Lec 8: Introduction to semiparametric models

Yaejun Han

Nov 7. 2023

Previous lectures: parametric models $Y_1, \cdots, Y_n \sim P_\theta$.
$\theta \in \mathbb{R}^p$ is finite dimensional

This lecture: semiparametric models $Y_1, \cdots, Y_n \sim P_{\theta, \eta}$:

θ:	target parameter (typically finite-dimensional)
η:	nuisance parameter (could be infinite-dimensional)

Historic remark: symmetric location family
(by C. Stein, "efficient nonparametric testing & estimation", 1956)

Model: $Y_1, \cdots, Y_n \sim f(\cdot - \theta)$, where
- $\theta \in \mathbb{R}$: target location parameter
- f: unknown density symmetric around zero (i.e. $f(x) = f(-x)$)
(nuisance)

Estimators for θ:

1. sample mean $\bar{Y} = \frac{1}{n}\sum_{i=1}^n Y_i$

$$\mathbb{E}_\theta[Y_1] = \int y f(y-\theta)\,dy = \theta + \int (y-\theta) f(y-\theta)\,dy = \theta$$
$$\Rightarrow \quad \mathbb{E}_\theta[\bar{Y}] = \theta$$
$$\mathrm{Var}_\theta(\bar{Y}) = \frac{1}{n}\int y^2 f(y)\,dy.$$

2. efficient estimator with known f: **MLE**

$$\hat{\theta}^{MLE} = \operatorname*{argmax}_\theta \frac{1}{n}\sum_i \log f(Y_i - \theta) \Rightarrow \frac{1}{n}\sum_{i=1}^n \frac{f'(Y_i - \hat{\theta}^{MLE})}{f(Y_i - \hat{\theta}^{MLE})} = 0.$$

Using Fisher info, one can show that

$$\mathbb{E}_\theta[(\hat{\theta}^{MLE}-\theta)^2] = \frac{1+o_n(1)}{n}\left(\int \frac{f'(y)^2}{f(y)}dy\right)^{-1}$$

↑ asymptotically optimal MSE

3. What about unknown f?

Stein (1956) showed that if we use some nonparametric procedures to estimate f by $\hat{f}$, then estimate θ by

$$\hat{\theta}: \quad \frac{1}{n}\sum_{i=1}^{n}\frac{\hat{f}'(y_i-\hat{\theta})}{\hat{f}(y_i-\hat{\theta})} = 0 \quad \text{(plug-in approach)}$$

then

$$\mathbb{E}_\theta[(\hat{\theta}-\theta)^2] = \frac{1+o_n(1)}{n}\left(\int \frac{f'(y)^2}{f(y)}dy\right)^{-1}$$

semiparametric efficient!

(the same asymptotic efficiency can be achieved without knowing the nuisance; NOT all semiparametric problems admit semiparametric efficient estimators)

Key ideas behind semiparametric models:

- do not want to propose a restrictive model for the nuisance;
- hope that even if the nuisance estimation error is large, the target estimation error is still small;
- orthogonality will play a central role.

Examples.

1. Linear regression.

$$Y = X\theta_0 + \varepsilon, \quad \mathbb{E}[\varepsilon \mid X] = 0$$

Target: $\theta_0 \in \mathbb{R}^p$

Nuisance: distribution of ε

(Remark: we do not assume that $\varepsilon \sim N(0, \sigma^2)$, nor the independence of (X, ε))

2. Partial linear regression:

$$\begin{cases} Y = D\theta_0 + g_0(X) + \varepsilon_1, & \mathbb{E}[\varepsilon_1 \mid X, D] = 0 \\ D = m_0(X) + \varepsilon_2, & \mathbb{E}[\varepsilon_2 \mid X] = 0 \end{cases}$$

Data: (X_i, D_i, Y_i)

Target: θ_0

Nuisance: (g_0, m_0, distributions of $(\varepsilon_1, \varepsilon_2)$)

(closely related to the potential outcome model in causal inference next lecture)

3. Errors in variables:

$$\begin{cases} Y = \alpha + \beta Z + \varepsilon_1, & \varepsilon_1 \sim N(0, \sigma_1^2) \\ X = Z + \varepsilon_2, & \varepsilon_2 \sim N(0, \sigma_2^2) \end{cases}$$

Data: (X_i, Y_i)

Target: (α, β)

Nuisance: distribution of Z.

