Optimal Dynamic Intervention

The optimal rule $W \to d_0(W)$ is defined by

$$d_0 = \arg\min_d E_0 Y_d.$$

It is given by:

$$d_0(W) = I(B_0(W) > 0),$$

where

$$B_0(W) = \bar{Q}_0(1, W) - \bar{Q}_0(0, W)$$

is the conditional additive treatment effect.
Both the rule d_0 as well as its performance $E_0 Y_{d_0}$ are quantities of interest in precision medicine.

Stochastic Intervention

- Let G^* be a conditional probability distribution of A, given W.
- We could modify the structural equation model by replacing the equation $A = f_A(W, U_A)$ by drawing $A^* \sim G^*(\cdot \mid W)$. One can also define $A^* = d(W, U^*)$ for a rule d and random error U^*.
- This defines a counterfactual $Y_{G^*} = f_Y(W, A^*, U_Y)$.
- The mean outcome $E_0 Y_{G^*}$ is the quantity of interest.
- Under RA, it is identified by

$$E_0 Y_{G^*} = \int_{a,w} \bar{Q}_0(a, w) dG^*(a \mid w) dQ_{W,0}(w).$$

Examples of Stochastic Interventions

- $A^* \sim Bernoulli(p)$ for some known p.
- $A^* \sim Bernoulli(p(W))$ for some known $p(W)$.
- $A^* = A + \delta$ for a deterministic rule. This corresponds with first drawing A from the treatment mechanism, and subsequently evaluating $A + \delta$:

$$g^*(a^* \mid W) = g_0(a^* - \delta \mid W).$$

- More generally, if $A^* = d(A, W)$, then

$$g^*(a^* \mid W) = g_0(d_W^{-1}(a^*) \mid W),$$

where d_W^{-1} is the inverse function of $a \to d(a, W)$.

Average Causal Effect Among the Treated

Consider the following modified system of structural equations: $W = f_W(U_W)$, $A = f_A(W, U_A)$, $A^* = 1$, $Y_1 = f_Y(W, A^*, U_Y)$. Similarly, we can define this for $A^* = 0$. We can now define

$$E_0(Y_1 - Y_0 \mid A = 1).$$

This is called the effect among the treated. Under RA it is identified by:

$$\begin{aligned} E(Y_1 - Y_0 \mid A = 1) &= E_0(\bar{Q}_0(1, W) - \bar{Q}_0(0, W) \mid A = 1) \\ &= \int_w \{\bar{Q}_0(1, w) - \bar{Q}_0(0, w)\} \frac{g_0(1 \mid w)}{P_0(A = 1)} dP_0(w). \end{aligned}$$

Missing and Censoring Indicators can be included in SCM as endogenous variables

- For example, suppose that our observed data structure on one unit is $(W, A, \Delta, Y^* = \Delta Y) \sim P_0$.
- We define structural equation model: $W = f_W(U_W)$, $A = f_A(W, U_A)$, $\Delta = f_\Delta(W, A, U_\Delta)$, $Y = f_Y(W, A, U_Y)$, $Y^* = \Delta Y$.
- The counterfactual Y_1^* of interest is now the one corresponding with intervention $A = 1$ and $\Delta = 1$, and similarly Y_0^*.
- Under RA, the average causal effect $E_0 Y_1^* - E_0 Y_0^*$ is identified by

$$E_0\{E_0(Y^* \mid A = 1, \Delta = 1, W) - E_0(Y^* \mid A = 0, \Delta = 1, W)\}.$$

Identification of Intervention Specific Distribution for multiple time-point interventions: G-Computation Formula

Suppose $O = (L(0), A(0), L(1), A(1), L(2) = Y) \sim P_0$, where $A(t) = (A_1(t), \Delta(t))$ with $A_1(t)$ treatment and $\Delta(t)$ monitoring indicator. We can define an SCM:

$$\begin{aligned} L(0) &= f_{L(0)}(U_{L(0)}) \\ A(0) &= f_{A(0)}(L(0), U_{A(0)}) \\ L(1) &= f_{L(1)}(L(0), A(0), U_{L(1)}) \\ A(1) &= f_{A(1)}(L(0), A(0), L(1), U_{A(1)}) \\ Y &= f_Y(L(0), A(0), L(1), A(1), U_Y). \end{aligned}$$

Consider a stochastic intervention $g^*_{A(0)}, g^*_{A(1)}$ on $(A(0), A(1))$. This defines counterfactual $O_{g^*} = (L(0), A^*(0), L_{g^*}(1), A^*(1), L_{g^*}(2))$.

