Lec 8 : Introduction to semiparametric models

Yaojun Han

Nov 5. 2024

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
  (nuisance) (i.e. $f(x) = f(-x)$)

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: $\boxed{\text{MLE}}$

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

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

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

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

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

$$\hat\theta = \underset{\theta}{\operatorname{argmax}} \frac{1}{n}\sum_{i=1}^{n} \log p_\theta(y_i)$$

$$\overset{\text{F.O.C.}}{\Longrightarrow} \quad 0 = \nabla_\theta \left[\frac{1}{n}\sum_{i=1}^{n} \log p_\theta(y_i)\right]\Big|_{\theta=\hat\theta} = \frac{1}{n}\sum_{i=1}^{n} s_{\hat\theta}(y_i)$$

(estimating eqn. for MLE)

Another interpretation.

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

Pf. $\mathbb{E}_{\theta_0}[s_{\theta_0}(y)] = \mathbb{E}_{\theta_0}\left[\nabla_\theta \log p_\theta(y) \big|_{\theta=\theta_0} \right]$

$$= \mathbb{E}_{\theta_0}\left[\frac{\nabla_\theta p_\theta(y)\big|_{\theta=\theta_0}}{p_{\theta_0}(y)} \right]$$

$$= \int \cancel{p_{\theta_0}(y)} \frac{\nabla_\theta p_\theta(y)\big|_{\theta=\theta_0}}{\cancel{p_{\theta_0}(y)}} dy$$

$$= \nabla_\theta \underbrace{\int p_\theta(y)\, dy}_{=1} \Big|_{\theta=\theta_0} = 0 \qquad \square$$

View estimating equation in terms of score matching:

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

(intuition: solve for θ_0 from

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

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. $(x_1, y_1), \dots, (x_n, y_n) \sim N\left(\begin{bmatrix}\theta_0\\ \eta_0\end{bmatrix}, \begin{bmatrix}1 & \rho\\ \rho & 1\end{bmatrix}\right)$

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:

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

Example 2 $y_i = \langle \theta_0, x_i\rangle + \varepsilon_i$, $\mathbb{E}[\varepsilon_i | x_i] = 0$, $i = 1, \dots, n$.

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

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

$\Longrightarrow$ estimating eqn: $\frac{1}{n}\sum_{i=1}^n (y_i - \langle\hat\theta, x_i\rangle)x_i = 0$

$\Longrightarrow$ $\hat\theta = (X^T X)^{-1} X^T Y$ (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:

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

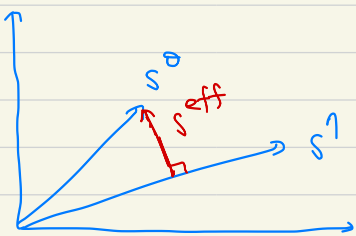
Estimating eqn. for θ_0: given $\hat{\eta}$, solve

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

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).

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

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

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

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

Estimating eqn. based on efficient score:

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

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

<u>Example 3</u> (Stein's symmetric location model)

$$y \sim f_{\eta_0}(\cdot - \theta_0)$$ (f symmetric around zero; assumed to be parametrized by η.)

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

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

anti-symmetric around θ_0

$$\mathbb{E}_{\theta_0,\eta_0}\left[s^{\theta}_{\theta_0,\eta_0}(y)\, s^{\eta}_{\theta_0,\eta_0}(y)\right] = \mathbb{E}_{\theta_0,\eta_0}\left[-\frac{f'_{\eta_0}(y-\theta_0)}{f_{\eta_0}(y-\theta_0)^2}\frac{\partial}{\partial\eta}f_{\eta}(y-\theta_0)\Big|_{\eta=\eta_0}\right] = 0$$

symmetric around θ_0

$$\implies s^{\mathrm{eff}}(y) = s^{\theta}_{\theta_0,\eta_0}(y) = -\frac{f'_{\eta_0}(y-\theta_0)}{f_{\eta_0}(y-\theta_0)}$$

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

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

Taylor expansion around $\hat{\eta}\approx\eta_0$

$$\Rightarrow\quad \mathbb{E}_{\theta_0,\eta_0}[f((\theta_0,\hat{\eta}),y)]\approx\mathbb{E}_{\theta_0,\eta_0}[f((\theta_0,\eta_0),y)]=0$$

requirement of estimating eqn.

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

Thm: efficient scores are Neyman orthogonal.

Pf (optional):

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

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

$$=\underbrace{\nabla_\eta s^\theta_{\theta,\eta}(y)-\alpha(\theta,\eta)\nabla_\eta s^\eta_{\theta,\eta}(y)}_{\mathbb{E}_{\theta,\eta}[\cdot]=0}\;-\;\underbrace{\underbrace{\nabla_\eta\alpha(\theta,\eta)}_{\text{does not depend on } y}\cdot\underbrace{s^\eta_{\theta,\eta}(y)}_{\text{has expectation zero}}}_{\mathbb{E}_{\theta,\eta}[\cdot]=0}$$

can show:

$$\left.\begin{aligned}\mathbb{E}_{\theta,\eta}[\nabla_\eta s^\theta_{\theta,\eta}(y)]&=-\mathbb{E}_{\theta,\eta}[s^\theta_{\theta,\eta}(y)\,s^\eta_{\theta,\eta}(y)]\\ \mathbb{E}_{\theta,\eta}[\nabla_\eta s^\eta_{\theta,\eta}(y)]&=-\mathbb{E}_{\theta,\eta}[s^\eta_{\theta,\eta}(y)^2]\end{aligned}\right\}\ (*)$$

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

Proof of $(*)$:

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

$\square$