Augmenting Statistical Inference with Machine Learning I

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

ENAR Spring Meeting 2026
16 March 2026

UNIVERSITY OF CAMBRIDGE

Machine learning methods are useful for statistics

Machine learning methods can be useful for statistical tasks:

- parameter estimation
- hypothesis testing

They can offer greater

- flexibility / robustness
- accuracy
- power

These lectures aim to cover some general tools for integrating machine learning with statistical thinking.

Outline

Lecture 1

- Parametric statistics
- Semiparametric statistics
  - Partially linear model

Lecture 2

- Conditional independence testing
- Optimal inference in semiparametric models
- Nonparametric models
  - Average partial effect
- Optimal 'robust' inference in semiparametric models
- Grouped data

Lecture 3

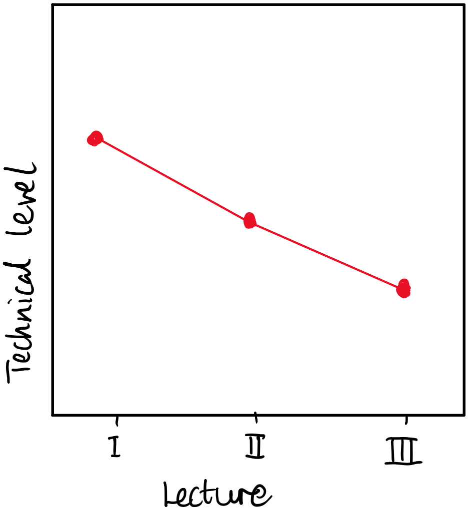
- Multiple sample-splitting
- Conditional mean independence an goodness-of-fit
- Nonparametric regression

Disclaimers

- Topics covered skew heavily towards my own interests
  - Important topics have been omitted
- Semiparametric statistics has a long history
  - Many excellent books available: Bickel et al., Tsiatis, van der Vaart, ...
- Many excellent tutorials available: Edward Kennedy, Oliver Hines et al., ...
- Will aim to highlight the main ideas, rather than focus on details

Disclaimers

- Topics covered skew heavily towards my own interests
  - Important topics have been omitted
- Semiparametric statistics has a long history
  - Many excellent books available: Bickel et al., Tsiatis, van der Vaart, ...
- Many excellent tutorials available: Edward Kennedy, Oliver Hines et al., ...
- Will aim to highlight the main ideas, rather than focus on details



Parametric models

Parametric models

- Y_i: number of visits to doctor
- X_i: smoking status
- $Z_i \in \mathbb{R}^p$: other health indicators

Model: $$Y_i \mid X_i, Z_i \sim \text{Poisson}\big(\exp(\mu + \theta X_i + \beta^\top Z_i)\big)\,,$$ independently for $i = 1, \dots, n$.

Parametric models

- Y_i: number of visits to doctor
- X_i: smoking status
- $Z_i \in \mathbb{R}^p$: other health indicators

Model: $Y_i \mid X_i, Z_i \sim \text{Poisson}\big(\exp(\mu + \theta X_i + \beta^\top Z_i)\big)$, independently for $i = 1, \dots, n$.

Maximise the log-likelihood

$$\ell_n(\mu, \theta, \beta) = \sum_{i=1}^{n} \log p(Y_i; \mu + \theta X_i + \beta^\top Z_i)$$

or equivalently, solve the score equations

$$\frac{1}{n} \sum_{i=1}^{n} S(Y_i; \mu + \theta X_i + \beta^\top Z_i) = 0.$$

Asymptotics

MLE $\hat{\gamma}$ of $\gamma := (\mu, \theta, \beta) \in \mathbb{R}^d$ satisfies

$$\sqrt{n}(\hat{\gamma} - \gamma_0) \xrightarrow{d} \mathcal{N}(0, i(\gamma_0)^{-1}),$$

where i is the (average) *Fisher information matrix*

$$i(\gamma) := \mathrm{Cov}_\gamma(S(Y; \mu + \theta X_i + \beta^\top Z_i)).$$

Asymptotics

MLE $\hat{\gamma}$ of $\gamma := (\mu, \theta, \beta) \in \mathbb{R}^d$ satisfies

$$\sqrt{n}(\hat{\gamma} - \gamma_0) \xrightarrow{d} \mathcal{N}(0, i(\gamma_0)^{-1}),$$

where i is the (average) *Fisher information matrix*

$$i(\gamma) := \mathrm{Cov}_\gamma(S(Y; \mu + \theta X_i + \beta^\top Z_i)).$$

By the delta method, for differentiable $\vartheta : \mathbb{R}^d \to \mathbb{R}$, have

$$\sqrt{n}\{\vartheta(\hat{\gamma}) - \vartheta(\gamma_0)\} \xrightarrow{d} \mathcal{N}(0, \nabla\vartheta(\gamma_0)^\top i(\gamma_0)^{-1} \nabla\vartheta(\gamma_0)).$$

In particular,

$$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} \mathcal{N}(0, \{i(\gamma_0)^{-1}\}_{\theta,\theta}).$$

MLE $\hat{\gamma}$ of $\gamma := (\mu, \theta, \beta) \in \mathbb{R}^d$ satisfies

$$\sqrt{n}(\hat{\gamma} - \gamma_0) \xrightarrow{d} \mathcal{N}(0, i(\gamma_0)^{-1}),$$

where i is the (average) *Fisher information matrix*

$$i(\gamma) := \mathrm{Cov}_\gamma(S(Y; \mu + \theta X_i + \beta^\top Z_i)).$$

By the delta method, for differentiable $\vartheta : \mathbb{R}^d \to \mathbb{R}$, have

$$\sqrt{n}\{\vartheta(\hat{\gamma}) - \vartheta(\gamma_0)\} \xrightarrow{d} \mathcal{N}(0, \nabla\vartheta(\gamma_0)^\top i(\gamma_0)^{-1} \nabla\vartheta(\gamma_0)).$$

In particular,

$$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} \mathcal{N}(0, \{i(\gamma_0)^{-1}\}_{\theta,\theta}).$$

Asymptotics

MLE $\hat{\gamma}$ of $\gamma := (\mu, \theta, \beta) \in \mathbb{R}^d$ satisfies

$$\sqrt{n}(\hat{\gamma} - \gamma_0) \xrightarrow{d} \mathcal{N}(0, i(\gamma_0)^{-1}),$$

where i is the (average) *Fisher information matrix*

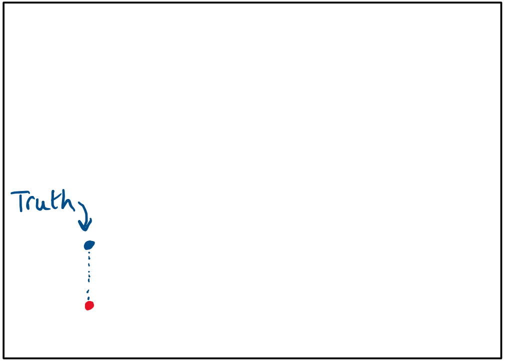
$$i(\gamma) := \mathrm{Cov}_\gamma(S(Y; \mu + \theta X_i + \beta^\top Z_i)).$$

By the delta method, for differentiable $\vartheta : \mathbb{R}^d \to \mathbb{R}$, have

$$\sqrt{n}\{\vartheta(\hat{\gamma}) - \vartheta(\gamma_0)\} \xrightarrow{d} \mathcal{N}(0, \nabla\vartheta(\gamma_0)^\top i(\gamma_0)^{-1} \nabla\vartheta(\gamma_0)).$$

