}(\hat{\theta}_n - \theta_0)$ is asymptotically normal with mean zero and variance I^*.

A score equation approach

By definition, M-estimators maximize an objective function

$$(\hat{\theta}_n, \hat{\eta}_n) = \arg\max \mathbb{P}_n m(\theta, \eta; X).$$

We have

$$\mathbb{P}_n m_1(\hat{\theta}_n, \hat{\eta}_n) = 0, \quad \mathbb{P}_n m_2(\hat{\theta}_n, \hat{\eta}_n)[a] = 0,$$

where a runs over $\mathbb{A}$. We can relax above equations to the following "nearly-maximizing" conditions:

$$\mathbb{P}_n m_1(\hat{\theta}_n, \hat{\eta}_n) = o_P(n^{-1/2}), \quad \mathbb{P}_n m_2(\hat{\theta}_n, \hat{\eta}_n)[a] = o_P(n^{-1/2}),$$

for all $a \in \mathbb{A}$.

A score equation approach

B3 (Stochastic equicontinuity) For any $\delta_n \downarrow 0$ and $C > 0$,

$$sup_{|\theta-\theta_0|\leq\delta_n, \|\eta-\eta_0\|\leq Cn^{-c_1}} |\sqrt{n}(\mathbb{P}_n - P)(m_1(\theta,\eta) - m_1(\theta_0,\eta_0))| = o_P(1),$$

$$sup_{|\theta-\theta_0|\leq\delta_n, \|\eta-\eta_0\|\leq Cn^{-c_1}} |\sqrt{n}(\mathbb{P}_n - P)(m_2(\theta,\eta) - m_2(\theta_0,\eta_0))[A^*]| = o_P(1)$$

where c_1 is as in Condition A1.

A score equation approach

B4 (Smoothness of the model) For some $c_2 > 1$ satisfying $c_1 c_2 > 1/2$ and for all (θ, η) satisfying $\{(\theta, \eta) : |\theta - \theta_0| \leq \delta_n, \|\eta - \eta_0\| \leq Cn^{-c_1}\}$,

$$
\begin{aligned}
&\left| P\left\{ m_1(\theta, \eta) - m_1(\theta_0, \eta_0) - m_{11}(\theta_0, \eta_0)(\theta - \theta_0) \right.\right. \\
&\left.\left. - m_{12}(\theta_0, \eta_0)[\frac{\eta - \eta_0}{\|\eta - \eta_0\|}]\|\eta - \eta_0\| \right\} \right| \\
=& o(|\theta - \theta_0|) + O(\|\eta - \eta_0\|^{c_2}),
\end{aligned}
$$

and

$$
\begin{aligned}
&\left| P\left\{ m_2(\theta, \eta)[A^*] - m_2(\theta_0, \eta_0)[A^*] - m_{21}(\theta_0, \eta_0)[A^*](\theta - \theta_0) \right.\right. \\
&\left.\left. - m_{22}(\theta_0, \eta_0)[A^*][\frac{\eta - \eta_0}{\|\eta - \eta_0\|}]\|\eta - \eta_0\| \right\} \right| \\
=& o(|\theta - \theta_0|) + O(\|\eta - \eta_0\|^{c_2}).
\end{aligned}
$$

A score equation approach

Corollary

Suppose that the estimator $(\hat{\theta}_n, \hat{\eta}_n)$ *satisfies*

$$\mathbb{P}_n m_1(\hat{\theta}_n, \hat{\eta}_n) = o_P(n^{-1/2}), \quad \mathbb{P}_n m_2(\hat{\theta}_n, \hat{\eta}_n)[a] = o_P(n^{-1/2}),$$

for all $a \in \mathbb{A}$, *and Conditions A1, A2, B3 and B4 all hold. Then*

$$\begin{aligned}\sqrt{n}(\hat{\theta}_n - \theta_0) = &- \sqrt{n}\{P(m_{11}(\theta_0, \eta_0) - m_{21}(\theta_0, \eta_0)[A^*])\}^{-1} \\ &\times \mathbb{P}_n(m_1(\theta_0, \eta_0) - m_2(\theta_0, \eta_0)[A^*]) + o_P(1).\end{aligned}$$

Hence $\sqrt{n}(\hat{\theta}_n - \theta_0)$ *is asymptotically normal with mean zero and variance* I^*.

Weighted M-Estimators and the Weighted Bootstrap

The weighted bootstrap is an effective and nearly universal inference tool for semiparametric M-estimation. We first study the unconditional behavior of weighted M-estimators and then use these results to establish conditional asymptotic validity of the weighted bootstrap.

Consider n i.i.d. observations $X_1, \ldots, X_n$ drawn from the true distribution P. Denote $\xi_i, i = 1, \ldots, n$ as n i.i.d. positive random weights, satisfying $E(\xi) = 1$ and $0 \leq var(\xi) = v_0 < \infty$ and which are independent of the data $\mathcal{X}_n = \sigma\{X_1, \ldots, X_n\}$.