Identification by G-computation formula under Sequential Randomization Assumption

- Assume SRA: $A(j)$ is independent of Y_{g^*}, given $\bar{L}(j), \bar{A}(j-1)$, $j = 0, 1$.
- The distribution P_{g^*} of L_{g^*} is identified by the density $p^{g^*} = q_{L(0)} g^*_{A(0)} q_{L(1)} g^*_{A(1)} q_{L(2)}$:

$$\begin{aligned} p_{g^*}(o) \quad = \quad & q_{L(0)}(l(0)) g^*_{A(0)}(a(0) \mid l(0)) \\ & q_{L(1)}(l(1) \mid l(0), a(0)) g^*_{A(1)}(a(1) \mid l(0), a(0), l(1)) \\ & q_{L(2)}(l(2) \mid l(0), a(0), l(1), a(1)). \end{aligned}$$

- The existence of this density relies on conditioning events having positive probability (the positivity assumption):

$$\frac{g^*_{A(j)}(a(j) \mid \bar{L}(j), \bar{A}(j-1))}{g_{0,A(j)}(a(j) \mid \bar{L}(j), \bar{A}(j-1)} < \infty,$$

across all possible histories $\bar{L}(j), \bar{A}(j-1)$. That is, if the probability that one assigns the value $a(j)$ to a unit with history $\bar{L}(j), \bar{A}(j-1)$

Outline

Kernel density estimation

Suppose we observe n iid observations $O_i \sim P_0$, $i = 1, \ldots, n$, and that the statistical model $\mathcal{M}$ consists of probability distributions dominated by a dominating measure μ, and satisfying some smoothness assumptions.

For simplicity, let's consider the case that O a univariate and that μ is the Lebesgue measure.

Suppose that our target parameter is $p_0 = dP_0/d\mu$ is the density w.r.t. Lebesgue. Consider a kernel density estimator

$$p_{n,h}(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{O_i - x}{h}\right),$$

where K is a kernel and h is a bandwidth. If O would be d-dimensional, then

$$p_{n,h}(x) = \frac{1}{nh^d} \sum_{i=1}^{n} K\left(\frac{O_i - x}{h}\right),$$

where now K is a d-variate real valued function and $(O - x)/h \equiv ((O_j - x_j)/h : j = 1, \ldots, d)$.

One could also naturally extend this latter definition to $h = (h_1, \ldots, h_d)$, a bandwidth for each dimension.

Variance of kernel density estimator

The variance of the kernel dentist estimator is given by:

$$\frac{1}{nh^2}\text{VAR}\left\{K\left(\frac{O-x}{h}\right)\right\}.$$

It follows that the variance of the kernel is $O(h)$, so that we can conclude that

$$\text{VAR}(p_{n,h}(x)) = O\left(\frac{1}{nh}\right).$$

The simple intuition is that the kernel density estimator at a point is essentially an empirical mean over $O(nh)$ iid observations, so that its variance is $O(1/(nh))$.

Bias of kernel density estimator

The expectation of the kernel density estimator is given by:

$$Ep_{n,h}(x) = \frac{1}{h} E\left\{K\left(\frac{O-x}{h}\right)\right\}.$$

It follows that

$$EK\left(\frac{O-x}{h}\right) = \int K(y)p_0(x+hy)dy.$$

So that the bias of the kernel density estimator is given by:

$$\text{Bias}(p_{n,h}(x)) = \int K(y)\left\{p_0(x+hy) - p_0(x)\right\}dy.$$

Tailor expansion of density at x to obtain alternative bias expression

Suppose that p_0 is m-times continuously differentiable at x. Then,

$$p_0(x+hy) - p_0(x) = \sum_{j=1}^{m-1} \frac{(hy)^j}{j!} p_0^{(j)}(x) + \frac{(hy)^m}{m!} p_0^{(m)}(\xi(x, hy)),$$

for a $\xi(x, hy)$ between x and $x + hy$. Thus, the bias of the kernel density estimator can be represented as:

$$\begin{aligned} \text{Bias}(p_{n,h}(x)) &= \int K(y) \sum_{j=1}^{m-1} \frac{(hy)^j}{j!} dy p_0^{(j)}(x) \\ &\quad + \int K(y) \frac{(hy)^m}{m!} p_0^{(m)}(\xi(x, hy)) dy. \end{aligned}$$

Bias under smoothness assumptions and orthogonal kernel

Suppose that p_0 is m-times continuously differentiable at x and that K satisfies $\int K(y)dy = 1$, $\int K(y)y^j dy = 0$ for $j = 1, \ldots, m-1$. Then it follows that

$$\text{Bias}(p_{n,h}(x)) = \int K(y)\frac{(hy)^m}{m!}p_0^{(m)}(\xi(x, hy))dy,$$

and thus that

$$\text{Bias}(p_{n,h}(x)) = O(h^m).$$

Selection of bandwidth and kernel:

If one would know the underlying smoothness m of the true density p_0 at x, then it would be good to select an m-orthogonal kernel. An optimal bandwidth could then be defined by

$$h_0 = \arg\min_h MSE(p_{n,h}(x)) = \arg\min_h E(p_{n,h}(x) - p_0(x))^2.$$

This MSE can be written as $\text{VAR}(p_{n,h}(x)) + \text{BIAS}^2(p_{n,h}(x))$.

