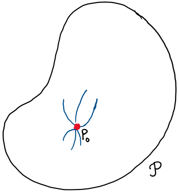
Augmenting Statistical Inference with Machine Learning II

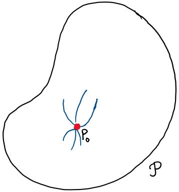
Rajen D. Shah (Statistical Laboratory, University of Cambridge)

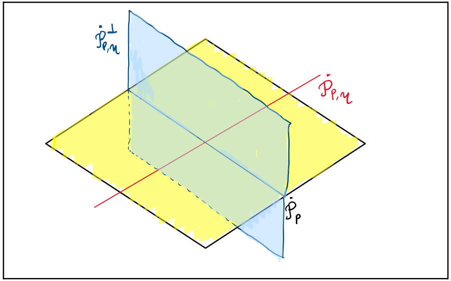
ENAR Spring Meeting 2026
17 March 2026

Recap



Recap





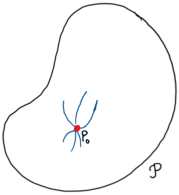
$\dot{\mathcal{P}}_{P,\eta}^{\perp}$
$\dot{\mathcal{P}}_{P,\eta}$
$\dot{\mathcal{P}}_{P}$

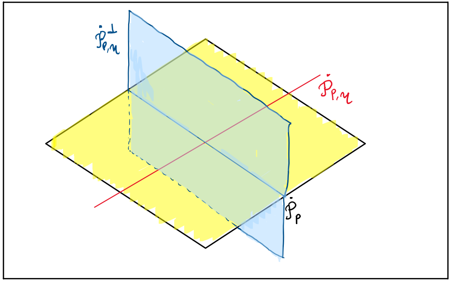
Estimate nuisance parameters $\hat{\eta}$.

For $\psi_{\theta,\eta}$ in the nuisance tangent space, solve the estimating equation

$$\sum_{i=1}^{n} \psi_{\theta,\hat{\eta}}(Y_i, X_i, Z_i).$$

Recap



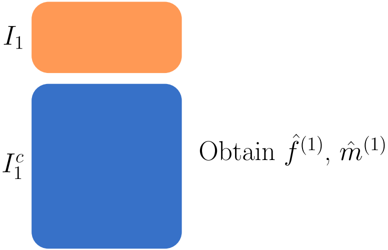


$\dot{\mathcal{P}}_{P,\eta}^{\perp}$
$\dot{\mathcal{P}}_{P,\eta}$
$\dot{\mathcal{P}}_{P}$

Estimate nuisance parameters $\hat{\eta}$.

For $\psi_{\theta,\eta}$ in the nuisance tangent space, solve the estimating equation

$$\sum_{i=1}^{n} \psi_{\theta,\hat{\eta}}(Y_i, X_i, Z_i).$$



Obtain $\hat{f}^{(1)}$, $\hat{m}^{(1)}$

Overview

Lecture 2

- Conditional independence testing
- Optimal inference in semiparametric models
- Nonparametric models
- Optimal ‘robust’ inference in semiparametric models
- Grouped data

Conditional independence testing

Revisiting the PLM

In the PLM $Y = X\theta + f(Z) + \varepsilon$, we can estimate θ via

$$\hat{\theta} = \frac{\sum_i \{Y_i - \hat{m}_Y(Z_i)\}\{X_i - \hat{m}_X(Z_i)\}}{\sum_i \{X_i - \hat{m}_X(Z_i)\}^2}.$$

Revisiting the PLM

In the PLM $Y = X\theta + f(Z) + \varepsilon$, we can estimate θ via

$$\hat{\theta} = \frac{\sum_i \{Y_i - \hat{m}_Y(Z_i)\}\{X_i - \hat{m}_X(Z_i)\}}{\sum_i \{X_i - \hat{m}_X(Z_i)\}^2}.$$

The same idea can be used to develop a test for the conditional independence $X \perp\!\!\!\perp Y \,|\, Z$.

Conditional independence

Recall $X \perp\!\!\!\perp Y \,|\, Z$ has the interpretation ‘knowing Z renders X irrelevant for learning Y’.

Nonparametric variable significance

Under mild conditions

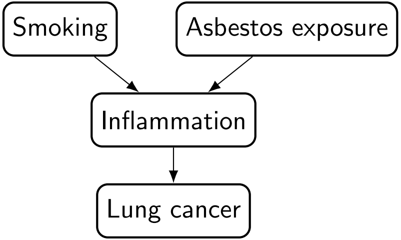
$$Y \perp\!\!\!\perp X_{S^c} \,|\, X_S$$

for

$$S := \{j : X_j \not\perp\!\!\!\perp Y \,|\, X_{-j}\}.$$

‘X_S contains all the information in X needed to learn Y.’

Causal structure



d-separation implies conditional independence

Connection to PLM

$X \perp\!\!\!\perp Y \,|\, Z \Rightarrow \theta = 0$ in PLM.

Connection to PLM

$X \perp\!\!\!\perp Y \mid Z \Rightarrow \theta = 0$ in PLM.

Generalised Covariance Measure (GCM) (Rajen D. Shah and Peters, 2020) tests $X \perp\!\!\!\perp Y \mid Z$ using

$$L_i := \{Y_i - \hat{m}_Y(Z_i)\}\{X_i - \hat{m}_X(Z_i)\}$$

$$\mathrm{GCM}_{Y,X|Z} := \sqrt{n}\frac{\frac{1}{n}\sum_{i=1}^n L_i}{\{\frac{1}{n}\sum_{i=1}^n (L_i - \bar{L})^2\}^{1/2}}.$$

Under conditions we have $\mathrm{GCM}_{Y,X|Z} \xrightarrow{d} \mathcal{N}(0,1)$ under the null. (No cross-fitting required.)

Connection to PLM

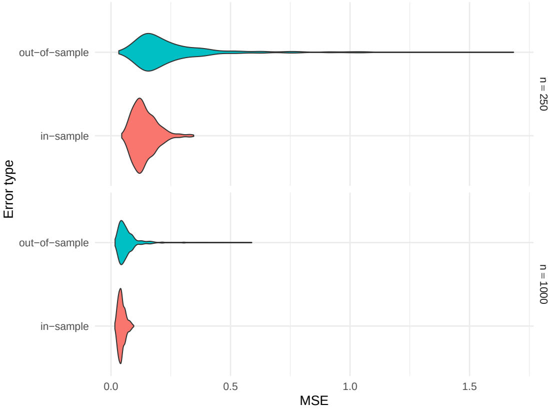
$X \perp\!\!\!\perp Y \mid Z \Rightarrow \theta = 0$ in PLM.

Generalised Covariance Measure (GCM) (Rajen D. Shah and Peters, 2020) tests $X \perp\!\!\!\perp Y \mid Z$ using

$$L_i := \{Y_i - \hat{m}_Y(Z_i)\}\{X_i - \hat{m}_X(Z_i)\}$$

$$\mathrm{GCM}_{Y,X|Z} := \sqrt{n}\frac{\frac{1}{n}\sum_{i=1}^n L_i}{\{\frac{1}{n}\sum_{i=1}^n (L_i - \bar{L})^2\}^{1/2}}.$$

Under conditions we have $\mathrm{GCM}_{Y,X|Z} \xrightarrow{d} \mathcal{N}(0,1)$ under the null. (No cross-fitting required.)



Require $\frac{1}{n}\sum_{i=1}^n\{m_Y(Z_i) - \hat{m}_Y(Z_i)\}^2 \times \frac{1}{n}\sum_{i=1}^n\{m_X(Z_i) - \hat{m}_X(Z_i)\}^2 = o_P(n^{-1})$.

