Introduction to semiparametric theory

Chris Holmes and Andrew Yiu

Department of Statistics, University of Oxford

Leuven, September 2024

DEPARTMENT OF
STATISTICS

Recap: general strategy for causal inference

General strategy

1. Determine a **causal quantity** that will answer the scientific question of interest;
2. Check that the quantity is **identifiable** from the data you have and assumptions you are willing to make;
3. Perform **statistical inference** for the identified quantity based on your data and assumptions.

Yesterday, we explored various concepts for tackling the second objective of *identification*. Our focus today is on the third objective of *inference*. In particular, for this session, we will study approaches based on semiparametric theory.

Objectives

Semiparametric theory has a long history and has garnered a reputation for being difficult to get to grips with at first.

Focus: intuition, motivation, big picture ideas, things that I wish were explained to me when I was a PhD student...

Objectives:

- Provide some introductory details towards demystifying concepts underpinning double-robustness, double machine learning, targeted maximum likelihood estimation etc.
- Make papers on semiparametric inference more accessible (especially the ones relating to causality).
- Understand the challenges of semiparametric Bayesian inference and how we might deal with them.

Frequentist Bayes

In these two lectures, we will work under a frequentist set-up, so a *"model"* is a collection of candidate distributions that could have generated the i.i.d. data $Z_1, \ldots, Z_n$.

This won't stop us from considering Bayesian estimation procedures—it's just that we will be evaluating the performance of our procedures with respect to frequentist metrics like *bias*, *root mean squared error*, *coverage*, etc. And we will study frequentist asymptotic properties like *asymptotic normality* and *efficiency*.

This is sometimes referred to as frequentist Bayes (e.g., by Aad van der Vaart and collaborators).

Frequentist Bayes

Why should we care about frequentist properties if we're Bayesians?

- A frequentist perspective is ubiquitous in causal inference: applications/collaborators will usually demand frequentist guarantees, e.g., can you test a treatment effect away from zero with Type I error control, can you provide confidence intervals with (approximately) nominal coverage?
- Arguably, causal inference offers a more natural interpretation of frequentism than many other fields. This is because we usually think about a *target superpopulation*.
- An assumption like overlap/positivity is inherently frequentist. Testing/handling overlap (or lack thereof) requires thinking from a frequentist point-of-view. (There will be more on this in the practical.)

What does "semiparametric" even mean?

So in the frequentist set-up, the data $Z_1, \ldots, Z_n$ are drawn iid from an unknown distribution belonging to a model $\mathscr{P}$.

Parametric inference

The model $\mathscr{P} = \{P_\theta : \theta \in \Theta \subset \mathbb{R}^d\}$ is indexed by a **finite-dimensional** parameter.
Example: logistic regression.

Nonparametric inference

The model $\mathscr{P} = \{P_\eta : \eta \in \mathscr{H}\}$ and the target estimand are both **infinite-dimensional**.
Example: density estimation with smoothness assumptions.

Semiparametric inference

The model is **infinite-dimensional**, but the target estimand is **finite-dimensional**.

Example 1: linear regression with heteroscedastic errors

Linear regression

Consider

$$Y = T\theta + \beta^{\mathrm{T}} X + U, \quad \mathbb{E}[U \mid T, X] = 0,$$

where Y is the outcome, T is the treatment, and X consists of remaining covariates. The target parameter is the treatment effect θ.

This is an example of a strict semiparametric model. The nuisance parameters are β and the conditional distribution of U given (T, X)—the latter is an infinite-dimensional parameter without further model restrictions.

Example 2: partially linear regression

Partially linear regression

Consider

$$Y = T\theta + g(X) + U, \quad \mathbb{E}[U \mid T, X] = 0,$$

where Y is the outcome, T is the treatment, and X consists of remaining covariates.

This is again a strict semiparametric model. We can partition the parameters into the finite-dimensional target θ and the infinite-dimensional nuisance parameters: $g(x)$ and the law U given (T, X).

For those familiar with survival analysis: many popular models are strict semiparametric models, e.g., proportional hazards, proportional odds, accelerated failure time.

Example 3: integrated squared density

Integrated squared density

Suppose $Z \sim f$, where f is a Lebesgue density. The target parameter is

$$\chi(f) = \int f(z)^2 \, dz = \mathbb{E}_f[f(Z)].$$

This is a semiparametric problem in a more general sense. We have an infinite-dimensional parameter f, and the target is a one-dimensional functional, i.e. a mapping $\chi : \mathscr{P} \to \mathbb{R}$ from the model to the real line.

The previous strict semiparametric problems are a special case of this: $\mathscr{P} = \{P_{\theta,\eta} : \underbrace{\theta \in \Theta}_{\text{target}}, \underbrace{\eta \in \mathscr{H}}_{\text{nuisance}}\}$ and $\chi(P_{\theta,\eta}) = \theta$.

Example 4: average treatment effect

Average treatment effect

Suppose the data takes the form $Z = (X, T, Y)$, where X is a vector of covariates, T is a binary treatment indicator, and Y is the outcome variable of interest.

The target estimand is the ATE $\chi(P) = \mathbb{E}_P[Y^1 - Y^0]$, which is identified by

$$\chi(P) = \mathbb{E}_P[\mathbb{E}_P(Y \mid T = 1, X)] - \mathbb{E}_P[\mathbb{E}_P(Y \mid T = 0, X)]$$

under the assumptions:

- (unconfoundedness) $Y^t \perp\!\!\!\perp T \mid X$ for $t = 0, 1$.
- (overlap/positivity) $0 < \pi(X) < 1$ with P-probability 1, where $\pi(x) = P(T = 1 \mid X = x)$ is the **propensity score**.

What's so special?

You might be thinking: *since we (usually) have to estimate the infinite-dimensional parameters anyways, what makes this different to nonparametric inference?*

Parametric inference

Often straightforward to obtain a $\sqrt{n}$-consistent and asymptotically normal estimator (e.g., MLE, parametric Bayes).

Nonparametric inference

Aside from very special cases (like estimating a CDF), the rate of convergence is always slower than $\sqrt{n}$, and there is very limited asymptotic distribution theory.

Semiparametric inference

Similar asymptotic theory to the parametric case if the functional is "differentiable".

Bridging statistics and machine learning

The main reason semiparametric theory currently gets so much attention in causal inference: use machine learning to estimate nuisance parameters and still obtain valid statistical guarantees for the target estimand!

But...

- Unlike the parametric setting, we can't expect to automatically get good inference for all estimands in one swoop. We need to carefully tailor our estimation towards the target.
- Naïve use of nonparametric algorithms/machine learning can be disastrous
  - Unclear whether the bootstrap is valid (or whether we have asymptotic normality at all).
  - We might not even have a $\sqrt{n}$ convergence rate.

