**STATS 314B** **Spring 2016**

# Lecture 2: March 30

*Lecturer:* *Scribe:*

**Disclaimer**: .

## 2.1 General Setup

$X_1, \ldots, X_n$ are a random sample from $P_\eta$ with $\eta \in \mathbb{H}$ where $\mathbb{H}$ is subset of a Hilbert space with norm $\|\cdot\|$. Assume also that $P_\eta$ has a density $p_\eta$ w.r.t. a $\sigma$-finite measure $\mu$ on the sample space $(\Omega, \mathcal{A})$. We are interested in estimating $\psi(P_\eta) = \chi(\eta) : \mathbb{H} \to \mathbb{R}$.

If we start with an initial estimator $\hat{\eta}$ of $\eta$, then typically $\chi(\hat{\eta})$ will be a consistent estimator of $\chi(\eta)$ for most "nice" functionals. However, one might do better by going a through a "one-step" kind of estimator as follows. Suppose $\chi$ admits a Fréchet-type Taylor expansion type of representation as follows:

$$\chi(\eta) = \chi(\hat{\eta}) + \chi'_{\hat{\eta}}(\hat{\eta} - \eta) + O(\|\hat{\eta} - \eta\|^2), \tag{2.1}$$

where $\chi'$ is a bounded linear functional on $\mathbb{H}$. For the sake of intuitive understanding, we will use such expansions freely without worrying about the technical details.

Expansion (2.1) implies that, unless the derivative term vanishes, the plug-in estimator $\chi(\hat{\eta})$ will have error $O_P(\|\hat{\eta} - \eta\|)$ (this follows from Banach-Steinhaus Theorem provided we assume things like the class $\chi'_\eta, \eta \in \mathbb{H}$ is point-wise bounded ). When $\mathbb{H}$ is "high-dimensional", this error can be potentially large.

However, the same expansion above also suggest a possibly better way to alleviate this concern– "estimate the derivative term". Of course, we need some form of the derivative to do this. One way of doing this is assuming a "Von-Mises" kind of representation of the derivative:

$$\begin{aligned} \chi'_{\hat{\eta}}(\hat{\eta} - \eta) &= \int \chi^1_{\hat{\eta}} (dP_\eta - dP_{\hat{\eta}}) \\ &= \int \chi^1_{\hat{\eta}} dP_\eta. \end{aligned} \tag{2.2}$$

Above we have assumed that $\int \chi^1_\eta dP_\eta = 0$ (which can be arranged since $\int 1 d(P_\eta - dP_{\hat{\eta}}) = 0$ and thereby defining $\widetilde{\chi^1_{\hat{\eta}}} = \chi^1_{\hat{\eta}} - \int \chi^1_{\hat{\eta}} dP_{\hat{\eta}}$). Now that the derivative term of (2.1) is an expectation with respect to the data generating mechanism, we can estimate it by the sample average $\mathbb{P}_n \chi^1_{\hat{\eta}}$ and have the resulting one-step estimator

$$\hat{\chi}_n = \chi(\hat{\eta}) + \mathbb{P}_n \chi^1_{\hat{\eta}}. \tag{2.3}$$

The error of this estimator can be informally analyzed as follows:

$$\begin{aligned} \hat{\chi}_n - \chi(\eta) &= \chi(\hat{\eta}) - \chi(\eta) + \mathbb{P}_n \chi^1_{\hat{\eta}} \\ &= O_p\left(\|\hat{\eta} - \eta\|^2\right) + (\mathbb{P}_n - P_\eta)\chi^1_{\hat{\eta}} \\ &= O_p\left(\|\hat{\eta} - \eta\|^2\right) + O_P(n^{-1/2}). \end{aligned} \tag{2.4}$$

Above we have assumed that the difference $(\mathbb{P}_n - P)\chi^1_{\hat{\eta}}$ is "centered" and "variance" $O(1/n)$. Note that the words "centered" and "variance" are in quotes. This is because the randomness in the initial estimator $\hat{\eta}$ prevents a simple calculation of mean and variance. One can use empirical process theory can be used to show that the effect of replacing $\chi^1_\eta$ by $\chi^1_{\hat{\eta}}$ is negligible, if the class of functions $\chi^1_\eta$ is not "too large" (e.g. Donsker type classes). If we care about orders of magnitude only, and then a simpler approach is to split the sample and use separate observations to construct $\hat{\eta}$ and $\mathbb{P}_n$. Then the order calculations above can be justified by reasoning conditionally on the first sample, and it suffices that $\chi^1_{\hat{\eta}} dP_\eta$ remains bounded in probability.