No cross-fitting

Write

$$\begin{aligned} Y_i &= m_Y(Z_i) + \varepsilon_i, \qquad & \mathbb{E}(\varepsilon_i \mid Z_i) = 0 \\ X_i &= m_X(Z_i) + \xi_i, \qquad & \mathbb{E}(\xi_i \mid Z_i) = 0. \end{aligned}$$

No cross-fitting

Write

$$\begin{aligned} Y_i &= m_Y(Z_i) + \varepsilon_i, \qquad && \mathbb{E}(\varepsilon_i \mid Z_i) = 0 \\ X_i &= m_X(Z_i) + \xi_i, \qquad && \mathbb{E}(\xi_i \mid Z_i) = 0. \end{aligned}$$

Substituting into $L_i = \{Y_i - \hat{m}_Y(Z_i)\}\{X_i - \hat{m}_X(Z_i)\}$, we have

$$\begin{aligned} L_i &= \{m_Y(Z_i) - \hat{m}_Y(Z_i) + \varepsilon_i\}\{m_X(Z_i) - \hat{m}_X(Z_i) + \xi_i\} \\ &= \{m_Y(Z_i) - \hat{m}_Y(Z_i)\}\{m_X(Z_i) - \hat{m}_X(Z_i)\} + \varepsilon_i \xi_i \\ &\qquad + \varepsilon_i\{m_X(Z_i) - \hat{m}_X(Z_i)\} + \xi_i\{m_Y(Z_i) - \hat{m}_Y(Z_i)\}. \end{aligned}$$

No cross-fitting

Write

$$\begin{aligned} Y_i &= m_Y(Z_i) + \varepsilon_i, \qquad \mathbb{E}(\varepsilon_i \,|\, Z_i) = 0 \\ X_i &= m_X(Z_i) + \xi_i, \qquad \mathbb{E}(\xi_i \,|\, Z_i) = 0. \end{aligned}$$

Substituting into $L_i = \{Y_i - \hat{m}_Y(Z_i)\}\{X_i - \hat{m}_X(Z_i)\}$, we have

$$\begin{aligned} L_i &= \{m_Y(Z_i) - \hat{m}_Y(Z_i) + \varepsilon_i\}\{m_X(Z_i) - \hat{m}_X(Z_i) + \xi_i\} \\ &= \{m_Y(Z_i) - \hat{m}_Y(Z_i)\}\{m_X(Z_i) - \hat{m}_X(Z_i)\} + \varepsilon_i\xi_i \\ &\qquad + \varepsilon_i\{m_X(Z_i) - \hat{m}_X(Z_i)\} + \xi_i\{m_Y(Z_i) - \hat{m}_Y(Z_i)\}. \end{aligned}$$

Under the null $X \perp\!\!\!\perp Y \,|\, Z$, $\mathbb{E}(\varepsilon_i \,|\, X_i, Z_i) = \mathbb{E}(\varepsilon_i \,|\, Z_i) = 0$, and also

$$\varepsilon_i \underbrace{\{m_X(Z_i) - \hat{m}_X(Z_i)\}}_{\text{Function of } (X_j, Z_j)_{j=1}^n}.$$

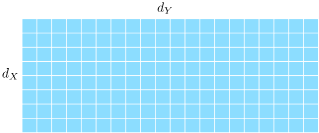
Generalised covariance measures

Multivariate X and Y

- Can apply the GCM componentwise to each (X_j, Y_k, Z).
- Can obtain a final test statistic via e.g. $\max_{j,k} |\mathrm{GCM}_{X_j,Y_k|Z}|$ can calibrate the test using a Gaussian multiplier bootstrap.
- Can also apply to versions of X and Y expanded using feature maps $\varphi_X(X,Z)$, $\varphi_Y(Y,Z)$ as under the null

$$\varphi_X(X,Z) \perp\!\!\!\perp \varphi_Y(Y,Z) \,|\, Z$$

for all functions φ_X, φ_Y.



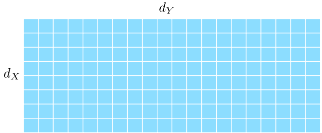
Generalised covariance measures

Multivariate X and Y

- Can apply the GCM componentwise to each (X_j, Y_k, Z).
- Can obtain a final test statistic via e.g. $\max_{j,k} |\mathrm{GCM}_{X_j,Y_k|Z}|$ can calibrate the test using a Gaussian multiplier bootstrap.
- Can also apply to versions of X and Y expanded using feature maps $\varphi_X(X,Z)$, $\varphi_Y(Y,Z)$ as under the null

$$\varphi_X(X,Z) \perp\!\!\!\perp \varphi_Y(Y,Z) \,|\, Z$$

for all functions φ_X, φ_Y.

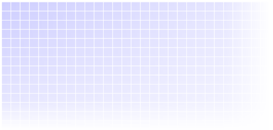


Functional X and Y (Lundborg, Rajen D Shah, and Peters, 2022)

- EEG data, height or weight curves over time, weather data over time...
- Take the squared Hilbert–Schmidt norm of the outer product of the residuals, or equivalently

$$\sum_{i=1}^{n}\sum_{j=1}^{n}\langle\hat{\varepsilon}_i,\hat{\varepsilon}_j\rangle\langle\hat{\xi}_i,\hat{\xi}_j\rangle.$$

- Asymptotically a weighted sum of χ^2 random variables



Hardness of conditional independence testing

- Null hypothesis $\mathcal{P}$: the collection of distributions for (X, Y, Z) absolutely continuous with respect to Lebesgue measure where $X \perp\!\!\!\perp Y \,|\, Z$.
- Alternative hypothesis $\mathcal{Q}$: as above but with $X \not\!\perp\!\!\!\perp Y \,|\, Z$.

A "good" α-level test ψ_n should have

$$\sup_{P \in \mathcal{P}} \mathbb{P}_P(\psi_n = 1) \leq \alpha \ \text{ and } \ \mathbb{P}_Q(\psi_n = 1) \gg \alpha \text{ for many } Q \in \mathcal{Q}$$

Hardness of conditional independence testing

- Null hypothesis $\mathcal{P}$: the collection of distributions for (X, Y, Z) absolutely continuous with respect to Lebesgue measure where $X \perp\!\!\!\perp Y \mid Z$.
- Alternative hypothesis $\mathcal{Q}$: as above but with $X \not\!\perp\!\!\!\perp Y \mid Z$.

A "good" α-level test ψ_n should have

$$\sup_{P \in \mathcal{P}} \mathbb{P}_P(\psi_n = 1) \leq \alpha \;\text{ and }\; \mathbb{P}_Q(\psi_n = 1) \gg \alpha \text{ for many } Q \in \mathcal{Q}$$

Theorem ((Rajen D. Shah and Peters, 2020))

Suppose ψ_n has size α. Then the power at each alternative $Q \in \mathcal{Q}$ is at most α.

Hardness of conditional independence testing

- Null hypothesis $\mathcal{P}$: the collection of distributions for (X, Y, Z) absolutely continuous with respect to Lebesgue measure where $X \perp\!\!\!\perp Y \,|\, Z$.
- Alternative hypothesis $\mathcal{Q}$: as above but with $X \not\!\perp\!\!\!\perp Y \,|\, Z$.

A "good" α-level test ψ_n should have

$$\sup_{P \in \mathcal{P}} \mathbb{P}_P(\psi_n = 1) \leq \alpha \;\text{ and }\; \mathbb{P}_Q(\psi_n = 1) \gg \alpha \text{ for many } Q \in \mathcal{Q}$$

Theorem ((Rajen D. Shah and Peters, 2020))

Suppose ψ_n has size α. Then the power at each alternative $Q \in \mathcal{Q}$ is at most α.

Suppose ψ_n has power β at an alternative $Q \in \mathcal{Q}$. Then there exists null distribution $P \in \mathcal{P}$ such that $\mathbb{P}_P(\psi_n = 1) \geq \beta$.

Hardness of conditional independence testing

- Null hypothesis $\mathcal{P}$: the collection of distributions for (X, Y, Z) absolutely continuous with respect to Lebesgue measure where $X \perp\!\!\!\perp Y \,|\, Z$.
- Alternative hypothesis $\mathcal{Q}$: as above but with $X \not\!\perp\!\!\!\perp Y \,|\, Z$.

A "good" α-level test ψ_n should have

$$\sup_{P \in \mathcal{P}} \mathbb{P}_P(\psi_n = 1) \leq \alpha \text{ and } \mathbb{P}_Q(\psi_n = 1) \gg \alpha \text{ for many } Q \in \mathcal{Q}$$

Theorem ((Rajen D. Shah and Peters, 2020) With great power comes... great Type I error)

Suppose ψ_n has size α. Then the power at each alternative $Q \in \mathcal{Q}$ is at most α.

Suppose ψ_n has power β at an alternative $Q \in \mathcal{Q}$. Then there exists null distribution $P \in \mathcal{P}$ such that $\mathbb{P}_P(\psi_n = 1) \geq \beta$.

Hardness of conditional independence testing

- Null hypothesis $\mathcal{P}$: the collection of distributions for (X, Y, Z) absolutely continuous with respect to Lebesgue measure where $X \perp\!\!\!\perp Y \,|\, Z$.
- Alternative hypothesis $\mathcal{Q}$: as above but with $X \not\!\perp\!\!\!\perp Y \,|\, Z$.

A "good" α-level test ψ_n should have

$$\sup_{P \in \mathcal{P}} \mathbb{P}_P(\psi_n = 1) \leq \alpha \;\text{ and }\; \mathbb{P}_Q(\psi_n = 1) \gg \alpha \text{ for many } Q \in \mathcal{Q}$$

Theorem ((Rajen D. Shah and Peters, 2020) With great power comes... great Type I error)

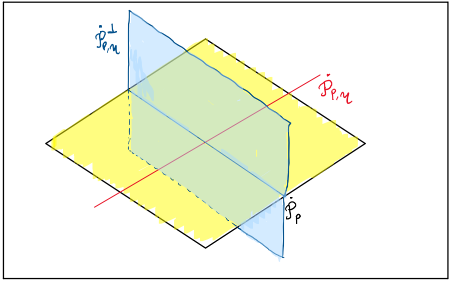
Suppose ψ_n has size α. Then the power at each alternative $Q \in \mathcal{Q}$ is at most α.