Illustration: integrated squared density

Integrated squared density

Recall: suppose $Z \sim f$, where f is a Lebesgue density. The target parameter is

$$\chi(f) = \int f(z)^2 \, dz = \mathbb{E}_f[f(Z)].$$

How might we go about this in using Bayesian nonparametrics? A popular choice for density estimation is to use Dirichlet process mixture models (DPMM's):

$$f_{G,\sigma}(z) = \int_{\mathbb{R}} \varphi_\sigma(z-\mu) dG(\mu, \sigma), \quad G \sim DP(M, G_0), \quad M > 0,$$

where $\varphi_\sigma(z-\mu)$ is the Gaussian density centred at μ with standard deviation σ, and G_0 is a (conjugate) normal-inverse Gamma distribution.

Simulation study: Laplace density

Suppose that the data are generated i.i.d. from a $\text{Laplace}(0, 1)$ distribution, which has density

$$f_0(z) = \frac{1}{2}\exp(-|z|).$$

For our Dirichlet process Gaussian mixture model, we used the default implementation from the R `dirichletprocess` package.

Numerical results

Sample size	Method	Bias	MAE	RMSE	Cov	Int. len.
$n = 200$	DPMM (plug-in)	-0.0113	0.0157	0.0212	87.7%	0.0685
$n = 400$	DPMM (plug-in)	-0.0100	0.0126	0.0165	84.6%	0.0503

Table: Monte Carlo numerical results across 1000 trials. Point estimation was evaluated using the posterior mean, and intervals were constructed using the 95% central credible region.

Numerical results

Sample size	Method	Bias	MAE	RMSE	Cov	Int. len.
$n = 200$	DPMM (plug-in)	-0.0113	0.0157	0.0212	87.7%	0.0685
$n = 400$	DPMM (plug-in)	-0.0100	0.0126	0.0165	84.6%	0.0503

Table: Monte Carlo numerical results across 1000 trials. Point estimation was evaluated using the posterior mean, and intervals were constructed using the 95% central credible region.

Numerical results

Sample size	Method	Bias	MAE	RMSE	Cov	Int. len.
$n = 200$	DPMM (plug-in)	-0.0113	0.0157	0.0212	87.7%	0.0685
$n = 400$	DPMM (plug-in)	-0.0100	0.0126	0.0165	84.6%	0.0503

Table: Monte Carlo numerical results across 1000 trials. Point estimation was evaluated using the posterior mean, and intervals were constructed using the 95% central credible region.

dpmm_old.png

What went wrong?

The asymptotic theory for DPMM's is well-understood: as we observe more and more data, the posterior will "contract" around the truth f_0. But to achieve a good contraction rate, it introduces bias to stop the variance from blowing up. It does this through oversmoothing.

The bias bleeds into the marginal posterior for $\chi(f)$ as a consequence of Bayesian inference being a "plug-in" method.

This is a general phenomenon for nonparametric statistics.
"A good bias-variance trade-off for the whole infinite-dimensional parameter doesn't necessarily translate into a good trade-off for the low-dimensional target estimand."

Partly for these reasons, Bayesian inference remains relatively unpopular for semiparametric problems and causal inference...

Why Bayes?

Some potential benefits of nonparametric Bayes:

- The prior offers a natural approach for regularization, as well as a route for incorporating expert knowledge to increase precision.
- Bayesian algorithms have been shown to provide competitive or even state-of-the-art predictive performance for many problems, e.g. BART, Gaussian processes.
- A Bayesian model can automatically adapt to unknown regularity/complexity parameters.

We want to take advantage of these attractive features of nonparametric Bayes while still obtaining good inference for our target estimands.

Semiparametric theory to the rescue! We will use the specific structure of the estimand to remove bias and enable rigorous statistical guarantees (e.g. coverage of confidence intervals, Type I error control etc.)

Structure

1. Semiparametric efficiency theory

- Parametric submodels
- Tangent spaces
- Pathwise differentiability
- The efficient influence function

2. Modern applications and bias correction

- The von Mises expansion
- One-step estimation
- Posterior corrections

Some historical context

Semiparametric efficiency theory was first conceived by Charles Stein in 1956.

EFFICIENT NONPARAMETRIC TESTING
AND ESTIMATION

CHARLES STEIN
STANFORD UNIVERSITY

The idea was to take existing efficiency theory for parametric models and generalize to semiparametric problems (though the word *"semiparametric"* wasn't coined until later).

Stein's ideas were formalized and extended in the 70's-90's.[1] The application of semiparametric theory to causal inference was pioneered by Robins and Rotnitzky.

[1] Including work by Levit, Koshevnik, Pfanzagl, Bickel, Begun, Klaassen, Wellner, Ritov, Tsiatis, van der Vaart, Murphy, et al.

Parametric theory

Suppose we have a parametric model $\{P_\theta : \theta \in \Theta\}$ and i.i.d. data $Z_1, \ldots, Z_n \sim P_{\theta_0}$. We are interested in estimating $\chi(P_{\theta_0})$ for some mapping $\chi : \mathscr{P} \to \mathbb{R}$.

Recall:

- Score function $s_\theta(z) = \frac{\partial}{\partial \theta} \log p_\theta(z)$, which has mean zero: $\mathbb{E}_\theta[s_\theta(Z)] = 0$.
- Fisher information $I_\theta = \mathbb{E}_\theta[s_\theta(Z) s_\theta(Z)^{\mathrm{T}}]$.

Cramér-Rao lower bound

Let $\theta \mapsto \chi(P_\theta)$ be differentiable with derivative χ'_θ, and let $\hat{\chi} = \hat{\chi}(Z_{1:n})$ be any unbiased estimator of $\chi(P_\theta)$. Then

$$n \mathrm{var}_{\theta_0}(\hat{\chi}) \geq \chi'_{\theta_0} I_{\theta_0}^{-1} \chi'^{\,\mathrm{T}}_{\theta_0} \quad \text{for all } \theta_0 \in \mathrm{int}(\Theta).$$

Asymptotic efficiency

The Cramér-Rao lower bound informally motivates an asymptotic notion of efficiency.

Asymptotic efficiency

An estimator sequence $\hat{\chi}$ is said to be **asymptotically efficient** for estimating $\chi(P_{\theta_0})$ if

$$\sqrt{n}(\hat{\chi} - \chi(\theta_0)) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \chi'_{\theta_0} I_{\theta_0}^{-1} s_{\theta_0}(Z_i) + o_{P_{\theta_0}}(1).$$

So the best limiting distribution is $\mathcal{N}(0, \chi'_{\theta_0} I_{\theta_0}^{-1} \chi'^{\mathrm{T}}_{\theta_0})$ by the CLT.

We say that $\chi'_{\theta_0} I_{\theta_0}^{-1} s_{\theta_0}(z)$ is the influence function of the estimator $\hat{\chi}$, *"the error $\hat{\chi} - \chi(P_{\theta_0})$ behaves like the sample average of the influence function"*.

Under regularity conditions, the "plug-in" $\hat{\chi} = \chi(P_{\hat{\theta}_{mle}})$ attains asymptotic efficiency.