The improvement of the estimator over the plug-in can be justified as follows. Whereas to achieve an error of $O_P(n^{-\frac{1}{2}})$ one would have required $\|\hat{\eta} - \eta\| = O_P(n^{-\frac{1}{2}})$ for the plug-in, one requires a more modest $\|\hat{\eta} - \eta\| = O_P(n^{-\frac{1}{4}})$ to achieve the same goal with $\hat{\chi}_n$.

**Quadratic Functional:** Consider $X_1, \ldots, X_n$ i.i.d. from density $\eta$ w.r.t. Lebsegue measure $\mu$ on $[0,1]^d$. Consider $\eta \in \mathbb{H} \subset L_2[0,1]^d$ and $\chi(\eta) = \int \eta^2 d\mu$. An expansion like (2.2) can be justified with $\chi^1_\eta(x) = 2(\eta(x) - \chi(\eta))$. The corresponding estimator $\hat{\chi}_n$ will required $\|\hat{\eta} - \eta\|_2 = O_p(n^{-\frac{1}{4}})$ whereas the plug-in $\int \hat{\eta}^2 d\mu$ requires $\|\hat{\eta} - \eta\|_2 = O_p(n^{-\frac{1}{2}})$. If $\eta$ is assumed to belong to a class of smoothness $\alpha$, then the former often demands $\frac{\alpha}{2\alpha+d} \geq \frac{1}{4}$ (*i.e.* $\alpha \geq \frac{d}{2}$) and the latter $\frac{\alpha}{2\alpha+d} \geq \frac{1}{2}$ (not possible).

The question now stands about how to obtain an expansion (2.1) satisfying (2.2). Interestingly, $\chi^1_\eta$ above is what is known as an "Influence Function" in semiparametric theory, developed significantly through 1980-1990. Informally the routine goes as follows. A Tangent Set at $P_\eta$ of the model $(P_\eta, \eta \in \mathbb{H})$ is the set of all score functions at $t = 0$ i.e.

$$\dot{g}_\eta = \left( \frac{\partial}{\partial t|_{t=0}} p_{\eta_t} \right) / p_\eta,$$

of (smooth) one dimensional sub-models $(P_{\eta_t}, t \in (-1,1))$ with $\eta_0 = \eta$. By smooth sub-models above we mean maps $t \to \eta_t$ such that the derivative above exists. An influence function of $\chi(\eta)$ at $P_\eta$ is a measurable map $x \to \chi^1_{\eta,*}(x)$ such that for all paths $t \to \eta_t$ as above

$$\frac{d}{dt|_{t=0}} \chi(\eta_t) = \frac{d}{dt|_{t=0}} P_{\eta_t} \chi^1_{\eta,*} = P_\eta \chi^1_{\eta,*} \dot{g}_\eta. \tag{2.5}$$

Let us intuitively convince ourselves that this influence function $\chi^1_{\eta,*}$ is the same as $\chi^1_\eta$ in the Von-Mises type expansion (2.2). To compare (2.5) and (2.2), identify $\eta$ with $\eta_t$ and $\hat{\eta}$ with $\eta$. Then the Von-Mises expansion suggests that

$$\begin{aligned}
\chi(\eta_t) &= \chi(\eta) + P_{\eta_t} \chi^1_\eta + O\left(\|\eta_t - \eta\|^2\right) \\
&= \chi(\eta) + \int \chi^1_\eta p_{\eta_t} d\mu + O\left(\|\eta_t - \eta\|^2\right) \\
&= \chi(\eta) + \int \chi^1_\eta (p_{\eta_t} - p_\eta) d\mu + \int \chi^1_\eta p_\eta d\mu + O\left(\|\eta_t - \eta\|^2\right) \\
&= \chi(\eta) + t \int \chi^1_\eta \dot{g}_\eta p_\eta d\mu + O(t^2). \quad \text{(Some kind of invertibility of the map: } \eta \to p_\eta)
\end{aligned}$$

