Chapter 1. Introduction

References for the section on semiparametric theory

1. Van der Vaart. (2000) *Asymptotic Statistics*. (this is a very complete and rigorous but hard to read book on asymptotics, which has one chapter, Chap 25 on semiparametric theory).

2. Tsiatis, A. (2006) *Semiparametric Theory and Missing Data*. (the first half of the book gives an accessible non-super technical introduction to semiparametric theory. Our treatment of semiparametric theory will be at a technical level somewhat in between the books of Tsiatis and van der vaart).

3. Newey, W. (1990) Semiparametric efficiency bounds. *Journal of Applied Econometrics*, vol 5, 99-135. (this is a GREAT introductory paper on semiparametric theory).

References for the section on semiparametric theory

4. Van der Vaart. (2002) Semiparametric Statistics in *Lectures on Probability and Statistics*, Ecole d'Ete de Probabilites de Saint Flour XXiX -1999. (this is a monograph with material expanded a bit more than Ch 25 of the asymptotic statistics book).

5. Bickel, Klaassen, Ritov and Wellner. (1993) *Efficient and adaptive inference in semiparametric models*. (this book provides a rigourous treatment of semiparametric theory but it is hard to read).

6. Kosorok, M. (2008) *Introduction to Empirical Processes and Semiparametric Inference*. Springer Series in Statistics. (this book covers essentially the same material as van der Vaart, 2000, at a slightly lower level, and often referrs to that book for proofs).

References for the section on semiparametric theory

7. van der Laan, M. and Robins, J. (2003) *Unified methods for censored and longitudinal data*. (this book starts with an intro of semiparametric Theory but then focus on inference in semiparametric models with missing and coarsened at random data. The book is a bit chaotic and disorganized).

8. Luenberger, D. G. (1969) *Optimization by Vector Space Methods*. Wiley, New York. (this is a fabulously clear book that contains all that you need to know about Hilbert space theory for this course).

Introduction to semiparametric models

- Semiparametric models.
- Examples.
- Questions of interest.

Recommended readings: Tsiatis, ch 1

Modern epidemiological and clinical studies routinely collect, on each of n subjects, high dimensional data (often comprised of many baseline and time dependent variables).

However, often, the scientific interest is on a low dimensional functional

$$\beta(F)$$

of the distribution F of the data, with little or no knowledge about F.

The methods that we will study in this course meet the analytic challenge in these circumstances because they give valid inferences under non or semiparametric models that make minimal assumptions about the parts of the law of the data that are not of scientific interest. As such, they are protected from misspecification of models for these secondary parts of the law of the data.

Parametric models for i.i.d data

Data are n i.i.d copies $Z_1, \cdots Z_n$ of a random structure Z whose cumulative distribution function F is assumed to belong to the family

$$\mathcal{F} = \{F_\theta : \theta \in \Theta\},$$

where Θ is a subset of Euclidean space, i.e. $\Theta \in \mathbb{R}^p$.

Semiparametric models for i.i.d data

Data are n i.i.d copies $Z_1, \cdots Z_n$ of a random structure Z whose cumulative distribution function F is assumed to belong to the family

$$\mathcal{F} = \{F_\theta : \theta \in \Theta\},$$

where Θ is a "massive" set, i.e. it is NOT the subset of any Euclidean space.

We aim at estimating the value of $\beta(F)$ some function

$$\beta : \mathcal{F} \to \mathbb{R}^k.$$

Notational remark: $\beta(\theta) \equiv \beta(F_\theta)$.

Technical note: though the definition of semiparametric models does not require that the distributions in the family $\mathcal{F}$ be dominated by a measure, say μ, we will assume so in this course.

We will use f to denote the density of F with respect to the dominating measure μ and in a slight abuse of notation, we will use dx to denote $d\mu(x)$.

Example 1.1 (Nonparametric model). The model assumes "nothing"about F. Then $\mathcal{F}$ is the collection of all probability distributions (on a given sample space). Here Θ is the collection of all probability distributions and each $\theta \in \Theta$ is a probability distribution.