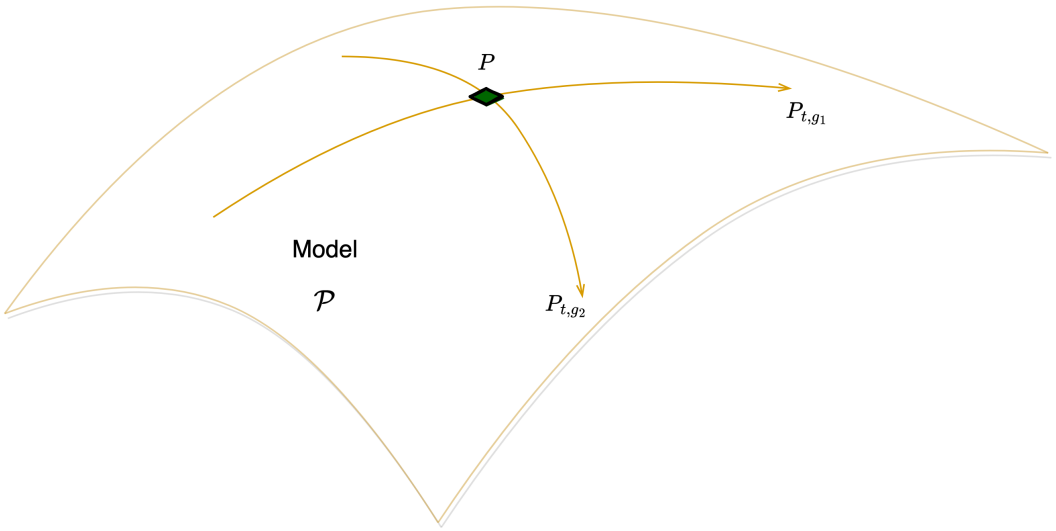
Semiparametric efficiency

Let's return to the more general setting where $\mathscr{P}$ is possibly infinite-dimensional. We want to estimate a functional $\chi : \mathscr{P} \to \mathbb{R}$. How do we define the lower bound at a distribution $P \in \mathscr{P}$?

Suppose $\{P_t\} \subset \mathscr{P}$ is a parametric model that is contained in our model and passes through P. Estimating $\chi(P)$ is at least as hard in $\mathscr{P}$ as it is in $\{P_t\}$.

Idea

Define the semiparametric lower bound to be the greatest Cramér-Rao lower bound across all parametric models that are contained in $\mathscr{P}$ and pass through P.


P
P_{t,g_1}
Model
$\mathcal{P}$
P_{t,g_2}

Parametric submodels

A **parametric submodel** $\{P_{t,g} : t \in (-,)\}$ is a smooth parametric model that passes through P at $t = 0$ with

$$g(z) = \frac{\partial}{\partial t} \log p_{t,g}(z)|_{t=0},$$

i.e. the score function at P is equal to g.

We only care about the value of the score function at $t = 0$ because this is sufficient to compute the C-R lower bound

$$\frac{(\frac{\partial}{\partial t}\chi(P_{t,g})|_{t=0})^2}{\mathbb{E}_P[g(Z)^2]}.$$

Important: a parametric submodel is not substantively meaningful; it's just a technical device that sets up the theoretical framework.

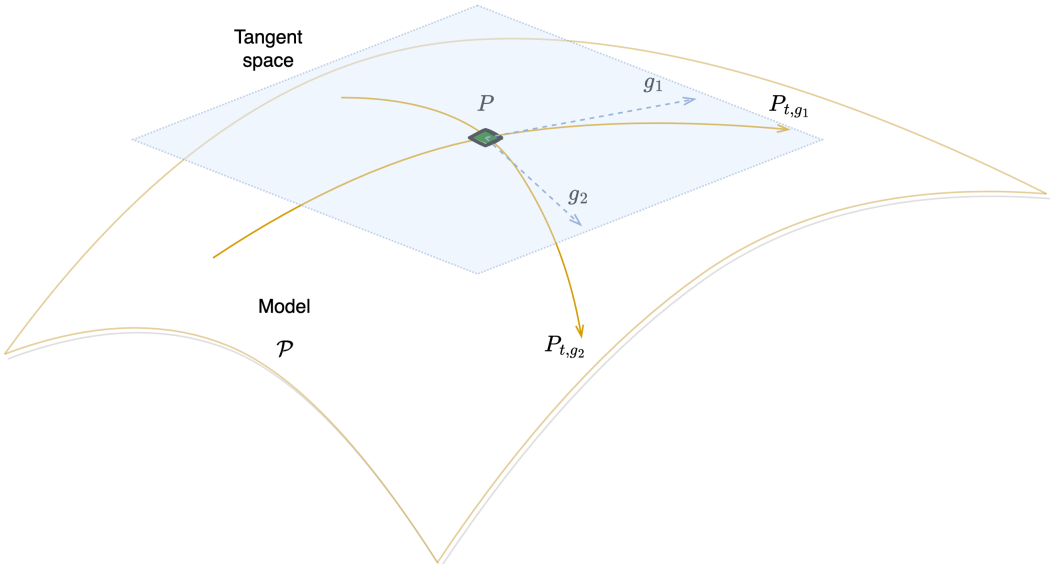
Tangent spaces

Tangent space

The **tangent space** $\dot{\mathscr{P}}_P$ is the collection of all score functions g across the parametric submodels contained in $\mathscr{P}$ that pass through P.

We can interpret a score function as the "direction" in which the submodel passes through P. Then the tangent space is the set of directions in which we can move an infinitesimal distance away from P and still remain in the model.

Thus, the larger the model, the larger the tangent space.


Tangent
space
P
g_1
P_{t,g_1}
g_2
Model
$\mathcal{P}$
P_{t,g_2}

Greatest lower bound

We can now write down our semiparametric lower bound at P:

$$\sup_{g \in \dot{\mathscr{P}}_P} \frac{(\frac{\partial}{\partial t}\chi(P_{t,g})|_{t=0})^2}{\mathbb{E}_P[g(Z)^2]}.$$

Recall that this is the greatest Cramér-Rao lower bound across all parametric submodels contained in $\mathscr{P}$ that pass through P.

Clearly, we need the mapping $t \mapsto \chi(P_{t,g})$ to be differentiable at $t = 0$ for any parametric submodel. But this is not sufficient; we need a stronger form of differentiability...

Pathwise differentiability

A functional χ is **(pathwise) differentiable** at P with respect to $\dot{\mathscr{P}}_P$ if:

(a) the mapping $t \mapsto \chi(P_{t,g})$ is differentiable at $t = 0$, and

(b) there exists a fixed function $\phi_P : \mathscr{Z} \to \mathbb{R}$ such that

$$\frac{\partial \chi(P_{t,g})}{\partial t}|_{t=0} = \mathbb{E}_P[\phi_P g]$$

for every $g \in \dot{\mathscr{P}}_P$ and any parametric submodel $\{P_{t,g}\}$ with score function g. We call ϕ_P a **gradient** of χ at P.

So not only do we need differentiability of $t \mapsto \chi(P_{t,g})$ in the ordinary sense, but we also require a special representation of the derivative.