4. Cox model:

$$h(t \mid x) = e^{\beta^T x} h(t)$$

Target: β

Nuisance: baseline hazard h.

Estimation. Joint/profile MLE: given $y_1, \dots, y_n \sim P_{\theta,\eta}(y)$, compute

$$(\hat\theta, \hat\eta) = \operatorname*{argmax}_{(\theta,\eta)} \frac{1}{n}\sum_{i=1}^n \log P_{\theta,\eta}(y_i)$$

or $$\hat\theta = \operatorname*{argmax}_\theta \left(\max_\eta \frac{1}{n}\sum_{i=1}^n \log P_{\theta,\eta}(y_i) \right).$$

Sometimes works (e.g. in Cox model), but in many cases computationally infeasible.

A simplified question:
Suppose we are given a (possibly coarse) estimator $\hat\eta$ of η. How should we use $\hat\eta$ to estimate θ?

Score function & estimating equation

Score. For $y \sim P_{\theta_0}$, the score of y at θ_0 is

$$s_{\theta_0}(y) = \nabla_\theta \log P_\theta(y)\Big|_{\theta=\theta_0}.$$

Relationships between score and MLE

For $y_1, \dots, y_n \sim P_\theta$, the MLE for θ is

$$\hat\theta = \operatorname*{argmax}_\theta \frac{1}{n}\sum_{i=1}^n \log P_\theta(y_i)$$

F.O.C. $\Longrightarrow$

$$0 = \nabla_\theta \left[\frac{1}{n}\sum_{i=1}^n \log P_\theta(y_i)\right]\Big|_{\theta=\hat\theta} = \frac{1}{n}\sum_{i=1}^n s_{\hat\theta}(y_i)$$

(estimating eqn. for MLE)

Another interpretation.

Lemma. $\mathbb{E}_{\theta_0}[s_{\theta_0}(y)] = 0$ for all θ_0.

Pf.

$$\begin{aligned}
\mathbb{E}_{\theta_0}[s_{\theta_0}(y)] &= \mathbb{E}_{\theta_0}\left[\nabla_\theta \log p_\theta(y)\big|_{\theta=\theta_0}\right] \\
&= \mathbb{E}_{\theta_0}\left[\frac{\nabla_\theta p_\theta(y)\big|_{\theta=\theta_0}}{p_{\theta_0}(y)}\right] \\
&= \int \cancel{p_{\theta_0}(y)} \frac{\nabla_\theta p_\theta(y)\big|_{\theta=\theta_0}}{\cancel{p_{\theta_0}(y)}}\, dy \\
&= \nabla_\theta \underbrace{\int p_\theta(y)\, dy}_{=1}\Big|_{\theta=\theta_0} = 0
\end{aligned}$$

□

View estimating equation in terms of score matching:

$$\underbrace{\frac{1}{n}\sum_{i=1}^n s_\theta(y_i)}_{\text{empirical score at } \theta} = \underbrace{0}_{= \mathbb{E}_{\theta_0}[s_{\theta_0}(y)] \text{ is true score at } \theta_0}$$

(intuition: solve for θ_0 from

$$0 = \mathbb{E}_{\theta_0}[s_{\theta_0}(y)] \approx \frac{1}{n}\sum_{i=1}^n s_{\theta_0}(y_i).$$)

General estimating equation.