We might be interested in estimating, for example, the mean of Z, i.e.

$$\beta(F) = \int z f(z) dz.$$

Example 1.2 (Symmetric distributions). Suppose that Z is a scalar continuous random variable with distribution F assumed to have a density f satisfying

$$f(z) = g(z - \beta^*),$$

for some unknown $\beta^* \in \mathbb{R}$ and some unknown function $g(u)$ which is symmetric around 0, i.e.

$$g(u) = g(-u) \text{ for all } u.$$

Note that F is determined by the center of location β^* and the function $g(u)$. Therefore, the family $\mathcal{F}$ can be indexed by

$$\theta = (\beta, g),$$

ranging in the set $\Theta = \mathbb{R} \times \mathcal{G}$ where $\mathcal{G}$ is the set of positive real valued functions on the reals that are symmetric around 0 and integrate to 1.

In this problem the parameter of interest is typically the center of location

$$\beta(F) = \text{ the value } \beta^* \text{ such that } f(z) = g(z - \beta^*).$$

Example 1.3 (Conditional mean model). $Z = (Y, X), Y$ is a response (which we assume here continuous), X is a vector of covariates.

The model assumes that

$$E(Y \mid X) = g(X; \beta^*),$$

where $g(X;\beta)$ is a known function of X and β, e.g. $g(X;\beta) = \exp\left(X^T\beta\right)$ and $\beta^* \in \mathbb{R}^k$ is unknown. Other popular example is linear model and Logistic regression model. No other assumptions are made.

Focusing on continuous Y, define

$$\varepsilon = Y - g(X; \beta^*).$$

Then β and the joint distribution of (ϵ, X) determine the joint distribution of (Y, X).

Specifically, an arbitrary distribution in the model has density,

$$\begin{aligned} f_{Y,X}(y, x; \beta, \eta_1, \eta_2) &= f_{Y|X}(y \mid x; \beta, \eta_1) f_X(x; \eta_2) \\ &= f_{\varepsilon|X}(y - g(x;\beta) \mid x; \eta_1) \eta_2(x) \\ &= \eta_1(y - g(x;\beta), x)\eta_2(x), \end{aligned}$$

Where $\eta_1(\varepsilon, x)$ and $\eta_2(x)$ are non-negative functions restricted only by

$$\text{for all } x: \int \eta_1(\varepsilon, x) d\varepsilon = 1 \text{ and } \int \varepsilon\eta_1(\varepsilon, x) d\varepsilon = 0, \tag{1}$$

$$\int \eta_2(x) dx = 1. \tag{2}$$

Thus the family $\mathcal{F}$ can be indexed by

$$\theta = (\beta, \eta_1, \eta_2),$$

ranging in the set

$$\Theta = \mathbb{R}^k \times \boldsymbol{\eta_1} \times \boldsymbol{\eta_2},$$

where $\boldsymbol{\eta_1}$ is the set of all non-negative functions of (ε, x) satisfying (1), and $\boldsymbol{\eta_2}$ is the set of all non-negative functions of x satisfying (2).

Example 1.4 (Cox proportional hazards model). $Z = (T, X)$, T time to an event, X vector of covariates. Model assumes only that

$$\lambda(t \mid X) = \lambda_0(t) \exp\left(\beta^T X\right)$$

where $\lambda(t \mid X)$ is the conditional hazard at time t,

$$\lambda(t \mid X) \equiv \lim_{h \to 0} \frac{1}{h} \Pr(t \leq T < t + h \mid T \geq t, X).$$

If T is continuous, then $\lambda(t \mid X) = \frac{f(t|X)}{1-F(t|X)} = \frac{f(t|X)}{S(t|X)}$ can be interpreted roughly as the "instantaneous probability"of experiencing an event at time t given that you have not experienced an event before t.

