STATS 314B **Spring 2016**

Lecture 1: March 28

Lecturer: *Scribe:*

Disclaimer: .

1.1 Introduction

This set of lecture notes explores certain inference problems in nonparametric statistics. Loosely speaking nonparametric models will mean classes of probability distributions for observed data which are indexed by infinite dimensional parameter spaces. Rather than simply trying to understand the estimation of the infinite dimensional objects parametrizing the distributions, we will focus on estimating real valued functions of the distribution. These functions will typically have as its domain some metric space and will be in general called functionals in this course. Apart from exploring "optimal" estimation of such functionals, we shall also seek connections of such estimation problems with statistical inference problems such as: goodness of fit testing and construction of confidence sets.

Historically, estimation of real valued functionals was initially rigorously viewed from a $\sqrt{n}$-rate of estimation point of view. This required one to find conditions on the kinds of functionals to be estimated under which the mean squared error of estimation of nice estimators scaled like n^{-1}, along with accompanied theory of efficiency of estimators, in the sense of Local Asymptotic Minimaxity type of results ([Aad2000]). In this course, we will typically go below the $\sqrt{n}$-rate of estimation– a common phenomenon while working with functionals and function classes of "low regularity".

1.2 Examples

Before going into the general set up, we begin with examples of some functionals of interest, which will be our go to objects for the rest of the course.

(i) **Mean and Variance Functional:** Two of the oldest and most classically studied functionals of data generating mechanism might be the population mean and variance. In general, they can be descried as follows. If P denotes the distribution of a typical observation X over sample space $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$, then the mean and variance functionals are (provided they exist)

$$\psi_{\text{mean}}(P) = \int x dP(x),$$

$$\psi_{\text{var}}(P) = \int (x - \psi_{\text{mean}}(P))^2 \, dP(x).$$

If P is dominated by a σ-finite measure μ yielding a density f, then the mean and variance functionals ca also be written as

$$\psi_{\text{mean}}(P) = \int x f(x) d\mu(x) = \chi_{\text{mean}}(f),$$

$$\psi_{\text{var}}(P) = \int (x - \chi_{\text{mean}}(f))^2 f(x) d\mu(x) = \chi_{\text{var}}(f).$$

"Typically" it is possible to provide $\sqrt{n}$-consistent estimators of these functionals.

(ii) **Quadratic Functionals:** Suppose that a typical observation X is a draw from a distribution on $[0,1]$ having a density f with respect to the Lebesgue measure. Thus the distribution P of an observation X is described by f. We wish to estimate the quadratic functional

$$\psi(P_f) = \int f^2 d\mu = \chi(f).$$

This functional arises in multiple contexts: confidence set construction for Hodges-Lehmann adaptive estimators of location, goodness of fit testing in L_2 norm, and constructions of L_2 confidence balls ([BR88], [BR90], [KP96]).

(iii) **Treatment Effetct Functionals:** Consider the estimation of a treatment effect on an outcome in presence of a vector Z of confounding variables. Specifically, for a binary treatment A and response Y, a convenient way of summarizing effect of treatment A on outcome Y is through the variance weighted average treatment effect ([CHIM2009]) defined as

$$\tau := \mathbb{E}\left(\frac{Var(A|Z)c(Z)}{\mathbb{E}(Var(A|Z)}\right) = \frac{\mathbb{E}(cov(Y,A|Z))}{\mathbb{E}(Var(A|Z)} \tag{1.1}$$

where

$$c(z) = \mathbb{E}(Y|A=1, Z=z) - \mathbb{E}(Y|A=0, Z=z). \tag{1.2}$$

The above follows from a simple calculation and $c(z)$ is called the average treatment effect among subjects with $Z = z$ under the assumption of no unmeasured confounding. A semiparametric constraint in this set up is

$$c(z) = \psi^* \text{ for all } z \tag{1.3}$$

or specifically the model

$$\mathbb{E}(Y|A,Z) = \psi^* A + b(Z), \tag{1.4}$$

where $b(Z) = \mathbb{E}(Y|Z)$. It turns out that under above model , τ equals ψ^*. Moreover, the inference on τ is closely related to the estimation $\mathbb{E}(Cov(Y,A|Z))$. Specifically, point and interval estimator for τ can be constructed from point and interval estimator of the numerator $\mathbb{E}(cov(Y,A|Z))$ of τ. In particular, for any fixed $\tau^* \in \mathbb{R}$, define $Y^*(\tau^*) = Y - \tau^* A$ and the corresponding functional

$$\psi(\tau^*) = \mathbb{E}((Y^*(\tau^*) - \mathbb{E}(Y^*(\tau^*)|Z))(A - \mathbb{E}(A|Z))) = \mathbb{E}(cov(Y^*(\tau^*), A|Z)).$$