In particular,

$$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} \mathcal{N}(0, \{i(\gamma_0)^{-1}\}_{\theta,\theta}).$$



Asymptotics

MLE $\hat{\gamma}$ of $\gamma := (\mu, \theta, \beta) \in \mathbb{R}^d$ satisfies

$$\sqrt{n}(\hat{\gamma} - \gamma_0) \xrightarrow{d} \mathcal{N}(0, i(\gamma_0)^{-1}),$$

where i is the (average) *Fisher information matrix*

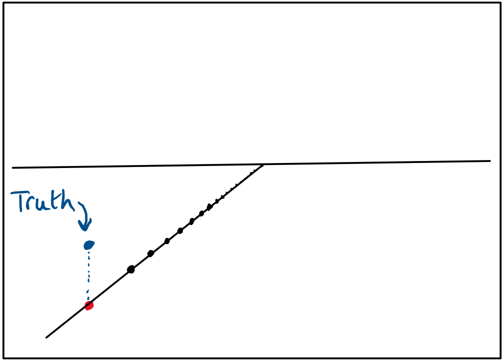
$$i(\gamma) := \mathrm{Cov}_\gamma(S(Y; \mu + \theta X_i + \beta^\top Z_i)).$$

By the delta method, for differentiable $\vartheta : \mathbb{R}^d \to \mathbb{R}$, have

$$\sqrt{n}\{\vartheta(\hat{\gamma}) - \vartheta(\gamma_0)\} \xrightarrow{d} \mathcal{N}(0, \nabla\vartheta(\gamma_0)^\top i(\gamma_0)^{-1} \nabla\vartheta(\gamma_0)).$$

In particular,

$$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} \mathcal{N}(0, \{i(\gamma_0)^{-1}\}_{\theta,\theta}).$$



Asymptotics

MLE $\hat{\gamma}$ of $\gamma := (\mu, \theta, \beta) \in \mathbb{R}^d$ satisfies

$$\sqrt{n}(\hat{\gamma} - \gamma_0) \xrightarrow{d} \mathcal{N}(0, i(\gamma_0)^{-1}),$$

where i is the (average) *Fisher information matrix*

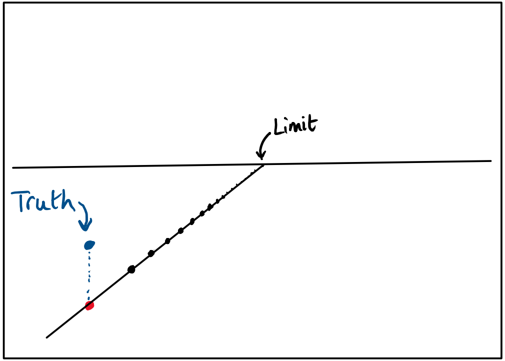
$$i(\gamma) := \mathrm{Cov}_\gamma(S(Y; \mu + \theta X_i + \beta^\top Z_i)).$$

By the delta method, for differentiable $\vartheta : \mathbb{R}^d \to \mathbb{R}$, have

$$\sqrt{n}\{\vartheta(\hat{\gamma}) - \vartheta(\gamma_0)\} \xrightarrow{d} \mathcal{N}(0, \nabla\vartheta(\gamma_0)^\top i(\gamma_0)^{-1} \nabla\vartheta(\gamma_0)).$$

In particular,

$$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} \mathcal{N}(0, \{i(\gamma_0)^{-1}\}_{\theta,\theta}).$$



The score

What is making this work is that

$$\mathbb{E}_\gamma S(Y_i; \mu + \theta X_i + \beta^\top Z_i) = 0.$$

The score

What is making this work is that

$$\mathbb{E}_\gamma S(Y_i; \mu + \theta X_i + \beta^\top Z_i) = 0.$$

More generally, in a parametric model, the score $S(Y; \gamma)$ satisfies

$$\mathbb{E}_\gamma(S(Y;\gamma)\, f(Y)) = \nabla_\gamma \mathbb{E}_\gamma(f(Y))$$

for any statistic $f(Y)$.

The score

What is making this work is that

$$\mathbb{E}_\gamma S(Y_i; \mu + \theta X_i + \beta^\top Z_i) = 0.$$

More generally, in a parametric model, the score $S(Y; \gamma)$ satisfies

$$\mathbb{E}_\gamma(S(Y;\gamma)\, f(Y)) = \nabla_\gamma \mathbb{E}_\gamma(f(Y))$$

for any statistic $f(Y)$.

Indeed, considering the case $\gamma \in \mathbb{R}$ for simplicity,

$$\mathbb{E}_\gamma(S(Y;\gamma)f(Y)) = \int \frac{d}{d\gamma} \log(p(y;\gamma))f(y)p(y;\gamma)dy$$

The score

What is making this work is that

$$\mathbb{E}_\gamma S(Y_i; \mu + \theta X_i + \beta^\top Z_i) = 0.$$

More generally, in a parametric model, the score $S(Y;\gamma)$ satisfies

$$\mathbb{E}_\gamma(S(Y;\gamma)\, f(Y)) = \nabla_\gamma \mathbb{E}_\gamma(f(Y))$$

for any statistic $f(Y)$.

Indeed, considering the case $\gamma \in \mathbb{R}$ for simplicity,

$$\begin{aligned}\mathbb{E}_\gamma(S(Y;\gamma)f(Y)) &= \int \frac{d}{d\gamma} \log(p(y;\gamma)) f(y)p(y;\gamma)dy \\ &= \int \frac{\frac{d}{d\gamma} p(y;\gamma)}{p(y;\gamma)} f(y)p(y;\gamma)dy\end{aligned}$$

The score

What is making this work is that

$$\mathbb{E}_\gamma S(Y_i; \mu + \theta X_i + \beta^\top Z_i) = 0.$$

More generally, in a parametric model, the score $S(Y;\gamma)$ satisfies

$$\mathbb{E}_\gamma(S(Y;\gamma)\, f(Y)) = \nabla_\gamma \mathbb{E}_\gamma(f(Y))$$

for any statistic $f(Y)$.

Indeed, considering the case $\gamma \in \mathbb{R}$ for simplicity,

$$\begin{aligned}
\mathbb{E}_\gamma(S(Y;\gamma) f(Y)) &= \int \frac{d}{d\gamma} \log(p(y;\gamma)) f(y) p(y;\gamma) dy \\
&= \int \frac{\frac{d}{d\gamma} p(y;\gamma)}{p(y;\gamma)} f(y) p(y;\gamma) dy \\
&= \int \frac{d}{d\gamma} p(y;\gamma) f(y) dy
\end{aligned}$$

The score

What is making this work is that

$$\mathbb{E}_\gamma S(Y_i; \mu + \theta X_i + \beta^\top Z_i) = 0.$$

More generally, in a parametric model, the score $S(Y;\gamma)$ satisfies

$$\mathbb{E}_\gamma(S(Y;\gamma)\, f(Y)) = \nabla_\gamma \mathbb{E}_\gamma(f(Y))$$

for any statistic $f(Y)$.