Suppose ψ_n has power β at an alternative $Q \in \mathcal{Q}$. Then there exists null distribution $P \in \mathcal{P}$ such that $\mathbb{P}_P(\psi_n = 1) \geq \beta$.

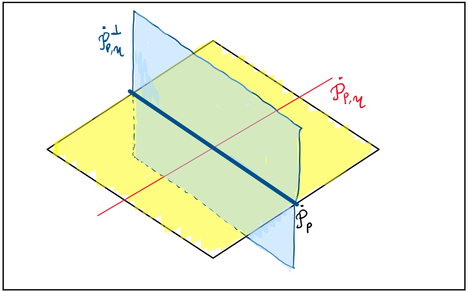
In particular, even the slow rates assumptions for ML methods should not be taken entirely for granted in general.

Optimal semiparametric inference

Optimality in semiparametric problems

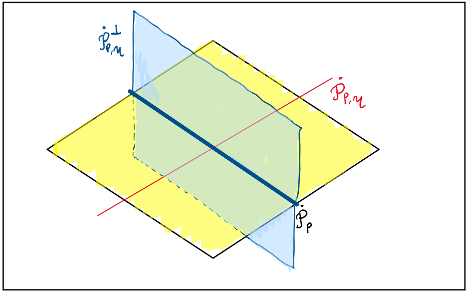

$\dot{\mathcal{P}}_{P,\mathcal{M}}^{\perp}$
$\dot{\mathcal{P}}_{P,\mathcal{M}}$
$\dot{\mathcal{P}}_{P}$

Optimality in semiparametric problems


$\dot{\mathcal{P}}_{P,\eta}^{\perp}$
$\dot{\mathcal{P}}_{P,\eta}$
$\dot{\mathcal{P}}_{P}$

Consider elements of the orthogonal complement of the nuisance tangent set $\dot{\mathcal{P}}_{P,\eta}^{\perp}$ that are sub-model scores.

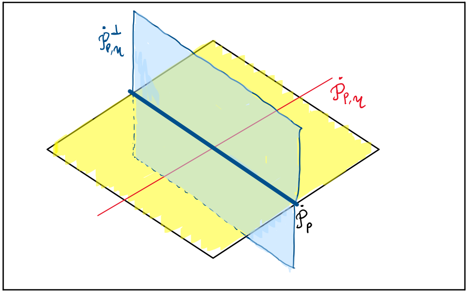
Optimality in semiparametric problems


$\dot{\mathcal{P}}^{\perp}_{P,\eta}$
$\dot{\mathcal{P}}_{P,\eta}$
$\dot{\mathcal{P}}_{P}$

Consider elements of the orthogonal complement of the nuisance tangent set $\dot{\mathcal{P}}^{\perp}_{P,\eta}$ that are sub-model scores.

These give estimating equations that are insensitive to nuisance function estimation *and* correspond to solving score equations, just as in maximum likelihood estimation.

Optimality in semiparametric problems


$\dot{\mathcal{P}}_{P,\eta}^{\perp}$
$\dot{\mathcal{P}}_{P,\eta}$
$\dot{\mathcal{P}}_{P}$

Consider elements of the orthogonal complement of the nuisance tangent set $\dot{\mathcal{P}}_{P,\eta}^{\perp}$ that are sub-model scores.

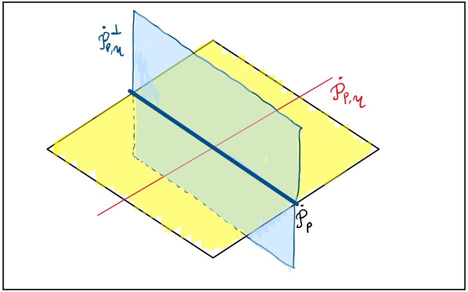
These give estimating equations that are insensitive to nuisance function estimation *and* correspond to solving score equations, just as in maximum likelihood estimation.

As scores in the nuisance tangent space satisfy a single linear constraint (θ is 1d),

$$\dot{\mathcal{P}}_{P,\eta}^{\perp} \cap \dot{\mathcal{P}}_{P}$$

is a 1d-subspace. i.e. there is a unique optimal score equation (up to scaling).

Another perspective

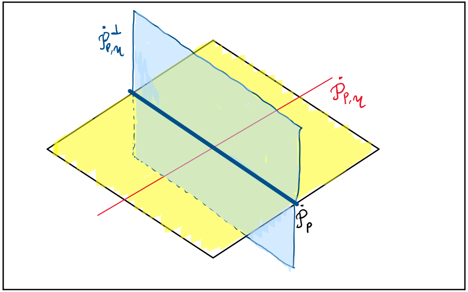

$\dot{\mathcal{P}}_{P,\mathcal{M}}^{\perp}$
$\dot{\mathcal{P}}_{P,\mathcal{M}}$
$\dot{\mathcal{P}}_{P}$

Recall pathwise differentiability: there exists an *influence function* $\tilde{\psi}_P$ such that for every sub-model $t \mapsto P_t$,

$$\frac{d}{dt}\theta(P_t)\Big|_{t=0} = \mathbb{E}_P(S\tilde{\psi}_P),$$

where $S \in \dot{\mathcal{P}}_P$ is a score at $t = 0$ for the path.

Another perspective


$\dot{\mathcal{P}}_{P,\eta}^{\perp}$
$\dot{\mathcal{P}}_{P,\eta}$
$\dot{\mathcal{P}}_P$

Recall pathwise differentiability: there exists an *influence function* $\tilde{\psi}_P$ such that for every sub-model $t \mapsto P_t$,

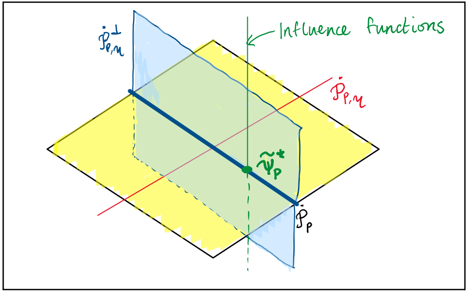
$$\frac{d}{dt}\theta(P_t)\Big|_{t=0} = \mathbb{E}_P(S\tilde{\psi}_P),$$

where $S \in \dot{\mathcal{P}}_P$ is a score at $t = 0$ for the path.

Considering $S \in \dot{\mathcal{P}}_{P,\eta}$, we see $\tilde{\psi}_P \in \dot{\mathcal{P}}_{P,\eta}^{\perp}$.

There always exists a unique influence function $\tilde{\psi}_P^* \in \dot{\mathcal{P}}_P$, the *efficient influence function*.

Another perspective


$\dot{\mathcal{P}}_{P,\eta}^{\perp}$
Influence functions
$\dot{\mathcal{P}}_{P,\eta}$
$\tilde{\psi}_P^*$
$\dot{\mathcal{P}}_P$

Recall pathwise differentiability: there exists an *influence function* $\tilde{\psi}_P$ such that for every sub-model $t \mapsto P_t$,

$$\frac{d}{dt}\theta(P_t)\Big|_{t=0} = \mathbb{E}_P(S\tilde{\psi}_P),$$

where $S \in \dot{\mathcal{P}}_P$ is a score at $t = 0$ for the path.

Considering $S \in \dot{\mathcal{P}}_{P,\eta}$, we see $\tilde{\psi}_P \in \dot{\mathcal{P}}_{P,\eta}^{\perp}$. There always exists a unique influence function $\tilde{\psi}_P^* \in \dot{\mathcal{P}}_P$, the *efficient influence function*.

All other influence functions are obtained by adding elements of $\dot{\mathcal{P}}_P^{\perp}$ to $\tilde{\psi}_P^*$.