Noting that (2.5) suggests the same expansion with $\chi^1_{\eta,*}$ implies equality of $\chi^1_{\eta,*}$ and $\chi^1_\eta$. It is easier to verify that a functional $\chi$ has an influence function as above rather than a Von-Mises expansion, which is a stronger requirement. Indeed, we use the Von-Mises type expansion only for intuitive understanding of the decay of error terms and for practical purposes of finding estimators we will essentially make use of (2.5) instead.

**Quadratic Functional:** Consider $X_1, \ldots, X_n$ i.i.d. from density $\eta$ w.r.t. Lebsegue measure $\mu$ on $[0,1]$. Consider $\eta \in \mathbb{H} \subset L_2[0,1]$ and $\chi(\eta) = \int \eta^2 d\mu$. Then $\chi^1_\eta(x) = 2(\eta(x) - \chi(\eta))$ is an influence function at $\eta$ as in (2.5). To check this take any "smooth" one-dimensional submodel $(\eta_t, t \in (-1,1))$ with $\eta_0 = \eta$ with score function $\dot{g}_\eta$. Then $\left(\frac{\partial}{\partial|_{t=0}}\eta_t\right)/\eta = \dot{g}_\eta$ and as a result $\frac{d}{dt|_{t=0}}\chi(\eta_t) = \frac{d}{dt|_{t=0}}\int \eta_t^2 d\mu = 2\int \eta \dot{g}_\eta \eta d\mu = 2P_\eta\left(\eta - \chi(\eta)\right)\dot{g}_{\eta^-}$ verifying the claim.

Note that, an influence function as defined in (2.5) is not necessarily unique. This is because only its inner products with elements $\dot{g}_\eta$ of the tangent set matter. An influence function that is contained in the closed linear span of the tangent set is called the efficient influence function, since it minimizes the variance $Var_{P_\eta}\mathbb{P}_n\chi^1_\eta$ over all influence functions. (This also happens to be the "influence function" of asymptotically efficient estimators)

Although using the intuition as above one can justify $\sqrt{n}$-consistent estimators of $\chi(\eta)$ if $\|\hat{\eta} - \eta\| = O_p(n^{-1/4})$, the class of $\eta$ might be too large in nonparametric problems to even guarantee this. For example, for the case of quadratic functional of density, one needed at least $d/2$ derivatives of the density. For even moderately large $d$, this might be a restrictive assumption to make for developing a reasonable theory. The idea indeed is to take higher order Von-Mises type expansion along with higher order versions of influence function and scores. Let us try to develop an heuristic version of this for a second order expansion.

Taking (2.2) one step further, suppose

$$\chi(\eta) = \chi(\hat{\eta}) + \chi'_{\hat{\eta}}(\hat{\eta} - \eta) + \frac{1}{2}\chi''_{\hat{\eta}}(\hat{\eta} - \eta, \hat{\eta} - \eta) + O(\|\hat{\eta} - \eta\|^3), \tag{2.6}$$

where $\chi''$ is a bounded bilinear functional on $\mathbb{H}$. A second order Von-Mises type expansion will require

$$\chi'_{\hat{\eta}}(\hat{\eta} - \eta) + \frac{1}{2}\chi''_{\hat{\eta}}(\hat{\eta} - \eta, \hat{\eta} - \eta) = \int \chi^1_{\hat{\eta}}\left(dP_\eta - dP_{\hat{\eta}}\right) + \frac{1}{2}\int\int \chi^2_{\hat{\eta}}\left(dP_\eta - dP_{\hat{\eta}}\right) \times \left(dP_\eta - dP_{\hat{\eta}}\right) \tag{2.7}$$