1. find $f(\theta, y) \in \mathbb{R}^p$ s.t. $\mathbb{E}_{\theta_0}[f(\theta_0, y)] = 0$
2. estimate θ_0 by $\hat{\theta}$ from the estimating eqn.

$$\frac{1}{n}\sum_{i=1}^n f(\hat{\theta}, y_i) = 0.$$

Example 1. $(X_1, Y_1), \cdots, (X_n, Y_n) \sim N\left(\begin{bmatrix}\theta_0 \\ \eta_0\end{bmatrix}, \begin{bmatrix}1 & \rho \\ \rho & 1\end{bmatrix}\right)$

unknown: (θ_0, η_0) known: ρ.

$$\log p_{\theta,\eta}(x,y) = \text{const} - \frac{(x-\theta)^2 + (y-\eta)^2 - 2\rho(x-\theta)(y-\eta)}{2(1-\rho^2)}$$

$$S_{\theta_0,\eta_0}(x,y) = \begin{bmatrix} \nabla_\theta \log p_{\theta,\eta}(x,y)\big|_{\substack{\theta=\theta_0 \\ \eta=\eta_0}} \\ \nabla_\eta \log p_{\theta,\eta}(x,y)\big|_{\substack{\theta=\theta_0 \\ \eta=\eta_0}} \end{bmatrix} = \frac{1}{1-\rho^2}\begin{bmatrix} x-\theta_0-\rho(y-\eta_0) \\ y-\eta_0-\rho(x-\theta_0) \end{bmatrix}$$

MLE estimating equation:

$$\begin{cases} \frac{1}{n}\sum_{i=1}^n [(x_i-\hat\theta)-\rho(y_i-\hat\eta)] = 0 \\ \frac{1}{n}\sum_{i=1}^n [(y_i-\hat\eta)-\rho(x_i-\hat\theta)] = 0 \end{cases}$$

Example 2 $Y_i = \langle \theta_0, X_i\rangle + \varepsilon_i, \quad \mathbb{E}[\varepsilon_i | X_i] = 0, \quad i = 1, \cdots, n,$

Let $f(\theta,(x,y)) = (y - \langle\theta, x\rangle)x \in \mathbb{R}^p$, then

$$\mathbb{E}_{\theta_0}[f(\theta_0,(x,y))] = \mathbb{E}_{\theta_0}[(y-\langle\theta_0,x\rangle)x]$$
$$= \mathbb{E}_{\theta_0}[\varepsilon x] = \mathbb{E}_{\theta_0}[\mathbb{E}[\varepsilon|x]x] = 0$$

$\Rightarrow$ estimating eqn:

$$\frac{1}{n}\sum_{i=1}^n (Y_i - \langle\hat\theta, X_i\rangle)X_i = 0$$
$$\Rightarrow \hat\theta = (X^TX)^{-1}X^TY \quad \text{(least squares)}$$

Question in semiparametric models for example 1:
If η_0 is a nuisance parameter and $\hat\eta$ is given to us, which estimating eqn. should we use?

Efficient score function.

Let $y \sim P_{\theta_0,\eta_0}$ in a semiparametric model with target θ_0 and nuisance η_0 (for simplicity we assume $\theta_0, \eta_0 \in \mathbb{R}$)

$$\text{Score function} \quad s_{\theta_0,\eta_0}(y) = \begin{bmatrix} s^{\theta}_{\theta_0,\eta_0}(y) \\ s^{\eta}_{\theta_0,\eta_0}(y) \end{bmatrix} = \begin{bmatrix} \nabla_\theta \log p_{\theta,\eta}(y) \\ \nabla_\eta \log p_{\theta,\eta}(y) \end{bmatrix}\Bigg|_{\substack{\theta=\theta_0 \\ \eta=\eta_0}}$$

Efficient score function for θ_0:

$$s^{\text{eff}}_{\theta_0,\eta_0}(y) = s^{\theta}_{\theta_0,\eta_0}(y) - \frac{\mathbb{E}_{\theta_0,\eta_0}\left[s^{\theta}_{\theta_0,\eta_0}(y)\, s^{\eta}_{\theta_0,\eta_0}(y)\right]}{\mathbb{E}_{\theta_0,\eta_0}\left[s^{\eta}_{\theta_0,\eta_0}(y)^2\right]} s^{\eta}_{\theta_0,\eta_0}(y)$$