Indeed, considering the case $\gamma \in \mathbb{R}$ for simplicity,

$$\begin{aligned}
\mathbb{E}_\gamma(S(Y;\gamma) f(Y)) &= \int \frac{d}{d\gamma} \log(p(y;\gamma)) f(y) p(y;\gamma) dy \\
&= \int \frac{\frac{d}{d\gamma} p(y;\gamma)}{p(y;\gamma)} f(y) p(y;\gamma) dy \\
&= \int \frac{d}{d\gamma} p(y;\gamma) f(y) dy \\
&= \frac{d}{d\gamma} \int p(y;\gamma) f(y) dy.
\end{aligned}$$

Optimality

For any estimator $\hat{\gamma} = \hat{\gamma}(Y)$: $$\frac{d}{d\gamma}\mathbb{E}_\gamma \hat{\gamma} = \mathbb{E}_\gamma(S(Y;\gamma)\hat{\gamma})$$

Optimality

For any estimator $\hat{\gamma} = \hat{\gamma}(Y)$: $$\left(\frac{d}{d\gamma}\mathbb{E}_\gamma\hat{\gamma}\right)^2 = (\mathbb{E}_\gamma(S(Y;\gamma)\hat{\gamma}))^2$$

Optimality

For any estimator $\hat{\gamma} = \hat{\gamma}(Y)$:

$$\left(\frac{d}{d\gamma}\mathbb{E}_\gamma \hat{\gamma}\right)^2 = (\mathbb{E}_\gamma(S(Y;\gamma)\hat{\gamma}))^2$$
$$= (\mathbb{E}_\gamma[S(Y;\gamma)\{\hat{\gamma} - \mathbb{E}_\gamma(\hat{\gamma})\}])^2$$

Optimality

For any estimator $\hat{\gamma} = \hat{\gamma}(Y)$:

$$\begin{aligned}\left(\frac{d}{d\gamma}\mathbb{E}_\gamma\hat{\gamma}\right)^2 &= (\mathbb{E}_\gamma(S(Y;\gamma)\hat{\gamma}))^2 \\ &= (\mathbb{E}_\gamma[S(Y;\gamma)\{\hat{\gamma} - \mathbb{E}_\gamma(\hat{\gamma})\}])^2 \\ &\leq \mathrm{Var}_\gamma(S(Y;\gamma))\mathrm{Var}(\hat{\gamma}).\end{aligned}$$

Optimality

For any estimator $\hat{\gamma} = \hat{\gamma}(Y)$:

$$\begin{aligned}\left(\frac{d}{d\gamma}\mathbb{E}_\gamma \hat{\gamma}\right)^2 &= (\mathbb{E}_\gamma(S(Y;\gamma)\hat{\gamma}))^2 \\ &= (\mathbb{E}_\gamma[S(Y;\gamma)\{\hat{\gamma} - \mathbb{E}_\gamma(\hat{\gamma})\}])^2 \\ &\leq \mathrm{Var}_\gamma(S(Y;\gamma))\mathrm{Var}(\hat{\gamma}).\end{aligned}$$

Cramér–Rao lower bound: Suppose $\gamma \in \mathbb{R}^p$ and $\hat{\vartheta}$ is an unbiased estimator of $\vartheta(\gamma) \in \mathbb{R}$ based on n i.i.d. observations. Then

$$n\mathrm{Var}_\gamma(\hat{\vartheta}) \geq \nabla\vartheta(\gamma)^\top i(\gamma)^{-1}\nabla\vartheta(\gamma).$$

Optimality

For any estimator $\hat{\gamma} = \hat{\gamma}(Y)$:

$$\left(\frac{d}{d\gamma}\mathbb{E}_\gamma \hat{\gamma}\right)^2 = (\mathbb{E}_\gamma(S(Y;\gamma)\hat{\gamma}))^2$$
$$= (\mathbb{E}_\gamma[S(Y;\gamma)\{\hat{\gamma} - \mathbb{E}_\gamma(\hat{\gamma})\}])^2$$
$$\leq \mathrm{Var}_\gamma(S(Y;\gamma))\mathrm{Var}(\hat{\gamma}).$$

Cramér–Rao lower bound: Suppose $\gamma \in \mathbb{R}^p$ and $\hat{\vartheta}$ is an unbiased estimator of $\vartheta(\gamma) \in \mathbb{R}$ based on n i.i.d. observations. Then

$$n\mathrm{Var}_\gamma(\hat{\vartheta}) \geq \nabla\vartheta(\gamma)^\top i(\gamma)^{-1}\nabla\vartheta(\gamma).$$

Locally asymptotically minimax bound: For any sequence of estimators $\hat{\vartheta}_n$ and any $\delta > 0$,

$$\liminf_{n\to\infty} \sup_{\|\gamma'-\gamma\|<\delta} \mathbb{E}_{\gamma'}[n\|\hat{\vartheta}_n - \vartheta(\gamma')\|^2] \geq \nabla\vartheta(\gamma)^\top i(\gamma)^{-1}\nabla\vartheta(\gamma).$$

Optimality

For any estimator $\hat{\gamma} = \hat{\gamma}(Y)$:

$$\left(\frac{d}{d\gamma}\mathbb{E}_\gamma\hat{\gamma}\right)^2 = (\mathbb{E}_\gamma(S(Y;\gamma)\hat{\gamma}))^2$$
$$= (\mathbb{E}_\gamma[S(Y;\gamma)\{\hat{\gamma} - \mathbb{E}_\gamma(\hat{\gamma})\}])^2$$
$$\leq \text{Var}_\gamma(S(Y;\gamma))\text{Var}(\hat{\gamma}).$$

Cramér–Rao lower bound: Suppose $\gamma \in \mathbb{R}^p$ and $\hat{\vartheta}$ is an unbiased estimator of $\vartheta(\gamma) \in \mathbb{R}$ based on n i.i.d. observations. Then

$$n\text{Var}_\gamma(\hat{\vartheta}) \geq \nabla\vartheta(\gamma)^\top i(\gamma)^{-1}\nabla\vartheta(\gamma).$$

Locally asymptotically minimax bound: For any sequence of estimators $\hat{\vartheta}_n$ and any $\delta > 0$,

$$\liminf_{n\to\infty} \sup_{\|\gamma'-\gamma\|<\delta} \mathbb{E}_{\gamma'}[n\|\hat{\vartheta}_n - \vartheta(\gamma')\|^2] \geq \nabla\vartheta(\gamma)^\top i(\gamma)^{-1}\nabla\vartheta(\gamma).$$

(Functions of) MLEs are asymptotically optimal for estimating (functions of) the true parameter.

Increasing model complexity

Returning to our example, what if the health indicators Z_i contribute nonlinearly to the outcome?

Increasing model complexity

Returning to our example, what if the health indicators Z_i contribute nonlinearly to the outcome?

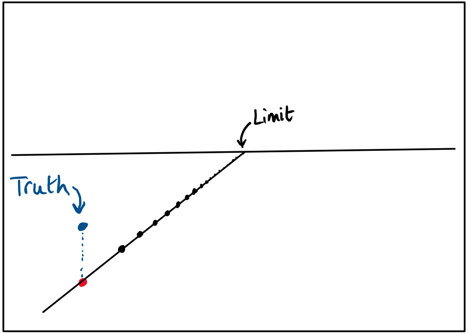
Can add quadratic / higher degree polynomial effects etc.

Increasing model complexity

Returning to our example, what if the health indicators Z_i contribute nonlinearly to the outcome?