Since the variance is order $1/(nh)$ and the bias is order h^m, it follows that $h_0 = O\left(n^{-1/(2m+1)}\right)$.

The MSE of the kernel density estimator using this bandwidth h_0 would be $O(n^{-2m/(2m+1)})$.

This is known to be an optimal minimax rate of convergence for the class of densities that are m-times continuously differentiable.

Note that if m gets larger and larger, this rate of convergence starts approximating the parametric model rate $1/n^{0.5}$.

Estimation of bias based on parametric working model and assuming smoothness

However, this is not very useful to know for the purpose of selecting a bandwidth for a particular sample. The variance can be estimated well, but as we noticed the bias depends on $p_0^{(m)}$ in the neighborhood of x.

So in order to estimate the bias we would need to estimate the m-th derivative of the density, a much harder problem than estimation of p_0 itself.

Therefore, accurate estimation of the MSE is extremely important making such bandwidth selection methods problematic.

A possible approach is to derive the estimate of the bias under a parametric working model and use this. In that manner one still uses a bandwidth that converges to zero at the optimal rate, and if the working model is a great approximation for the m-derivative, then it might also do a reasonable job on the constant.

Problem with bandwidth selection in this manner

Firstly, even when we know the smoothness m of p_0 at x, then we still have the problem of how to estimate the m-th derivative and thereby the bias term. Secondly, why would one know the underlying smoothness? So this whole approach is theoretically fun to talk about but is not practical.

Adaptive estimation of the choice of kernel and bandwidth

This outstanding problem can be beautifully solved with cross-validation in the following manner.

Let $p_{n,h,m}$ be a kernel density estimator using an m-orthogonal kernel K_m and bandwidth h, where h varies over an interval and $m = 0, 1, 2, \ldots, M$ for some large M.

Suppose that we divide the sample of n observations into V equal size subgroups, which defines V splits of the sample into a training sample of size $n(V-1)/V$ and complementary validation sample VAL_v of size n/V defined as one of the V subgroups.

For a given $v = 1, \ldots, V$, let $p^v_{n,h,m}$ be the kernel density estimator applied to the v-th training sample.

Loss function for density

Let $(p, O) \to L(p)(O)$ be the so called log-likelihood loss defined as:

$$L(p)(O) = -\log p(O).$$

We have that

$$p_0 = \arg \min_{p \in \mathcal{M}} P_0 L(p).$$

$P_0 L(p) = \int L(p)(o) dP_0(o)$ is called the risk of candidate p and p_0 is the unique density that minimizes this risk.

Cross-validated log-likelihood

One can now define the cross-validated empirical risk w.r.t this loss function:

$$CV_n(h, m) \equiv \sum_{v=1}^{V} \sum_{i \in \text{VAL}_v} L(p^v_{n,h,m})(O_i).$$

Since we use the log-likelihood loss, one calls this the cross-validated empirical log-likelihood. The cross-validation selector of the bandwidth and kernel is now given by:

$$(h_n, m_n) \equiv \arg\min_{h,m} CV_n(h, m).$$

The resulting density estimator is now defined as:

$$p_n = p_{n,h_n,m_n}.$$

This super-learner adapts to underlying smoothness

This estimator p_n is an example of a super-learner of the true density p_0, where the candidate estimators are indexed by the choice of kernel m and the bandwidth h.
In the next chapter we discuss the theoretical oracle inequality this cross-validation selector which shows that the density estimator p_n is asymptotically equivalent with the kernel density estimator $p_{n,h_{0,n},m_{0,n}}$ using an oracle selector $(h_{0,n}, m_{0,n})$ that for the given sample minimizes the average over v of the Kullback-Leibner dissimilarity between the candidate kernel density estimator applied to the v-th training sample and the true density p_0.

Thus this oracle selector is doing the perfect trade-off between bias and variance, and somehow this data adaptive density estimator does not only converge at the same optimal rate but cannot even be distinguished up till the constant.

The only assumptions under which this holds is that the loss function at the candidate density estimators needs to be bounded by some universal $M < \infty$, that the number of candidate estimators is bounded by some polynomial power in sample size n.

Suppose that p_0 is m_0-times continuously differentiable, but not $m_0 + 1$ at x, but that this maximal smoothness is unknown, as it will be in any practical application.

Since this super-learner is asymptotically equivalent with the oracle selector, it follows that it will achieve the rate of convergence of the kernel density estimator using the m_0-orthogonal kernel and the corresponding oracle bandwidth.

So we achieve the same performance as if we would have known the underlying smoothness and we would have been told the best possible bandwidth choice of the given sample.