The canonical gradient

So we need

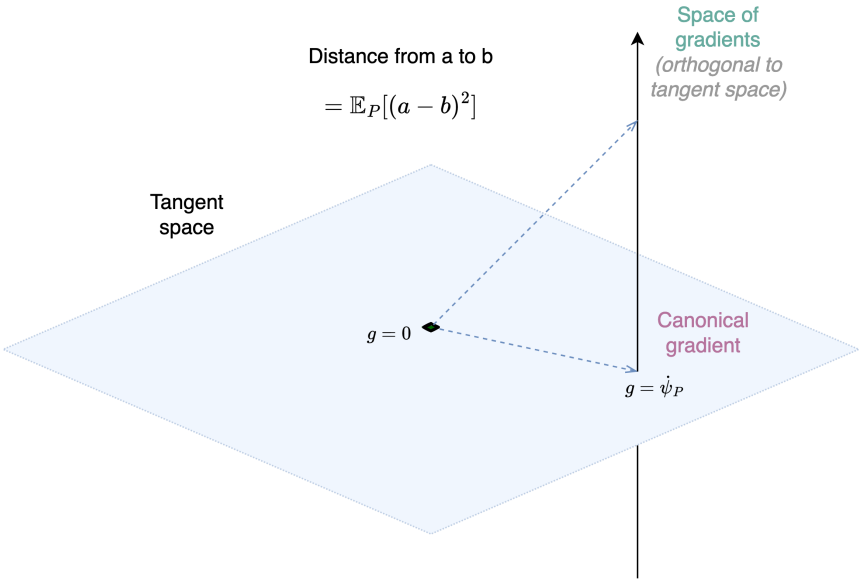
$$\frac{\partial \chi(P_{t,g})}{\partial t}|_{t=0} = \mathbb{E}_P[\phi_P g].$$

The gradient ϕ_P is not unique; we can always replace ϕ_P by $\phi_P + h$, where h satisfies $\mathbb{E}_P[hg] = 0$ for all $g \in \dot{\mathscr{P}}_P$ (i.e. h is "orthogonal" to the tangent space).

Canonical gradient/Efficient influence function

There is a unique gradient that has minimum variance amongst all mean-zero gradients. We call it the **canonical gradient** or **efficient influence function**, denoted by $\dot{\chi}_P$. It is the *"$L_2(P)$-projection"* of any gradient onto the tangent space, i.e.

$$\dot{\chi}_P = \underset{g \in \dot{\mathscr{P}}_P}{\operatorname{argmin}} \mathbb{E}_P[(\phi_P - g)^2].$$


Space of gradients
(orthogonal to tangent space)
Distance from a to b
$= \mathbb{E}_P[(a-b)^2]$
Tangent space
$g = 0$
Canonical gradient
$g = \dot{\psi}_P$

Variance of the canonical gradient

Returning to the lower bound:

$$\sup_{g \in \dot{\mathscr{P}}_P} \frac{(\frac{\partial}{\partial t}\chi(P_{t,g})|_{t=0})^2}{\mathbb{E}_P[g(Z)^2]} = \sup_{g \in \dot{\mathscr{P}}_P} \frac{\mathbb{E}_P[\dot{\chi}_P(Z)g(Z)]^2}{\mathbb{E}_P[g(Z)^2]} \leq \mathbb{E}_P[\dot{\chi}_P(Z)^2]$$

by the Cauchy-Schwarz inequality.

But by definition, $\dot{\chi}_P$ lies within the tangent space $\dot{\mathscr{P}}_P$, so we also have the reverse inequality by taking $g = \dot{\chi}_P$, i.e.

$$\mathbb{E}_P[\dot{\chi}_P(Z)^2] \leq \sup_{g \in \dot{\mathscr{P}}_P} \frac{\mathbb{E}_P[\dot{\chi}_P(Z)g(Z)]^2}{\mathbb{E}_P[g(Z)^2]}.$$

Thus, the variance lower bound is the variance of the canonical gradient.

Asymptotic efficiency

Semiparametric efficiency

An estimator sequence $\hat{\chi}$ is said to be **asymptotically efficient** for estimating $\chi(P_0)$ if

$$\sqrt{n}(\hat{\chi} - \chi(P_0)) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \dot{\chi}_{P_0}(Z_i) + o_P(1).$$

So the best limiting distribution is $\mathcal{N}(0, \mathrm{var}_{P_0}[\dot{\chi}_{P_0}])$ by the CLT.

This is why $\dot{\chi}_{P_0}$ is called the efficient influence function at P_0.

And we call $\mathrm{var}_{P_0}[\dot{\chi}_{P_0}]$ the *"efficient variance"* or *"efficiency bound"* for the estimation problem.

Example 4: average treatment effect

Average treatment effect

Suppose the data takes the form $Z = (X, T, Y)$, where X is a vector of covariates, T is a binary treatment indicator, and Y is the outcome variable of interest.

The target estimand is $\chi(P) = \mathbb{E}_P[\mathbb{E}_P(Y \mid T = 1, X)] - \mathbb{E}_P[\mathbb{E}_P(Y \mid T = 0, X)]$.

Denote:

- $\pi(x) = P(T = 1 \mid X = x)$ (propensity score)
- $\mu^t(x) = \mathbb{E}[Y \mid T = t, X = x]$.

The efficient influence function is

$$\dot{\chi}_P(Z) = \frac{T(Y - \mu^1(X))}{\pi(X)} - \frac{(1-T)(Y - \mu^0(X))}{1 - \pi(X)} + \mu^1(X) - \mu^0(X) - \chi(P).$$

Note the propensity score in the denominators: this is called inverse probability weighting.

Summary of semiparametric efficiency theory

- In parametric theory, we quantify the hardness of estimating a parameter at a point using the *Cramér-Rao lower bound*.
- For the infinite-dimensional case, we can generalize this by taking the supremum of the lower bounds across all *parametric submodels* that pass through the distribution.
- We need the functional to be *"pathwise differentiable"*, which means that there exists a pathwise derivative (a *"gradient"*) that takes the score function as input and then outputs the corresponding slope in χ.
- The unique gradient that lies in the tangent space is called the *"canonical gradient"* or *"efficient influence function"*. Its variance is equal to the semiparametric Cramér-Rao lower bound.

Next steps: how do we use this theory to help us construct good estimation procedures?

Some notation

Following convention, we use the shorthand (operator) notation

$$P[h] = \int h(z)\, dP(z).$$

This is different to taking the expectation when we have a random function $\hat{h}$ (e.g. an estimator $\hat{h}$ constructed from the data).

An important case is the empirical distribution $\mathbb{P}_n = n^{-1} \sum_{i=1}^{n} \delta_{Z_i}$, where we have

$$\mathbb{P}_n[h] = n^{-1} \sum_{i=1}^{n} h(Z_i).$$

Estimation problem

Suppose we have observed iid data $Z_1, \ldots, Z_n$ from an unknown distribution P_0 belonging to a model $\mathscr{P}$, and the target estimand is $\chi(P_0)$, where $\chi : \mathscr{P} \to \mathbb{R}$ is a functional.

The obvious thing to do is plug-in estimation, i.e.

(i) Construct a point estimate $\hat{P}$ and then use $\chi(\hat{P})$ to estimate $\chi(P_0)$, or;

(ii) Construct a posterior for P and then use the marginal posterior for $\chi(P)$ to perform inference.

But we saw earlier that this can be problematic when we use nonparametric procedures: we get biased inference and interval estimates fail to attain nominal coverage.