Assuming without loss of generality that $\chi^1_\eta$ and $\chi^2_\eta$ are degenerate with respect to $P_\eta$, we then have a two-step estimator as follows:

$$\hat{\chi}_n = \chi(\hat{\eta}) + \mathbb{P}_n(\chi^1_{\hat{\eta}}) + \frac{1}{2}\mathbb{U}_n\chi^2_{\hat{\eta}} = \chi(\hat{\eta}) + \mathbb{U}_n\left(\chi^1_{\hat{\eta}} + \frac{1}{2}\chi^2_{\hat{\eta}}\right)$$

Because the variance of a U-statistic is of order $O(1/n)$, this estimator ought to have an error of the order $O_p\left(\|\hat{\eta} - \eta\|^3\right) + O_P(n^{-1/2})$.

To operationalize such an expansion, one can go through definitions of higher order scores and influence functions as follows. Note that higher order influence functions should have the right product with higher order scores representing th higher order derivatives of $\chi(\eta_t)$ for smooth one dimensional submodels.

Connecting with (2.7), this should roughly imply

$$\begin{aligned}
\chi(\eta_t) &\approx \chi(\eta) + \int \chi^1_\eta(x_1)p_{\eta_t}(x_1)d\mu(x_1) + \int\int \chi^2_\eta(x_1,x_2)p_{\eta_t}(x_1)p_{\eta_t}(x_2)d\mu(x_1)d\mu(x_2) \\
&= \chi(\eta) + \int\int \chi^1_\eta(x_1)p_{\eta_t}(x_1)p_{\eta_t}(x_2)d\mu(x_1)d\mu(x_2) + \int\int \chi^2_\eta(x_1,x_2)p_{\eta_t}(x_1)p_{\eta_t}(x_2)d\mu(x_1)d\mu(x_2) \\
&\approx \chi(\eta) + t\int\int \left(\chi^1_\eta(x_1) + \frac{1}{2}\chi^2_\eta(x_1,x_2)\right)\left(\frac{\partial}{\partial t|_{t=0}}\prod_{i=1}^2 p_{\eta_t}(x_i)\right)/\left(\prod_{i=1}^2 p_\eta(x_i)\right)dP_\eta(x_1)dP_\eta(x_2) \\
&+ \frac{t^2}{2}\int\int \left(\chi^1_\eta(x_1) + \frac{1}{2}\chi^2_\eta(x_1,x_2)\right)\left(\frac{\partial^2}{\partial t^2|_{t=0}}\prod_{i=1}^2 p_{\eta_t}(x_i)\right)/\left(\prod_{i=1}^2 p_\eta(x_i)\right)dP_\eta(x_1)dP_\eta(x_2)
\end{aligned}$$

In particular, a *tangent set of order* 2 at $P_\eta$ is the set of all derivatives

$$\dot{g}(x_1, x_2) = \left( \frac{\partial}{\partial t|_{t=0}} \prod_{i=1}^{2} p_{\eta_t}(x_i) \right) / \left( \prod_{i=1}^{2} p_\eta(x_i) \right)$$

$$\ddot{g}_\eta(x_1, x_2) = \left( \frac{\partial^2}{\partial t^2|_{t=0}} \prod_{i=1}^{2} p_{\eta_t}(x_i) \right) / \left( \prod_{i=1}^{2} p_\eta(x_i) \right)$$

which arise from submodels $(P_{\eta_t}, t \in (-1, 1, ))$. This implies defining an *influence function of order* 2 to be $(x_1, x_2) \to \chi^1_\eta(x_1) + \frac{1}{2}\chi^2_\eta(x_1, x_2)$ satisfying

$$\frac{d}{dt|_{t=0}} \chi(\eta_t) = \frac{d}{dt|_{t=0}} P^2_{\eta_t} \left( \chi^1_\eta + \frac{1}{2}\chi^2_\eta \right) = P^2_\eta \left( \chi^1_\eta + \frac{1}{2}\chi^2_\eta \right) \dot{g}_\eta,$$

$$\frac{d^2}{dt^2|_{t=0}} \chi(\eta_t) = \frac{d^2}{dt^2|_{t=0}} P^2_{\eta_t} \left( \chi^1_\eta + \frac{1}{2}\chi^2_\eta \right) = P^2_\eta \left( \chi^1_\eta + \frac{1}{2}\chi^2_\eta \right) \ddot{g}_\eta. \tag{2.8}$$