Optimal asymptotic variance

Recall that the Cramér–Rao lower bound for estimating θ in the sub-model $t \mapsto P_t$ with score S at $t = 0$ is

$$\frac{\left(\frac{d}{dt}\theta(P_t)\right)^2}{\mathbb{E}S^2}$$

Optimal asymptotic variance

Recall that the Cramér–Rao lower bound for estimating θ in the sub-model $t \mapsto P_t$ with score S at $t = 0$ is

$$\frac{\left(\frac{d}{dt}\theta(P_t)\right)^2}{\mathbb{E}S^2} = \frac{\{\mathbb{E}(S\tilde{\psi}_P^*)\}^2}{\mathbb{E}S^2}$$

Optimal asymptotic variance

Recall that the Cramér–Rao lower bound for estimating θ in the sub-model $t \mapsto P_t$ with score S at $t = 0$ is

$$\begin{aligned} \frac{\left(\frac{d}{dt}\theta(P_t)\right)^2}{\mathbb{E}S^2} &= \frac{\{\mathbb{E}(S\tilde{\psi}_P^*)\}^2}{\mathbb{E}S^2} \\ &\leq \frac{\mathbb{E}(S^2)\mathbb{E}(\tilde{\psi}_P^{*,2})}{\mathbb{E}S^2} \qquad \text{by Cauchy–Schwarz} \\ &= \mathbb{E}(\tilde{\psi}_P^{*,2}) \end{aligned}$$

Optimal asymptotic variance

Recall that the Cramér–Rao lower bound for estimating θ in the sub-model $t \mapsto P_t$ with score S at $t = 0$ is

$$\begin{aligned}\frac{\left(\frac{d}{dt}\theta(P_t)\right)^2}{\mathbb{E}S^2} &= \frac{\{\mathbb{E}(S\tilde{\psi}_P^*)\}^2}{\mathbb{E}S^2} \\ &\leq \frac{\mathbb{E}(S^2)\mathbb{E}(\tilde{\psi}_P^{*,2})}{\mathbb{E}S^2} \qquad \text{by Cauchy–Schwarz} \\ &= \mathbb{E}(\tilde{\psi}_P^{*,2})\end{aligned}$$

The variance of the efficient influence function gives the optimal asymptotic variance.

Nonparametric models

Nonparametric models vs semiparametric models

The PLM is an example of a *strictly semiparametric model*: it places some structural restrictions on the possible distributions.

An alternative strategy considers *nonparametric models*, which make no such restrictions, and targets a functional that addresses the question of interest.

Nonparametric models vs semiparametric models

The PLM is an example of a *strictly semiparametric model*: it places some structural restrictions on the possible distributions.

An alternative strategy considers *nonparametric models*, which make no such restrictions, and targets a functional that addresses the question of interest.

In nonparametric models, $\dot{\mathcal{P}}_P$ is the entire space of mean-zero (square-integrable) functions.

E.g. Consider $t \mapsto p_0(w)(1 + ta(w))$ with $\int ap_0 = 0$.

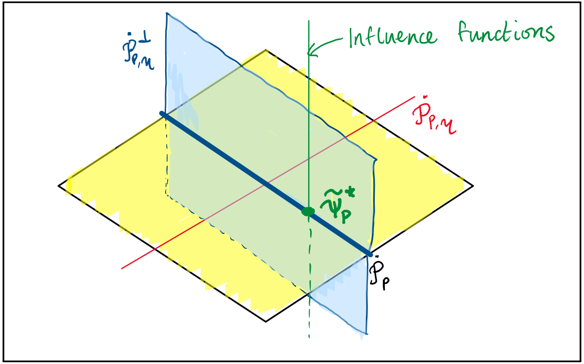
Nonparametric models vs semiparametric models

The PLM is an example of a *strictly semiparametric model*: it places some structural restrictions on the possible distributions.

An alternative strategy considers *nonparametric models*, which make no such restrictions, and targets a functional that addresses the question of interest.

In nonparametric models, $\dot{\mathcal{P}}_P$ is the entire space of mean-zero (square-integrable) functions.

E.g. Consider $t \mapsto p_0(w)(1 + ta(w))$ with $\int ap_0 = 0$.


$\dot{\mathcal{P}}_{P,\mathcal{M}}^{\perp}$
Influence functions
$\dot{\mathcal{P}}_{P,\mathcal{M}}$
$\tilde{\psi}_P^*$
$\dot{\mathcal{P}}_P$

Nonparametric models vs semiparametric models

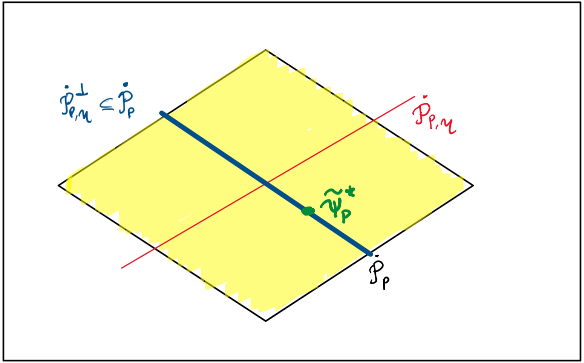
The PLM is an example of a *strictly semiparametric model*: it places some structural restrictions on the possible distributions.

An alternative strategy considers *nonparametric models*, which make no such restrictions, and targets a functional that addresses the question of interest.

In nonparametric models, $\dot{\mathcal{P}}_P$ is the entire space of mean-zero (square-integrable) functions.

E.g. Consider $t \mapsto p_0(w)(1 + ta(w))$ with $\int ap_0 = 0$.

In particular, the efficient influence function is the *only* influence function.


$\dot{\mathcal{P}}_{P,\mathcal{M}}^{\perp} \subseteq \dot{\mathcal{P}}_P$
$\dot{\mathcal{P}}_{P,\mathcal{M}}$
$\tilde{\psi}_P^*$
$\dot{\mathcal{P}}_P$

Nonparametric models vs semiparametric models

The PLM is an example of a *strictly semiparametric model*: it places some structural restrictions on the possible distributions.

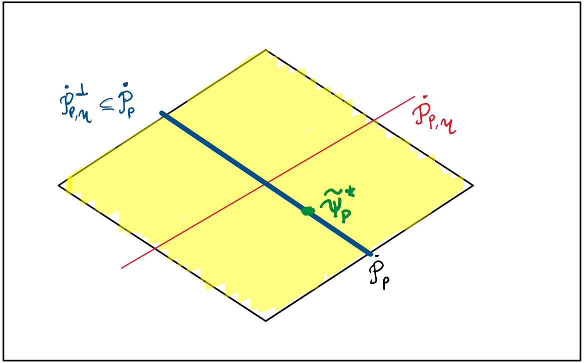
An alternative strategy considers *nonparametric models*, which make no such restrictions, and targets a functional that addresses the question of interest.

In nonparametric models, $\dot{\mathcal{P}}_P$ is the entire space of mean-zero (square-integrable) functions.

E.g. Consider $t \mapsto p_0(w)(1 + ta(w))$ with $\int ap_0 = 0$.

In particular, the efficient influence function is the *only* influence function.

$\Rightarrow$ other recipes for finding $\tilde{\psi}_P^*$ (see e.g. Kennedy, 2022).


$\dot{\mathcal{P}}_{P,\mathcal{M}}^{\perp} \subseteq \dot{\mathcal{P}}_P$
$\dot{\mathcal{P}}_{P,\mathcal{M}}$
$\tilde{\psi}_P^*$
$\dot{\mathcal{P}}_P$

Average partial effect

E.g. Consider the *average partial effect* (APE)

$$\theta := \mathbb{E}\left[\frac{\partial}{\partial x} f(X, Z)\right] \qquad \text{where} \qquad f(x, z) := \mathbb{E}(Y \mid X = x, Z = z).$$

Average partial effect

E.g. Consider the *average partial effect* (APE)

$$\theta := \mathbb{E}\left[\frac{\partial}{\partial x} f(X, Z)\right] \qquad \text{where} \qquad f(x, z) := \mathbb{E}(Y \mid X = x, Z = z).$$

For example, if f is additive so

$$f(x, z) = f_X(x) + f_Z(z),$$

the APE θ is the average slope of f_X, with the average being w.r.t. the marginal distribution of X.

Average partial effect

E.g. Consider the *average partial effect* (APE)

$$\theta := \mathbb{E}\left[\frac{\partial}{\partial x} f(X, Z)\right] \qquad \text{where} \qquad f(x, z) := \mathbb{E}(Y \,|\, X = x, Z = z).$$

For example, if f is additive so

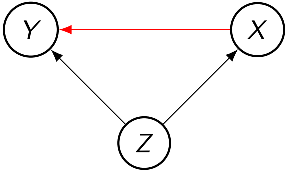
$$f(x, z) = f_X(x) + f_Z(z),$$

the APE θ is the average slope of f_X, with the average being w.r.t. the marginal distribution of X.