We will start with the non-Bayesian case (i) and then return to the Bayesian case (ii) later.

The von Mises expansion

So the problem is that the plug-in error $\chi(\hat{P}) - \chi(P_0)$ might be large. The idea is that we will use semiparametric theory to help us approximate this error and remove it from the plug-in estimate to (hopefully) obtain a better estimator.

Earlier, we discussed how the efficient influence function $\dot{\chi}_P$ could be interpreted as a kind of *first-order derivative* of the functional. So we want to use it to construct a first-order (i.e. linear) approxmation to the estimation error:

$$\chi(\hat{P}) - \chi(P_0) = (\hat{P} - P_0)[\dot{\chi}_{\hat{P}}] - r_2(P_0, \hat{P}).$$

This looks a bit like a Taylor expansion—in fact, it's called a von Mises expansion. The error $r_2(P_0, \hat{P})$ that is left over is called the second-order remainder/bias.

The one-step estimator

Let's try to use the von Mises expansion to help us out. With a bit of rearranging, we get:

$$\chi(P_0) = \chi(\hat{P}) + P_0[\dot{\chi}_{\hat{P}}] + r_2(P_0, \hat{P}).$$

We hope that r_2 is "small" relative to the initial plug-in estimation error (more on this later).

The above suggests that $\chi(\hat{P}) + P_0[\dot{\chi}_{\hat{P}}]$ would improve on the plug-in, but of course, P_0 is unknown. So we replace P_0 with the empirical distribution $\mathbb{P}_n$.

One-step estimator

The **one-step estimator** is defined as

$$\hat{\chi}_{\text{1-step}} = \chi(\hat{P}) + \mathbb{P}_n[\dot{\chi}_{\hat{P}}].$$

Asymptotic efficiency revisited

As we discussed earlier, the gold-standard asymptotic result is to establish

$$\sqrt{n}(\hat{\chi} - \chi(P_0)) \rightsquigarrow \mathcal{N}(0, V).$$

for some point estimator $\hat{\chi}$. The best possible variance V in the semiparametric case is $V = \mathrm{var}_{P_0}(\dot{\chi}_{P_0})$.

We want to

- Provide sufficient conditions for this to hold for our one-step estimator $\hat{\chi}_{\text{1-step}}$.
- Be able to estimate V so that we can construct confidence intervals.

The one-step estimator

First we study sufficient conditions for the one-step estimator to be asymptotically efficient.

$$\begin{aligned}\sqrt{n}(\hat{\chi}_{\text{1-step}} - \chi(P_0)) &= \sqrt{n}(\chi(\hat{P}) + \mathbb{P}_n[\dot{\chi}_{\hat{P}}] - \chi(P_0)) \\ &= \underbrace{\sqrt{n}\mathbb{P}_n[\dot{\chi}_{\hat{P}}]}_{(1)} + \underbrace{\sqrt{n}(\mathbb{P}_n - P_0)[\dot{\chi}_{\hat{P}} - \dot{\chi}_{P_0}]}_{(2)} - \underbrace{\sqrt{n}r_2(P_0, \hat{P})}_{(3)}\end{aligned}$$

after a bit of manipulation.

- Term (1): this is exactly what we want—it converges to the efficient normal limiting distribution $\mathcal{N}(0, \text{var}_{P_0}(\dot{\chi}_{P_0}))$ by the central limit theorem.
- So we just need terms (2) and (3) to be negligible, i.e. $o_{P_0}(1)$.

Second-order bias

For term (3) to be $o_{P_0}(1)$, we need the second-order bias $r_2(P_0, \hat{P})$ to converge to zero at a rate faster than $1/\sqrt{n}$.

We started with the intuition that r_2 should be in some sense quadratic in the error between $\hat{P}$ and P_0. In the extra slides, we confirm that this intuition is in fact implied by the definition of pathwise differentiability.

The form of r_2 needs to be checked on a case-by-case basis for each functional of interest.

Integrated squared density revisited

Consider again the integrated squared density $\chi(f) = \int f(z)^2 \, dz$. The efficient influence function is $\dot{\chi}_f(Z) = 2\{f(Z) - \chi(f)\}$, and it is straightforward to show that

$$r_2(f_0, \hat{f}) = \|\hat{f} - f_0\|_2^2 = \int (\hat{f}(z) - f_0(z))^2 \, dz,$$

where $\|\cdot\|_2$ denotes the L^2 distance wrt to the Lebesgue measure.

So we need $\|\hat{f} - f_0\|_2 = o_{P_0}(n^{-1/4})$.

ATE example revisited

For the ATE example, we need an estimate $\hat{\pi}$ of the propensity score, and estimates $\hat{\mu}^t$ of the outcome regression functions.

Under some standard assumptions, is possible to show that

$$r_2(P_0, \hat{P}) \leq \|\hat{\pi} - \pi_0\| \left(\|\hat{\mu}^1 - \mu_0^1\| + \|\hat{\mu}^0 - \mu_0^0\| \right).$$

up to a constant, where $\|\cdot\|$ is the $L^2(P_0)$ norm.

So r_2 depends on the product of the errors for $\hat{\pi}$ and $\hat{\mu}^t$. In particular, it's sufficient if both $\hat{\pi}$ and $\hat{\mu}^t$ converge faster than $n^{-1/4}$.

But one of them can converge slower if it's compensated by the other, e.g. if we estimate π_0 using a correctly specified logistic regression model, then $\hat{\mu}^t$ is permitted to converge arbitrarily slowly. This is called double-robustness.

The empirical process term

Let's now look at term (2)

$$\sqrt{n}(\mathbb{P}_n - P_0)[\dot{\chi}_{\hat{P}} - \dot{\chi}_{P_0}].$$

The stochastic process $\mathbb{G}_n = \sqrt{n}(\mathbb{P}_n - P_0)$ is called the empirical process, so term (2) is often called the "empirical process term".

There are two sources of error in term (2):

- The empirical distribution $\mathbb{P}_n$ approximates P_0.
- The estimate $\dot{\chi}_{\hat{P}}$ of the efficient influence function $\dot{\chi}_{P_0}$.

Both $\mathbb{P}_n$ and $\dot{\chi}_{\hat{P}}$ depend on the data—we're kind of using the data twice to construct two different estimates $\mathbb{P}_n$ and $\hat{P}$ of the data-generating distribution P_0.

Roughly speaking, if there is a strong association between them, then term (2) might not be negligible. The literature calls this overfitting bias.

Handling the empirical process term

There are two options for handling the empirical process term:

1. We must ensure that our initial estimate $\hat{P}$ is *"not too complex"*, such that we can get away with double-dipping the data without inducing too much overfitting bias.

   A more precise statement requires *empirical process theory*, which is beyond the scope of our course. For example, in the literature, you may come across authors talking about Donsker classes.

2. We can avoid complexity conditions by implementing sample-splitting and cross-fitting. Sample-splitting means that we split our data, fitting $\dot{\chi}_{\hat{P}}$ on one fold of the data and then fitting $\mathbb{P}_n$ on the remaining data.