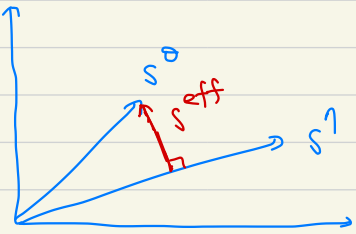
Estimating eqn. for θ_0: given $\hat{\eta}$, solve

$$\frac{1}{n}\sum_{i=1}^{n} s^{\text{eff}}_{\theta,\hat{\eta}}(y_i) = 0 \implies \hat{\theta}$$

Geometric interpretation of s^{eff}:

Gram-Schmidt orthogonalization of s^{θ} with respect to s^{η} in $L^2(P_{\theta_0,\eta_0})$

("orthogonalization")



Example 1 (continued).

$$s^{\theta}(x,y) = \frac{1}{1-\rho^2}\left[(x-\theta_0) - \rho(y-\eta_0)\right]$$
$$s^{\eta}(x,y) = \frac{1}{1-\rho^2}\left[(y-\eta_0) - \rho(x-\theta_0)\right]$$

$$\mathbb{E}_{\theta_0,\eta_0}\left[s^{\theta}(x,y)\, s^{\eta}(x,y)\right] = \frac{1}{(1-\rho^2)^2}\left[(1+\rho^2)\rho - 2\rho\right] = -\frac{\rho}{1-\rho^2}$$

$$\mathbb{E}_{\theta_0,\eta_0}\left[s^{\eta}(x,y)^2\right] = \frac{1}{(1-\rho^2)^2}\left[1+\rho^2-2\rho^2\right] = \frac{1}{1-\rho^2}$$

$$
\begin{aligned}
s^{\text{eff}}(x,y) &= s^{\theta}(x,y) - \frac{-\frac{\rho}{1-\rho^2}}{1/(1-\rho^2)} s^{\eta}(x,y) \\
&= s^{\theta}(x,y) + \rho s^{\eta}(x,y) \\
&= \frac{1}{1-\rho^2}\left[(x-\theta_0) - \rho(y-\eta_0) + \rho(y-\eta_0) - \rho^2(x-\theta_0)\right] \\
&= x - \theta_0
\end{aligned}
$$

Estimating eqn. based on efficient score:

$$
\frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{\theta}) = 0 \quad \Longrightarrow \quad \hat{\theta} = \frac{1}{n}\sum_{i=1}^{n} x_i
$$

(independent of $\hat{\eta}$)

Example 3 (Stein's symmetric location model)

$$
y \sim f_{\eta_0}(\cdot - \theta_0)
$$

(f symmetric around zero; assumed to be parametrized by η_0)

$$
s^{\theta}_{\theta_0,\eta_0}(y) = \frac{\partial}{\partial\theta}\log f_{\eta}(y-\theta)\Big|_{\substack{\theta=\theta_0\\ \eta=\eta_0}} = -\frac{f'_{\eta_0}(y-\theta_0)}{f_{\eta_0}(y-\theta_0)}
$$

$$
s^{\eta}_{\theta_0,\eta_0}(y) = \frac{\partial}{\partial\eta}\log f_{\eta}(y-\theta)\Big|_{\substack{\theta=\theta_0\\ \eta=\eta_0}} = \frac{1}{f_{\eta_0}(y-\theta_0)}\cdot\frac{\partial}{\partial\eta} f_{\eta}(y-\theta_0)\Big|_{\eta=\eta_0}
$$

$$
\mathbb{E}_{\theta_0,\eta_0}\left[s^{\theta}_{\theta_0,\eta_0}(y)\, s^{\eta}_{\theta_0,\eta_0}(y)\right] = \mathbb{E}_{\theta_0,\eta_0}\left[-\frac{f'_{\eta_0}(y-\theta_0)}{f_{\eta_0}(y-\theta_0)^2}\frac{\partial}{\partial\eta} f_{\eta}(y-\theta_0)\Big|_{\eta=\eta_0}\right]
$$

($\frac{f'_{\eta_0}(y-\theta_0)}{f_{\eta_0}(y-\theta_0)^2}$: symmetric around θ_0; $\frac{\partial}{\partial\eta} f_{\eta}(y-\theta_0)\big|_{\eta=\eta_0}$: anti-symmetric around θ_0)

$$
\Longrightarrow \quad s^{\text{eff}}(y) = s^{\theta}_{\theta_0,\eta_0}(y) = -\frac{f'_{\eta_0}(y-\theta_0)}{f_{\eta_0}(y-\theta_0)} \qquad (= 0)
$$

Estimating eqn: based on $\hat{f} = f_{\hat{\eta}}$, solve $\hat{\theta}$ from

$$
\frac{1}{n}\sum_{i=1}^{n}\frac{\hat{f}'(y_i-\hat{\theta})}{\hat{f}(y_i-\hat{\theta})} = 0 \quad \text{(Stein's estimator)}
$$

Why efficient score?