Combining the basic relationships in (2.5) and (2.8), we can define $m^{\text{th}}$ order influence function of $\chi(\eta)$ at $P_\eta$ to be $\dot{\chi}^{(m)}_\eta$, which is a measurable function of $m$ variables $x_1, \ldots, x_m$ satisfying,

$$\frac{d^j}{dt^j|_{t=0}} \chi(\eta_t) = \frac{d^j}{dt^j|_{t=0}} P^m_{\eta_t} \dot{\chi}^{(m)}_\eta, \quad j = 1, \ldots, m. \tag{2.9}$$

This can indeed be motivated by a corresponding $m^{\text{th}}$ order Von-Mises type representation of $\chi(\eta)$, which in turn suggests a $m$-step estimator of $\chi(\eta)$ as

$$\hat{\chi}^{(m)}_n = \chi(\hat{\eta}) + \mathbb{U}_n \dot{\chi}^{(m)}_{\hat{\eta}}. \tag{2.10}$$

The first and second order calculations above also suggest that

$$\dot{\chi}^{(m)}_\eta = \chi^1_\eta + \frac{1}{2}\chi^2_\eta + \ldots + \frac{1}{m!}\chi^m_\eta,$$

where $\chi^j_\eta$, $j = 1, \ldots, m$ corresponds to the first $m$ terms of a Von-Mises type expansion.

For computation in examples the defining equations (2.9) of a higher order influence function can be tedious. Alternatively, it is usually easier to apply (and justify) the rule that a higher order derivative is the derivative of the previous order derivative. One computes the first order influence function $x_1 \to \chi^1_\eta(x_1)$ of the functional $\chi(\eta)$ at $P_\eta$ as usual. Next recursively for $j = 2, 3, \ldots, m$, one determines influence functions $x_j \to \chi^j_\eta(x_1, \ldots, x_j)$ as influence functions of the functionals $\eta \to \chi^{j-1}_\eta(x_1, \ldots, x_{j-1})$, for fixed $(x_1, ..., x_{j1})$. The function $\chi^j_\eta$ can then be made degenerate by subtracting its projection on the linear span of all functions of one argument less. We shall prove later that such a computation actually yields an influence function according to the definition (2.9).

Coming back to our estimator in (2.10), our intuition above also suggests that we might expect that for any fixed $m$,

$$\hat{\chi}^{(m)}_n - \chi(\eta) = \left( \mathbb{U}_n - P^m_\eta \right) \dot{\chi}^{(m)}_\eta + o_P(n^{-1/2}) + O_p\left( \|\hat{\eta} - \eta\|^{m+1} \right). \tag{2.11}$$

In words, the bias of the plug-in estimator $\chi(\hat{\eta})$ would be corrected to the order $O_p\left( \|\hat{\eta} - \eta\|^{m+1} \right)$, and therefore $\sqrt{n}$-estimators for $\chi(\eta)$ exist even in situations where $\eta$ is estimable only with low precision. The only cost would be a slightly larger variance in the U-statistic relative to the empirical measure.