   This incurs a loss of efficiency because of the reduction in sample size, but the efficiency can be regained by swapping the roles of the splits and combining the estimators. More details can be found in the extra slides.

Variance estimation

Suppose the previous conditions are all satisfied, so $\hat{\chi}_{\text{1-step}}$ is asymptotically efficient. We still need to estimate the asymptotic variance $V = \text{var}_{P_0}(\dot{\chi}_{P_0})$.

Remarkably, we can do this by just evaluating the sample variance of the estimated efficient influence function:

$$\hat{V} = \frac{1}{n} \sum_{i=1}^{n} \left(\dot{\chi}_{\hat{P}}(Z_i) - \left[\sum_{j=1}^{n} \dot{\chi}_{\hat{P}}(Z_j) \right] \right)^2.$$

So we don't need to take into account the uncertainty in $\hat{P}$! This is invaluable because we'd like to use black-box machine learning algorithms, and they may not have well-understood/well-behaved distributional properties.

Also, we can't bootstrap the one-step estimator in general because the bootstrap can fail for flexible, data-adaptive procedures.

Summary

- The efficient influence function can be interpreted as a kind of first-order distributional derivative. In particular, we can use it as part of the linear term in a first-order "von-Mises" expansion.
- This expansion gives us a way of debiasing smoothed nonparametric estimators to obtain efficient estimators of our target functional. We can implement this using **one-step estimators**, though there are other methods such as *"double machine learning"* and *"targeted maximum likelihood estimation"*.

Now we will proceed to introduce a method for correcting Bayesian posteriors. This is based on our work in `arXiv:2306.06059`.

The one-step posterior

For our proposal, the starting point is any Bayesian posterior $\Pi(\cdot \mid Z_{1:n})$—we look to correct an entire probability measure, rather than just a point estimator.

Comparing with the one-step estimator, we replace the initial estimator $\hat{P}$ with the parameter P distributed according to the posterior:

$$\hat{P} \longmapsto P \sim \Pi(\cdot \mid Z_{1:n}).$$

Then, the empirical distribution is replaced by the Bayesian bootstrap posterior:

$$\mathbb{P}_n \longmapsto \tilde{P} \sim \Pi_{BB}(\cdot \mid Z_{1:n}) = \sum_{i=1}^{n} W_i \delta_{Z_i}, \quad (W_1, \ldots, W_n) \sim \mathrm{Dir}(n; 1, \ldots, 1)$$

One-step corrected parameter

The **one-step corrected parameter** is defined by

$$\tilde{\chi}(P, \tilde{P}) = \chi(P) + \tilde{P}[\dot{\chi}_P]$$

Correction algorithm

1. Run your posterior computation algorithm of choice to obtain a sample $\{P^{(1)}, \ldots, P^{(B)}\}$ from the initial posterior $\Pi(P \mid Z_{1:n})$.
2. Pair each posterior sample $P^{(b)}$ of P with an independent draw of $(W_1^{(b)}, \ldots, W_n^{(b)})$ from $\mathrm{Dir}(n; 1, \ldots, 1)$.
3. Compute
$$\tilde{\chi}^{(b)} = \chi(P^{(b)}) + \sum_{i=1}^{n} W_i^{(b)} \dot{\chi}_{P^{(b)}}(Z_i)$$
for every posterior sample.
4. Output $(\tilde{\chi}^{(1)}, \ldots, \tilde{\chi}^{(B)})$, which is a sample from the one-step posterior.

Geometric interpretation

Let's try to build some geometric intuition for the one-step posterior correction.[2]

Consider the following mixture distributions constructed from an arbitrary $P \in \mathscr{P}$ (representing a draw from our initial posterior) and the true distribution P_0:

$$P^\varepsilon = (1-\varepsilon)P + \varepsilon P_0$$

for $0 \leq \varepsilon \leq 1$. Within our model space $\mathscr{P}$, we could interpret $\{P^\varepsilon\}_{\varepsilon \in [0,1]}$ as a line segment linking P and P_0.

This induces a path $\{\chi(P^\varepsilon)\}_{\varepsilon \in [0,1]}$ in the parameter space that reaches the true parameter value $\chi(P_0)$ at $\varepsilon = 1$. Starting at $\varepsilon = 0$, we might hope to get closer to the truth by constructing a good approximation to this path.

[2] Based on an idea from Fisher & Kennedy (2021). Visually communicating and teaching intuition for influence functions. *The American Statistician*, 70:162–172.

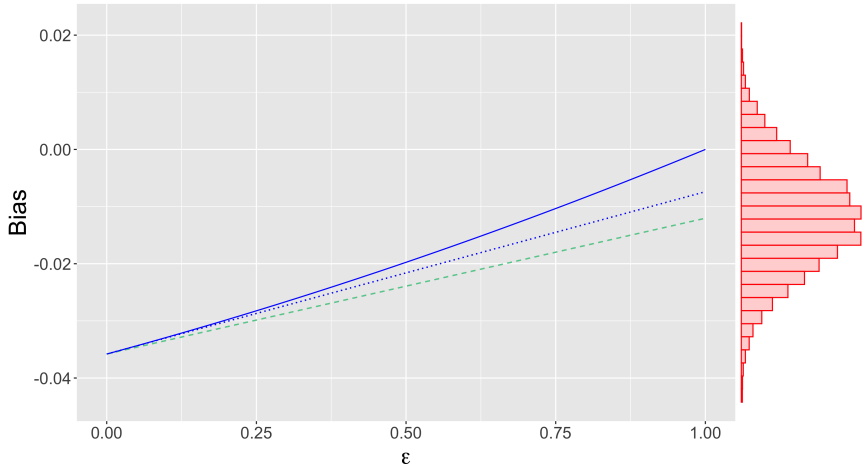


Figure: The blue solid curve () is $\chi(P^{\varepsilon}) - \chi(P_0)$, whose tangent at $\varepsilon = 0$ is given by the blue dotted line (). The green dashed line () estimates the tangent with the empirical distribution, and the histogram (red) on the right corresponds to the distribution of the tangent drawn from the Bayesian bootstrap, representing the uncertainty in the estimand conditional on the starting point P.

Geometric interpretation

To summarize, the one-step posterior combines the uncertainty from two unknown quantities.

First, we have the uncertainty in the functional $\chi(P)$, which arises from our initial posterior.

Given a "starting point" P, we also have uncertainty in $P_0[\dot{\chi}_P]$, which is the slope of the tangent to the curve connecting $\chi(P)$ to the truth $\chi(P_0)$. This latter source of uncertainty is modelled by replacing the unknown P_0 with the Bayesian bootstrap parameter $\tilde{P}$.

Bayesian asymptotics

Now we touch briefly on asymptotic theory. Recall that in the non-Bayesian case, we want

$$\sqrt{n}(\hat{\chi} - \chi(P_0)) \rightsquigarrow \mathcal{N}(0, V).$$

For the one-step posterior we want instead

$$\sqrt{n}(\tilde{\chi} - \hat{\chi}) \mid Z_{1:n} \rightsquigarrow \mathcal{N}(0, V)$$

for any efficient estimator $\hat{\chi}$.