It can be shown that an arbitrary distribution in the model has density

$$
\begin{aligned}
f_{T,X}(t, x; \beta, \lambda_0, \eta) &= f_{T|X}(t \mid x; \beta, \lambda_0) f_X(x; \eta) \\
&= \lambda(t \mid x; \beta, \lambda_0) \exp\left\{-\int_0^t \lambda(u \mid x; \beta, \lambda_0)\, du\right\} f_X(x; \eta) \\
&= \lambda_0(t) \exp\left(\beta^T x\right) \exp\left\{-\int_0^t \lambda_0(u) \exp\left(\beta^T x\right) du\right\} \eta(x),
\end{aligned}
$$

where $\lambda_0(t)$ is a positive but otherwise unrestricted function of t, and $\eta(x)$ is a positive function restricted only by

$$\int \eta(x) dx = 1, \tag{3}$$

where we have used the fact that

$$S(t \mid X) = \exp\left\{-\int_0^t \lambda(u \mid X) du\right\} \text{ and } f(t \mid X) = \lambda(t \mid X) S(t \mid X).$$

Thus the family $\mathcal{F}$ can be indexed by

$$\theta = (\beta, \lambda_0, \eta),$$

ranging in the set

$$\Theta = \mathbb{R}^k \times \mathbf{\Gamma} \times \boldsymbol{\eta},$$

where $\mathbf{\Gamma}$ is the set of all non-negative functions of t, and $\boldsymbol{\eta}$ is the set of all non-negative functions of x satisfying (3).

Example 1.5 (Partially linear regression model). Data $Z = (R, V, Y), Y$ is a real valued continuous response vector, R and V are vectors of covariates.
The model assumes that

$$E(Y \mid R, V) = h(V) + \beta^{*T} R,$$

where $h(V)$ is an unknown and unrestricted function of V, β^* is an unknown $k \times 1$ vector.
This model is like the conditional mean model with one additional infinite dimensional parameter, namely, the function $h(\cdot)$. It is sometimes referred to as semiparametric regression with identity link function.

Example 1.6 (Generalized partially linear regression model). Data $Z = (R, V, Y), Y$ is either binary or a count, R and V are vectors of covariates. The model assumes that

$$g\{E(Y \mid R, V)\} = h(V) + \beta^{*T} R,$$

where g is either a log or logit link function, $h(V)$ is an unknown and unrestricted function of V, β^* is an unknown $k \times 1$ vector.
This model is like a generalized conditional mean model with one additional infinite dimensional parameter, namely, the function $h(\cdot)$. It is sometimes referred to as semiparametric regression with link function g.

Example 1.7 (Single index binary choice model). Data $Z = (V, Y), Y$ is a binary variable, V are vectors of covariates.
The model assumes that

$$E(Y \mid V) = g\left(\beta^{*T} V\right),$$

where g is an unknown CDF, β^* is an unknown $k \times 1$ vector.
This model is like a generalized conditional mean model with the link function being unrestricted.

Example 1.8 (Semiparametric additive instrumental variable model). Data $Z = (V, R, A, Y), Y$ is a continuous outcome, V are vectors of covariates, R is a treatment variable, and A is an instrumenal variable.
The model assumes that

$$E\left(Y - \beta^* R \mid A, V\right) = E\left(Y - \beta^* R \mid V\right), \tag{4}$$

where β^* is an unknown treatment effect on the additive scale. Identification requires that $E(R \mid A = a, V = v)$ depends on a for at least one value of v. Note that β^* is not identified by standard regression of Y on R, whether or not one conditions on V and A, basically because the necessary condition for identification by such an approach requrire $E\left(Y - \beta^* R \mid R, A, V\right) = E\left(Y - \beta^* R \mid A, V\right)$ is not implied by (4). The treatment effect is said to be confounded.

Causal interpretation of Semiparametric regression

Let Y_r denote the potential outcome one would observe if one could intervene to set the treatment variable R to r. In a randomized experiment, randomization ensures that such intervention can be done as the treatment is under control of the experimenter. In an observational study, we try to mimic a randomized trial by assuming that within levels of covariates, it is as if nature performed a randomized trial, i.e. there is no unmeasured confounder. Then

$$E\left(Y_r \mid V\right) = E(Y \mid R = r, V) = h(V) + \beta^* r,$$

implies constant average causal effect of the treatment conditional on V, $E\left(Y_{r=1} - Y_{r=0} \mid V\right) = \beta^*$.