Can add quadratic / higher degree polynomial effects etc.

When the number of parameters becomes large, the asymptotic arguments become increasingly shaky...

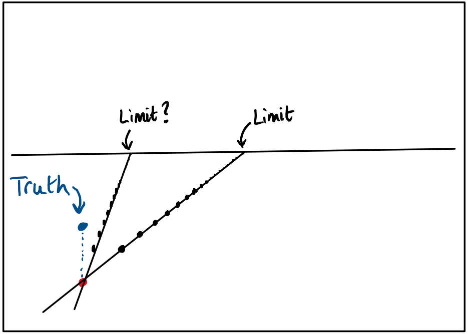


Increasing model complexity

Returning to our example, what if the health indicators Z_i contribute nonlinearly to the outcome?

Can add quadratic / higher degree polynomial effects etc.

When the number of parameters becomes large, the asymptotic arguments become increasingly shaky...

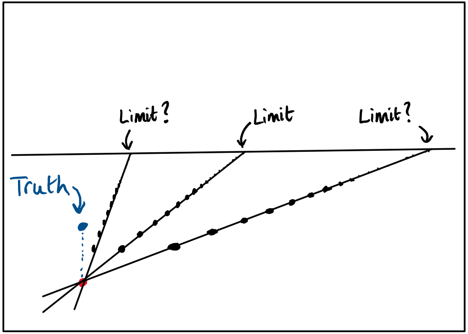


Increasing model complexity

Returning to our example, what if the health indicators Z_i contribute nonlinearly to the outcome?

Can add quadratic / higher degree polynomial effects etc.

When the number of parameters becomes large, the asymptotic arguments become increasingly shaky...



Semiparametric statistics

Partially linear model

Instead, can consider a model rich enough to include the models we might have considered e.g. a *generalised partially linear model*.

For simplicity, we will discuss a *partially linear model*

$$Y_i = \theta X_i + f(Z_i) + \varepsilon_i$$

where $\mathbb{E}(\varepsilon_i \,|\, X_i, Z_i) = 0$ and f is an unknown function.

Our interest continues to centre on θ, which describes the contribution of X_i after accounting for Z_i.

This is an example of a *semiparametric model*: the model cannot be parametrised by a finite-dimensional vector.

Plug-in approach

$$\text{Model:} \qquad Y_i = \theta X_i + f(Z_i) + \varepsilon_i$$

Naive approach: estimate θ and f directly using ML.

Plug-in approach

$$\text{Model:} \qquad Y_i = \theta X_i + f(Z_i) + \varepsilon_i$$

Naive approach: estimate θ and f directly using ML.

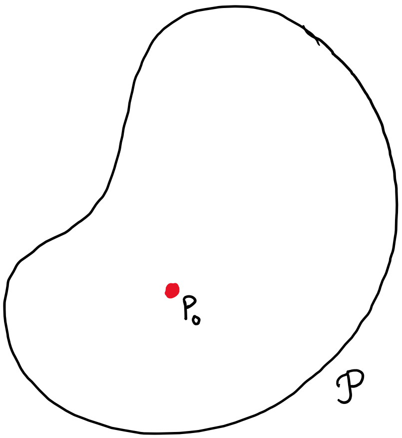
Typically have slower than $1/\sqrt{n}$ rate for estimating f.

E.g. If $\mathcal{F}$ is a class of β-Hölder smooth functions, then

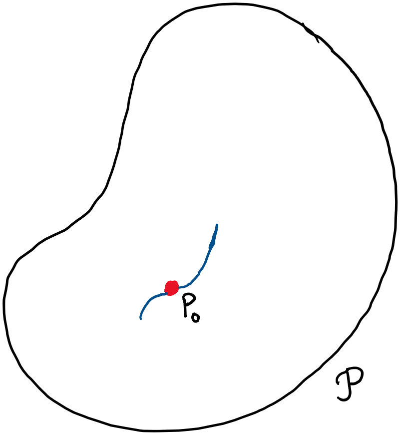
$$\inf_{\text{estimators } \hat{f}} \sup_{f \in \mathcal{F}} \left(\mathbb{E}\{f(Z) - \hat{f}(Z)\}^2 \right)^{1/2} \geq C n^{-\frac{\beta}{2\beta + p}}.$$

May not work well: slower than $1/\sqrt{n}$ rate for estimating f can propagate to estimation error of θ.

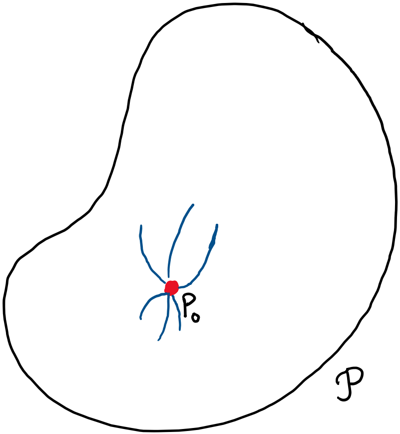
Parametric sub-models



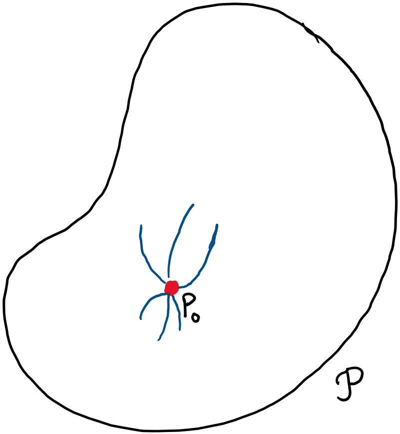
Parametric sub-models



Parametric sub-models



Parametric sub-models



For each $P \in \mathcal{P}$, let $\theta(P)$ be our parameter of interest.

Consider parametric sub-model $t \mapsto P_t$ with score S at $t = 0$.

The collection of all sub-model scores at $P \in \mathcal{P}$ is known as the *tangent space* $\dot{\mathcal{P}}_P$ at P.

A Cramér–Rao lower bound at P_0 for the sub-model is

$$\frac{\left(\frac{d}{dt}\theta(P_t)|_{t=0}\right)^2}{\mathrm{Var}_{P_0}(S)}$$

Can we achieve the sup over all such C–R lower bounds?

Parametric sub-models in the PLM

Consider paths:

$$t \mapsto p_t(y|x,z)p_t(x,z)$$

- Given bounded function $a(x,z)$ satisfying $\int ap_0 = 0$,

$$p_t(x,z) := p_0(x,z)(1 + ta(x,z))$$

- Similarly we can take

$$p_t(y|x,z) := p_0(y|x,z)(1 + tb(y,x,z))$$

  where b is such that $\exists$ function b_2 where $\forall\ x,z$:

$$\int b(y,x,z)p_0(y|x,z)dy = 0$$
$$\int yb(y,x,z)p_0(y|x,z)dy = b_1 x + b_2(z).$$

Parametric sub-models in the PLM

Consider paths:

$$t \mapsto p_t(y|x,z)p_t(x,z)$$

- Given bounded function $a(x,z)$ satisfying $\int ap_0 = 0$,

$$p_t(x,z) := p_0(x,z)(1 + ta(x,z))$$

- Similarly we can take

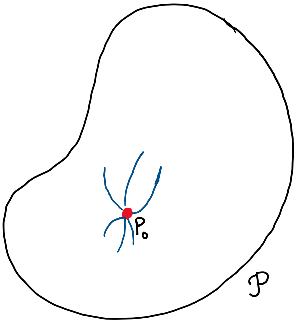
$$p_t(y|x,z) := p_0(y|x,z)(1 + tb(y,x,z))$$

  where b is such that $\exists$ function b_2 where $\forall\ x, z$:

$$\int b(y,x,z)p_0(y|x,z)dy = 0$$
$$\int yb(y,x,z)p_0(y|x,z)dy = b_1 x + b_2(z).$$

The sub-model score will be $a(x,z) + b(y,x,z)$.