Roughly speaking, this states that the one-step posterior is asymptotically normal, centred at an efficient estimator, and has asymptotic variance equal to $V = P_0[\dot{\chi}^2_{P_0}]$.

This is called a semiparametric Bernstein-von Mises theorem (see extra slides for further details).

Conditions for BvM

If the Bernstein-von Mises theorem holds:

- The one-step posterior contracts to the truth at the optimal rate.
- Central credible intervals will have approximately nominal coverage, e.g. the posterior probability region between the 2.5% to 97.5% quantiles will be approx. a 95% confidence set.

In the paper, we provide some sufficient conditions for this to occur:

- Our initial posterior for P must contract sufficiently quickly around P_0 (cf. convergence rates in point estimation theory and double-robustness). The usual threshold is $n^{-1/4}$, similar to the non-Bayesian theory.
- The posterior must also put almost all of its probability mass on a subspace of the model that is not too big/complex (needs to be studied using empirical process theory).

Simulation study: Laplace density

Let's return to the integrated squared density functional $\chi(f) = \int f^2(z)\,dz$ from earlier.

Suppose that the data are generated i.i.d. from a $\mathrm{Laplace}(0, 1)$ distribution, which has density

$$f_0(z) = \frac{1}{2}\exp(-|z|).$$

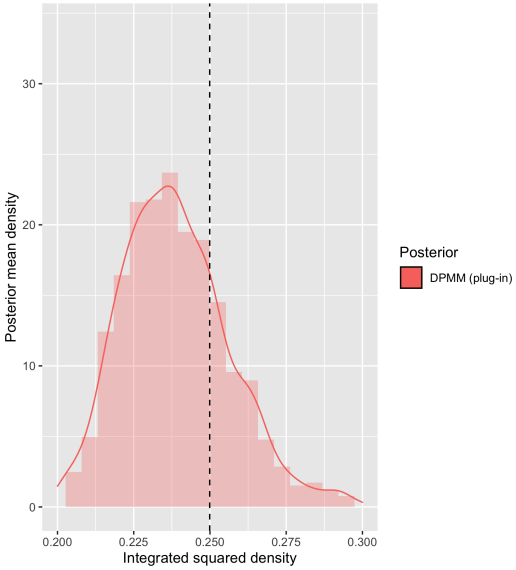
For our Dirichlet process Gaussian mixture model, we used the default implementation from the R `dirichletprocess` package.

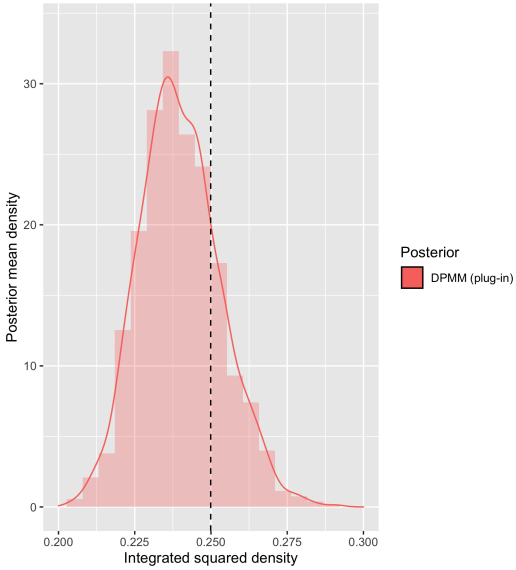
Integrated squared density revisited

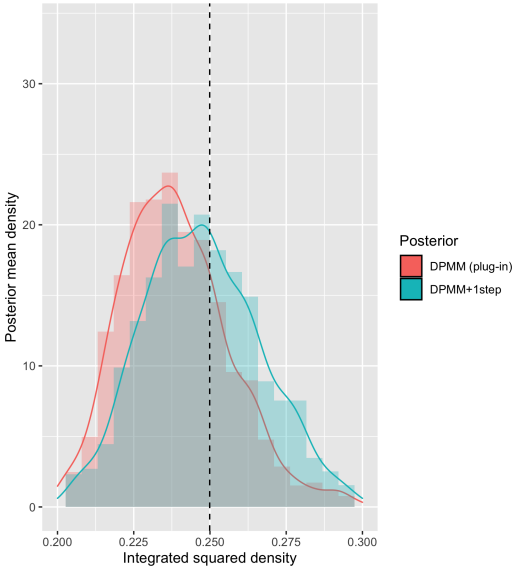
The efficient influence function is $\dot{\chi}_f(Z) = 2\{f(Z) - \chi(f)\}$. So the one-step corrected parameter is

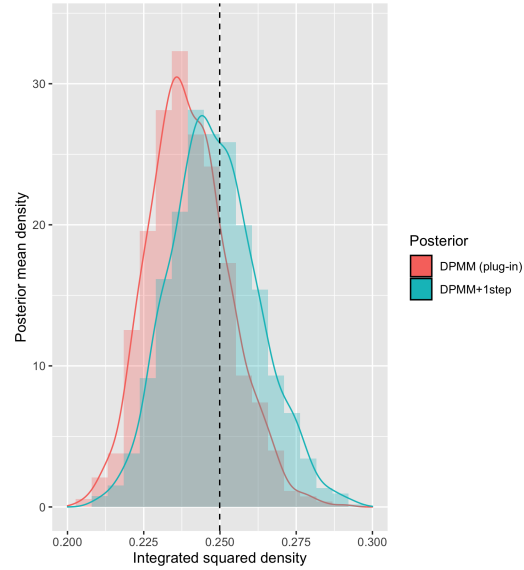
$$\tilde{\chi} = 2\tilde{P}[f(Z)] - \chi(f).$$

1. Sample $\{f^{(1)}, \ldots, f^{(B)}\}$ from our DPMM posterior.
2. Pair each posterior sample $f^{(b)}$ with an independent draw of $(W_1^{(b)}, \ldots, W_n^{(b)})$ from $\text{Dir}(1, \ldots, 1)$.
3. Compute $\tilde{\chi}^{(b)} = 2\sum_{i=1}^n W_i^{(b)} f^{(b)}(Z_i) - \chi(f^{(b)})$ for every posterior sample.
4. Output $(\tilde{\chi}^{(1)}, \ldots, \tilde{\chi}^{(B)})$.


Posterior mean density
30
20
10
0
0.200
0.225
0.250
0.275
0.300
Integrated squared density
Posterior
DPMM (plug-in)


Posterior mean density
30
20
10
0
0.200
0.225
0.250
0.275
0.300
Integrated squared density
Posterior
DPMM (plug-in)


Posterior mean density
30
20
10
0
0.200
0.225
0.250
0.275
0.300
Integrated squared density
Posterior
DPMM (plug-in)
DPMM+1step


Posterior mean density
30
20
10
0
0.200
0.225
0.250
0.275
0.300
Integrated squared density
Posterior
DPMM (plug-in)
DPMM+1step

Numerical results with correction

Sample size	Method	Bias	MAE	RMSE	Cov	Int. len.
$n = 200$	DPMM (plug-in)	-0.0113	0.0157	0.0212	87.7%	0.0685
	DPMM+1step	-0.0015	0.0141	0.0205	97.0%	0.0903
$n = 400$	DPMM (plug-in)	-0.0100	0.0126	0.0165	84.6%	0.0503
	DPMM+1step	-0.0016	0.0099	0.0145	97.0%	0.0655

Table: Monte Carlo numerical results comparing the integrated squared density posteriors with and without correction.