Causal interpretation of Semiparametric regression

Under no unmeasured confounding, we have that $R \perp Y_r \mid V$ for all r, this assumption in of itself is not testable without an additional assumption, i.e. does not place any restriction on the observed data distribution, however under the semiparametric structural model $E\left(Y_{r=1} - Y_{r=0} \mid V\right) = \beta^*$, one can show that

$$E\left(Y_R - \beta^* R \mid V, R\right) = E\left(Y_{r=0} \mid V, R\right) = E\left(Y_{r=0} \mid V\right).$$

By no unmeasured confounding assumption, that is the structural model of no treatment by V interaction on the additive scale together with no unmeasured confounding imply semiparametric regression model.

Causal interpretation of semiparametric additive IV model

Suppose that instead of assumption of no unmeasured confounding, one has access to a randomized instrumental variable, that is instead of $R \perp Y_r \mid V$ for all r, the most one can assume is that we have observed a variable A such that $A \perp Y_r \mid V$ for all r.

This assumption is often reasonable in the health and social sciences (e.g. Mendelian Randomization in Epidemiology, or Randomization in randomized clinical or encouragement trials in Biostatistics).

Causal interpretation of semiparametric additive IV model

Then the semiparametric structural model is given by

$$E\left(Y_r - Y_{r=0} \mid V, A, R = r\right) = \beta^* r,$$

which implies that

$$E\left(Y_R - \beta^* R \mid V, A\right) = E\left(Y_{r=0} \mid V, A\right) = E\left(Y_{r=0} \mid V\right).$$

Thus the IV assumption and the semiparametric structural model implies the observed data semiparametric model instrumental variable model

$$E\left(Y - \beta^* R \mid A, V\right) = E\left(Y - \beta^* R \mid V\right),$$

for the observed data distribution.

Example 1.9 (Missing data model). Consider the full data restricted mean model

$$E(Y \mid X) = g\left(X; \beta^*\right),$$

where $X = (X_1, X_2)$. Now suppose we observe iid samples on (Y, R, RX_1, X_2) where $R = 1$ if X_1 is observed and $R = 0$ if X_1 is missing for a person. One can show that the restricted mean model is not in general point identified from the observed data distribution without placing a restriction on the missing data mechanism. A common assumption is that data are missing at random, $R \perp X_1 \mid Y, X_2$.

The semiparametric on the observed data distribution has likelihood for a single realization

$$\begin{aligned}&\left\{\int f\left(\varepsilon\left(\beta^*\right) \mid X; \eta_1\right) f\left(X; \eta_2\right) dX_1\right\}^{1-R} f\left(R \mid X_2, Y; \eta_3\right)\\ &\times \left\{f\left(\varepsilon\left(\beta^*\right) \mid X; \eta_1\right) f\left(X; \eta_2\right)\right\}^R,\end{aligned}$$

which introduces a new infinite dimensional nuisance parameter η_3 indexing the missing data mechanism that now needs to be accounted for.

Semiparametric estimators

Loosely speaking, a semiparametric estimator $\widehat{\beta}_n$ of $\beta(\mathrm{F})$ is one which satisfies

$$\sqrt{n}\left\{\widehat{\beta}_n - \beta(F)\right\} \overset{D(F)}{\to} N(0, \Sigma(F)),$$

for all distributions F in the family $\mathcal{F}$.

For example, the solution to the GEE equations

$$\sum_{i=1}^{n} d\left(X_i; \widehat{\beta}_n\right)\left\{Y_i - g\left(X_i; \widehat{\beta}_n\right)\right\} = 0,$$

is (under regularity conditions) a semiparametric estimator in the conditional mean model.

Questions of interest

- For which semiparametric $\beta(F)$ can we hope to find semiparametric estimators?
- When they exist, how do we construct them?
- How do we define an analogous to the Cramer-Rao variance bound?
- For which functionals are there "optimal"semiparametric estimators whose variance is the smallest possible, i.e. that they achieve the semiparametric C-R bound?
- If an "optimal"estimator exists, how do we construct it?

In order to answer the previous questions we will need to review notions of asymptotic efficiency in parametric models.

We will investigate these notions from a geometric perspective as this is the natural approach for dealing with the infinite dimensional case.

This perspective benefits and relies on some basic results from Hilbert space theory that we will summarize next.