Historically, interest in the APE has been due to the fact that it recovers the coefficients in a single index model (Stoker, 1986; Powell, Stock, and Stoker, 1989):

$$\text{APE} = \mathbb{E}\left[\frac{\partial}{\partial x} g(X\theta + \beta^\top Z)\right] = \theta\mathbb{E}[g'(X\theta + \beta^\top Z)].$$

Causal interpretation



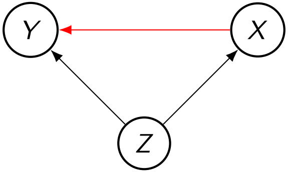
Consider intervening by $X \mapsto X + \delta$, so the in terms of the joint density

$$p(y \mid x, z)p(x \mid z)p(z) \mapsto p(y \mid x, z)p(x - \delta \mid z)p(z) =: q_\delta(x, y, z).$$

Then

$$\lim_{\delta \to 0} \frac{\mathbb{E}_{q_\delta}(Y) - \mathbb{E}_p(Y)}{\delta} = \mathbb{E}[f'(X, Z)] = \theta.$$

Causal interpretation



Consider intervening by $X \mapsto X + \delta$, so the in terms of the joint density

$$p(y \mid x, z)p(x \mid z)p(z) \mapsto p(y \mid x, z)p(x - \delta \mid z)p(z) =: q_\delta(x, y, z).$$

Then

$$\lim_{\delta \to 0} \frac{\mathbb{E}_{q_\delta}(Y) - \mathbb{E}_p(Y)}{\delta} = \mathbb{E}[f'(X, Z)] = \theta.$$

Thus $\theta \cdot \delta$ reflects the change in the mean of Y under an infinitessimal distribution shift of size δ (Rothenhäusler and Yu, 2020).

- The effect on consumption of an infinitessimal increase in log(income)
- The effect on a given health outcome of an increase in log(exercise)...

Semiparametric approach

The naive estimate $\tilde{\theta} := \frac{1}{n}\sum_{i=1}^{n} \hat{f}'(X_i, Z_i)$ will typically inherit the slow rate of convergence of the nonparametric estimate $\hat{f}$.

Semiparametric approach

The naive estimate $\tilde{\theta} := \frac{1}{n}\sum_{i=1}^n \hat{f}'(X_i, Z_i)$ will typically inherit the slow rate of convergence of the nonparametric estimate $\hat{f}$.

Instead, we can use (see e.g. Bickel et al. (1993))

$$\hat{\theta} := \frac{1}{n}\sum_{i=1}^n \hat{f}'(X_i, Z_i) - \hat{\rho}(X_i, Z_i)\{Y_i - \hat{f}(X_i, Z_i)\},$$

where $\hat{\rho}$ estimates the *conditional score*

$$\rho(x, z) := \frac{\partial}{\partial x} \log p(x \,|\, z) = \frac{p'(x \,|\, z)}{p(x \,|\, z)}.$$

Double machine learning

$$\hat{\theta} := \frac{1}{n}\sum_{i=1}^{n} \hat{f}'(X_i, Z_i) - \hat{\rho}(X_i, Z_i)\{Y_i - \hat{f}(X_i, Z_i)\}.$$

Suppose for simplicity that $\hat{f}$ and $\hat{\rho}$ are trained on an auxiliary dataset D (in reality, we use cross-fitting).

Good news: When

$$\begin{aligned} \mathbb{E}[\{f(X,Z) - \hat{f}(X,Z)\}^2 \mid D] \cdot \mathbb{E}[\{\rho(X,Z) - \hat{\rho}(X,Z)\}^2 \mid D] &= o_P(n^{-1}) \\ \mathbb{E}[\{f'(X,Z) - \hat{f}'(X,Z)\}^2 \mid D] &= o_P(1), \end{aligned}$$

we have that $\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} \mathcal{N}(0, \sigma^2)$, where σ^2 is the semiparametric efficient variance bound.

Double machine learning

$$\hat{\theta} := \frac{1}{n}\sum_{i=1}^{n} \hat{f}'(X_i, Z_i) - \hat{\rho}(X_i, Z_i)\{Y_i - \hat{f}(X_i, Z_i)\}.$$

Suppose for simplicity that $\hat{f}$ and $\hat{\rho}$ are trained on an auxiliary dataset D (in reality, we use cross-fitting).

Good news: When

$$\begin{aligned} \mathbb{E}[\{f(X,Z) - \hat{f}(X,Z)\}^2 \mid D] \cdot \mathbb{E}[\{\rho(X,Z) - \hat{\rho}(X,Z)\}^2 \mid D] &= o_P(n^{-1}) \\ \mathbb{E}[\{f'(X,Z) - \hat{f}'(X,Z)\}^2 \mid D] &= o_P(1), \end{aligned}$$

we have that $\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} \mathcal{N}(0, \sigma^2)$, where σ^2 is the semiparametric efficient variance bound.

Bad news:

- Our favourite $\hat{f}$ may be poor for estimating f', e.g. tree-based estimators.
- Estimating $\rho(x, z) = p'(x \mid z)/p(x \mid z)$ may be very challenging.

Our approach (Klyne and Rajen D Shah, 2026)

Our favourite $\hat{f}$ may be poor for estimating f', e.g. tree-based estimators.

Our approach (Klyne and Rajen D Shah, 2026)

Our favourite $\hat{f}$ may be poor for estimating f', e.g. tree-based estimators.

- We show that if $\hat{f}$ is good at prediction, so is its convolution with a Gaussian.
- The derivative of the convolution will additionally estimate f' sufficiently well.

Our approach (Klyne and Rajen D Shah, 2026)

Our favourite $\hat{f}$ may be poor for estimating f', e.g. tree-based estimators.

- We show that if $\hat{f}$ is good at prediction, so is its convolution with a Gaussian.
- The derivative of the convolution will additionally estimate f' sufficiently well.

Estimating $\rho(X, Z) = p'(x \mid z)/p(x \mid z)$ may be very challenging.

Our approach (Klyne and Rajen D Shah, 2026)

Our favourite $\hat{f}$ may be poor for estimating f', e.g. tree-based estimators.

- We show that if $\hat{f}$ is good at prediction, so is its convolution with a Gaussian.
- The derivative of the convolution will additionally estimate f' sufficiently well.

Estimating $\rho(X, Z) = p'(x \mid z)/p(x \mid z)$ may be very challenging.

- We make use of a location–scale model

$$X = m(Z) + \sigma(Z)\varepsilon$$

  - $m(z) := \mathbb{E}(X \mid Z = z)$,
  - $\sigma^2(z) := \mathrm{Var}(X \mid Z = z)$,
  - $\varepsilon \perp\!\!\!\perp Z$ with $\mathbb{E}(\varepsilon) = 0$, $\mathrm{Var}(\varepsilon) = 1$.

- Show that provided we can estimate m and σ sufficiently well, then only the univariate score for ε needs to be estimated well in order for the overall procedure to work.

Our approach (Klyne and Rajen D Shah, 2026)

Our favourite $\hat{f}$ may be poor for estimating f', e.g. tree-based estimators.

- We show that if $\hat{f}$ is good at prediction, so is its convolution with a Gaussian.
- The derivative of the convolution will additionally estimate f' sufficiently well.

Estimating $\rho(X, Z) = p'(x \mid z)/p(x \mid z)$ may be very challenging.

- We make use of a location–scale model

$$X = m(Z) + \sigma(Z)\varepsilon$$

  - $m(z) := \mathbb{E}(X \mid Z = z)$,
  - $\sigma^2(z) := \mathrm{Var}(X \mid Z = z)$,
  - $\varepsilon \perp\!\!\!\perp Z$ with $\mathbb{E}(\varepsilon) = 0$, $\mathrm{Var}(\varepsilon) = 1$.

- Show that provided we can estimate m and σ sufficiently well, then only the univariate score for ε needs to be estimated well in order for the overall procedure to work.

Although a fully nonparametric approach may be appealing, in practice the nuisance functions involved may be too difficult to estimate.

Expected conditional covariance

Consider once more the partially linear model

$$Y = \theta X + f(Z) + \varepsilon.$$

An alternative nonparametric estimand that coincides with β is

$$\frac{\mathbb{E}\mathrm{Cov}(Y, X \mid Z)}{\mathbb{E}\mathrm{Var}(X \mid Z)}.$$

This is exactly what our (reparametrised) $\hat{\theta}$ targeted.

Vansteelandt and Dukes, 2022 consider nonparametric estimands related to generalised partially linear models.

Robust inference in semiparametric models

Optimality in the PLM

In the case of the PLM, the optimal estimating equation is (see e.g. Ma et al. (2006))

$$\psi(Y, X, Z; \theta, f, h) := \frac{1}{v(X, Z)}(X - h(Z))(Y - X\theta - f(Z))$$

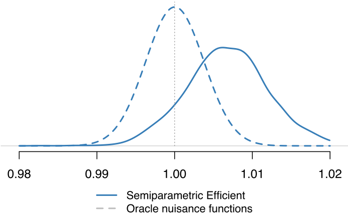
where $v(X, Z) := \mathrm{Var}(Y \mid X, Z)$ and $h(Z) := \mathbb{E}\left(\frac{1}{v(X,Z)} \,\middle|\, Z\right)^{-1} \mathbb{E}\left(\frac{X}{v(X,Z)} \,\middle|\, Z\right)$.