Conclusions

- Nonparametric Bayesian procedures for low-dimensional estimands will generally suffer from bias and poor coverage.
- We have introduced a simple posterior correction that not only debiases the posterior but also corrects its shape. The structure of the correction is informed by semiparametric efficiency theory.
- The method is computationally efficient and attaches onto any existing posterior sampling implementation without the need to modify the original algorithm.
- Preprint: *"Semiparametric posterior corrections"*
  `https://arxiv.org/abs/2306.06059`

Accessible introductions

Kennedy (2022) "Semiparametric doubly robust..."

Nice general introduction on functional estimation, leaning towards causal problems.

Hines et al. (2022) "Demystifying statistical learning..."

Lots of examples on deriving efficient influence functions.

Newey (1990) "Semiparametric efficiency bounds"

Readable introduction to semiparametric efficiency theory.

More in-depth

van der Vaart (2002) "Semiparametric statistics"

Detailed overview of semiparametric efficiency theory with some basic tools from empirical process theory. Includes material on strict semiparametric models, survival analysis, mixture models, and nonparametric maximum likelihood.

Tsiatis (2006) "Semiparametric theory and missing data"

Lots of geometric intuition provided, with a focus on missing and coarsened data problems.

Unfortunately, there's very little on semiparametric Bayesian inference at the moment!

Thanks for listening! Any questions?

The von Mises expansion

Let's take another look at pathwise differentiability. Recall that we need

$$\frac{\partial \chi(P_{t,g})}{\partial t}|_{t=0} = P[\dot{\chi}_P g]$$

for every $g \in \dot{\mathscr{P}}_P$ and any parametric submodel $\{P_{t,g}\}$ with score function g.

Suppose we try to form a distributional Taylor expansion (von Mises expansion):

$$\chi(P_{t,g}) - \chi(P) = (P_{t,g} - P)[\dot{\chi}_P] + r_2(P_{t,g}, P),$$

where $r_2(P_{t,g}, P)$ is simply the LHS minus the first term on the RHS.

If this is a real Taylor expansion, then we need $\lim_{t \downarrow 0} \frac{r_2(P_{t,g}, P)}{t} = 0$.

The von Mises expansion

For the first term on the right-hand side, consider

$$\begin{aligned}
\frac{\partial}{\partial t}(P_{t,g} - P)[\dot{\chi}_P]|_{t=0} &= \frac{\partial}{\partial t} P_{t,g}[\dot{\chi}_P]|_{t=0} \qquad (\dot{\chi}_P \text{ has mean zero under } P) \\
&= \int \dot{\chi}_P \frac{\partial}{\partial t} p_{t,g}|_{t=0} \qquad (\text{bring } \frac{\partial}{\partial t} \text{ inside the integral}) \\
&= \int \dot{\chi}_P \underbrace{\left\{ \frac{\partial}{\partial t} \log p_{t,g} \right\}}_{\text{score at } t} p_{t,g}|_{t=0} \\
&= P[\dot{\chi}_P g].
\end{aligned}$$

This is the "RHS" of the pathwise differentiability condition.

The von Mises expansion

Returning to the expansion, we now have

$$\underbrace{\frac{\chi(P_{t,g})-\chi(P)}{t}}_{\to \frac{\partial \chi(P_{t,g})}{\partial t}|_{t=0}} = \underbrace{\frac{(P_{t,g}-P)[\dot{\chi}_P]}{t}}_{\to P[\dot{\chi}_P g]} + \frac{r_2(P_{t,g},P)}{t},$$

So pathwise differentiability is exactly saying that

$$\lim_{t\downarrow 0} \frac{r_2(P_{t,g},P)}{t} = \frac{\partial}{\partial t} r_2(P_{t,g},P)|_{t=0} = 0.$$

So $r_2(P_{t,g},P)$ can indeed be interpreted like a second-order remainder. This is the essence of concepts like double robustness and Neyman orthogonality.

Cross-fitting and Donsker classes

There's a simple way to break the association between $\mathbb{P}_n$ and $\hat{P}$—fit them on separate splits of the data! Sample-splitting reduces our sample size—we can recover efficiency by cross-fitting.

If we swap the roles of the splits, we can obtain a second estimator $\hat{\chi}^{(2)}$. Then we take

$$\hat{\chi} = \frac{1}{2}(\hat{\chi}^{(1)} + \hat{\chi}^{(2)}).$$

Open question: This can be generalized to more than two splits. What is the optimal number of splits? Some authors recommend using most of the data to estimate $\hat{P}$.

We might still be ok without sample-splitting and cross-fitting if our class of estimators $\dot{\chi}_{\hat{P}}$ is not too complex. This is framed in terms of empirical process theory, e.g. $\dot{\chi}_{\hat{P}}$ lies in a Donsker class.

Double machine learning

Double machine learning takes a slightly different approach to one-step estimation. It basically involves using an influence function to form a set of estimating equations.

Suppose we have a "score" function $\varphi(Z; \chi, \eta)$ satisfying

$$\mathbb{E}_P[\varphi(Z; \chi, \eta)] = 0,$$

where χ is the target parameter, and η is the nuisance parameter.

First we construct an estimate $\hat{\eta}$ (using machine learning perhaps), and then we plug-and-solve for χ:

$$\mathbb{P}_n[\varphi(Z; \chi, \hat{\eta})] = 0.$$

Neyman orthogonality

The score function is required to satisfy Neyman orthogonality, which means that the Gateaux derivative of the expected score is zero with respect to η.

Neyman orthogonality

For any $\tilde{\eta}$,

$$\lim_{t \downarrow 0} \frac{\mathbb{E}_P[\varphi(Z; \chi, \eta + t(\tilde{\eta} - \eta)]}{t} = 0.$$

Interpretation: the expectation of the score should be insensitive to changes in η.

In the statistics literature, this is called the "no-bias condition".[3]

[3] See Klaassen (1987), van der Vaart (1998), Murphy and van der Vaart (2000).

Semiparametric Bernstein-von Mises theorem

Definition

Let $\mathscr{L}_{\Pi \times \Pi_{BB}}(\sqrt{n}(\tilde{\chi} - \hat{\chi}_n) \mid Z^{(n)})$ denote the posterior law of $\sqrt{n}(\tilde{\chi} - \hat{\chi}_n)$, where $\hat{\chi}_n$ is any sequence of asymptotically efficient estimators.

Semiparametric Bernstein-von Mises theorem

We say that the one-step posterior satisfies the **semiparametric Bernstein-von Mises theorem** if $\mathscr{L}_{\Pi \times \Pi_{BB}}(\sqrt{n}(\tilde{\chi} - \hat{\chi}_n) \mid Z^{(n)})$ converges conditionally in distribution to $\mathscr{N}(0, P_0[\dot{\chi}^2_{P_0}])$ in probability.