Then τ is the unique solution of $\psi(\tau^*) = 0$. Suppose we can construct point estimators $\hat{\psi}(\tau^*)$ and $(1-\alpha)$ interval estimator of $\psi(\tau^*)$. Then $\hat{\tau}$ satisfying $\psi(\hat{\tau}) = 0$ is an estimator of τ with similar properties. Further a $(1-\alpha)$ confidence set for τ is the set of τ^* for which $(1-\alpha)$ interval estimator of $\psi(\tau^*)$ contains 0. Considering the inference on $\mathbb{E}(Cov(Y,Y|Z))$ we note that

$$\mathbb{E}(Cov(Y,A|Z)) = \mathbb{E}(AY) - \mathbb{E}(\mathbb{E}(A|Z)\mathbb{E}(Y|Z)).$$

Above, $\mathbb{E}(AY)$ is easy to estimate by sample average. The crux of estimating $\mathbb{E}(Cov(Y,A|Z))$ lies in $\mathbb{E}(\mathbb{E}(A|Z)\mathbb{E}(Y|Z))$. This will be our functional of interest.

Suppose that a typical observation is distributed as $X = (Y,A,Z)$ for A taking values in the two-point set $\{0,1\}$. We think of Y as a response variable, A as a treatment variable and Z other covariate information collected on subjects under study. The covariate Z is chosen such one can assume the condition of *no unmeasured confounders*. The model can be parameterized by the marginal density

f of Z (relative to some dominating measure μ), $b(z) = \mathbb{E}(Y \mid Z = z)$, $a(z) = \mathbb{P}(A = 1 \mid Z = z)$, and $c(z) = \mathbb{E}(Y|A = 1, Z = z) - \mathbb{E}(Y|A = 0, Z = z)$. Thus the distribution P of an observation X is described by the triple (a, b, c, f). We wish to estimate

$$\psi(P_{(a,b,c,f)}) = \int abf d\mu = \chi((a, b, c, f)).$$

(iv) **Missing Data Models and Mean Functionals:** Suppose that a typical observation is distributed as $X = (YA, A, Z)$ for Y and A taking values in the two-point set $\{0, 1\}$ and conditionally independent given Z. We think of Y as a response variable, which is observed only if the indicator A takes the value 1. The covariate Z is chosen such that it contains all information on the dependence between response and missingness indicator, thus making the response *missing at random*. Alternatively, we think of Y as a "counterfactual" outcome if a treatment were given ($A = 1$) and estimate (half) the treatment effect under the assumption of *no unmeasured confounders*. The model can be parameterized by the marginal density f of Z (relative to some dominating measure μ) and the probabilities $b(z) = \mathbb{P}(Y = 1 \mid Z = z)$ and $a(z)^{-1} = \mathbb{P}(A = 1 \mid Z = z)$. (Using a for the inverse probability simplifies later formulas.) Alternatively, the model can be parameterized by the pair (a, b) and the function $g = f/a$, which is the conditional density of Z given $A = 1$, up to the norming factor $\mathbb{P}(A = 1)$. Thus the distribution P of an observation X is described by the triple (a, b, f), or equivalently the triple (a, b, g). We wish to estimate the mean response $\mathbb{E}Y$, i.e. the functional

$$\psi(P_{(a,b,g)}) = \int bf\, d\mu = \int abg\, d\mu = \chi((a, b, g)).$$

Estimators that are $\sqrt{n}$-consistent and asymptotically efficient in the semiparametric sense have been constructed using a variety of methods (e.g. [RR95], [T2007]) but only if a and b are restricted to sufficiently small regularity classes.

1.3 Notations and Asymptotic Framework

Let $X_1, \ldots, X_n$ is a random sample from a distribution P on a sample space $(\Omega, \mathcal{A})$. Let $\mathbb{U}_n$ denote the *empirical U-statistic* measure, viewed as an operator on functions. For given $k \leq n$ and a function $f : \Omega^k \to \mathbb{R}$ on the sample space this is defined by

$$\mathbb{U}_n f = \frac{1}{n(n-1)\cdots(n-k+1)} \sum\sum \cdots \sum_{1 \leq i_1 \neq i_2 \neq \cdots \neq i_k \leq n} f(X_{i_1}, X_{i_2}, \cdots, X_{i_k}).$$

We do not let the order k show up in the notation $\mathbb{U}_n f$. This is unnecessary, as the notation is consistent in the following sense: if a function $f : \Omega^l \to \mathbb{R}$ of $l < k$ arguments is considered a function of k arguments that is constant in its last $k - l$ arguments, then the right side of the preceding display is well defined and is exactly the corresponding U-statistic of order l. In particular, $\mathbb{U}_n f$ is the *empirical distribution* $\mathbb{P}_n$ applied to f if $f : \Omega \to \mathbb{R}$ depends on only one argument.