Weighted M-Estimators and the Weighted Bootstrap

The weighted M-estimator $(\hat{\theta}_n^\circ, \hat{\eta}_n^\circ)$ satisfies

$$(\hat{\theta}_n^\circ, \hat{\eta}_n^\circ) = \arg\max \mathbb{P}_n\{\xi m(\theta, \eta; X)\}.$$

Since we assume the random weights are independent of $\mathcal{X}_n$, the consistency and convergence rate for the estimators of all parameters can be established using previous theorems in Chapter 2 and Chapter 14.

Weighted M-Estimators and the Weighted Bootstrap

Assume that the estimator $(\hat{\theta}_n^\circ, \hat{\eta}_n^\circ)$ satisfies

$$\mathbb{P}_n^\circ \tilde{m}(\hat{\theta}_n^\circ, \hat{\eta}_n^\circ) = \mathbb{P}_n\{\xi \tilde{m}(\hat{\theta}_n^\circ, \hat{\eta}_n^\circ)\} = o_P(n^{-1/2}).$$

We now investigate the unconditional limiting distribution of $\hat{\theta}_n^\circ$:

Corollary

Replace all $\tilde{m}$ in the previous theorem with $\xi \tilde{m}$. Suppose Conditions A1-A4 hold, then

$$\sqrt{n}(\hat{\theta}_n^\circ - \theta_0^\circ) = -\sqrt{n}\{P(m_{11}(\theta_0, \eta_0) - m_{21}(\theta_0, \eta_0)[A^*])\}^{-1} \times \mathbb{P}_n^\circ(m_1(\theta_0, \eta_0) - m_2(\theta_0, \eta_0)[A^*]) + o_P(1).$$

Thus $\sqrt{n}(\hat{\theta}_n^\circ - \theta)$ is asymptotically normal with variance $(1 + v_0)I^*$.

Weighted M-Estimators and the Weighted Bootstrap

Corollary

Suppose that the estimator $(\hat{\theta}_n^\circ, \hat{\eta}_n^\circ)$ *satisfies*

$$\mathbb{P}_n^\circ m_1(\hat{\theta}_n^\circ, \hat{\eta}_n^\circ) = o_P(n^{-1/2}), \quad \mathbb{P}_n^\circ m_2(\hat{\theta}_n^\circ, \hat{\eta}_n^\circ)[a] = o_P(n^{-1/2}),$$

for all $a \in \mathbb{A}$*, and Conditions A1, A2, B3 and B4 all hold. Then*

$$\begin{aligned}\sqrt{n}(\hat{\theta}_n^\circ - \theta_0^\circ) = &- \sqrt{n}\{P(m_{11}(\theta_0, \eta_0) - m_{21}(\theta_0, \eta_0)[A^*])\}^{-1} \\ &\times \mathbb{P}_n^\circ(m_1(\theta_0, \eta_0) - m_2(\theta_0, \eta_0)[A^*]) + o_P(1).\end{aligned}$$

Thus $\sqrt{n}(\hat{\theta}_n^\circ - \theta)$ *is asymptotically normal with variance* $(1 + v_0)I^*$*.*

Weighted M-Estimators and the Weighted Bootstrap

The above results can be used to justify the use of weighted bootstrap for general M-estimators. The following theorem shows that the weighted bootstrap is asymptotically valid for inference on $\hat{\theta}_n$.

Theorem

Suppose the M-estimator $\hat{\theta}_n$, and the weighted M-estimator $\hat{\theta}_n^\circ$ satisfy:

$$\sqrt{n}(\hat{\theta}_n - \theta_0) = \tilde{I}_0^{-1}\sqrt{n}\mathbb{P}_n\tilde{m} + o_P(1), \sqrt{n}(\hat{\theta}_n^\circ - \theta_0) = \tilde{I}_0^{-1}\sqrt{n}\mathbb{P}_n^\circ\tilde{m} + o_P(1).$$

Assume that the conclusions of previous theorem and corollary hold. Then we have $\sqrt{n}(\hat{\theta}_n^\circ - \hat{\theta}_n) = \tilde{I}_0^{-1}\sqrt{n}(\mathbb{P}_n^\circ - \mathbb{P}_n)\tilde{m} + o_P(1)$. Since $E(\xi) = 1$ and ξ is independent of $\mathcal{X}_n$, $\sqrt{n/v_0}(\hat{\theta}_n^\circ - \hat{\theta}_n) \underset{\xi}{\overset{P}{\rightsquigarrow}} Z_0$, where $\underset{\xi}{\overset{P}{\rightsquigarrow}}$ denotes conditional convergence given the data $\mathcal{X}_n$, and Z_0 is mean zero Gaussian with covariance I^*.