Unfortunately, there is no such free lunch. In particular, one cannot keep on correcting the bias without seriously increasing the variance. Although the preceding heuristics are correct, they are not applicable, since higher order influence functions typically do not exist. The main reason being that when $L_2(P_\eta)$ is not finite dimensional, multilinear maps from $L_2(P_\eta)^j \to L_2(P_\eta)$ cannot be represented as repeated integral of the type demanded by Von-Mises type expansions. To see this note that, a second order Von-Mises expansion demands that a particular bounded bilinear form $B(h_1, h_2) : L_2(P_\eta) \times L_2(P_\eta) \to \mathbb{R}$ to be representable through some measurable map $\ddot{\chi} : \Omega^2 \to \mathbb{R}$ as

$$B(h_1, h_2) = \int \int h_1(x_1)\ddot{\chi}(x_1, x_2)h_2(x_2)dP_\eta(x_1)dP_\eta(x_2)$$

Such a bilinear functional $B(h_1, h_2)$ can always we represented as $\int h_1(x_1)Ah_2(x_1)dP_\eta(x_1)$ for some continuous linear functional $A : L_2(P_\eta) \to L_2(P_\eta)$. However, it is not in general true that a bounded linear functional $A : L_2(P_\eta) \to L_2(P_\eta)$ can always be represented as $Ah(\cdot) = \int \ddot{\chi}(\cdot, x')h(x')dP_\eta(x')$. For example consider $Ah = h$ to be the identity map, which is of course a continuous linear functional. If one was able to represent this as a kernel operation as above, then $h(x) = \int \ddot{\chi}(x, x')h(x')dP_\eta(x')$ for all $x$. But if for fixed $x$, $\ddot{\chi}(x, x') \in L_2(P_\eta)$ , then $\int \ddot{\chi}(x, x')h(x')dP_\eta(x')$ is a bounded linear functional. This means that $h \to h(x)$ is a bounded linear functional $L_2(P_\eta) \to \mathbb{R}$, which is not true.

Failure of existence of second order influence functions in the above sense does not mean that the idea to use a quadratic expansion for improved estimation is not fruitful. We can still try and estimate the higher order terms as well as possible, and still improve on the linear estimator. Focusing on the second order expansion, a key observation to attain this is that a bilinear map on a finite-dimensional subspace $L \times L \subset L_2(P_\eta) \times L_2(P_\eta)$ is always representable by a kernel. This is captured by the next result.

**Theorem 2.1.** *Suppose $L \subset L_2(P_\eta)$ is a finite-dimensional subspace and $B : L \times L \to \mathbb{R}$ is continuous and bilinear. Then there exists a function $\ddot{\chi} \in L_2(P_\eta \times P_\eta)$ such that*

$$B(h_1, h_2) = \int \int h_1(x_1)\ddot{\chi}(x_1, x_2)h_2(x_2)dP_\eta(x_1)dP_\eta(x_2)$$

*for all $(h_1, h_2) \in L_2(P_\eta) \times L_2(P_\eta)$.*

*Proof.* Let $dim(L) = k < \infty$ and $\{e_1, \ldots, e_k\}$ be an orthonormal basis of $L$. Then

$$\ddot{\chi}(x_1, x_2) = \sum_{i=1}^{k} e_i(x_1)e_i(x_2)$$

is the required kernel. □

This theorem implies that we can therefore always represent and estimate the $m^{\text{th}}$ order derivative at differences $\eta - \hat{\eta}$ within any given finite-dimensional linear space $L$. The resulting estimator looks like

$$\chi(\hat{\eta}) + \mathbb{U}_n \left( \chi^1_{\hat{\eta}} + \sum_{j=2}^{m} \frac{1}{j!} \chi^j_{\hat{\eta},L} \right)$$

where $\chi^j_{\hat{\eta},L}$'s are the different pieces of the partial $m^{\text{th}}$ order influence function based on an approximating space $L$. The error in nonrepresented directions however remains. The main challenge is to determine the directions of non-representation so that we can balance the following three terms.

(i) The bias in the non-represented directions.

(ii) The always present estimation bias: $O_P(\|\hat{\eta} - \eta\|^{m+1})$.

(iii) The variance of the U-statistic arising from the representing subspaces $L$.

Let us pay some attention to the third component. Although the variance of a U-statistic with a fixed kernel is dominated by its linear term (of the Hoeffding's Decomposition) and is therefore of $O(1/n)$, since we need to represent the functionals in more and more directions given larger sample size $n$ (to reduce the representation bias), we end up with kernels that become more and more complex with $n$. The resulting variance of is therefore typically larger than $O(1/n)$.

# References