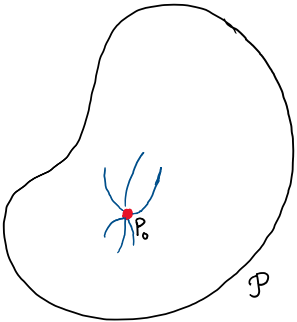
Nuisance tangent space



Consider paths $t \mapsto P_t$ where θ is locally constant:

$$\frac{d}{dt}\theta(P_t)\Big|_{t=0} = \lim_{t\to 0} \frac{\theta(P_t) - \theta(P_0)}{t} = 0.$$

Nuisance tangent space

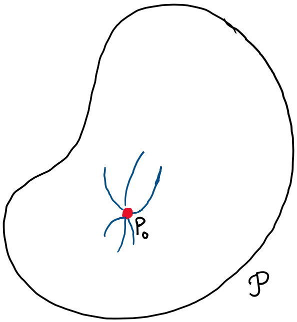


Consider paths $t \mapsto P_t$ where θ is locally constant:

$$\frac{d}{dt}\theta(P_t)\Big|_{t=0} = \lim_{t\to 0} \frac{\theta(P_t) - \theta(P_0)}{t} = 0.$$

The set of scores associated with these paths is known as the *nuisance tangent set* $\dot{\mathcal{P}}_{P_0,\eta}$.

Nuisance tangent space



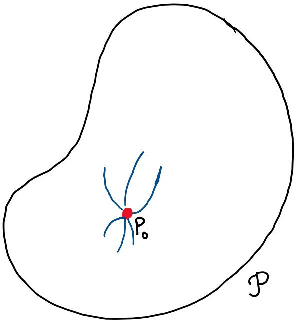
Consider paths $t \mapsto P_t$ where θ is locally constant:

$$\frac{d}{dt}\theta(P_t)\Big|_{t=0} = \lim_{t\to 0}\frac{\theta(P_t) - \theta(P_0)}{t} = 0.$$

The set of scores associated with these paths is known as the *nuisance tangent set* $\dot{\mathcal{P}}_{P_0,\eta}$.

For this notion to make sense, we need some regularity.

Nuisance tangent space



Consider paths $t \mapsto P_t$ where θ is locally constant:

$$\frac{d}{dt}\theta(P_t)\Big|_{t=0} = \lim_{t\to 0} \frac{\theta(P_t) - \theta(P_0)}{t} = 0.$$

The set of scores associated with these paths is known as the *nuisance tangent set* $\dot{\mathcal{P}}_{P_0,\eta}$.

For this notion to make sense, we need some regularity.

The correct form of this is *pathwise differentiability*. This asks for the existence of a function $\tilde{\psi}_{P_0}$ such that for every sub-model $t \mapsto P_t$,

$$\frac{d}{dt}\theta(P_t)\Big|_{t=0} = \mathbb{E}_{P_0}(S\tilde{\psi}_{P_0}).$$

Here S is a score at $t = 0$ for the path.

Orthogonal complement of the nuisance tangent space

Pathwise differentiability: there exists a function $\tilde{\psi}_{P_0}$ such that for every sub-model $t \mapsto P_t$,

$$\frac{d}{dt}\theta(P_t)\Big|_{t=0} = \mathbb{E}_{P_0}(S\tilde{\psi}_{P_0}).$$

Here S is a score at $t = 0$ for the path.

Key point: $\frac{d}{dt}\theta(P_t)\Big|_{t=0} = 0$ represents a linear constraint on the scores.

Orthogonal complement of the nuisance tangent space

Pathwise differentiability: there exists a function $\tilde{\psi}_{P_0}$ such that for every sub-model $t \mapsto P_t$,

$$\frac{d}{dt}\theta(P_t)\Big|_{t=0} = \mathbb{E}_{P_0}(S\tilde{\psi}_{P_0}).$$

Here S is a score at $t = 0$ for the path.

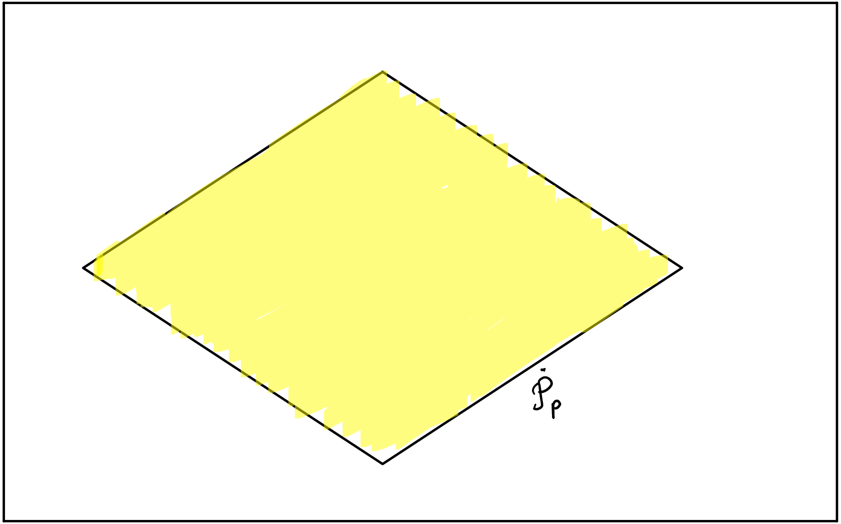
Key point: $\frac{d}{dt}\theta(P_t)\Big|_{t=0} = 0$ represents a linear constraint on the scores.

Key idea: Consider mean-zero functions ψ_{P_0} orthogonal to the nuisance tangent space i.e.

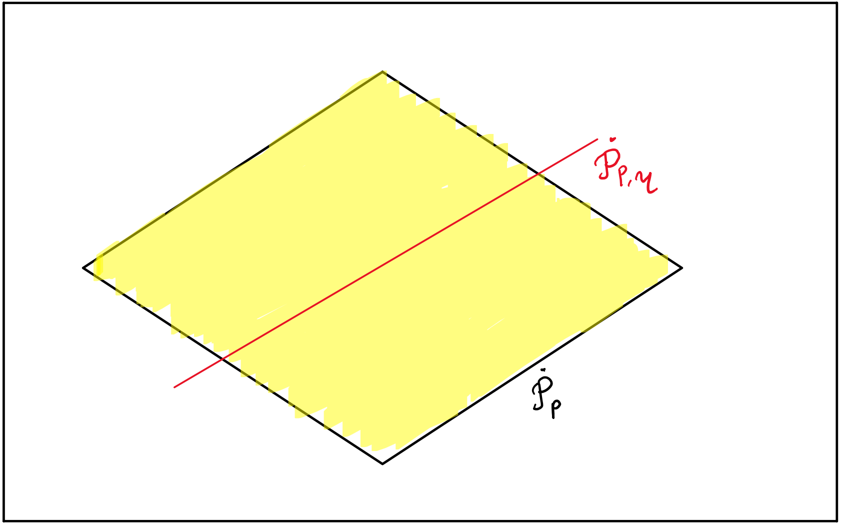
$$\mathbb{E}_{P_0}(S\psi_{P_0}) = 0$$

for all S in the nuisance tangent space.