We write $P^n \mathbb{U}_n f = P^k f$ for the expectation of $\mathbb{U}_n f$ if $X_1, \ldots, X_n$ are distributed according to the probability measure P. We also use this operator notation for the expectations of statistics in general. We call f *degenerate* relative to P if $\int f(x_1, \ldots, x_k)\, dP(x_i) = 0$ for every i and $(x_j : j \neq i)$, and we call it *symmetric* if its values are invariant under permutations of its arguments. Given an arbitrary measurable function $f : \Omega^k \to \mathbb{R}$ we can form a function that is degenerate relative to P by subtracting the orthogonal projection in $L_2(P^k)$ onto the functions of at most $k - 1$ variables. This degenerate function can be written in the form

$$(D_P f)(X_1, \ldots, X_k) = \sum_{A \subset \{1,\ldots,k\}} (-1)^{k-|A|} \mathbb{E}_P \left(f(X_1, \ldots, X_k) \mid X_i : i \in A\right), \tag{1.5}$$

where the sum if over all subsets A of $\{1,\ldots,k\}$, including the empty set, for which the conditional expectation is understood to be $P^k f$. If the function f is symmetric, then so is the function $D_P f$.

Given two functions $g,h:\Omega\to\mathbb{R}$ we write $g\times h$ for the function $(x,y)\mapsto g(x)h(y)$. More generally, given k functions $g_1,\ldots,g_k$ we write $g_1\times\cdots\times g_k$ for the tensor product of these functions. Such product functions are degenerate iff all functions in the product have mean zero.

Our framework of evaluating inference procedures will typically be asymptotic in nature i.e. we will let $n\to\infty$. We set up some basic conventions that will be followed throughout in this respect. If a_n and b_n are two sequences of real numbers then $a_n\gg b_n$ (and $a_n\ll b_n$) implies that $a_n/b_n\to\infty$ (respectively $a_n/b_n\to 0$) as $n\to\infty$. Similarly $a_n\gtrsim b_n$ (and $a_n\lesssim b_n$) implies that $\liminf a_n/b_n = C$ for some $C\in(0,\infty]$ (and $\limsup a_n/b_n = C$ for some $C\in[0,\infty)$). Alternatively, $a_n = o(b_n)$ will also imply $a_n\ll b_n$ and $a_n = O(b_n)$ will imply that $\limsup a_n/b_n = C$ for some $C\in[0,\infty)$).

Our optimality criterion will be that of a asymptotic rate minimaxity, defined in the following sense. Consider, $X_1,\ldots,X_n$ to be random sample from distribution P where P varies over a class of probability distributions $\mathcal{P}$. Consider estimating a functional $\psi:\mathcal{P}\to\mathbb{R}$ with a measurable map $T_n:(\Omega^n,\mathcal{A}^n)\to\mathbb{R}$ and evaluate its performance by the worst case mean squared error as follows:

$$Risk(T_n,\mathcal{P}) = \sup_{P\in\mathcal{P}} \mathbb{E}_P\left(T_n-\psi(P)\right)^2.$$

We will typically interested in understanding sequences $\psi_n(\mathcal{P})\to 0$, when they exist, such that

$$0 < C_1(\mathcal{P}) \le \sup_{T_n} \frac{Risk(T_n,\mathcal{P})}{\psi_n(\mathcal{P})} \le C_2(\mathcal{P}) \le \infty.$$

In the above sense, a $\sqrt{n}$-rate of estimation will mean $\psi_n(\mathcal{P}) = \frac{1}{n}$.

References

[BR88] P.J. Bickel, and Y. Ritov, Estimating integrated squared density derivatives: sharp best order of convergence estimates, *Sankhyā: The Indian Journal of Statistics, Series A* (1988), pp. 381–393.

[BR90] P.J. Bickel, and Y. Ritov, Achieving information bounds in non and semiparametric models, *The Annals of Statistics* (1990), pp. 925–938.

[CHIM2009] R. Crump, V.J. Hotz, G.W. Imbens, and O.A. Mitnik, Dealing with limited overlap in estimation of average treatment effects, *Biometrika* (2009).

[KP96] G. Kerkyacharian, and D. Picard, Estimating nonquadratic functionals of a density using Haar wavelets, *The Annals of Statistics* (1996), pp. 485–507.

[RR95] A. Rotnitzky, and J. Robins, Semi-parametric estimation of models for means and covariances in the presence of missing data, *Scandinavian Journal of Statistics* (1995), pp. 323–333.

[RA2006] J. Robins, and A.W. Van der Vaart, Adaptive nonparametric confidence sets, *The Annals of Statistics* (2006), pp. 229–253.

[T2007] A. Tsiatis, Semiparametric theory and missing data, *Springer Science & Business Media* (2007).

[Aad200] A.W. Van der Vaart,Asymptotic Statistics, *Cambridge university press* (2000).