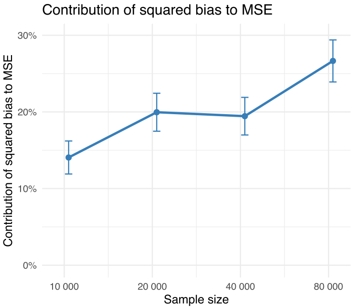
Optimality in the PLM

In the case of the PLM, the optimal estimating equation is (see e.g. Ma et al. (2006))

$$\psi(Y, X, Z; \theta, f, h) := \frac{1}{v(X,Z)}(X - h(Z))(Y - X\theta - f(Z))$$

where $v(X,Z) := \mathrm{Var}(Y \,|\, X, Z)$ and $h(Z) := \mathbb{E}\left(\frac{1}{v(X,Z)} \,\middle|\, Z\right)^{-1} \mathbb{E}\left(\frac{X}{v(X,Z)} \,\middle|\, Z\right)$.



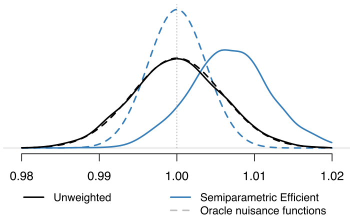


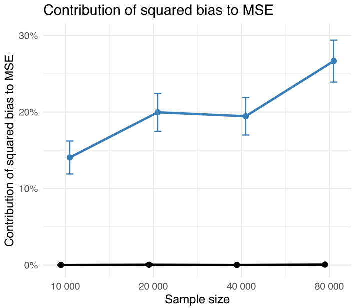
Optimality in the PLM

In the case of the PLM, the optimal estimating equation is (see e.g. Ma et al. (2006))

$$\psi(Y,X,Z;\theta,f,h) := \frac{1}{v(X,Z)}(X - h(Z))(Y - X\theta - f(Z))$$

where $v(X,Z) := \mathrm{Var}(Y \mid X,Z)$ and $h(Z) := \mathbb{E}\left(\frac{1}{v(X,Z)} \,\middle|\, Z\right)^{-1} \mathbb{E}\left(\frac{X}{v(X,Z)} \,\middle|\, Z\right)$.





Our strategy (Elliot H Young and Rajen D Shah, 2024)

The efficient influence function has strong theoretical justification, but may not be as well-founded from a practical perspective. Can we try to bridge this gap between theory and practice?

Our strategy (Elliot H Young and Rajen D Shah, 2024)

The efficient influence function has strong theoretical justification, but may not be as well-founded from a practical perspective. Can we try to bridge this gap between theory and practice?

Recall that influence functions in the PLM take the form

$$\{Y - X\theta - f(Z)\}\{\phi(X, Z) - \mathbb{E}(\phi(X, Z) \mid Z)\}.$$

Our strategy (Elliot H Young and Rajen D Shah, 2024)

The efficient influence function has strong theoretical justification, but may not be as well-founded from a practical perspective. Can we try to bridge this gap between theory and practice?

Recall that influence functions in the PLM take the form

$$\{Y - X\theta - f(Z)\}\{\phi(X, Z) - \mathbb{E}(\phi(X, Z) \mid Z)\}.$$

1. Specify nuisance functions that can be estimated well.

Our strategy (Elliot H Young and Rajen D Shah, 2024)

The efficient influence function has strong theoretical justification, but may not be as well-founded from a practical perspective. Can we try to bridge this gap between theory and practice?

Recall that influence functions in the PLM take the form

$$\{Y - X\theta - f(Z)\}\{\phi(X, Z) - \mathbb{E}(\phi(X, Z) \mid Z)\}.$$

1. Specify nuisance functions that can be estimated well.
   Specify functions $M_1, \ldots, M_J$ such that each of $\mathbb{E}(M_j(X) \mid Z)$ can be estimated well, e.g. $M_1(X) = X$.

Our strategy (Elliot H Young and Rajen D Shah, 2024)

The efficient influence function has strong theoretical justification, but may not be as well-founded from a practical perspective. Can we try to bridge this gap between theory and practice?

Recall that influence functions in the PLM take the form

$$\{Y - X\theta - f(Z)\}\{\phi(X, Z) - \mathbb{E}(\phi(X, Z) \mid Z)\}.$$

1. Specify nuisance functions that can be estimated well.
   Specify functions $M_1, \ldots, M_J$ such that each of $\mathbb{E}(M_j(X) \mid Z)$ can be estimated well, e.g. $M_1(X) = X$.
2. Find the most efficient estimator requiring only the above and f to be estimated well.

Our strategy (Elliot H Young and Rajen D Shah, 2024)

The efficient influence function has strong theoretical justification, but may not be as well-founded from a practical perspective. Can we try to bridge this gap between theory and practice?

Recall that influence functions in the PLM take the form

$$\{Y - X\theta - f(Z)\}\{\phi(X, Z) - \mathbb{E}(\phi(X, Z) \,|\, Z)\}.$$

1. Specify nuisance functions that can be estimated well.
   Specify functions $M_1, \ldots, M_J$ such that each of $\mathbb{E}(M_j(X) \,|\, Z)$ can be estimated well, e.g. $M_1(X) = X$.
2. Find the most efficient estimator requiring only the above and f to be estimated well.
   - A. Finding robust influence functions;
   - B. Attaining robust semiparametric efficiency.

Robust semiparametric efficiency

Can show that the set of permissible estimating equations is:

$$\psi(X, Y, Z; \theta, f, (m_j)_{j=1}^J) = (Y - X\theta - f(Z)) \sum_{j=1}^J w_j(Z)(M_j(X) - m_j(Z)).$$

Set $J = 1$ for notational simplicity. Each weight function w yields an estimator $\hat{\theta}_w$ via solving

$$\sum_{i=1}^n w(Z_i)\{Y_i - X_i\theta - f(Z_i)\}\{M(X_i) - m(Z_i)\} = 0.$$

Robust semiparametric efficiency

Can show that the set of permissible estimating equations is:

$$\psi(X, Y, Z; \theta, f, (m_j)_{j=1}^J) = (Y - X\theta - f(Z)) \sum_{j=1}^J w_j(Z)(M_j(X) - m_j(Z)).$$

Set $J = 1$ for notational simplicity. Each weight function w yields an estimator $\hat{\theta}_w$ via solving

$$\sum_{i=1}^n w(Z_i)\{Y_i - X_i\theta - f(Z_i)\}\{M(X_i) - m(Z_i)\} = 0.$$

We should seek w^* such that $\hat{\theta}_{w^*}$ has minimal asymptotic variance (see e.g. Rubin and Laan, 2008; Qu, Lindsay, and Li, 2000).

Robust semiparametric efficiency

Can show that the set of permissible estimating equations is:

$$\psi(X, Y, Z; \theta, f, (m_j)_{j=1}^J) = (Y - X\theta - f(Z)) \sum_{j=1}^J w_j(Z)(M_j(X) - m_j(Z)).$$

Set $J = 1$ for notational simplicity. Each weight function w yields an estimator $\hat{\theta}_w$ via solving

$$\sum_{i=1}^n w(Z_i)\{Y_i - X_i\theta - f(Z_i)\}\{M(X_i) - m(Z_i)\} = 0.$$

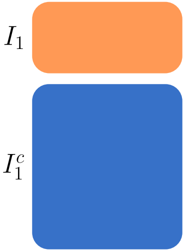
We should seek w^* such that $\hat{\theta}_{w^*}$ has minimal asymptotic variance (see e.g. Rubin and Laan, 2008; Qu, Lindsay, and Li, 2000).

Use the 'sandwich loss':

$$\hat{V}_w = \frac{\frac{1}{n}\sum_{i=1}^n \left[w(Z_i)\{Y_i - X_i\theta - f(Z_i)\}\{M(X_i) - m(Z_i)\}\right]^2}{\left(\frac{1}{n}\sum_{i=1}^n w(Z_i)\{M(X_i) - m(Z_i)\}\right)^2} =: \hat{L}_{\text{SL}}(w).$$

Importantly, we only need a consistent estimate of w^* to attain 'robust semiparametric efficiency'.