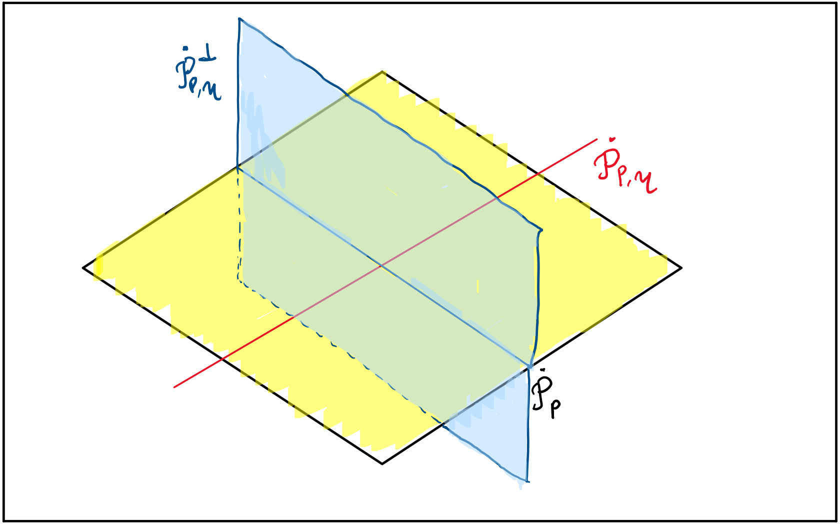
Orthogonal complement of the nuisance tangent space



Orthogonal complement of the nuisance tangent space



Orthogonal complement of the nuisance tangent space



Orthogonal complement of the nuisance tangent space

Then

$$0 = \mathbb{E}_{P_0}(S\psi_{P_0}) = \frac{d}{dt}\mathbb{E}_{P_t}(\psi_{P_0}).$$

Here $t \mapsto P_t$ is any sub-model where $\theta(P_t)$ is locally constant.

Orthogonal complement of the nuisance tangent space

Then

$$0 = \mathbb{E}_{P_0}(S\psi_{P_0}) = \frac{d}{dt}\mathbb{E}_{P_t}(\psi_{P_0}).$$

Here $t \mapsto P_t$ is any sub-model where $\theta(P_t)$ is locally constant.

In other words, the mean of ψ_{P_0} is insensitive to changes in the nuisance parameter.

If a distribution $\hat{P} \in \mathcal{P}$ is such that its associated nuisance parameters are close to that of P_0, then we can expect

$$\mathbb{E}_{\hat{P}}\psi_{P_0} \approx \mathbb{E}_{P_0}\psi_{P_0} = 0.$$

Orthogonal complement of the nuisance tangent space

Then

$$0 = \mathbb{E}_{P_0}(S\psi_{P_0}) = \frac{d}{dt}\mathbb{E}_{P_t}(\psi_{P_0}).$$

Here $t \mapsto P_t$ is any sub-model where $\theta(P_t)$ is locally constant.

In other words, the mean of ψ_{P_0} is insensitive to changes in the nuisance parameter.

If a distribution $\hat{P} \in \mathcal{P}$ is such that its associated nuisance parameters are close to that of P_0, then we can expect

$$\mathbb{E}_{\hat{P}}\psi_{P_0} \approx \mathbb{E}_{P_0}\psi_{P_0} = 0.$$

More importantly, reversing the roles of $\hat{P}$ and P_0 in the above, we can expect

$$\mathbb{E}_{P_0}(\psi_{\hat{P}}) \approx \mathbb{E}_{P_0}(\psi_{P_0}) = 0.$$

Orthogonal complement of the nuisance tangent space

More importantly, reversing the roles of $\hat{P}$ and P_0 in the above, we can expect

$$\mathbb{E}_{P_0}\psi_{\hat{P}} \approx \mathbb{E}_{P_0}\psi_{P_0} = 0.$$

What is this telling us?

- Typically ψ_P will depend on P through $\theta(P)$ and nuisance parameters η (this will become concrete when we return the the PLM example shortly)
  $\Rightarrow$ can write ψ_P as $\psi_{\theta,\eta}$.
- If $\hat{\eta}$ is an estimate of η, then

$$\sum_i \psi_{\theta,\hat{\eta}}(Y_i, X_i, Z_i) = 0$$

  should be an approximately unbiased estimating equation for θ.

Nuisance tangent space in the PLM

Recall that the tangent space of the PLM consisted of functions of the form

$$a(X, Z) + b(Y, X, Z)$$

where $\mathbb{E}a(X, Z) = 0$, $\mathbb{E}(b(Y, X, Z) \,|\, X, Z) = 0$ and $\mathbb{E}(Yb(Y, X, Z) \,|\, X, Z) = b_1 X + b_2(Z)$.

The nuisance tangent space then additionally requires

$$\mathbb{E}(Yb(Y, X, Z) \,|\, X, Z) = b_2(Z).$$

Nuisance tangent space in the PLM

Recall that the tangent space of the PLM consisted of functions of the form

$$a(X, Z) + b(Y, X, Z)$$

where $\mathbb{E}a(X, Z) = 0$, $\mathbb{E}(b(Y, X, Z) \,|\, X, Z) = 0$ and $\mathbb{E}(Yb(Y, X, Z) \,|\, X, Z) = b_1 X + b_2(Z)$.

The nuisance tangent space then additionally requires

$$\mathbb{E}(Yb(Y, X, Z) \,|\, X, Z) = b_2(Z).$$

Can show that the orthogonal complement of the nuisance tangent space consists of functions of the form

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

'Double machine learning' estimator

Taking $\phi(X, Z) = X$, and writing $m(z) := \mathbb{E}(X \mid Z = z)$, can form estimating equation

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

That is, we can take

$$\hat{\theta} = \frac{\sum_i \{Y_i - \hat{f}(Z_i)\}\{X_i - \hat{m}(Z_i)\}}{\sum_i X_i\{X_i - \hat{m}(X_i)\}}.$$

'Double machine learning' estimator

Taking $\phi(X, Z) = X$, and writing $m(z) := \mathbb{E}(X \mid Z = z)$, can form estimating equation

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

That is, we can take

$$\hat{\theta} = \frac{\sum_i \{Y_i - \hat{f}(Z_i)\}\{X_i - \hat{m}(Z_i)\}}{\sum_i X_i\{X_i - \hat{m}(X_i)\}}.$$

From our discussion, this should be relatively insensitive to the quality of the estimators $\hat{\eta} := (\hat{m}, \hat{f})$.