Neyman orthogonality: an estimating eqn. $f((\theta,\eta), y)$ is **Neyman orthogonal** iff

$$\mathbb{E}_{\theta_0,\eta_0}\left[\nabla_\eta f((\theta_0,\eta), y)\Big|_{\eta=\eta_0}\right] = 0$$

Insights: Neyman orthogonal

$$\Rightarrow \mathbb{E}_{\theta_0,\eta_0}\left[f((\theta_0,\hat{\eta}), y)\right] \overset{\text{Taylor expansion around } \hat{\eta}\approx\eta_0}{\approx} \mathbb{E}_{\theta_0,\eta_0}\left[f((\theta_0,\eta_0), y)\right] = 0$$

↳ requirement of estimating eqn.

(i.e. nuisance estimation error has second-order effects).

Thm: efficient scores are Neyman orthogonal.

Pf (optional): $\nabla_\eta s^{\text{eff}}_{\theta,\eta}(y) = \nabla_\eta\left[s^\theta_{\theta,\eta}(y) - \alpha(\theta,\eta)\, s^\eta_{\theta,\eta}(y)\right]$

$$\left(\alpha(\theta,\eta) = \frac{\mathbb{E}_{\theta,\eta}\left[s^\theta_{\theta,\eta}(y)\, s^\eta_{\theta,\eta}(y)\right]}{\mathbb{E}_{\theta,\eta}\left[s^\eta_{\theta,\eta}(y)^2\right]}\right)$$

$$= \underbrace{\nabla_\eta s^\theta_{\theta,\eta}(y) - \alpha(\theta,\eta)\nabla_\eta s^\eta_{\theta,\eta}(y)}_{(*)} - \underbrace{\nabla_\eta\alpha(\theta,\eta)}_{\text{does not depend on } y} \cdot \underbrace{s^\eta_{\theta,\eta}(y)}_{\text{has expectation zero}}$$

$\mathbb{E}_{\theta,\eta}[\cdot] = 0$

can show:

$$\mathbb{E}_{\theta,\eta}\left[\nabla_\eta s^\theta_{\theta,\eta}(y)\right] = -\mathbb{E}_{\theta,\eta}\left[s^\theta_{\theta,\eta}(y)\, s^\eta_{\theta,\eta}(y)\right]$$

$$\mathbb{E}_{\theta,\eta}\left[\nabla_\eta s^\eta_{\theta,\eta}(y)\right] = -\mathbb{E}_{\theta,\eta}\left[s^\eta_{\theta,\eta}(y)^2\right]$$

$\Rightarrow \mathbb{E}_{\theta,\eta}[\cdot] = 0$ by defn. of $\alpha(\theta,\eta)$

Proof of (*):

$$0 = \nabla_\eta \mathbb{E}_{\theta,\eta}\left[s^\theta_{\theta,\eta}(y)\right]$$
$$= \nabla_\eta \int p_{\theta,\eta}(y)\, s^\theta_{\theta,\eta}(y)\, dy$$
$$= \int\left(\nabla_\eta p_{\theta,\eta}(y)\cdot s^\theta_{\theta,\eta}(y) + p_{\theta,\eta}(y)\cdot\nabla_\eta s^\theta_{\theta,\eta}(y)\right) dy$$
$$= \mathbb{E}_{\theta,\eta}\left[s^\theta_{\theta,\eta}(y)\, s^\eta_{\theta,\eta}(y) + \nabla_\eta s^\theta_{\theta,\eta}(y)\right].$$

□