K-fold cross-fitting

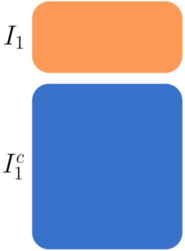


Obtain $\hat{f}^{(1)}$, $\hat{m}^{(1)}$
and estimated weight function $\hat{w}^{(1)}$

Choose $\hat{\theta}$ to solve

$$\sum_{k=1}^{K} \sum_{i \in I_k} \hat{w}^{(k)}(Z_i)\{Y_i - X_i\theta - \hat{f}^{(k)}(Z_i)\}\{M(X_i) - \hat{m}^{(k)}(Z_i)\} = 0.$$

K-fold cross-fitting



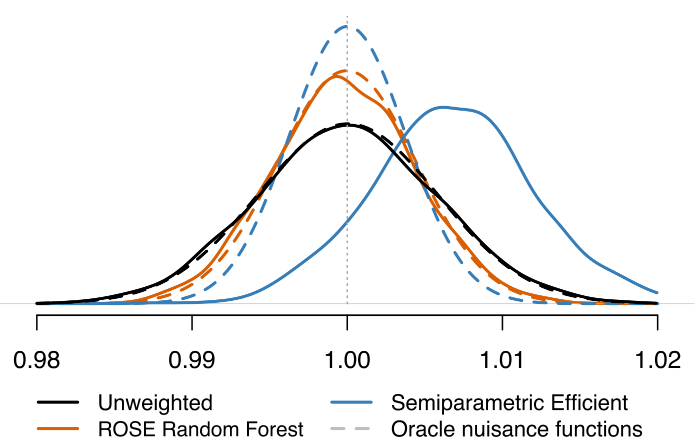
Obtain $\hat{f}^{(1)}$, $\hat{m}^{(1)}$
and estimated weight function $\hat{w}^{(1)}$

Choose $\hat{\theta}$ to solve

$$\sum_{k=1}^{K} \sum_{i \in I_k} \hat{w}^{(k)}(Z_i)\{Y_i - X_i\theta - \hat{f}^{(k)}(Z_i)\}\{M(X_i) - \hat{m}^{(k)}(Z_i)\} = 0.$$

We use a modified random forest with split points in trees chosen to minimise $\hat{L}_{\mathsf{SL}}$. (R package `RoseRF`.)

Simulation example

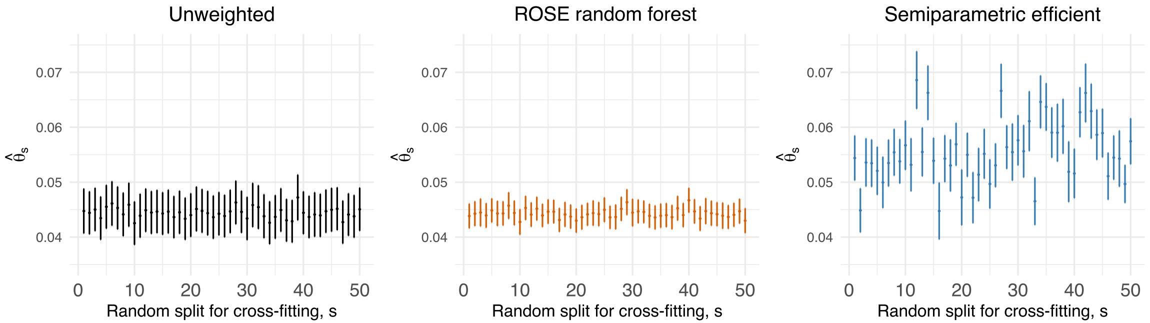


Bike sharing data

Data consisting of hourly count data of rental bikes in Seoul over a one year period.

Take Y = Bike Count, X = Temp, Z = other predictors and use generalised PLM

$$\log\left(\mathbb{E}(Y \mid X, Z)\right) = \theta X + f(Z).$$



Next time...

- (Multiple) sample splitting
- Nonparametric variable significance testing; goodness-of-fit testing
- Nonparametric regression

Thank you for listening.

Grouped data

Grouped data

- E.g. data on student performance from different schools, patient data from different hospitals;
- longitudinal data where we have multiple measurements on individual subjects collected over time.

Linear model

$$\underbrace{Y_i}_{\in \mathbb{R}^{n_i}} = \underbrace{X_i}_{\in \mathbb{R}^{n_i \times p}} \underbrace{\beta}_{\in \mathbb{R}^p} + \underbrace{\varepsilon_i}_{\in \mathbb{R}^{n_i}}.$$

Grouped data

- E.g. data on student performance from different schools, patient data from different hospitals;
- longitudinal data where we have multiple measurements on individual subjects collected over time.

Linear model

$$\underbrace{Y_i}_{\in\mathbb{R}^{n_i}} = \underbrace{X_i}_{\in\mathbb{R}^{n_i\times p}} \underbrace{\beta}_{\in\mathbb{R}^p} + \underbrace{\varepsilon_i}_{\in\mathbb{R}^{n_i}}.$$

Random effects can be used to model the conditional covariances $\mathrm{Cov}(\varepsilon_i \,|\, X_i)$.
E.g. $Y_{ij} = \beta^\top (X_i)_{j,\cdot} + \gamma_i + \xi_{ij}$.

When such models are well-specified, (Re)ML can deliver efficient estimates of β taking the form

$$\hat{\beta}(\rho) = \left(\sum_i X_i^\top W_i(\rho) X_i\right)^{-1} \left(\sum_i X_i^\top W_i(\rho) Y_i\right).$$

The maximum likelihood estimate $\hat{\rho}$ satisfies $W_i(\hat{\rho}) \approx \mathrm{Cov}(\varepsilon_i \,|\, X_i)^{-1}$.

Conditional variance misspecification

If our model for $\mathrm{Var}(\varepsilon_i \,|\, X_i)$ is misspecified then our estimate may no longer be efficient.

GEE philosophy (see e.g. Ziegler, 2011; Tsiatis, 2006, ...):

1. specify a *working* model for the conditional covariance;
2. estimate parameter ρ by minimising

$$\sum_i \| W_i(\rho)^{-1} - \hat{\varepsilon}_i \hat{\varepsilon}_i^\top \|^2,$$

   where $\hat{\varepsilon}_i$ is a pilot estimate of the residuals obtained e.g. from an unweighted estimator.

If the working conditional conditional model is well-specified, both the MLE under a Gaussian likelihood and the GEE least squares minimisation deliver efficiency.

Conditional variance misspecification

If our model for $\mathrm{Var}(\varepsilon_i \mid X_i)$ is misspecified then our estimate may no longer be efficient.

GEE philosophy (see e.g. Ziegler, 2011; Tsiatis, 2006, ...):

1. specify a *working* model for the conditional covariance;
2. estimate parameter ρ by minimising

$$\sum_i \|W_i(\rho)^{-1} - \hat{\varepsilon}_i \hat{\varepsilon}_i^\top\|^2,$$

where $\hat{\varepsilon}_i$ is a pilot estimate of the residuals obtained e.g. from an unweighted estimator.

If the working conditional conditional model is well-specified, both the MLE under a Gaussian likelihood and the GEE least squares minimisation deliver efficiency.

Which (if any) behaves well under misspecification?

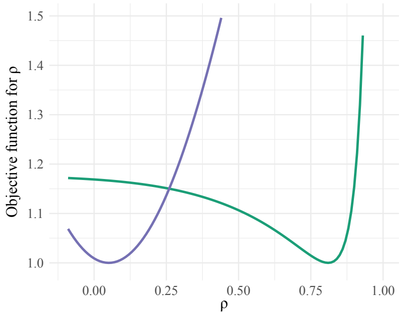
Longitudinal data example

Truth: ARMA(2, 1) covariance structure **Model:** AR(1) covariance structure.

Estimator: $\hat{\beta}(\rho) = \left(\sum_i X_i^\top W_i(\rho) X_i\right)^{-1} \left(\sum_i X_i^\top W_i(\rho) Y_i\right)$.

▬ Gaussian negative log likelihood $L_{\mathrm{ML}}(\rho) = \mathbb{E}\big(-\log\det W(\rho) + \mathrm{tr}\{W(\rho)\mathrm{Cov}(Y|X)\}\big)$

▬ Generalised estimating equation $L_{\mathrm{GEE}}(\rho) = \mathbb{E}\big(\|W(\rho)^{-1} - \mathrm{Cov}(Y|X)\|^2\big)$

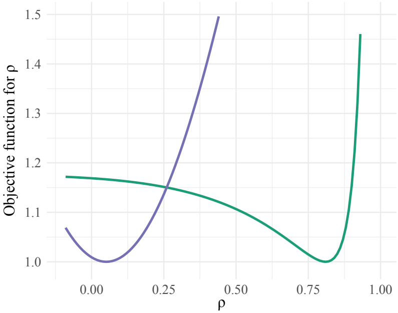


Longitudinal data example

Truth: ARMA(2, 1) covariance structure **Model:** AR(1) covariance structure.

Estimator: $\hat{\beta}(\rho) = \left(\sum_i X_i^\top W_i(\rho) X_i\right)^{-1} \left(\sum_i X_i^\top W_i(\rho) Y_i\right)$.