'Double machine learning' estimator

In fact

$$\sqrt{n}(\hat\theta - \theta) = \sqrt{n}\frac{\frac{1}{n}\sum_i\{Y_i - \hat f(Z_i)\}\{X_i - \hat m(Z_i)\} - \theta X_i\{X_i - \hat m(Z_i)\}}{\frac{1}{n}\sum_i X_i\{X_i - \hat m(X_i)\}}$$

$$= \frac{\frac{1}{\sqrt{n}}\sum_i\{Y_i - \theta X_i - \hat f(Z_i)\}\{X_i - \hat m(Z_i)\}}{\frac{1}{n}\sum_i X_i\{X_i - \hat m(X_i)\}}.$$

‘Double machine learning’ estimator

In fact

$$
\begin{aligned}
\sqrt{n}(\hat{\theta} - \theta) &= \sqrt{n}\frac{\frac{1}{n}\sum_i\{Y_i - \hat{f}(Z_i)\}\{X_i - \hat{m}(Z_i)\} - \theta X_i\{X_i - \hat{m}(Z_i)\}}{\frac{1}{n}\sum_i X_i\{X_i - \hat{m}(X_i)\}} \\
&= \frac{\frac{1}{\sqrt{n}}\sum_i\{Y_i - \theta X_i - \hat{f}(Z_i)\}\{X_i - \hat{m}(Z_i)\}}{\frac{1}{n}\sum_i X_i\{X_i - \hat{m}(X_i)\}}.
\end{aligned}
$$

Recall

- $Y_i = \theta X_i + f(Z_i) + \varepsilon_i$ where $\mathbb{E}(\varepsilon_i \,|\, X_i, Z_i) = 0$,
- $X_i = m(Z_i) + \xi_i$ where $\mathbb{E}(\xi_i \,|\, Z_i) = 0$.

'Double machine learning' estimator

In fact

$$\sqrt{n}(\hat{\theta} - \theta) = \sqrt{n}\frac{\frac{1}{n}\sum_i \{Y_i - \hat{f}(Z_i)\}\{X_i - \hat{m}(Z_i)\} - \theta X_i\{X_i - \hat{m}(Z_i)\}}{\frac{1}{n}\sum_i X_i\{X_i - \hat{m}(X_i)\}}$$

$$= \frac{\frac{1}{\sqrt{n}}\sum_i \{Y_i - \theta X_i - \hat{f}(Z_i)\}\{X_i - \hat{m}(Z_i)\}}{\frac{1}{n}\sum_i X_i\{X_i - \hat{m}(X_i)\}}.$$

Recall

- $Y_i = \theta X_i + f(Z_i) + \varepsilon_i$ where $\mathbb{E}(\varepsilon_i \,|\, X_i, Z_i) = 0$,
- $X_i = m(Z_i) + \xi_i$ where $\mathbb{E}(\xi_i \,|\, Z_i) = 0$.

Substituting in, numerator gives an ith summand of the form

$$\{f(Z_i) - \hat{f}(Z_i) + \varepsilon_i\}\{m(Z_i) - \hat{m}(Z_i) + \xi_i\} = \underbrace{\varepsilon_i \xi_i}_{\text{CLT}} + \underbrace{\{f(Z_i) - \hat{f}(Z_i)\}\{m(Z_i) - \hat{m}(Z_i)\}}_{\text{Product of errors}}$$

$$+ \underbrace{\varepsilon_i\{m(Z_i) - \hat{m}(Z_i)\}}_{\text{mean zero}} + \underbrace{\xi_i\{\hat{f}(Z_i) - f(Z_i)\}}_{\text{small?}}.$$

'Double machine learning' estimator

In fact

$$\sqrt{n}(\hat\theta - \theta) = \sqrt{n}\frac{\frac{1}{n}\sum_i \{Y_i - \hat f(Z_i)\}\{X_i - \hat m(Z_i)\} - \theta X_i\{X_i - \hat m(Z_i)\}}{\frac{1}{n}\sum_i X_i\{X_i - \hat m(X_i)\}}$$
$$= \frac{\frac{1}{\sqrt{n}}\sum_i \{Y_i - \theta X_i - \hat f(Z_i)\}\{X_i - \hat m(Z_i)\}}{\frac{1}{n}\sum_i X_i\{X_i - \hat m(X_i)\}}.$$

Recall

- $Y_i = \theta X_i + f(Z_i) + \varepsilon_i$ where $\mathbb{E}(\varepsilon_i \,|\, X_i, Z_i) = 0$,
- $X_i = m(Z_i) + \xi_i$ where $\mathbb{E}(\xi_i \,|\, Z_i) = 0$.

Substituting in, numerator gives an ith summand of the form

$$\{f(Z_i) - \hat f(Z_i) + \varepsilon_i\}\{m(Z_i) - \hat m(Z_i) + \xi_i\} = \underbrace{\varepsilon_i \xi_i}_{\text{CLT}} + \underbrace{\{f(Z_i) - \hat f(Z_i)\}\{m(Z_i) - \hat m(Z_i)\}}_{\text{Product of errors}}$$
$$+ \underbrace{\varepsilon_i\{m(Z_i) - \hat m(Z_i)\}}_{\text{mean zero}} + \underbrace{\xi_i\{\hat f(Z_i) - f(Z_i)\}}_{\text{small?}}.$$

Product of errors

Using the Cauchy–Schwarz inequality

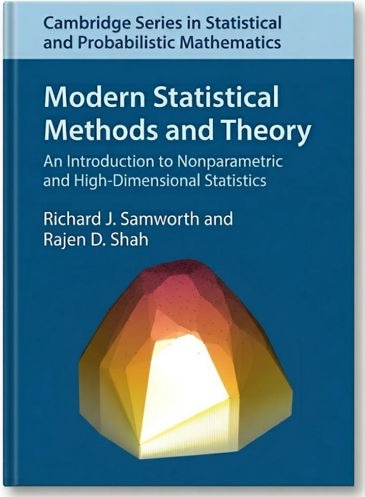
$$
\begin{aligned}
&\mathbb{E}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^n |f(Z_i) - \hat{f}(Z_i)||m(Z_i) - \hat{m}(Z_i)|\right) \\
\leq &\mathbb{E}\left\{\frac{1}{\sqrt{n}}\left(\sum_{i=1}^n \{f(Z_i) - \hat{f}(Z_i)\}^2\right)^{1/2}\left(\sum_{i=1}^n \{m(Z_i) - \hat{m}(Z_i)\}^2\right)^{1/2}\right\} \\
\leq &\sqrt{n}\left\{\mathbb{E}\left(\frac{1}{n}\sum_{i=1}^n \{f(Z_i) - \hat{f}(Z_i)\}^2\right)\right\}^{1/2}\left\{\mathbb{E}\left(\frac{1}{n}\sum_{i=1}^n \{m(Z_i) - \hat{m}(Z_i)\}^2\right)\right\}^{1/2}
\end{aligned}
$$

Product of errors

$$\sqrt{n}\left\{\mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n}\{f(Z_i)-\hat{f}(Z_i)\}^2\right)\right\}^{1/2}$$
$$\times\left\{\mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n}\{m(Z_i)-\hat{m}(Z_i)\}^2\right)\right\}^{1/2}$$

Not unreasonable to expect each of the above MSPEs to be $\ll \sqrt{n}$, so the product is $\ll \sqrt{n}$ as required for it to be negligible.

E.g. True when performing kernel ridge regression and the regression function is in an RKHS.