▬	Gaussian negative log likelihood	$L_{\text{ML}}(\rho) = \mathbb{E}\left(-\log\det W(\rho) + \operatorname{tr}\{W(\rho)\operatorname{Cov}(Y\|X)\}\right)$
▬	Generalised estimating equation	$L_{\text{GEE}}(\rho) = \mathbb{E}\left(\\|W(\rho)^{-1} - \operatorname{Cov}(Y\|X)\\|^2\right)$
▬	Goal	Minimal $\operatorname{Var}\hat{\beta}(\rho)$



Longitudinal data example

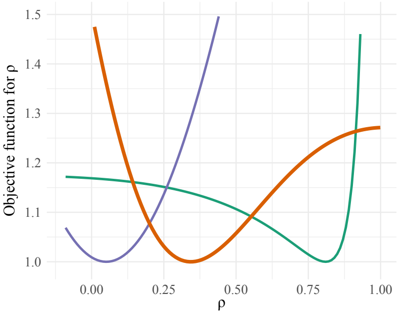
Truth: ARMA(2, 1) covariance structure **Model:** AR(1) covariance structure.

Estimator: $\hat{\beta}(\rho) = \left(\sum_i X_i^\top W_i(\rho) X_i\right)^{-1} \left(\sum_i X_i^\top W_i(\rho) Y_i\right)$.

▬ Gaussian negative log likelihood $L_{\mathrm{ML}}(\rho) = \mathbb{E}\left(-\log\det W(\rho) + \mathrm{tr}\{W(\rho)\mathrm{Cov}(Y|X)\}\right)$

▬ Generalised estimating equation $L_{\mathrm{GEE}}(\rho) = \mathbb{E}\left(\|W(\rho)^{-1} - \mathrm{Cov}(Y|X)\|^2\right)$

▬ Goal Minimal $\mathrm{Var}\hat{\beta}(\rho)$

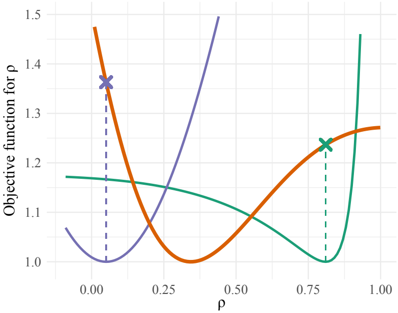


Longitudinal data example

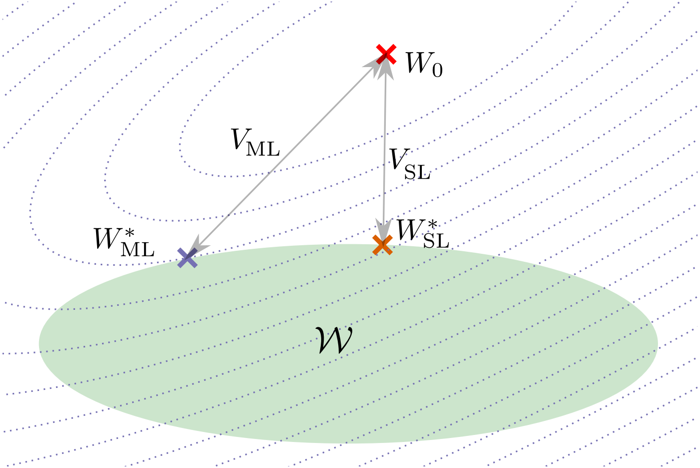
Truth: ARMA(2, 1) covariance structure **Model:** AR(1) covariance structure.

Estimator: $\hat{\beta}(\rho) = \left(\sum_i X_i^\top W_i(\rho) X_i\right)^{-1} \left(\sum_i X_i^\top W_i(\rho) Y_i\right)$.

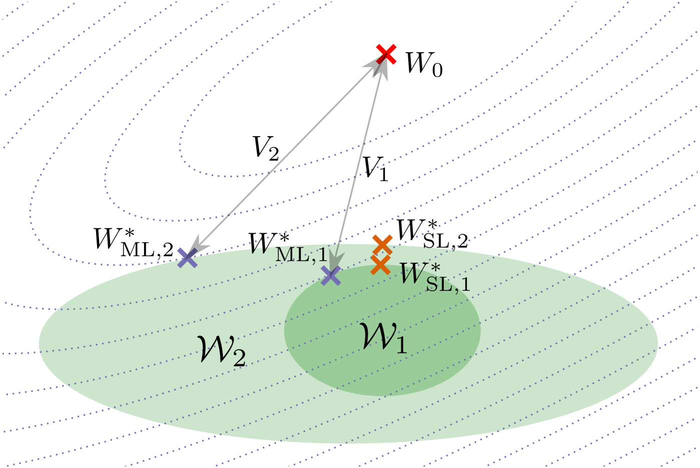
▬	Gaussian negative log likelihood	$L_{\mathrm{ML}}(\rho) = \mathbb{E}\big(-\log\det W(\rho) + \mathrm{tr}\{W(\rho)\mathrm{Cov}(Y\|X)\}\big)$
▬	Generalised estimating equation	$L_{\mathrm{GEE}}(\rho) = \mathbb{E}\big(\\|W(\rho)^{-1} - \mathrm{Cov}(Y\|X)\\|^2\big)$
▬	Goal	Minimal $\mathrm{Var}\hat{\beta}(\rho)$



Behaviour under misspecification



Behaviour under misspecification



Sandwich boosting for grouped data (E. H. Young and R. D. Shah, 2024)

Recall that in the grouped setting, the weight functions become weight *matrices*. In Young & Shah (2024) we consider a class of such matrices of the form

$$W(Z_i) := \left\{\mathrm{diag}(\sigma(Z_{i1}), \ldots, \sigma(Z_{in_i}))\ \underbrace{C_\theta(Z_i)}_{\text{parametric}}\ \mathrm{diag}(\sigma(Z_{i1}), \ldots, \sigma(Z_{in_i}))\right\}^{-1}.$$

We ‘estimate’ θ and the function σ jointly via gradient boosting (with the latter using decision trees as a base learner) minimising the sandwich loss

$$\hat{L}_{\mathrm{SL}}(W) = \frac{\frac{1}{n}\sum_{i=1}^n \left(\{Y_i - X_i\theta - f(Z_i)\}^\top W(Z_i)\{M(X_i) - m(Z_i)\}\right)^2}{\left(\frac{1}{n}\sum_{i=1}^n \{M(X_i) - m(Z_i)\}^\top W(Z_i)\{M(X_i) - m(Z_i)\}\right)^2}.$$

References I

- Bickel, Peter J et al. (1993). *Efficient and adaptive estimation for semiparametric models*. Vol. 4. Springer.
- Kennedy, Edward H (2022). "Semiparametric doubly robust targeted double machine learning: a review". In: *arXiv preprint arXiv:2203.06469*.
- Klyne, Harvey and Rajen D Shah (2026). "Average partial effect estimation using double machine learning". In: *The Annals of Statistics* 54.1, pp. 176–200.
- Lundborg, Anton Rask, Rajen D Shah, and Jonas Peters (2022). "Conditional independence testing in Hilbert spaces with applications to functional data analysis". In: *Journal of the Royal Statistical Society Series B: Statistical Methodology* 84.5, pp. 1821–1850.
- Powell, James L., James H. Stock, and Thomas M. Stoker (1989). "Semiparametric estimation of index coefficients". In: *Econometrica* 57.6, pp. 1403–1430.
- Qu, Annie, Bruce G Lindsay, and Bing Li (2000). "Improving generalised estimating equations using quadratic inference functions". In: *Biometrika* 87.4, pp. 823–836.
- Rothenhäusler, Dominik and Bin Yu (2020). "Incremental causal effects". In: *arXiv*, p. 1907.13258v4.

References II

- Rubin, D. B. and M. J. van der Laan (2008). "Empirical efficiency maximization: improved locally efficient covariate adjustment in randomized experiments and survival analysis". In: *The International Journal of Biostatistics* 4.1, pp. 4–5.
- Shah, Rajen D. and Jonas Peters (2020). "The hardness of conditional independence testing and the generalised covariance measure". In: *The Annals of Statistics* 48.3, pp. 1514–1538.
- Stoker, Thomas M. (1986). "Consistent Estimation of Scaled Coefficients". In: *Econometrica* 54.6, pp. 1461–1481.
- Tsiatis, Anastasios A (2006). *Semiparametric theory and missing data*. Springer.
- Vansteelandt, Stijn and Oliver Dukes (2022). "Assumption-lean inference for generalised linear model parameters". In: *Journal of the Royal Statistical Society Series B: Statistical Methodology* 84.3, pp. 657–685.
- Young, E. H. and R. D. Shah (May 2024). "Sandwich boosting for accurate estimation in partially linear models for grouped data". In: *Journal of the Royal Statistical Society Series B: Statistical Methodology*.
- Young, Elliot H and Rajen D Shah (2024). "ROSE Random Forests for Robust Semiparametric Efficient Estimation". In: *arXiv preprint arXiv:2410.03471*.

References III

Ziegler, Andreas (2011). *Generalized estimating equations*. Lecture notes in statistics. New York: Springer.