'Double machine learning' estimator

In fact

$$\sqrt{n}(\hat{\theta} - \theta) = \sqrt{n}\frac{\frac{1}{n}\sum_i \{Y_i - \hat{f}(Z_i)\}\{X_i - \hat{m}(Z_i)\} - \theta X_i\{X_i - \hat{m}(Z_i)\}}{\frac{1}{n}\sum_i X_i\{X_i - \hat{m}(X_i)\}}$$

$$= \frac{\frac{1}{\sqrt{n}}\sum_i \{Y_i - \theta X_i - \hat{f}(Z_i)\}\{X_i - \hat{m}(Z_i)\}}{\frac{1}{n}\sum_i X_i\{X_i - \hat{m}(X_i)\}}.$$

Recall

- $Y_i = \theta X_i + f(Z_i) + \varepsilon_i$ where $\mathbb{E}(\varepsilon_i \mid X_i, Z_i) = 0$,
- $X_i = m(Z_i) + \xi_i$ where $\mathbb{E}(\xi_i \mid Z_i) = 0$.

Substituting in, numerator gives an ith summand of the form

$$\{f(Z_i) - \hat{f}(Z_i) + \varepsilon_i\}\{m(Z_i) - \hat{m}(Z_i) + \xi_i\} = \underbrace{\varepsilon_i \xi_i}_{\text{CLT}} + \underbrace{\{f(Z_i) - \hat{f}(Z_i)\}\{m(Z_i) - \hat{m}(Z_i)\}}_{\text{Product of errors}}$$

$$+ \underbrace{\varepsilon_i\{m(Z_i) - \hat{m}(Z_i)\}}_{\text{mean zero}} + \underbrace{\xi_i\{\hat{f}(Z_i) - f(Z_i)\}}_{\text{small?}}.$$

'Double machine learning' estimator

In fact

$$\sqrt{n}(\hat{\theta} - \theta) = \sqrt{n}\frac{\frac{1}{n}\sum_i \{Y_i - \hat{f}(Z_i)\}\{X_i - \hat{m}(Z_i)\} - \theta X_i\{X_i - \hat{m}(Z_i)\}}{\frac{1}{n}\sum_i X_i\{X_i - \hat{m}(X_i)\}}$$
$$= \frac{\frac{1}{\sqrt{n}}\sum_i \{Y_i - \theta X_i - \hat{f}(Z_i)\}\{X_i - \hat{m}(Z_i)\}}{\frac{1}{n}\sum_i X_i\{X_i - \hat{m}(X_i)\}}.$$

Recall

- $Y_i = \theta X_i + f(Z_i) + \varepsilon_i$ where $\mathbb{E}(\varepsilon_i \,|\, X_i, Z_i) = 0$,
- $X_i = m(Z_i) + \xi_i$ where $\mathbb{E}(\xi_i \,|\, Z_i) = 0$.

Substituting in, numerator gives an ith summand of the form

$$\{f(Z_i) - \hat{f}(Z_i) + \varepsilon_i\}\{m(Z_i) - \hat{m}(Z_i) + \xi_i\} = \underbrace{\varepsilon_i\xi_i}_{\text{CLT}} + \underbrace{\{f(Z_i) - \hat{f}(Z_i)\}\{m(Z_i) - \hat{m}(Z_i)\}}_{\text{Product of errors}}$$
$$+ \underbrace{\varepsilon_i\{m(Z_i) - \hat{m}(Z_i)\}}_{\text{mean zero}} + \underbrace{\xi_i\{\hat{f}(Z_i) - f(Z_i)\}}_{\text{small?}}.$$

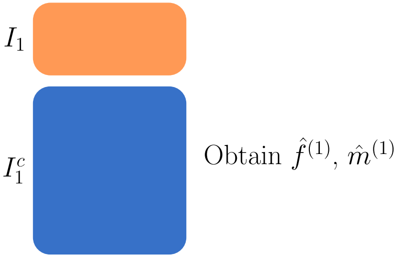
Cross-fitting

If $\hat{f}$ were estimated based on independent data, we would have $\mathbb{E}[\xi_i\{\hat{f}(Z_i) - f(Z_i)\}] = 0$.

Cross-fitting

If $\hat{f}$ were estimated based on independent data, we would have $\mathbb{E}[\xi_i\{\hat{f}(Z_i) - f(Z_i)\}] = 0$.

Cross-fitting aims to mimic this by splitting data into K folds $I_1, \ldots, I_K$:



Choose $\hat{\theta}$ to solve

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

Aside: Reparametrisation

Note that we can reparametrise the estimating equation we obtained in the PLM. We had

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

Noting that $\mathbb{E}(Y \mid Z) = \theta\mathbb{E}(X \mid Z) + f(Z)$, we can write the above as

$$\{X - \mathbb{E}(X \mid Z)\}[Y - \underbrace{\mathbb{E}(Y \mid Z)}_{m_Y(Z)} - \theta\{X - \underbrace{\mathbb{E}(X \mid Z)}_{m_X(Z)}\}].$$

Aside: Reparametrisation

Note that we can reparametrise the estimating equation we obtained in the PLM. We had

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

Noting that $\mathbb{E}(Y \mid Z) = \theta\mathbb{E}(X \mid Z) + f(Z)$, we can write the above as

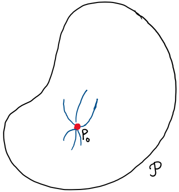
$$\{X - \mathbb{E}(X \mid Z)\}[Y - \underbrace{\mathbb{E}(Y \mid Z)}_{m_Y(Z)} - \theta\{X - \underbrace{\mathbb{E}(X \mid Z)}_{m_X(Z)}\}].$$

This gives the estimator

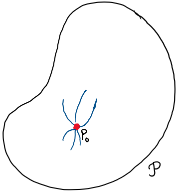
$$\hat{\theta} = \frac{\sum_i \{Y_i - \hat{m}_Y(Z_i)\}\{X_i - \hat{m}_X(Z_i)\}}{\sum_i \{X_i - \hat{m}_X(Z_i)\}^2}.$$

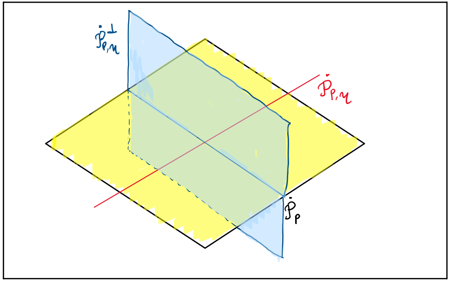
One advantage is that there is no need to estimate f directly: any machine learning method of choice may be used to estimate the regression functions m_Y and m_X.

Summary



Summary



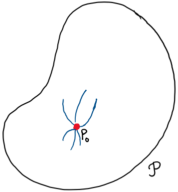


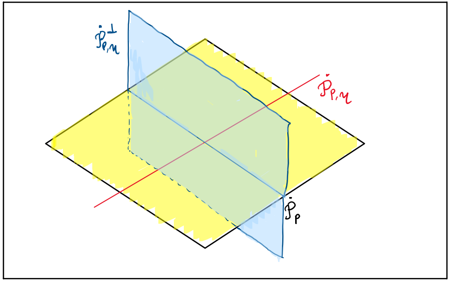
Estimate nuisance parameters $\hat{\eta}$.

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

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

Summary

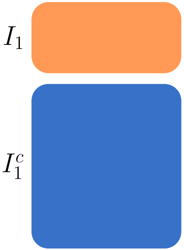




Estimate nuisance parameters $\hat{\eta}$.

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

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



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

Next time...

- In the PLM, we had a choice of estimating equations we could use. Which should we use?
- Optimality: can we achieve the sup of all the C–R lower bounds coming from sub-models?
- Nonparametric models

Thank you for listening.